Junior Management Science

www.jums.academy
ISSN: 2942-1861

Impact of Audit Assurance on the Quality of Sustainability Reporting

Alexander Grommes

Catholic University of Eichstätt-Ingolstadt

Abstract

The subject of sustainability reporting is becoming increasingly important. In consequence of the implementation of the Corporate Sustainability Reporting Directive, a substantial number of companies will be required to have their sustainability reports audited beginning from financial year 2024. This paper examines the influence of external assurance on the quality of those sustainability reports. Therefore, the reports of all DAX and MDAX companies for financial year 2022 are examined using a novel textual analysis approach, to determine the individual report quality. The results demonstrate that there is no statistically significant relationship between assurance level and the quality of sustainability reports. Conversely, it was found that companies that are acting sustainable disclose a higher quantity of information and are more likely to demand voluntary assurance of their reports. These findings offer insights into the implications of assurance on sustainability reporting. Furthermore, the detailed overview of traditional and state-of-the-art textual analysis methods offers researchers a valuable resource for identifying the most appropriate methods to address their individual research questions.

Keywords: audit assurance; CSRD; natural language processing; sustainability reporting; textual analysis

1. Introduction

1.1. Motivation

Sustainability has been a topic of interest in business and academic research for some time, but now more than ever. The number of companies reporting on sustainability-related issues is growing rapidly, as is the number of scientific publications (e.g. Amel-Zadeh and Serafeim, 2018, p. 87, Lucarelli et al., 2020, p. 5, Guidry and Patten, 2012, p. 81). This is due to an intrinsic interest in sustainability on the part of companies and their stakeholders (Tworzydło et al., 2022, p. 144), but also to regulatory requirements that have been newly imposed and increasingly refined in recent years (H. Christensen et al., 2021, pp. 1178–1179).

The Global Reporting Initiative (GRI) reporting framework has emerged as the leading standard for sustainability reporting. As an autonomous entity, the GRI has created guidelines with input from stakeholders across the board, fostering a reliable framework for reporting. Companies are not required to adhere to these guidelines by national lawmakers. Rather, they serve as a common ground for reporting. If adopted, the GRI standards enable standardized reporting and facilitate comparison between companies, regardless of their size, sector, or country of operation (Christofi et al., 2012, pp. 163–164).

However, there is more than just voluntary guidelines. In fact, the Directive 2014/95/EU created by the European Union (EU) requires companies to report non-financial information. As a result, public interest entities with over 500 employees must comply with the Non-Financial Reporting Directive (NFRD). Starting in the 2017 fiscal year, these companies are required to disclose information regarding environmental, social, and employee-related matters in their management reports. The purpose of this requirement is to provide stakeholders with a clear understanding of the current state of development and position of companies in these areas (European Union (EU), 2014, pp. 4–5, 8).

Not long after the NFRD took effect, the EU revised its sustainability reporting guidelines through the implementation of Directive 2022/2464/EU, also known as the Corporate Sustainability Reporting Directive (CSRD). This directive was introduced to address significant deficiencies in the pre-

DOI: https://doi.org/10.5282/jums/v10i1pp201-235

vious requirements, which lacked sufficient depth and scope, and to consider issues such as data comparability and reliability. However, one of the main drivers for change is the limited number of reporting companies. The CSRD will make sustainability reporting mandatory not only for public interest entities but also small and medium-sized companies in the future (European Union (EU), 2022, pp. 19–20).

In addition to the new reporting requirements and increased coverage, the CSRD mandates an audit process. Specifically, companies are required to undergo a limited assurance review of their sustainability reports by an external auditor. Under the NFRD, auditors only had to confirm that the related information was published at all. Furthermore, Member States had the option to impose a substantive audit requirement at the national level. The EU aims to establish a consistent link between financial reporting and sustainability reporting by requiring a substantive audit of sustainability reporting conducted by an external auditor, as financial reporting is already subject to a statutory audit. The Commission also reserves the option to take a decision by 2028 to adjust the assurance level from limited assurance to reasonable assurance (European Union (EU), 2022, pp. 34–35; Velte, 2023, p. 4). The mandatory implementation of sustainability report auditing has the potential to aid the EU Commission's objectives and enhance the general compliance of sustainability reports with regulatory requirements. There may be more benefits to consider, but the mandatory audit could also create an additional burden. Similarly, elevating the level of assurance from limited to reasonable could either positively impact reporting or cause unnecessary expenses.

Sustainability reporting is not limited to the European area. Of the world's 250 largest companies by revenue (G250), 96 % disclose sustainability information in the form of reports (KPMG, 2022, p. 13). Only 38 of the G250 are companies from EU members. China represents 30 % of the G250 companies and is showing a positive trend towards reporting on sustainability (KPMG, 2022, p. 18). The majority of the remaining non-EU G250 are located in the United States (69), Japan (26) and the United Kingdom (9) (KPMG, 2022, p. 75). All of these states have extensive, but varying, reporting requirements. On a global level, the International Financial Reporting Standards (IFRS) Foundation addresses the issue of sustainability through the implementation of new standards. In this manner, the standards are designed to meet the needs of the stakeholders of reporting companies, such as customers, employees and investors or the natural environment, which can be considered a stakeholder itself (Technical Readiness Working Group (IFRS Foundation), 2021). Two first two exposure drafts have already been issued. IFRS S1*General Requirements for Disclosure of Sustainability-related Financial Information* is intended to provide general requirements for the disclosure of sustainability-related financial information. IFRS S2 *Climate-related Disclosures* covers the disclosure of climate-related risks and opportunities. Both drafts relate to information that is meaningful to the cash flows of companies and thus to the valuation of those companies (International Sustainability Standards Board, 2022a, 2022b).

With different accounting standards, some mandatory, some voluntary, some national, some international, and with different scope and materiality levels, sustainability reporting is highly diverse. Attempts for homogenization are confronted with ongoing substantive and regulatory dynamics. The implementation of the CSRD regulations could be a crucial step towards improving and harmonizing the reporting landscape.

1.2. Problem definition and objective

While the NFRD is currently in effect, the CSRD will find application for the first companies as early as 2024 which means it will impact the fiscal year of 2023 (European Union (EU), 2022, p. 77). It is likely that the application of the Directive will not fully achieve the desired results at first. Similarly, even after the implementation of the NFRD, there remained potential for further improvement (Busco et al., 2022, p. 95), which is one of the reasons why the CSRD was created. In the context of identified weaknesses of the NFRD, the EU Commission directly mentions the role of the audit of reporting and that this should ensure the reliability of the reports (European Union (EU), 2022, p. 19). Additionally, external audits may also increase compliance with the CSRD and other regulations.

On the other hand, there are expenses associated with the engagement of audit firms. The increased scope of the audit beyond the financial reporting has a direct impact on the total audit costs (Zaman et al., 2011, p. 190). At the same time, a mandatory auditing requirement does not guarantee audit quality. Previous studies have shown a variety of weaknesses that can occur in the area of auditing (B. Christensen et al., 2016, p. 1671).

The purpose of this paper is to investigate the extent to which the audit of sustainability reports can improve the quality of these reports, where quality is primarily expressed in terms of the reports' compliance with regulatory requirements.

Non-financial reporting is heterogeneous and thus provides numerous opportunities for academic research. Due to its actuality, the domain still has some research gaps that can be closed. Since non-financial reporting essentially consists of qualitative reporting in text form, textual analysis methods are particularly helpful in filling these research gaps. This thesis makes several contributions. First, it contributes to the literature in the field of auditing, specifically the auditing of non-financial reporting. This area of auditing, while not entirely new, is considerably less investigated than the area of financial reporting. Second, the thesis also contributes to the literature on European financial reporting requirements. In particular, it connects those two streams of literature. Third, it offers a methodological contribution by providing an up-to-date review of textual analysis methods. Finally, a contribution is made by providing evidence on whether auditing improves the quality of non-financial reporting. Stakeholders and other recipients of non-financial reports can assess the value that auditing provides when making investment decisions. The findings can also support the EU Commission's decision on future assurance level increases.

1.3. Procedure of the work

This thesis is organized as follows. Chapter 2 presents the relevant theoretical background. First, it consists of the regulatory framework with a focus on European accounting, in particular the EU taxonomy. Second, the established theories of sustainability reporting and auditing in general are presented. The theoretical application of textual analysis, which is the central instrument of this thesis, concludes the chapter on the theoretical background.

The following chapters form the first of the two main parts of this thesis. Chapter 3 discusses the relevant literature in the domains of finance and accounting, while Chapter 4 presents the various methods of textual analysis in terms of their functionality and applicability. These methods are not only distinguished according to their field of application, but also between traditional methods and state-of-the-art methods, which utilize the recent technical developments in machine learning and artificial intelligence. On the one hand, the comprehensive presentation of all currently available methods forms the necessary groundwork for the second part of this thesis. On the other hand, it offers a contribution in itself, since it can assist the audience of this thesis in identifying appropriate methods for own research projects in the field of textual analysis.

The second part of the thesis involves utilizing textual analysis techniques to assess corporate non-financial reporting. For this purpose, Chapter 5 formulates three related hypotheses. Chapter 6 describes the methodology by discussing the data set, the research design, and the processing of relevant variables trough textual analysis. The results of the analyses are presented in Chapter 7. Finally, Chapter 8 concludes the thesis and discusses the limitations of the work.

2. Theoretical background

2.1. Development of the regulatory framework

The climate crisis is one of the greatest challenges of our time. Its negative effects are already being experienced today and will get worse as they become more difficult to mitigate in the future (United Nations, 2022). The majority of United Nations member states are committed to addressing the climate crisis through the Paris Agreement, which aims to limit the increase in global temperature to a maximum of two degrees Celsius above pre-industrial levels, raise overall adaptation capabilities to the impacts of climate change, and shift capital flows to a climate-friendly development (European Union (EU), 2016, p. 5). Europe is contributing through the European Green Deal. This framework includes a program of measures for the necessary transformation. The European Commission has set the goal of making Europe the first continent to become climate neutral by 2050 by reducing greenhouse gas emissions to net zero. The interim target is a reduction of emissions by 55 % until 2030 compared to 1990 levels. The European Green Deal also covers issues such as the sustainable use of consumer goods, with specifications for producers to enable consumers to repair products more easily and make them last longer so that the goods do not have to be replaced. Other parts of the program cover the fields of technology, mobility, food, energy and biodiversity, and various other subjects (European Commission, 2019).

Corporate governance is also specifically addressed in the European Green Deal. Companies are still too focused on short-term financial performance rather than sustainable development. Therefore, companies must increasingly disclose their information on sustainability-related issues alongside their annual reporting in order to inform investors about their development in these areas (European Commission, 2019, p. 17).

The EU taxonomy is part of the European Green Deal and is designed to accompany and support the transition of the environment to the target state. The taxonomy introduces various instruments to achieve this goal and is also supposed to support the financing of the transition by directing capital flows in a way that is conducive to the transition. Another integral part of the EU taxonomy is corporate disclosure (European Commission, 2020, p. 8). Regulations require two groups of companies in particular to address this issue: Financial market participants[1] offering financial products within the EU and companies meeting the size criteria of the NFRD. The required content, especially for the second group of companies, is discussed in more detail in Chapter 6.2 of this thesis. The information must be published either in the non-financial section of the consolidated or annual financial statements or as a stand-alone non-financial reporting or sustainability reporting (European Commission, 2020, p. 27). The EU taxonomy encourages, but does not require, companies to obtain assurance from external auditors (European Commission, 2020, p. 37).

Even before the introduction of the EU taxonomy, many researchers addressed these substantive issues (Lucarelli et al., 2020, p. 6). More recent research identifies the benefits of the taxonomy mainly in the area of harmonization and investment decision support (Dumrose et al., 2022, p. 2), which is consistent with the reporting objectives of the IFRS. It is also important to consider the full scope of the taxonomy. Beyond the entities directly impacted by the EU taxonomy, other entities are also indirectly affected (Dusík & Bond, 2022, p. 92). Suppliers and customers which do not meet the thresholds for mandatory NFRD reporting do not have to collect environmental data for themselves, but may need to be able to provide it to companies covered by the NFRD for their reporting.

In principle, the requirement for more comprehensive reporting also leads to a reduction in information asymmetries. Although this relationship exists in theory, it should not be blindly assumed without evidence and needs to be further

[1] e.g. Equity funds, exchange-traded funds, real estate funds, pension schemes, venture capital and private equity funds.

investigated (Breijer and Orij, 2022, p. 350; H. Christensen et al., 2021, p. 1231).

The EU taxonomy has one particular strength. By prescribing the narrative that sustainable activities can only be considered as sustainable if they do not harm other sustainable activities, trade-offs between different areas of development cannot be used as a loophole. The benefits of the EU taxonomy presented in academic literature may also extend beyond the European area. The EU serves as a prominent model for implementing regulations, making it probable that legislators outside of Europe will adopt these or raise similar requirements (Bloomberg, 2021; Dusík and Bond, 2022, pp. 93, 96).

2.2. Fundamental theories on sustainability reporting

Before dealing with the methodology, the theoretical principles need to be defined. Essentially, the publication of company data is crucial for capital market participants and their investment decisions, with information content and timeliness being particularly important (Ball & Brown, 1968, p. 176). For this thesis, theories that consider voluntary disclosure are most relevant. For a long time, sustainability reporting within non-financial reporting has been on a largely voluntary basis, and companies only disclosed data when the benefits exceeded the costs associated with the disclosure. The NFRD made sustainability reporting mandatory for some companies, but the regulation still allows a great amount of flexibility in the nature and extent of disclosure, which is why voluntary disclosure theories are especially relevant.

Voluntary disclosure refers to a company's decision to publish supplementary information beyond what is required by law. There are various determinants that influence whether and how much voluntary disclosure is made, including firm characteristics, ownership structure, or country-specific factors (Zamil et al., 2023, pp. 232–235). For the purposes of this thesis, however, the general theories, on which voluntary disclosure is based, are crucial.

Agency theory, which is most often applied in the context of voluntary disclosure (Zamil et al. 2002: 239), is closely linked to the well-known principal-agent problem from economics, which is primarily founded on information asymmetries between parties (Arrow, 1963, p. 967). According to agency theory, firms voluntarily disclose information in order to reduce information asymmetries between themselves and their stakeholders and thus facilitate business relationships or capital flows.

In addition to the agency theory, the next two theories most commonly used in this context are legitimacy theory and stakeholder theory. Legitimacy theory is concerned with the interaction between companies and their social environment and postulates that companies strive to shape their actions, decisions, and practices so that they are viewed as legitimate and acceptable by the society. Through voluntary disclosure, companies seek to achieve the necessary legitimacy and gain the trust of stakeholders. The theory is founded on the premise that there is a social contract between companies and society. Recently, increased awareness of Corporate Social Responsibility (CSR) concerns has influenced corporate practices in sustainability reporting, and companies have used CSR disclosure to gain legitimacy (Lepore & Pisano, 2023, pp. 56–57). Stakeholder theory, on the other hand, emphasizes that companies are not only beholden to the interests of their owners, but should also take into account the interests of a broader group of stakeholders who are affected by the company's activities. This theory emphasizes that companies should recognize the expectations, values, and needs of their various stakeholders. Therefore, voluntary disclosure can be seen as an effort to increase transparency and address stakeholder interests and concerns. In addition, there are several other theories on the basis of which voluntary disclosures can be useful for companies (Zamil et al., 2023, p. 239).

These theories explain different incentives for companies to voluntarily disclose information. Furthermore, the voluntary disclosure theory considers the costs of disclosure and suggests that information will be voluntarily provided only if the benefits for the company outweigh the costs of disclosure. According to this principle, information that is insignificant or disadvantageous to a company will not be disclosed (Verrecchia, 1983, pp. 179, 192).

2.3. Fundamental theories on audit

In addition to voluntary disclosure theories, the principal theories of auditing are particularly relevant for this thesis. Auditing is one of the central areas of accounting. The external verification of financial or non-financial information by an auditor can ensure the reliability of reporting for external stakeholders. In very simplified terms, this is achieved by the audit firm determining the actual financial position and performance of a company through various audit procedures and comparing these with the figures reported in the financial statements (Wagenhofer & Ewert, 2015, pp. 410–411).

The exact procedure and structure of the audit is not the focus of this thesis. Instead, the basic theories and related concerns are addressed in order to understand how they relate to the audit of sustainability reporting. Again, the principal-agent theory is a fundamental theory with significant importance. This behavioral theory can be applied to audit firms and their client companies. The principal, representing the company to be audited, hires an agent, the audit firm, to perform an audit of the company's disclosures. With a predetermined audit fee, the auditing company lacks in motivation to undertake high costs in the form of a detailed audit. Instead, the auditing company seeks to maximize its own benefit by minimizing the audit effort, since the compensation remains the same. The client and other parties seeking audit assurance suffer as a result (Antle, 1982, pp. 503, 508, 512). However, this opportunistic behavior exists only in theory. In practice, other factors also influence audit intensity. For example, the audit result itself is reviewed by other entities, and insufficient audit actions can be sanctioned. Nevertheless, it is useful to keep in mind the fundamental problems that arise in auditing and the application of flat fees.

Audit firms use various forms of auditing procedures to detect accounting manipulation or unintentional misstatements. The model structure distinguishes between substantive and systematic audit procedures. While substantive audit procedures provide assurance on specific balance sheet items, for instance by sampling, systematic audit procedures provide broader assurance, for example by testing the functionality of internal control systems. In most cases, the desired level of assurance is achieved through a combination of both types of procedures (Wagenhofer & Ewert, 2015, pp. 432–435). Auditing sustainability reporting is unique in that it concerns non-financial reporting. Essentially, non-financial reporting provides more qualitative information rather than actual numbers, as it would be in the case for financial reporting. Almost all audit firms refer to the International Standard on Assurance Engagements (ISAE) 3000 (Revised) when performing sustainability reporting audits. This standard specifically covers the audit of information that can be classified as non-financial information (International Auditing and Assurance Standards Board, 2013, p. 5) and is therefore considered an *umbrella standard*. As the ISAE 3000 (Revised) is applied to a wide range of disclosures, it does not contain explicit audit procedures. Rather, it describes general requirements for audit firms, such as integrity, independence, and professionalism, which are also required for financial audits. It further provides detailed information on the content and scope of the audit firm's reporting on its engagement. For the actual audit, the standard primarily requires auditors to review the content of the qualitative disclosures for material inconsistencies (International Auditing and Assurance Standards Board, 2013, pp. 20–21). However, the standard does not specify how materiality has to be determined or how audit procedures should be performed.

There is no dispute that the audit in itself is a valuable tool. In general, auditing increases the credibility of the information disclosed, as shown, for example, by the fact that firms with audited financial statements pay lower interest rates than comparable firms with unaudited financial statements (Blackwell et al., 1998, pp. 58, 68). It should be noted, however, that the magnitude of such an effect varies depending on whether the information disclosed is favorable or unfavorable for a company. Another variable of particular importance for this thesis is the voluntariness of the audit. The following 2x2 matrix illustrates four possible conditions that financial or non-financial reporting can adopt.

According to the attribution theory, financial statement users challenge positive information because it is consistent with the company's interests. Negative information, on the other hand, is less challenged since it would not be reasonable for companies to misrepresent information that is not in their best interest. This theory is confirmed by practice. In experiments, Coram et al. show that voluntary audit assurance of positive sustainability disclosures has a significant positive effect on the share price. In contrast, no significant results were found for negative disclosures. It can therefore be concluded that financial statement readers have reliability concerns mainly when the disclosed information is positive. These concerns are consistent with attribution theory, which suggests that it is beneficial for firms to voluntarily undergo an external audit when published information is positive, in order to increase reliability of those results, while negative information already carries a higher level of reliability (Coram et al., 2009, pp. 145–148).

These results are relevant for the research of this thesis, as the EU member states' option right and the implementation within Germany allow companies to voluntarily subject their non-financial reporting to an audit. Accordingly, such a voluntary submission can have different motivations: The counteraction of the principal-agent relationship, the creation of a higher reliability of the information, especially if it is positive and therefore, according to the attribution theory, more likely to be doubted by readers of the report, or simply the satisfaction of stakeholders in order not to be at a comparative disadvantage to other companies (Bradbury, 1990, p. 33).

The presented theories provide the basis for multiple research streams. They also form the basis for the development of the hypotheses for this thesis presented in Chapter 5.

2.4. Textual analysis in research

Research in business economics heavily relies on quantitative methods for gaining new findings. The rationale is clear: countless amounts of data exist in numerical form. Financial statements containing balance sheets and profit and loss statements, stock prices and a vast range of related financial indicators, as well as statistical information on companies, industries, regions, and countries. The amount of numerical data is enormous. When this data is effectively contextualized, new insights can be uncovered. However, how do researchers handle data that is not numbers, but letters? In addition to the balance sheet, every financial statement provides notes. The income statement enables insight into earnings, but the management report covers even more. Stock prices and financial ratios are paired with analysts' recommendations and company announcements, both written and verbal. Each statistical survey is accompanied by a corresponding text. All of this information is easily overlooked.

However, it would be incorrect to state that textual data is not a topic of interest in research at all. In fact, this field of research has been growing in importance for some time. As a result, both earlier and more recent papers cover not only results in this context, but also the methodology on its own (e.g. Bae et al. (2023), Bochkay et al. (2023), Gentzkow et al. (2019), and Loughran and McDonald (2016, 2020)).

The EU taxonomy and especially the NFRD requirements have greatly increased the volume of non-financial reporting. This new information in form of textual data provides potential for research using textual analysis methods. However, it is important to ensure that the methodology does not take precedence over the actual research question (Bochkay et al., 2023, p. 792; Bae et al., 2023, p. 3). Therefore, before addressing the hypotheses, the methodology of textual analysis will be examined in detail based on the existing literature.

Table 1: Effects of audit on reliability based on the experiment of Coram et al. (2009, pp. 142–145)

	Positive information content	Negative information content
Audit assurance	High reliability	High reliability
No audit assurance	Low reliability	High reliability

This review provides a summary as well as an explanation of current methods and highlights their advantages, disadvantages, and areas of application. This will not only identify the appropriate methods to apply to the research purpose of this thesis. It also provides a valuable contribution as it summarizes the leading research in various fields, particularly in the domain of finance and accounting.

Textual analysis has already been used to find evidence in several areas. Chen et al. found that both stock returns and earnings surprises can be predicted by peer-based knowledge on social media. They used one of the simplest methods imaginable: counting negative connoted words in articles written by individual investors on the social media platform Seeking Alpha. The ratio of negative connoted words[2] to the overall number of words was utilized to determine first a negative sentiment and then a decline in stock returns and even in earnings surprises. This effect increased as the number of negative words increased (H. Chen et al., 2014, pp. 1368–1369, 1382, 1400).

In a more recent study, Sautner et al. measured the extent of corporate exposure to climate change by using an algorithm to count key words[3] in earning call transcripts that are directly related to this topic. This method captured climate change exposure from the perspective of all key stakeholders, as the earning call transcripts included both shareholder and stakeholder questions as well as management responses (Sautner et al., 2023, pp. 1450–1451, 1492–1493).

Using a comparable methodology, Chen and Srinivasan analyzed the 10-K reports of non-tech firms to investigate the relationship between digital activities, firm value, and performance. Specifically, they measured the frequency of digital terms[4] in the description section of these reports. The authors discovered that non-tech firms have generally increased their digital activities over time, and that greater involvement in digital activities has a positive impact on firm value and stock performance. These findings were made possible by quantifying the degree of digitalization within firms through textual analysis (W. Chen & Srinivasan, 2023, pp. 2, 10, 29, 35).

As a final example, in a 2014 study, Purda and Skilliorn analyzed quarterly and annual financial statements for fraudulent activities. A multilevel textual analysis process first sorted words within the sample by frequency, then tested their predictive power using a decision tree-based approach, and finally concluded from the reports the probability that the statements were completely true and did not contain fraud. The algorithm based on textual analysis was able to confirm the presence or absence of fraudulent activity in over 82 % of the reports (Purda & Skillicorn, 2015, pp. 1194, 1197–1200, 1218).

The listed research is illustrative for the wide range of possible applications for textual analysis. Ranging from meaningful, market-relevant results in the area of finance, to risk exposure in the example of climate change, to opportunities for companies in the example of technology adaptation, to relevant accounting issues like fraud detection, the application possibilities are unlimited. The textual data examined ranges from individual social media posts to transcribed communications between companies and stakeholders to official corporate disclosures in annual and quarterly financial statements. This demonstrates that textual data can contain relevant information in any conceivable form, regardless of its type and nature.

In quantitative research, the approach is usually relatively straightforward. With the evaluation of a sample using statistical methods in direct relation to a hypothesis, researchers intend to obtain significant findings. The procedure in textual analysis is not as simple, as the database is initially qualitative. The information required for research can only be obtained in an exploitable form by means of an appropriate transformation (Loughran & McDonald, 2016, p. 1191).

The major difficulty is not that textual data is less structured or presented in a different way, but rather that it has a high dimensionality. The base of the dimensions is defined by the number of different words in a language, and the exponent by the number of words in the text, as shown in *Equation 1*. When taking a text that consists of only ten words, and it is written in a fictional language that also only possesses ten different words, then this text can have ten billion different dimensions, each of which is different from the others. In reality, texts are much longer than ten words, and languages consist of more than ten different words, so both the exponent and the base, and thus the total number of dimensions, take on an unimaginably high degree (Gentzkow et al., 2019, pp. 535–536).

$$t = n^l \tag{1}$$

$t = textual\ data\ dimensions$

$l = language\ word\ options$

$n = length\ of\ text\ in\ words$

[2] Examples for negative connoted words are *loss, termination, against or impairment.*

[3] The key words with the highest frequency were *renewable energy, electric vehicle, clean energy, new energy, climate change and wind power* (Sautner et al., 2023, p. 1466).

[4] Examples for digital terms are *analytics, virtual reality, automation, artificial intelligence, big data, data science or digitalization* (W. Chen & Srinivasan, 2023, p. 36).

Humans can handle the high number of dimensions because they are not important when reading text. Words are perceived and interpreted in the context of other words. Accordingly, sentences do not represent a sequence of independent variables. In textual analysis, however, these dimensions are important and must be addressed. At the beginning of any textual analysis, the number of dimensions considered should be drastically reduced in order to deal with the enormous amount of data. This reduction is usually performed within three steps. The first step is to divide the total volume of text into sections suitable for research. In the later course of this work, non-financial reports from different companies are analyzed. It is not necessary to examine the reports of all companies together. Rather it is sufficient to extract information from each report separately. This information can then be used to draw conclusions by applying other research techniques. By analyzing individual texts separately instead of performing one single overall analysis, the number of dimensions can be drastically reduced. In a second step, certain parts of the text can be excluded from the analysis. These are first of all frequently occurring words that maintain the grammatical structure of a text. Such words are important for human readers of a text, but contain little or no information that will emerge in the textual analysis. In addition, for some research it can also be useful to exclude words that occur very rarely in a text. Although these words may contain relevant information, the benefit of gaining this information could be outweighed by the additional effort involved in analyzing these words. In a final step, the stemming method can be used to adjust all words that have the same meaning but are spelled differently. This method unifies differently conjugated words by removing their suffixes. For example, the words *connected, connecting, connection, and connections* have the same informational meaning. By replacing them with their stem word *connect*, the dimensional base of the overall text is reduced once again (Porter, 1980, p. 130). The important aspect is to decrease the number of different words with identical information content. With these three steps, the dimensions of a text can be drastically reduced by lowering the base n of the dimension equation (Gentzkow et al., 2019, pp. 537–538).

The efforts for simplification are addressing the base n of *Equation 1*. From a purely mathematical point of view, a reduction of the exponent would be more effective. However, it is not as easy to reduce the exponent. Textual data from a sample can be decreased or simplified at the expense of information loss. But the vocabulary of the language in which a text is written is exogenous.

In textual analysis, instead of trying to reduce the exponent, often the entire formula gets modified. The so-called *bag of words* method ignores the position of words in a text. Alternatively, it only counts whether and how often individual words occur. This implies that the number of possible dimensions only result from the multiplication of the number of *words n* in the text by the number of possible words of a *language l*. Whereas before, a text of ten words length in a language containing only ten different words already had ten billion dimensions. Using the bag of word method a text in English language[5] with the same amount of dimensions can contain more than 100,000 words. Thus, ignoring the order of words in texts leads to a massive reduction of the dimensionality of a text (Gentzkow et al., 2019, pp. 539–540).

After presenting a general overview of how textual analysis works, the next step is to specify its application areas. Text contains some information, but what kind of valuable information is included and how can it be extracted? Prior research has established various applications of textual analysis for information acquisition, which will be presented in the following literature review for the finance and accounting domain, which does not present all the literature, but the most important in terms of the objective of this thesis.

3. Literature review

3.1. Readability

Readability is one of the main areas of use in textual analysis. Depending on the context, the definition of readability varies. Either way, it should somehow determine if text is designed in a way that readers can recognize and comprehend the underlying message (Loughran & McDonald, 2016, p. 1188). More specifically, readability represents the relationship between a text and the cognitive load required to understand it (Martinc et al., 2021, p. 143). Even if a text is generally comprehensible to a reader, a high cognitive load, or in simple terms, a text that is challenging to read, may indicate poor readability. The readability of a text always should be considered in the context of the target audience (Loughran & McDonald, 2020, p. 28). The United States Securities and Exchange Commission (SEC) supports this position (United States Securities and Exchange Commission (SEC), 1998, p. 9). The *Plain English Handbook* published by the SEC describes the linguistic form in which publications should be made. This includes the annual financial statements and non-financial reporting components. The handbook recommends, among other things, the use of everyday language words and short sentences. It further recommends to perform automatic readability checks by using formulas developed for this purpose, but also manual testing by simple human proofreading of own publications (United States Securities and Exchange Commission (SEC), 1998, pp. 18, 57). This thesis examines non-financial reporting with focus on sustainability reporting within the German market, although European or other international regulators have objectives for publication requirements similar to the SEC.

The readability of text has been subject of many studies. Li's widely cited paper examines the relationship between the readability of annual reports and company earnings. Here, readability was measured by two variables, the so-called *Fog Index* and the length of the reports. The Fog Index is closely

[5] There are several underlying bases for determining the number of English words. The following example is calculated using the number of 88,500 English words (Nagy & Anderson, 1984, p. 320).

related to the topic of readability and will be discussed in Chapter 4.2.1 in more detail. Li found that companies with a low readability score on their reports had worse earnings persistence. On the other hand, companies with easily readable annual reports are more persistent. This correlation could suggest that management is hiding negative information about its company by making the information in the reports more difficult to access (Li, 2008, pp. 222, 225–226, 244–245).

Reporting quality can also influence investment allocation. Biddle et al. found that companies with higher reporting quality are less likely to be over- or underinvested. Over- or underinvestment occurs when the marginal benefit of a capital investment is lower than the marginal cost of that investment. Information asymmetries between management and investors in capital markets are one main reason for such inefficient allocations. Higher reporting quality can reduce the occurrence of adverse selection and its effects. Biddle et al. use the Fog Index to measure reporting quality and find a positive relationship between increasing reporting quality and decreasing over- or underinvestment (Biddle et al., 2009, pp. 113–114).

Not only the general allocation efficiency can be related to readability, but also individual trading behavior. Miller finds that shares of companies with less readable reports are traded less frequently. This effect is most evident among small investors and can be explained by the higher cost of information acquisition, which is particularly important for such investor groups (Miller, 2010, pp. 2108, 2114, 2138). Lawrence found similar results in that investors are more likely to hold on to shares of companies with more readable financial disclosures. For the increase of one standard deviation in readability, stock returns increase by 91 basis points on average (Lawrence, 2013, pp. 131, 135, 141–142, 144). These two studies use the Fog Index and text length as readability measures.

Readability research has also provided insights in the domain of accounting. Chychyla et al. found a relationship between reporting complexity and accounting expertise within companies. They approximate accounting complexity by various parameters, including firm characteristics such as firm size and number of segments covered, as well as financial reporting variables such as the number of words in 10-Ks and their readability. The level of accounting expertise is approximated through boards of directors and audit committees. More specifically, the number of accounting experts[6] in these functions represent the level of accounting expertise within a company. Chychyla et al. argue that companies with a high level of reporting complexity also have a higher level of accounting expertise. This expertise should counteract the negative effects of accounting complexity and is expected to actively manage accounting complexity (Chychyla et al., 2019, pp. 227–229, 233–236, 247–248).

Another example of how companies can influence their own reporting is the study by Chakrabarty et al. on the relationship between executive compensation and disclosure transparency. The study concludes that firms with managers receiving higher risk incentives, measured by the stock option compensation, produce less comprehensible 10-K reports. This is because these incentives encourage managers to undergo risky projects with higher rewards, which may not be in line with the company's strategy. Management attempts to camouflage the undertaking of such projects by making the reporting less readable. Here, readability is assessed through the size of 10-K reports, as evidenced by Loughran and McDonald (2014). The result shows that companies in the top quartile of the stock option vegas[7] publish reports that are 15.4 % larger. The results were tested for robustness via variables such as firm complexity, while other testing, such as measuring readability via the Fog Index also supported the results (Chakrabarty et al., 2018, pp. 3, 5–7, 10–11, 13, 25).

A recent study by Dorfleitner et al. examines readability, among other issues, in a setting similar to the one in this thesis. Dorfleitner et al. investigate the impact of the General Data Protection Regulation (GDPR) on privacy statements. Just like the NFRD, the GDPR was published as a directive in the EU and became binding law in the form of a regulation in 2018. Using methods similar to the Fog method, privacy statements were tested for readability. The result was a worsening of readability due to the introduction of the GDPR. Even when considering the number of words as a readability measure, there was a decrease in readability due to an increase in the length of privacy statements (Dorfleitner et al., 2023, pp. 1–2, 4, 10–12).

The papers presented round off the literature review in the field of finance and accounting for the application area of readability, with the research area being constantly expanded and, above all, newer state-of-the-art methods being increasingly used.

3.2. Sentiment analysis

In the context of textual analysis in accounting, sentiment analysis finds even more interest than the topic of readability (Bochkay et al., 2023, p. 797). In sentiment analysis, texts are examined to determine whether they have a positive or a negative tone. This has already produced findings in a wide variety of research areas, even in far unrelated fields.

For example, Chevalier and Mayzlin examine the sentiment of customer reviews for books in the two largest online bookstores using a differences-in-differences approach and find that reviews with positive (negative) sentiment lead to significantly higher (lower) sales on the respective site. They also find that reviews with negative sentiment have a

[6] An individual is considered an expert if he or she is a certified public accountant (or similar) or has professional experience in relevant areas such as treasury or auditing.

[7] Vega measures how the value of a stock option changes as the volatility of the underlying asset changes (Black & Schloes, 1973, pp. 638–639). The vega parameter is used because it is expected that as volatility increases, management will take on riskier projects in order to increase the value of their own options.

stronger impact than reviews with positive sentiment (Chevalier & Mayzlin, 2006, pp. 345–346, 350).

In financial context, sentiment analysis has been successfully used to predict price movements in capital markets. Chen et al. find that the sentiment of opinion articles on the social media platform Seeking Alpha can be used to draw conclusions not only about stock performance, but also about the development of company earnings. With the help of sentiment analysis, they were able to identify crowd knowledge that contains not only capital market-relevant information but also real economic information that has not yet been included in the stock price (H. Chen et al., 2014, pp. 1368, 1382–1383, 1391–1392, 1400). Long et al. found similar results in predicting the price movements of so-called *meme stocks* by developing a sentiment score dictionary customized for the social media platform *Reddit*. They assigned weightings to words that have a particular meaning in a given context, so that meaningful words have a disproportionately strong influence on the sentiment of a sentence or text (Long et al., 2023, pp. 22, 25–27, 33–34).

In addition to prices, sentiment can also be used to determine other metrics such as liquidity. Agrawal et al. found that social media activity on platforms such as Twitter and StockTwits can be used to determine liquidity developments. For example, peaks and troughs in intraday liquidity can be identified. Analogous to the results of Chevalier and Mayzlin in a different context, they also show that negative sentiment has a much stronger influence than positive sentiment (Agrawal et al., 2018, pp. 86, 89, 93).

Next to social media, other sources can be used for sentiment analysis. Kothari et al. perform sentiment analysis on reports from corporate management, analysts, and news services to determine the impact on firms' cost of capital, stock return volatility, and analysts' earnings forecasts. The results confirm that positive or negative sentiment has an impact on each of the indicators examined, with the result depending on the credibility of the source. Since management is interested in presenting itself well, the impact of positive news from management is significantly weaker than that of negative news or positive news from other sources (Kothari et al., 2009, pp. 1641–1642, 1653, 1657, 1664). This result is of particular interest for this thesis, since sustainability reports are written by the companies themselves and are therefore subject to a fundamental sense of skepticism.

Huang et al. further illustrate that the words of management should not be taken at face value. By analyzing earnings press releases, they examine the tone of these releases and find that an abnormally positive tone predicts negative future earnings. The same content in an earnings release can be expressed in different ways. Sentiment analysis can be applied to distinguish whether earnings releases are expressed in a positive or negative tone, by measuring the ratio of positive to negative words. A neutral tone corresponds to describing the available fundamental information about a company, while an abnormally positive tone is indicated by too many positive words not related to the fundamentals. A profitable, growing company is not assumed to have an abnormally positive tone despite a high number of positive words, whereas poorly performing companies with the same number of positive words would be assigned such a classification. Huang et al. find that not only is an abnormally positive tone inconsistent with fundamentals, it is actually an indicator of negative future earnings and cash flows. An unusually positive tone is most likely to occur when a company is about to meet or exceed the previous year's results or analysts' forecasts. Conversely, earnings releases can also have an abnormally negative tone, especially if a significant portion of executive compensation consists of stock options, giving them an incentive to reduce the share price in the short term in order to buy shares at a favorable price later on. In both cases, management is distorting the company's situation for its own benefit and to the disadvantage of investors. This behavior can be revealed by sentiment analysis of earnings press releases (X. Huang et al., 2014, pp. 1084–1086, 1090–1091, 1094, 1103, 1111).

3.3. Disclosure quantity and similarity

Readability and sentiment analysis account for over 60 % of the textual analysis papers published in the top accounting journals[8] over the last decade (Bochkay et al., 2023, pp. 795, 797). Other important topics are disclosure quantity, text similarity, and topic discovery. These areas can be considered separately, depending on the research question. However, it can also be reasonable to consolidate them. The measurement of the quantity of content, the discovery of topics or the assignment of textual data to a topic is typically performed under the overriding interest of similarity and comparability studies. These areas of textual analysis also yield various findings.

Frankel et al. examine the characteristics of earnings calls, in which public companies discuss their financial performance, and find that earnings calls of companies that miss analysts' expectations by just a penny are disproportionately long. The average length of a call in the study sample is just over 50 minutes. Calls of companies that miss earnings expectations by a penny last 1.1 minutes longer. This deviation is significantly larger than, for example, a one-penny overshoot of expectations or the delta from a one-penny miss to a two-penny miss. Frankel et al. were thus able to demonstrate that the quantity of information is abnormally deviant in such cases. However, this result is particularly interesting because a parallel study of the sentiment of earnings calls found that these unusually long calls did not show any noticeable shifts in their tone. In general, calls that miss analysts' earnings expectations have a more negative tone than those that meet expectations. However, a one-penny miss does not lead to a noticeable difference in tone. Thus, in this case, the sentiment analysis of the data is inconclusive, but the quantity analysis does reveal new evidence (Frankel et al., 2010, pp. 221–222, 227, 230, 240).

[8] The papers which have been under consideration are *The Accounting Review, Journal of Accounting Research, Journal of Accounting and Economics, Contemporary Accounting Research, Review of Accounting Studies, Accounting, Organizations, and Society, and Management Science*.

In another study, Huang et al. address earning calls, but instead of examining the information quantity of those calls, they focus on the closely related topic of similarity. Huang et al. use textual analysis to compare earning calls with analyst reports based on those calls. In this context, analysts are expected to gather information from earnings calls and interpret them in a way so that investors can utilize their reports to make investment decisions. Latent Dirichlet Allocation (LDA) is used to determine which content of analyst reports corresponds to the topics of the earnings calls. LDA is a textual analysis method that identifies topics within text by capturing statistical correlations of words. Using their method, Huang et al. discovered that 31 % of the analyst reports only briefly addressed topics that appeared in the earnings calls, while the remaining 69 % directly corresponded to the topics of the earnings calls. Using similarity analysis and topic discovery within textual analysis, it was therefore possible to quantify the value added by analysts. The 31 % of analyst reports content that is not included in the earnings calls can be argued to offer added value in terms of information discovery (A. Huang et al., 2018, pp. 2833–2835, 2840, 2848).

However, similarity can be examined not only for earning calls. Gaulin and Peng use textual similarity analysis to compare compensation disclosure. Compensation disclosure is a highly regulated area. It is therefore appealing to explore the extent to which the disclosures of companies in different domains vary and where there is a high degree of similarity. Gaulin and Peng disregard traditional textual analysis methods and apply innovative machine learning techniques instead. A novel algorithm, trained on company disclosures, is designed to detect relationships that remain undiscovered by traditional methods. These traditional methods, which are examined in more detail in Chapter 4.2 of this thesis, have a significant weakness in that the similarity outcome is heavily reliant on the text length of the disclosure, which in turn is strongly correlated to the size of the company. New natural language processing (NLP) methods (described in Chapter 4.3 of this thesis) such as the novel algorithm can overcome these weaknesses. Gaulin and Peng also apply the traditional methods and recognize that their state-of-the-art methodology outperforms those methods. The study differs from many others in this domain in that it explicitly focuses on the narrative of the disclosures rather than the quantity of them. The results show that companies of comparable size, companies in the same industry or companies using the same compensation consultants show similar disclosures (Gaulin & Peng, 2022, pp. 1, 3, 4, 6, 21, 25–26). Although these findings are not relevant to this thesis, the results of the methodology highlight an important aspect and valuable research applications can be drawn from them.

In the accounting domain, Brown and Tucker perform a similarity analysis on Management Discussion and Analysis (MD&A) disclosures and examine how they are changing relative to real economic changes. MD&A disclosures are a mandatory part of reporting for U.S. companies, in which business-related topics such as earnings, liquidity, and material risks are presented from the perspective of management according to the management approach. Although the content of the topics is predefined, the form in which they are disclosed offers a huge degree of freedom, which is why an examination of these disclosures can provide interesting findings. Brown and Tucker develop a modification score for MD&A disclosures to measure the similarity of such documents across time and find that the score decreases over time as the length of disclosures increases. These results suggest that companies are increasingly using boilerplate language, i.e., standardized language that is not firm-specific or adapted to economic circumstances. At the same time, they also find that MD&A disclosures change more when economic changes are more pronounced. Brown and Tucker's similarity analysis is a methodological approach that complements other textual analysis methods such as readability and sentiment analysis and has made it possible to determine the usefulness of MD&A disclosures for investors or other stakeholders over time (S. Brown & Tucker, 2011, pp. 310, 312–313, 315–317, 327, 341).

Lang and Stice-Lawrence conduct similar studies, but instead of focusing on MD&A disclosures, they examine entire annual reports using a variety of analytical techniques. They find that reporting quality affects real economic variables such as liquidity, analyst coverage, and ownership structure. Their study covers several domains: reporting quantity in terms of the length of the annual report, readability as measured by the Fog Index, the proportion of boilerplate language, and the domains of comparability and similarity. The authors argue that the information content increases with the length of the annual reports, while a higher proportion of boilerplate language reduces the information content. Comparability between annual reports is measured not only on the basis of visual similarity, but also on the basis of content similarity. For this purpose, the cosine similarity is used as a measure, which was also used by Brown and Tucker for the similarity score and is discussed in more detail in Chapter 4.2.3 of this thesis. In their research, Lang and Stice-Lawrence observed the adoption of IFRS accounting by some companies and, in addition to the findings that reporting quality has an impact on real economic variables, they also found that companies with high reporting quality in particular benefit from the adoption of IFRS standards (Lang & Stice-Lawrence, 2015, pp. 110–113, 131). A similar application is also conceivable in the area of non-financial reporting and can help to address the research objectives of this thesis.

4. Presentation of methods in textual analysis

4.1. Overview

After the previous chapter has presented the relevant literature categorized according to the different fields of application, the focus now shifts towards the functionality of the subject areas. First, the traditional methods that form the foundation of textual analysis are reviewed. Before then introducing the state of the art methods, the functionality of NLP and machine learning in the field of textual analysis will be thoroughly covered. These methodological principles help to better comprehend the new methods. This is followed by a presentation of those state-of-the-art methods, which have been used especially in recent academic research and which complement and further develop the traditional methods. The detailed consideration of all methods on the one hand represents an up-to-date overview of the possibilities of textual analysis in the field of accounting. On the other hand, it ensures that the appropriate methods are identified for the research questions of this thesis and that the risk of overlooking potential applications is minimized. The methods analyzed are listed in *Table 2* for ease of reference.

4.2. Traditional methods

4.2.1. Readability

Beginning with the subject of readability, the Fog Index is indispensable. This index, developed by Robert Gunning in 1952, is not only one of the older readability measures, but is still used in academic research today. The Fog Index has been used in well over 100 accounting studies, mostly in the highest ranked papers (Efretuei & Hussainey, 2023, p. 322). Its simplicity and straightforward nature make it one of the most widely used readability measures. The result of the Fog Index equation is a simple number that indicates how many years of education a reader of average intelligence needs to comprehend a certain text.

$$Fog = (words\ per\ sentence + complex\ words\ in\ \%) *0.4 \tag{2}$$

The only two variables of *Equation 2* are the average number of words in a sentence within a text and the percentage of complex words that are defined as words with three or more syllables. The sum of these two factors is multiplied by the constant 0.4. The result represents the number of years of education (Li, 2008, p. 225). Depending on the educational system, a Fog Index < 12 indicates that schooling up to high school graduation is sufficient to understand a text, a Fog of up to 15 requires a bachelor's degree, a Fog of up to 17 requires a master's degree, while an even higher Fog indicates that the text is incomprehensible.

The Fog Index is widely used in research and is considered by regulators in the context of corporate disclosure (Efretuei and Hussainey, 2023, p. 319; Loughran and McDonald, 2014, p. 1644). However, it is not the only readability measure. Guay et al. investigate the complexity of financial statements in the context of the voluntariness of reporting elements and measure readability with a merged index, which consists of six individual measures, one of which is the Fog Index. The other measures are the Flesch-Kindcaid Index, the LIX Index, the RIX Index, the ARI Index and the SMOG Index. Each of these measures readability as a function of the average number of words, syllables, or characters in a text using similar calculation methods (Guay et al., 2016, p. 18). As a result, the measures are strongly positive correlated with each other. In their study, Guay et al. measured individual correlations of at least 0.84 to 0.99 (Guay et al., 2016, p. 70).

The Fog Index, and by default all similar indices, has been subject to severe criticism. For one thing, the first component of the Fog Index, the average number of words in a sentence, is less accurate for measuring readability in a business context than it is for general language texts, but it is the second component, which deals with complex words, that has been criticized in its application to financial and accounting texts in particular. Half of the weight of the Fog Index falls on complex words, which are defined only by their number of syllables. However, in a business context, multi-syllable words are quite common, which automatically inflates the result of the Fog Index. In addition, many of these words are not particularly difficult to read, even though they are long. Having a few different long words that make up a high proportion of the absolute complex words also argues against the composition of the Fog Index, because the repetition of these words makes them easy to read. Nevertheless, these words increase the index (Loughran & McDonald, 2014, pp. 1644–1645).

To overcome these problems, Bonsall et al. develop the Bog Index, which considers not only word and sentence length, but also more deeply the language characteristics. The Bog Index consists of three variables:

$$Bog = sentence\ Bog + word\ Bog\ -\ Pep \tag{3}$$

The first variable *Equation 3* is similar to the first variable of the Fog Index. The average sentence length of a text is squared and divided by a fixed long sentence limit of 35. Thus, longer sentences increase the Bog Index, analogous to the Fog Index. The second variable *word Bog* differs significantly. Here, not the number of syllables in a word determine complexity, but several parameters such as word difficulty, abbreviations, wordiness, passive verbs and stylistic issues that are taken into account. A set word list awards penalty points between zero and four to complex words depending on the parameters. These are multiplied by the factor 250 and divided by the total number of words. Thus, words only affect the index if they are actually classified as complex according to the parameters and not just according to the number of syllables. The third variable *Pep* has the special feature that it decreases the Bog Index. The Fog Index does not contain a comparable variable. *Pep* rewards certain characteristics of a text that make it easier to read. These include names, interesting words and personal pronouns. The sum of these words is multiplied by the factor 25 and divided by the total number of words in the text. This means that supporting words that make a text more readable are weighted less heavily than complex words, but they still have an impact on the index. Finally, a variation in sentence length has a positive effect on the *Pep* variable. Alternating short and long sentences improves the reading flow. Higher standard deviation is accounted by an additive factor in the Bog Index (Bonsall et al., 2017, pp. 12–14). The Bog Index is based on the functionality of the software program StyleWriter[9], which among

[9] The exact composition of the Bog Index is published on the website of

Table 2: Readability measures overview

	Readability	**Sentiment analysis**	**Disclosure quantity/ similarity**
Traditional methods	Fog Index Bog Index Similar indices[a]	Word book approach	Vector Space Model
State-of-the-art methods	File Size Machine Learning Key Figure Combination Large Language Models	Traditional Machine Learning (supervised) Deep Learning (Artificial Neural Networks) Large Language Models	Traditional Machine Learning (bag of words) Advanced Machine Learning (word embedding) Large Language Models

[a]Flesch-Kindcaid Index, LIX Index, RIX Index, ARI Index and SMOG Index.

other things is designed to improve text readability. The software works with various applications, for example a list of 200,000 words that are used in the punishment of complex words in the functionality of the Bog Index. The software has also been used in other studies to investigate the readability of texts (Miller, 2010, pp. 2114, 2140).

The Bog Index overcomes some of the weaknesses of the Fog Index, but it also has some shortcomings of its own. For instance, the Bog Index is fundamentally more complex and less easily replicable, mainly because the word list used is not publicly available. In addition, the Bog Index focuses on writing style rather than readability, which are very similar but not identical fields of application (Loughran & McDonald, 2020, pp. 25–26).

In summary, there are several methods to determine the readability of a text. The Fog Index is the most used (Loughran & McDonald, 2014, p. 1645), although other indices can complement and improve the Fog Index. However, these methods suffer from weaknesses, which is why additional methods are discussed in Chapter 4.4.1 of this thesis.

4.2.2. Sentiment analysis

Sentiment analysis has become more prevalent in textual analysis than readability research (Bochkay et al., 2023, p. 797). The majority of sentiment analysis studies use the bag of words method, which represents the information of a text as a vector containing the count of each occurring word (Loughran & McDonald, 2011, pp. 36–37). As explained in Chapter 2.4, this technique results in a substantial reduction of text dimensionality, which simplifies text processing, but also results in a loss of information. To determine the sentiment of a text, the vector is mapped to a word list that assigns a sentiment score to each word (McGurk et al., 2020, p. 463). For example, to determine whether the tone of a text is positive or negative, all positive and negative words in a text are aggregated. Thus, the number of different positive or negative words affects sentiment, as does the frequency of those words. The total number of positive and negative words is then contrasted and an overhang indicates the type and strength of the sentiment.

In this type of study, the vector is derived from the text under investigation and is thus exogenous. The world list used, on the other hand, can influence the result of the analysis. It is in the eye of the observer whether a word expresses a positive, negative or no sentiment at all. Many studies use the Harvard Psychosociological Dictionary or Harvard-IV-4 TagNeg (H4N). This list of words was designed for research in sociology and psychology and contains an extensive division of words into 182 different categories such as *strong, weak, active, pleasure* and *pain*. However, the most commonly used categories are positive and negative words. For example, the study by Tetlock et al. found that a higher number of H4N negative words in the press predicted lower company earnings and that such words also have a significant impact on stock returns (Tetlock et al., 2008, pp. 1464–1465).

The H4N word list is frequently used in sentiment analysis given the fact that it is publicly available and not proprietary (Loughran & McDonald, 2011, p. 38). By utilizing the dictionary, researchers have no control over the sentiment to which each word is assigned. This makes the word list an exogenous variable (Loughran & McDonald, 2011, p. 38). However, the H4N was not intended for use in business contexts, resulting in poor word classification. Loughran and McDonald discovered that almost three quarters of all negative words in the H4N word list are not considered negative in a financial context. They found this by counting negative words in corporate financial statements and running a regression with corporate earnings. For words like *tax, cost, capital, board, liability, foreign*, or *vice* there is no correlation between increasing word count and decreasing returns. In common language, the negative connotation of these words is clear. However, within financial statements, such words are used without negative implications (Loughran & McDonald, 2011, pp. 35–36).

To address this issue, Loughran and McDonald created wordbooks specifically designed for financial and account-

the company behind StyleWriter (https://www.stylewriter-usa.com/stylewriter-editing-readability.php).

ing studies, generated from textual data of over 50,000 financial statements. These are six separate lists covering the following sentiments: *negative words, positive words, uncertainty words, litigious words, strong modal words* and *weak modal words*. Most of these lists were created by including all words that occur in at least 5 % of the sample and are therefore part of the more frequently used language. The correlation between frequency and return was then measured for these words. In addition, Loughran and McDonald considered negations for positive words, so that words were not considered positive if they were preceded by negation words[10]. This consideration was not applied to negative words, since the authors assume that negative negations rarely occur. As a result, the list of negative words consists of 2,337 words, of which 1,121 match the H4N list. A total of 2.5 billion words are classified in the study, of which 3.79 % are classified as negative by H4N and only 1.39 % are classified as negative by Loughran and McDonald's negative word list. Many of the words from the H4N list are either not negative in the financial context (e.g., *claims* or *litigation*) or because they appear in the financial statements of particular industries (e.g., *cancer* in the pharmaceutical industry or *tires* in the automobile industry) (Loughran & McDonald, 2011, pp. 49–50). The appropriateness of Loughran and McDonald's word list for financial statements is demonstrated in *Figure 1*.

The sample is divided into five quintiles according to the number of negative words occurring. It is expected that companies with more negative words on average have a greater negative access return. This relationship is clearly visible in the word list of Loughran and McDonald (Fin-Neg), while the H4N word list (H4N-Inf) does not show the expected monotonically decreasing trend. Further, the companies from the quantile with the most negative words from the H4N word list also show the almost lowest negative excess returns, which is the opposite of what would be expected (Loughran & McDonald, 2011, p. 51).

4.2.3. Disclosure quantity and similarity

The subject of disclosure quantity and similarity has been the focus in a variety of studies, such as analyzing earnings calls (Frankel et al., 2010), identifying analyst value-added for such earning calls (A. Huang et al., 2018), comparing disclosures in general (Gaulin & Peng, 2022; Lang & Stice-Lawrence, 2015), or in specific areas like MD&A (S. Brown & Tucker, 2011). The consideration of these areas may also be relevant to this thesis. For instance, S. Brown and Knechel (2016) have used textual analysis methods to examine certain similarities between companies and their auditors.

Many of these and other studies determine similarity using the Vector Space Model (VSM). This model is also utilized by search engines, for example, to respond to search requests with matching websites. In the VSM, a text is represented as an n-dimensional vector, where n is the number of different words and the value of the vector represents how often each word occurs in the text (S. Brown & Tucker, 2011, p. 315). This procedure is equivalent to the bag of words method, which represents the information of a text in the same way.

Similarity is calculated as the cosine of the angle θ (therefore, the model is also referred to as cosine similarity) between two vectors v_1 and v_2, using *Equation 4* (S. Brown & Tucker, 2011, p. 316):

$$Similarity = cos\theta \;=\; \frac{v_1}{\|v_1\|}\cdot\frac{v_2}{\|v_2\|} = \frac{v_1\cdot v_2}{\|v_1\|\,\|v_2\|} \tag{4}$$

$v_1,\ v_2$ $= vectors$

$\|v_1\|,\ \|v_2\|$ $= vector\ length$

θ $= angle\ between\ vectors$

The vectors and the vector length determine the similarity of two sentences. The vector length is calculated by the square root of the sum of all squared dimensions of a vector. The dot product of the vectors forms the numerator, while the product of the vector lengths forms the denominator. The following example, which consists of two short sentences, illustrates the concept (word counts are listed in Table 3).

> *Sentence 1: Sustainability is as important as company earnings, therefore we focus on sustainability.*
>
> *Sentence 2: Sustainability impacts our earnings and thus sustainability is important for our company.*

$$Similarity_{example} = cos\theta \;=\; \frac{v_1}{4}\cdot\frac{v_2}{3,606} = \frac{8}{14,4224} = 0,5547 \tag{5}$$

Vektor v_1 *= (0,2,1,1,1,0,0,1,1,1,0,2,1,0,1)*

Vektor v_2 *= (1,0,1,1,0,1,1,1,1,0,1,2,0,1,0)*

Vektor length $\|v_1\| = 4$

Vektor length $\|v_2\| = 3{,}6056$

At first glance, the sentences seem to have some similarity. Certain words appear in both sentences, while others appear in only one. In both sentences, the word *sustainability* appears twice. A cosine similarity score of 0.55 is calculated for the sentences. Scores between 0 and 1 are possible, where two vectors of the same direction result in a score of 1 and two orthogonal vectors result in a score of 0 (A. Huang et al., 2018, p. 2853). The result itself is not conclusive for such a small data set and without any reference results. Nevertheless, it is interesting to see how the score is calculated. For example, half of the similarity score can be attributed to the word *sustainability*, since this word increased the dot product of the vectors from 4 to 8.

Cosine similarity also finds other applications such as plagiarism detection (Bochkay et al., 2023, p. 771). In order to

[10] Negation words considered are *no, not, none, neither, never, nobody.*

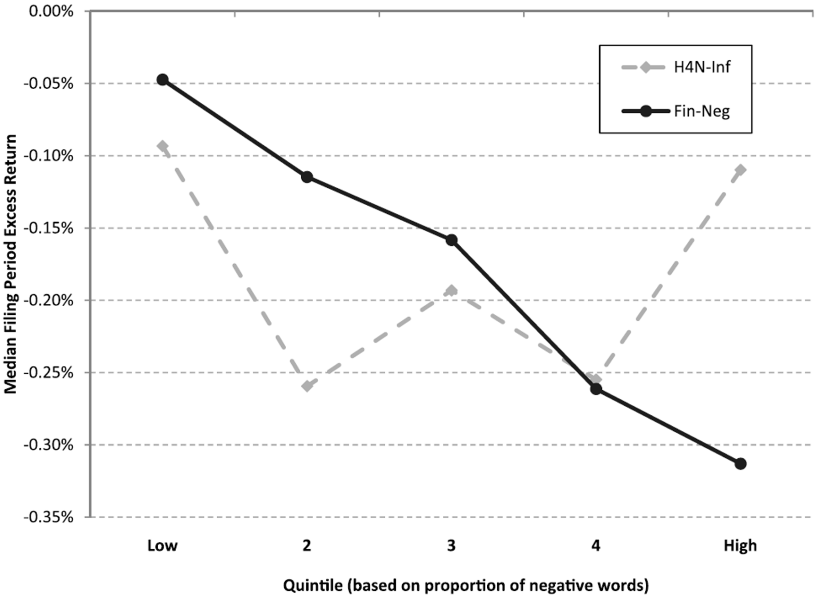

Figure 1: Median excess returns by quintile based on proportion of negative words (Loughran & McDonald, 2011, p. 51)

Table 3: Word count of sentences

word	v_1 count	v_2 count
and	0	1
as	2	0
company	1	1
earnings	1	1
focus	1	0
for	0	1
impacts	0	1
important	1	1
is	1	1
on	1	0
our	0	1
sustainability	2	2
therefore	1	0
thus	0	1
we	1	0
Total	**12**	**12**

open the model to further research and to improve it, several adjustments may be made. A first step is usually stemming, where words are reduced to their stem to lower the dimensionality of the vectors (Porter, 1980, p. 130), as described in Chapter 2.4. Reducing the dimensionality leads to more efficient processing of the data (S. Brown & Knechel, 2016, p. 770). Most importantly, this simplification is also helpful because in the vast majority of contexts, variations of the root word do not change the content. Therefore, it is reasonable that different variations of the same word increase text similarity.

Besides stemming, word weighting is another important customization option. Without any form of weighting, the VSM does not differentiate between words. One possible approach is to underweight common words and overweight less common words, as these are considered to be more relevant to the content. Often, the dimensions of the vectors, which reflect the number of individual words in the text, are multiplied by the logarithm of the function M/m. M represents the number of all texts compared in the sample and m represents the number of texts containing the corresponding word. This reduces the weight of frequently occurring words. Since $log(1) = 0$, words that occur in all texts (M=m) have a weight of zero and do not influence the similarity analysis (S. Brown & Tucker, 2011, pp. 316–317). Some studies even go a step further and not only reduce the weight of common words, but discard them altogether if they occur in a larger fraction of the sample (Hoberg & Phillips, 2010, p. 3782).

Finally, it is crucial to consider the text length. The greater the length of the texts to be compared, the higher the likelihood of common words occurring in both texts, thus increasing their similarity (S. Brown & Tucker, 2011, p. 317). Therefore, longer texts will naturally have a higher similarity than shorter ones. This can be counteracted by integrating a correction variable adapted to the respective study (S. Brown & Tucker, 2011, pp. 343–344) or by normalizing the vectors so that all vectors of a study have the same length (Hoberg & Phillips, 2010, p. 3809). Text length itself can also be used as an indicator of similarity. Brown et al. examine spillover effects in qualitative corporate disclosures. Specifically, they find that companies change their disclosures more if the industry leader, direct competitors, or an industry peer with the same auditor receives comments from regulators, even if their own disclosures were not criticized (S. Brown et al., 2018, pp. 623–625). To get to this finding, Brown et al. compared the absolute change in the number of words in risk disclosures from one year to the previous year, assuming that a change in the number of words also indicates a change in the information contained (S. Brown et al., 2018, pp. 631, 636).

Chapter 4.2 outlined the fundamental practices of traditional textual analysis. The following chapter establishes the methodological principles of NLP and machine learning for a deeper understanding of the subject before a similar presentation of the state-of-the-art methods is provided in Chapter 4.4.

4.3. Development of natural language processing and machine learning

Textual analysis has been utilized in academic research for a long time, but the advantages offered by NLP have brought the field to an advanced stage. Not all state-of-the-art methods make use of NLP, but it forms the basis for many newer methods. Therefore, before listing the state-of-the-art methods in the areas of readability, sentiment analysis, and disclosure quantity and similarity, the framework around NLP and related concepts will be further examined.

NLP offers the possibility to work with text in various ways. In general, NLP models function by taking given textual data as an input, transforming it, and presenting a new output as a result. In this transformation, NLP models differ in the way they operate. A first category includes rule-based models. The transformation of a text in these models is performed by manually developed rules. For instance, these rules may count the number of certain words or text elements. Keywords can be counted to identify specific contents of a text. Next to them, words defined as complex or the number of sentences can be counted. It can also be helpful to count words that have been previously assigned to categories, such as words with positive or negative sentiment (Bochkay et al., 2023, p. 769). These simple transformations are thus structured similar to traditional models like the Fog Index or traditional sentiment analysis methods. More complex are models like the VSM. Here, the input is modified by representing textual data as a vector, which is then transformed in a rule-based manner in a second step, for example to perform a similarity analysis (Bochkay et al., 2023, p. 771). However, NLP is not restricted to these specific applications, but also facilitates innovative applications utilizing machine learning models that exceed the current ones presented.

Machine learning is a process where an algorithm generates a model from pre-existing data. The input data and desired output data are defined, and the model attempts to match the input data to the corresponding output data using defined rules. Once trained on the given training data, the model is expected to be capable of applying this process to new data sets. The training data is typically formed of a small but representative portion of the overall data set to ensure that the machine learning algorithm works well on all data (Zhou, 2021, pp. 3–4). Machine learning is thus classified as artificial intelligence, although, unlike many other artificial intelligence applications, it relies on historical data or training data. From this data, it is possible to identify patterns that can be used for event prediction or classification tasks (Alloghani et al., 2020, pp. 3–4).

There are various machine learning models that are applied in the domain of textual analysis. One of them is the Naive Bayes model. The Naive Bayes model is a probabilistic generative model that calculates outputs based on conditional probabilities. It is mainly used for text classification or topic detection, which is an area of analysis that was not mentioned in the traditional methods, listed in Chapter 4.2, but is still used in text analysis. For example, Brown et al. use topic detection to determine whether it is possible to make assumptions about the probability of fraud cases based on company disclosures. They do not refer to measures such as readability or sentiment analysis, but specifically to the content of the disclosures and whether this can be exploited with the help of machine learning techniques to generate robust probabilities for the existence of fraud (N. Brown et al., 2020, pp. 238–239).

In the Naive Bayes model, similar to the bag of words method, a text is considered as a vector. However, in this model, the text is represented as a binary vector, so the frequency of one and the same word is not relevant. The model now requires texts as training data. From this data, the model is able to determine the probability of certain parameters, in this case words, occurring in a category. The probabilities determined in the training phase can be transferred to the subsequent prediction phase in order to assign uncategorized texts to the category that the model considers most likely. This is the category of texts representing most similar vectors. The Naive Bayes model gets its name because it is founded on the *naive* assumption that each probability of the occurrence of a word is independent from the occurrence of any other word. Thus, the joint probability of several specific words occurring is equal to the product of the probabilities of the individual words, which would not be the case in a real setting (Aggarwal, 2018, pp. 123–125).

Another example of machine learning is the nearest neighbor classifier. This model classifies variables such as text or separate text components into categories, similar to the Naive Bayes model. The major difference is that the nearest neighbor classifier does not require a learning phase, but performs its classification decisions based on the training data only. A variable that is not included in the training data set gets assigned to the category in which an already categorized variable is located, which in the case of textual analysis would be the text from the training data set that was previously categorized and has the smallest deviation, thus representing the nearest neighbor. The deviation is calculated using cosine similarity (Aggarwal, 2018, p. 133).

Text regression is another popular machine learning application in textual analysis. Due to the high dimensionality of textual data, as discussed in Chapter 2.4, common regression methods, such as ordinary least squares regression, are not suitable. The large number of different English words, which represents the number of parameters in a regression, usually even exceeds the number of observations in a sample. Instead, nonlinear regressions can be performed. The so called classification and regression trees model is constructed by iterating a text through all available branches of the decision tree. The text is stripped down to the most informative features, which are the words identified as most relevant. This can be done with the help of specific dictionaries. The features are then grouped and iterated through the decision tree, where the algorithm identifies more features that can help with the categorization. At each level of the tree, the algorithm selects features that best contribute to the separation of categories, resulting in increasingly specific criteria for classification. When a new data point, in this case a text outside of the training data, is added for classification, it travels along the branches based on the presence or absence of certain criteria, in this case specific words, and ends up at the end of a branch that determines the classification. In some models, the data point can also land at multiple ends, which are then weighted (Bochkay et al., 2023, pp. 771–772). Such classification and regression trees can be used in sentiment analysis, for example. The classification and regression tree model enables the detection of correlations in the form of a nonlinear regression by dividing variables into domains in which they are homogeneous. Relationships are not measured along one variable, but rather by discrete features. In the case of textual analysis, these features are words and are evaluated within the context of an association, such as sentiment.

The models presented so far belong to the supervised machine learning group. This means that the models operate in such a way that the training data is already correctly labeled and the models therefore have a blueprint to which they can refer. In contrast, there is unsupervised machine learning, in which case the model has to recognize patterns without classified training data (Alloghani et al., 2020, p. 4).

Topic modeling is one of the unsupervised machine learning models used in textual analysis. The most popular subset is LDA, which was briefly introduced in Chapter 3.3. In this unsupervised machine learning model, the algorithm detects topics within a text by using probabilistic methods to identify words that are related to a topic. The algorithm is therefore used in topic detection within texts, but also for similarity analysis, since texts with similar identified topics are assumed to have a higher similarity (Bochkay et al. 2022: 772).

All of these supervised and unsupervised models are traditional models of machine learning. They offer advantages over some other textual analysis methods, but also show their own shortcomings. For example, traditional machine learning models have trouble recognizing complex contexts in the learning process. This can result in an incorrect or insufficient algorithm that does not generate valuable outputs. Another weakness is that users manually define the investigated features. For example, for readability, the number of characters or syllables are defined as examination variables. Finally, it is necessary to train the models, which requires both time and the availability of a suitable training data set. Deep learning methods can overcome these weaknesses (Bochkay et al. 2022: 772-773).

At the beginning of this chapter, the functionality of NLP models was described, in which input data, such as textual data, is transformed and subsequently represented in a modified form as output data. The models therefore have three distinct layers: an input layer for the input of the data, a second layer in which a transformation process is applied according to the methods of the models described, and a third output layer for the display of the data.

The process of deep learning is comparable, but varies in the middle of the model structure, between the input layer and the output layer. Instead of a straightforward transformation function, a so-called hidden layer is used. This layer, in turn, can consist of several layers that are interconnected (Aggarwal, 2018, p. 326). The hidden layer represents a mathematical function that computes output values from input values. This function itself is composed of several simpler functions, the individual layers. It is the *depth* of the layers within the model that gives deep learning its name. The more layers a model has, the more complex tasks it can solve. Each layer performs a unique function in the overall transformation (Goodfellow et al., 2016, pp. 5, 8). *Figure 2* shows a simplified deep learning model where a specific output parameter is determined from various integer variables. An example use case would be the categorization of textual data. The hidden layer here has a depth of two and can be extended many fold in more complex models.

Such models are also called Artificial Neural Networks (ANN) because they resemble the neural network of the human brain (Bochkay et al., 2023, p. 773). The intersections between layers are connected to form a network through which information passes. Deep learning methods consisting of ANNs find application in the processing of visual material such as images and videos, audio material such as audio tracks, and also In the processing of text and speech. Especially in textual analysis, Deep Learning offers enormous possibilities for the future, by considering words and sentences

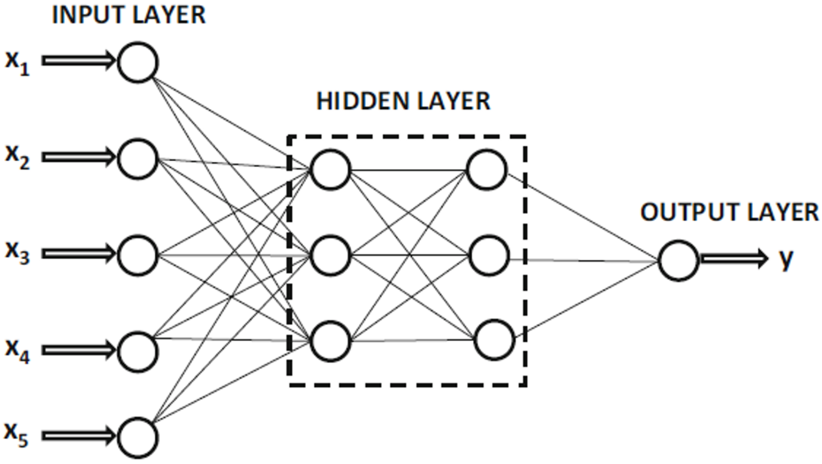

Figure 2: Multilayer neural network example (Aggarwal, 2018, p. 327)

in the context of the whole text (LeCun et al., 2015, pp. 436, 442).

The ANNs can be designed to process textual data as a simple vector. A more reliable use, taking into account the context of individual words, can be achieved by integrating loops into the structure of the ANNs. Inputs and outputs are not considered as single variables (words), but as dependent sequential variables (sentences or text segments). Such models are useful, for example, in translation applications or the in design of an artificial intelligence which is able to answer questions (Aggarwal, 2018, pp. 342–343, 350–351). In addition, there are models that calculate output via so-called *attention*. Here, the similarity of the input to different vector series which contain information is calculated first. The higher the similarity, the higher the assigned weight. The sum of the weighted vectors then provides information about the measure of attention, allowing the model to focus on certain parts of the text depending on which information is important for the given input. Vaswani et al. propose that attention weighting provides better output predictions than using loops in recurrent neural networks (Vaswani et al., 2017, pp. 2, 3, 5, 9).

The attention mechanism is used in state-of-the-art large language models (LLM). LLMs are NLP models that have large neural network architectures and are trained on large sets of textual data. Open AI's generative pre-training (GPT) language model *ChatGPT* is receiving tremendous attention. Immediately after its launch in November 2022, it was the fastest growing consumer application at the time and is now among the top 20 visited websites worldwide with 1.5 billion monthly users (Reuters, 2023, p. 1). Such GPT models are able to perform similarity assessments or to classify texts, but most of all they are known for their ability to answer questions through text generation. The models are built using a combination of unsupervised pre-training and supervised fine-tuning. The unsupervised pre-training is performed by taking a large amount of unlabeled texts as training data and transforming them into outputs using a multi-layer transformer decoder. The multi-layer transformer decoder works according to the attention mechanisms described before, where the transformation takes place over several layers in order to perform a step-by-step refinement of the interpretation of the text, from simple to complex contextual relationships. Finally, the fine-tuning is carried out with labeled textual data for the different tasks of such a model. The combination of unsupervised pre-training using the attention mechanism and supervised learning at the task level helps to bring the performance of GPT models to a new level (Radford et al., 2018, pp. 2–4, 8).

Next to ChatGPT is the Bidirectional Encoder Representations from Transformers (BERT). BERT is an LLM with versatile application possibilities. Like ChatGPT, BERT is built in two steps. The first step of pre-training is done using a multi-layer bidirectional transformer. The main difference between the models is the *direction* in which they generate text. In ChatGPT, the text is generated token[11] by token or word by word from left to right, as a human would read it. BERT uses the Masked Language Model (MLM) instead. In the MLM, random tokens or words are *masked* in pre-training, making them unrecognizable to the model, so that they can be predicted based on the surrounding words. The advantage of MLM is that texts are not only generated from left to right so that only the preceding words in the pre-training influence the predictions, but also the following words. The second step of fine-tuning is performed as in the GPT models, using labeled textual data that is required for the particular application. The pre-trained version of BERT is designed so that fine-tuning can be performed by adding a single additional layer to the ANN, allowing it to be built on top of the base model with minimal effort (Devlin et al., 2019, pp. 4171–4174).

Both models are pre-trained with a very large amount of data and thus can be referred to as LLMs. There is no

[11] Tokens are sequences of letters that are somewhat similar to words. On average, 100 tokens correspond to approximately 75 English-language words (https://platform.openai.com/tokenizer).

clear boundary at which NLP models are considered as a LLM. Chelba et al. have designed a benchmark for language modeling that contains one billion words (Chelba et al., 2013, pp. 1–5). When using such a huge benchmark, it is reasonable to describe a model as an LLM. ChatGPT uses the BooksCorpus[12] database for pre-training, which contains about one billion words, comparable to the Chelba et al. benchmark. BERT is pre-trained with the BooksCorpus database as well as with English Wikipedia and thus has access to well over three billion words. Both models do not use the Chelba et al. benchmark because it provides a shuffled sentence-level corpus. The sentences and words within the database have been randomly distributed, so that the models that access the benchmark only consider the actual relationship between the words, and not the natural order of those words (Chelba et al., 2013, p. 2). However, this order is fundamentally relevant in the case of ChatGPT and BERT. Throughout the use of BookCorpus and Wikipedia, it is possible for the models to investigate long contiguous sequences (Radford et al., 2018, pp. 4, 5; Devlin et al., 2019, p. 4175).

4.4. State-of-the-art methods

4.4.1. Readability

The previous chapter introduced the principles of NLP models. First, rule-based models were briefly presented, followed by machine learning models, including supervised and unsupervised models. Finally, ANNs were discussed and it was shown how they are applied in popular LLMs. Based on this foundation, the state-of-the-art models in the areas of readability, sentiment analysis, and disclosure quantity and similarity are presented now to complete the detailed overview of textual analysis methods. A comparison with the traditional models as well as an in-depth methodological insight then allows to identify the methods suitable for the research questions of this thesis.

In the area of readability, the Fog Index has been very popular. Weaknesses such as insufficient applicability to business texts or the classification of words as complex, when they actually tend to be easily readable, have been identified and countered by alternative measures such as the Bog index.

Loughran and McDonald find another measure that outperforms traditional readability indices: the 10-K document file size. The authors look at the file size of 10-Ks filed with the SEC. These files are highly standardized and presented in HTML format. Loughran and McDonald find that larger file sizes are associated with lower readability. The results are both strongly correlated with traditional readability measures and consistent with Loughran and McDonald's definition of readability that higher readability leads to less ambiguity in valuation, which the authors demonstrate in their research.

File size as a readability measure is as simple as one can imagine and requires very little adjustment before use, making it less prone to errors than other measures. However, it is questionable whether this measure can be applied to other texts such as sustainability reports. Given that readability is a measure of the ease with which value-relevant information can be extracted, file size works mainly on the simplistic premise that a higher quantity of textual data makes it more difficult to extract relevant information (Loughran & McDonald, 2014, pp. 1646, 1650, 1658–1658, 1667–1668).

Sustainability reports are very heterogeneous. The large variation in report length could yield significant results in the readability analysis, but it is debatable whether a longer sustainability report increases the difficulty of extracting value-relevant information or whether it just contains more information. It is also possible for the same information to be presented differently in two reports, one with a brief version and one with a more detailed version. While the longer version may be easier for the reader to understand, it would reduce readability in this model due to the increased file size (Bochkay et al., 2023, p. 779).

The approach of Loughran and McDonald is a method that does not take advantage of the developments in the field of NLP and machine learning that are described in Chapter 4.3. However, other state-of-the-art readability methods increasingly rely on them. Machine learning can be utilized to build an accurate model from available components that best fit a particular research question. *Table 4* lists different features that can be used in readability analysis, divided into categories.

Shallow features, such as the length of words or sentences, the ratio of simple words, or the ratio of different terms, are used in traditional models and form the basis of readability analysis. Morphological features capture how words are used in relation to their stem words, and thus can capture complexity and grammatical features. These features get lost in many traditional models because stemming is used to reduce the dimensionality of such words back to their root. Syntactic features capture the frequency of certain word combinations or sentence structure, which also often get lost because of the use of the bag of words approach. Semantic features evaluate readability at the content level, going beyond other methods. For example, the use of many synonyms decreases readability, as a low reading level audience is more likely to require a simplified vocabulary. Another semantic feature is cohesion, which uses similarities to measure how well sentences merge into each other. An easy transition through a higher similarity of the last words of one sentence to the first words of another sentence increases the readability of a text (Madrazo & Pera, 2020, pp. 4–6).

Once the analysis methods have been grouped, machine learning can be applied to determine the most effective methods for a particular use case. This can be done by training the model with only one category of analysis methods at a time and then comparing the results. Such a comparison can show, for example, that the shallow features are more accurate than the morphological features, or that the semantic

[12] BookCorpus is a database of novel books written by unpublished authors and contains over 11,000 books in genres like romance, history, and adventure (BookCorpus, 2023).

Table 4: Textual features (for definitions of terms see Madrazo and Pera (2020, pp. 4–6)

Shallow Features	Morphological Features	Syntactic Features	Semantic Features
Word length Sentence length Ratio of simple terms Ratio of different terms	Inflection ratio Morphological phenomena frequencies	Part of speech ratio Dependency tree complexity	Synonym usage Semantic closeness Cohesion

features are less accurate than any other group of features. Furthermore, it can also be determined which features within the categories have the greatest influence, so that, for example, the outcome of the shallow features predominantly depends on the result of the individual features word length and sentence length (Madrazo & Pera, 2020, pp. 4–6). Accordingly, the machine learning process does not consist of a single process in which the textual data passes through a very large number of layers. Instead, it is divided into several sub processes, each with a fewer amount of layers. On the one hand, ineffective criteria can be filtered out to reduce the computational effort compared to an all-encompassing process. At the same time, the accuracy of the analysis can be improved by eliminating features that take the wrong path in the learning process and lead to uninformative results.

Another way to apply machine learning in readability analysis is to compare not only criteria, but also entire models. Such comparisons would not be feasible without the automation possibilities offered by machine learning due to the high engineering costs. Comparing models works similarly to comparing analysis features. All models are fed the same textual data to ensure comparability, the results are compared with each other in terms of the output or classification accuracy, and only the best model is used in actual research beyond the training data set. To go one step further, the models can then be combined with each other. For example, it may be found that one model, e.g. BERT, has a word prediction accuracy of 90 %, and another model, e.g. GPT, has an accuracy of 85 %, but using both models in combination achieves an even higher accuracy than either model separately, because the strengths of one model compensate for the weaknesses of the other. By having multiple models work together and combining their predictions, an overall more robust and powerful result can be achieved. Combining models can be performed through different approaches, such as categorizing the input text into the category predicted by the majority of models (if more than two models are combined), or weighting the results of the models according to their individual accuracy (Filighera et al., 2019, pp. 335–336, 338–340, 344–345).

Readability models are being criticized for not being transferable between different types of text (Bochkay et al., 2023, p. 780). It is not worthwhile evaluating metrics that are of little importance to a text's target audience (Loughran & McDonald, 2020, p. 28). Schoolbook texts should be accessible to students and adapted to their experience and reading ability. However, financial statements or earning call transcripts are not likely to be read by lower-level students, but by more experienced readers who can be expected to comprehend a certain level of complexity.

Martinc et al. show that supervised and unsupervised NLP models are able to assess readability across different audiences. Using methods similar to those of Madrazo et al. and Filighera et al., they prove this by using diverse training data sets. Martinc et al. use text sources, such as educational materials, which are classified by reading ability, age group, or grade, but also large databases, such as Wikipedia. These databases are classified into different readability levels, such as simple, balanced, or normal. Readability can then be measured by an adjustable score that takes into account the reading skills of the respective audience (Martinc et al., 2021, pp. 241, 152, 166–169, 172–175).

Technological advances in NLP and machine learning allow for multifaceted readability research, partly through a more efficient evaluation of traditional metrics and partly through new developments. Measuring readability can provide valuable information to companies. It allows them to assess their qualitative disclosures to determine if those disclosures are comprehensible for their stakeholders. At the same time, legislators, internal and external regulators and other readers of disclosures can benefit from the results of such measures.

However, the level of readability must be relevant to the research question. Otherwise, it represents nothing more than an indicator without any particular meaning, which at worst is a reflection of the complexity of the company (Loughran & McDonald, 2020, p. 28).

4.4.2. Sentiment analysis

Various applications of sentiment analysis were introduced in Chapter 4.2.2. Traditional methods measure sentiment by counting the number of words in a text, which are classified into sentiment categories using dictionaries. The greatest potential for improvement in these methods lies in the improvement of these dictionaries. The creation of a dictionary specifically for the finance and accounting domain, instead of the commonly used H4N, has had a major impact on sentiment analysis research (Loughran & McDonald, 2011, pp. 61–62). The application of machine learning to the field is likely to be even more significant. Machine learning methods eliminate the need for dictionaries to assess sentiment. Instead, the classification of texts or text segments is performed by an algorithm that is trained on data samples to detect the tone of a text. This approach promises a more accurate classification than methods that rely on dictionaries

(Hartmann et al., 2023, pp. 76, 78).

Traditional machine learning works in the same way as supervised learning described in Chapter 4.3. The algorithm receives texts as training data that are pre-labeled with the corresponding sentiment. This approach is used, for example, by Azimi and Agrawal to extract information from the sentiment of 10-Ks. They find that both positive and negative sentiment can predict abnormal returns. Chen et al. already had similar findings in 2014 when analyzing SeekingAlpha articles (see Chapter 2.4), but Azimi and Agrawal's results differ in that they look at a different dataset with 10-Ks and that they analyze a much larger sample with over 200 million sentences. At the same time, their results are also significant for positive sentiment, in contrast to those of Chen et al. who could not find significant results for other sentiments besides negative tone (Azimi and Agrawal, 2021, pp. 2, 10, 20–21, 32; H. Chen et al., 2014, p. 1337). This might be because the machine learning approach is capturing relationships that are not apparent through the wordbook approach, but there also may be other reasons for this.

Machine learning has a significant advantage over dictionary approaches, as it is capable of consistently capturing sentiment over multiple periods. This is possible due to the volume and timeliness of the training data. Dictionaries can only capture a status quo and may have a lack of actuality.[13] Therefore, state-of-the-art methods are often preferable to dictionary methods. Nevertheless, traditional dictionary methods can be used if the temporal context does not matter and it is only the occurrence of individual words that is crucial for the research question, or if the cost of the more complex implementation of machine learning exceeds the benefit of more accurate classification (Frankel et al., 2022, pp. 5515, 5522–5524, 5529).

Even more accurate than traditional machine learning methods is the application of contextual deep learning to sentiment analysis. Although classical machine learning outperforms the dictionary approach, these methods, such as VSM or Naive Bayes, represent texts as bag of words and are thus subject to the problems described earlier. Algorithms that capture contextual information from word embedding, as in ANNs, can also capture the surrounding context and associated sentiment (Heitermann et al. 2023: 79). This is facilitated through the attention mechanism discussed in Chapter 4.3.

Sentiment analysis is also being used in areas other than financial and accounting, such as economics, political science, and medical research. A multidisciplinary study by Colón-Ruiz and Segura-Bedmar finds that LLMs have the highest accuracy for sentiment analysis. While traditional machine learning models, such as VSM, perform well especially with a large amount of training data, LLMs dominate the field, in particular the BERT algorithm. It delivers slightly better results than competing models, but at the expense of higher computational costs (Colón-Ruiz & Segura-Bedmar, 2020, pp. 1, 5–6, 9–10).

Since BERT is a pioneer in the field of LLMs, the model will now be the subject of a more detailed discussion. BERT is a pre-trained model, simplifying its use for users by eliminating the need to navigate through the underlying complexity. In order for BERT to be able to precisely adapt the analysis to research questions, only the fine-tuning of the model has to be performed. Here, BERT can achieve better results than traditional approaches even with only a few hundred training samples (Siano & Wysocki, 2021, pp. 6, 27–28). BERT is pre-trained according to MLM, enabling the model to predict missing *masked* words or to predict subsequent sentences. BERT is publicly available at no cost. While the algorithm requires high computational power, Google allows free use of online graphics processing units to operate BERT, so the model has few barriers for usage (Siano & Wysocki, 2021, pp. 7, 17, 22).

In contrast to dictionary models, the operation of LLMs is more complex and difficult to comprehend. However, it is feasible to verify that such models actually capture sentiment from the context of information in a text, as they are intended to do, by deleting or changing words in manual tests. Siano and Wysocki have performed such tests and found that BERT still performs better than traditional models even when key words that would have influenced sentiment in the wordbook approach are deleted. Although the accuracy decreases, this evaluation indicates that BERT generates its predictions based more on the context of a text than on the word count, as in the case of wordbook approaches. Furthermore, BERT loses much of its predictive power when words in a text get randomized, which again suggests that the model delivers on its promise and, unlike bag of words models, extracts information from the structural organization of a text (Siano & Wysocki, 2021, pp. 20, 25–26).

In addition to its many advantages, BERT also has some limitations. The biggest one is probably the limitation of tokens. Currently, texts with a maximum of 512 tokens can be analyzed. A token usually corresponds to a word or, depending on the tokenization, to only a fraction of a word. Therefore, lengthy texts, which would certainly include sustainability reports, cannot be analyzed as easily with BERT. Researchers can apply workarounds by selectively or randomly analyzing individual text components, or by analyzing each text component one at a time in a scrolling pattern. However, advances in machine learning and general technological progress offer hope that computational power will increase and these limitations will fade (Siano & Wysocki, 2021, pp. 9–10, 30).

BERT has been used and developed by researchers in various fields. One of the most important developments is FinBERT, a fine-tuned version of BERT specialized for the financial domain. Loughran and McDonald have already identified that the financial domain language differs significantly from general language and have revolutionized textual analysis in this field with their own dictionary (Loughran & McDonald, 2011, pp. 49–50). Huang et al. follow this example

[13] An example of this are the results of Long et al. (2023) presented in chapter 3.2, which could only be achieved by the authors creating a new dictionary adapted to time and context.

by adapting the new state-of-the-art to the financial domain. They do this by pre-training BERT with a large number of texts with financial context like corporate disclosures, financial analyst reports and earnings conference call transcripts. These texts help FinBERT to better process tasks related to financial information. In total, 4.9 billion tokens are used for fine-tuning, which even exceeds the population of the pre-training for the plain BERT version. FinBERT has been compared to other LLMs as well as traditional text analysis methods and outperforms them, as well as untrained BERT, when applied to finance-specific texts, but also when applied to texts related to the Environmental, Social and Governance (ESG) domain (A. Huang et al., 2022, pp. 8–9, 19).

Sentiment analysis has benefited from machine learning and NLP developments, which have led to new techniques and models that can significantly improve the accuracy in this field. The third main area of textual analysis considered in this thesis, disclosure quantity and similarity, also benefits from these developments.

4.4.3. Disclosure quantity and similarity

While traditional methods determine similarity, using the bag of words approach, researchers are now increasingly utilizing machine learning techniques to determine similarities between texts. In traditional methods, similarity is mainly assessed by overlap in word usage. Matching words or tokens in two texts increase the similarity score. Instead of single words, sequences of words can also be considered. Thus, the similarity score increases when word sequences, usually consisting of two to four words, appear in the texts to be compared. These traditional techniques can be further adapted, for example by applying frequency weighting. Here, less frequent words are given a higher weight under the assumption that they contain more information. The similarity between documents increases, especially when rare words or word sequences overlap (Gaulin & Peng, 2022, pp. 2–3, 12–13).

With the help of deep learning, word embedding algorithms are able to recognize similarity in texts without being limited to the occurrence of individual words or word sequences. For this purpose, the algorithm uses a sophisticated method in which it scans the text for predefined words and builds vectors from the surrounding words that are *near* the searched word. This *nearness* can be defined by a certain distance, e.g. up to ten words before or after the target word. These words are context words and are used to capture the relationship of the target word to its environment. The vectors of these context words are placed in the same vector space as the searched word, so that both grammatical and semantic relations between words can be captured. This allows for a deeper and more nuanced representation of textual content (Gaulin & Peng, 2022, p. 34).

When used in combination with cosine similarity (described in Chapter 4.1.3), word embedding algorithms are a powerful tool for accurately measuring disclosure similarity. Traditional methods based on the bag of words approach can still provide decent results if the research question is primarily based on word choice and less on the context of the texts (Bochkay et al., 2023, p. 781). In summary, however, this area of textual analysis also benefits from the new technical possibilities offered by NLP.

The literature review of the finance and accounting domain in Chapter 3 and the comprehensive presentation of traditional and state-of-the-art methods in Chapter 4 form the first main part of this thesis. The insights obtained from this study are significant for the following second main part of the thesis. This section will present a textual analysis application to a case in the accounting domain. This case is covered in the following chapters.

5. Hypotheses development

5.1. The relationship between auditing and compliance with regulatory requirements

At the outset of this thesis, it was noted that sustainability reporting varies greatly from company to company. Furthermore, according to the NFRD, EU member states still have the opportunity to opt out of mandatory external audits for sustainability reporting. In addition to heterogeneity in terms of content and structure, the audit of sustainability reports is another criterion for differences in some EU countries, as many companies voluntarily have their reports externally audited with limited assurance, some even with the higher assurance level of reasonable assurance.

To address the issue in more detail, the fundamental theories in the areas of sustainability reporting and auditing were examined first, followed by an extensive discussion of the textual analysis methodology. This included a summary of the principles of the methodology and a literature review in the financial and accounting domain. The fundamental theories considered in the area of auditing have been narrowed to the essential principles and have further emphasized the effects of the audit on the reliability of disclosures.

Literature indicates that auditors can also serve as intermediaries to support company compliance with regulations. This is because they possess comprehensive knowledge through their activities and their diverse structure (Walker, 2014, pp. 214–215). Therefore, the role of the auditor can not only increase regulatory compliance, but also increase its effectiveness in general (King, 2007, p. 213).

In qualitative research, Walker found that companies in the Australian trucking sector that participated in a voluntary compliance program achieved better performance and generated higher social value if they involved auditors in this process (Walker, 2014, pp. 215–216, 221). This example is not directly related to the thesis content wise, but the underlying structural relation is the same. Similar as in the Australian trucking sector case, European companies have the option to involve external auditors in a process voluntarily. Submitting reporting components from the non-financial reporting to an audit imposes additional auditing fees for companies, but can also provide benefits such as potentially increasing the reliability of the reporting in accordance to attribution theory or achieving a higher level of compliance with legal reporting requirements, which is desirable for both the company and its stakeholders. From the related example and the theories presented, the following hypothesis can be stated regarding the impact of an external audit on the quality of sustainability reporting:

> *H1: Audit assurance for sustainability reports increases their compliance with regulatory requirements.*

In Chapter 6.2, this is addressed by a more detailed examination of the requirements of the EU taxonomy as well as currently applicable and forthcoming auditing standards and the determination of the relevant dependent variables.

5.2. The relationship between the extend of corporate sustainability and reporting quantity and audit demand

In addition to the influence of the audit on the quality of sustainability reporting, there may be other factors that influence both reporting and the circumstances whether an external audit takes place. Chapter 2.2 discussed the major theories that determine whether and to what extent companies are willing to engage in voluntary reporting. The voluntary disclosure theory suggests that companies will only voluntarily disclose information if their benefits outweigh their costs. There is already much evidence within this theory. For instance, research has shown that firm characteristics significantly drive corporate disclosure. Firm size, for example, has a generally positive impact on disclosure as reporting expertise usually increases with growing firm size, but also because larger firms are subject to greater public exposure and have to legitimate themselves to a greater extent. Other firm characteristics, such as the degree of internationalization, board size, or media exposure, also affect voluntary corporate disclosure (Zamil et al., 2023, pp. 247, 249, 252).

Corporate sustainability is a comparatively less studied driver in this context. However, it is possible that sustainable companies report on their sustainable activities overproportionally in order to benefit from it. To address this research gap, the following hypothesis is posed:

> *H2: Companies that are acting sustainable disclose a higher quantity of information in their sustainability reports.*

The degree of sustainability of companies' actions here is defined in a simple manner using existing sustainability ratings. The quantity is measured using the file size of sustainability reports in an adjusted form, based on the methodology of Loughran and McDonald (2014). The detailed research design and and use of the variables is presented in section 6.3.2.

Reporting requirements have an undeniable influence on this as well. When regulatory bodies impose mandatory disclosures, these disclosures are more likely to be made (Duran & Rodrigo, 2018, p. 14). The implementation of voluntary requirements, such as the GRI, can be a significantly driver in the reporting landscape as well (Dissanayake et al., 2019, pp. 102–103).

A less studied influence is the impact of auditing. As voluntary disclosure is performed only if a company's benefits outweigh the related costs, this theory could also be applied to voluntary audits of disclosures. According to attribution theory, positive information is more likely to be doubted than negative information. Therefore, it would be more reasonable for companies to undergo a voluntary audit if the information contained in their disclosure is predominantly positive, which leads to the following final hypothesis:

> *H3: Companies that are acting sustainable are more likely to demand voluntary assurance of their reports.*

This hypothesis expands on the voluntary disclosure theory by shifting the focus from disclosure itself to the voluntary submission of voluntary disclosures as well as non-voluntary disclosures to an external audit.

H1 represents the central hypothesis of this thesis. The secondary hypotheses H2 and H3 are indirectly related to it and can provide further insights into the research area of sustainability reporting. However, they will only be addressed to a more limited extent.

The next chapter first describes the data gathered to address the hypotheses. This is followed by an exposition of the underlying research design. Next, the focus shifts to the determination of the dependent variables in regard to the hypotheses under consideration. Finally, the results are discussed.

6. Data and research design

6.1. Sample data

The sustainability reports of a subset of companies required to report under the NFRD, which are large companies within the EU with an average number of at least 500 employees (European Union (EU), 2014, p. 4), are now examined in order to investigate the three hypotheses of this thesis. In total, the requirements of the NFRD affect approximately 10,000 companies. The CSRD will extend the scope of application by including medium-sized companies to approximately 50,000 companies (KPMG, 2022, p. 37), beginning from financial year 2024. Also crucial for the verification of the hypotheses is the distinction that an audit with limited assurance is mandatory under the CSRD, whereas under the NFRD there is still an option at EU member state level to exempt companies from this requirement. Due to time constraints, this thesis does not examine all companies affected by the NFRD, but only a subsample, which consists of all companies listed in the German Stock Index (DAX) and the Midcap DAX (MDAX). This subsample contains 90 companies, which corresponds to about one percent of the overall affected companies, so that the results may not be unconditional replicable at the EU level. Germany was chosen as the country of analysis, as it is the country with the most

G250 companies within the EU[14] (KPMG, 2022, p. 75) and could therefore take on a pioneering role in reporting issues. Furthermore, Germany has exercised its right to opt out of the mandatory audit of sustainability reporting. Only in this way is it possible to verify the hypotheses presented. The assurance rate in the sample is 77 %, which is slightly higher than the G250 overall (63 %) (KMPG 2022: 24). *Appendix B* presents additional descriptive statistics on the sample data.

The dependent variables are gathered with the help of textual analysis methods outlined earlier in this thesis. Chapter 6.2 discusses the research design and the associated alignment of the dependent variables, while the actual gathering of the variables is described in Chapter 6.3.

6.2. Research design

To analyze H1, it is first necessary to define how compliance with regulatory requirements can be assessed. It is then important to determine which of the textual analysis methods, presented in Chapter 4, are appropriate for the assessment.

There is no clearly defined benchmark for reporting requirements compliance. This is because the reporting landscape itself is complex, ambiguous and sometimes even contradictory. Interregional standards such as the GRI Reporting Framework are opposed to the first drafts of sustainability standards from the International Sustainability Standards Board, supplemented by national regulations within the individual countries. The EU's attempt at harmonization further adds to the complexity. The requirements of the NFRD could not provide the desired effects. The inadequate specification of the directive has resulted in a lack of information within the reported data. The options of the EU member states, such as the requirement for an audit, but also the disclosure options that allow to disclosure sustainability reports within or outside the management report, make it difficult to compare information between companies. The recently enacted EU taxonomy regulation imposes further requirements on companies (Velte, 2023, pp. 1–2).

Reporting quality cannot only be derived from the regulatory requirements themselves. The relevant auditing standards may also be informative. Auditing standards extensively discuss regulatory requirements and provide guidance to auditors on how they can perform audit procedures.

While there are numerous auditing standards covering a variety of areas in financial reporting, the ISAE 3000 (Revised) in particular provides comprehensive coverage of the subject of non-financial reporting. The majority of companies in the sample refer to the standard in various places within their sustainability reports. The ISAE 3000 (Revised) explicitly covers all assurance engagements that do not include historical financial information, which also goes for the non-financial reporting, and has the objective of providing reasonable or limited assurance on that information (International Auditing and Assurance Standards Board, 2013, pp. 5–6). The auditing standard is extensive. It describes the requirements for complying with the standard, including areas such as audit planning, the determination of materiality and the required content of the auditor's report. Because the ISAE 3000 (Revised) covers such a wide range of topics, it does not provide many specific audit guidelines or requirements in terms of the content of the auditor's report. However, it does give some guidance for determining when reporting can be considered compliant. First, in its objectives, the standard states that limited or reasonable assurance can be obtained when the subject matter information is free of material misstatement (International Auditing and Assurance Standards Board, 2013, p. 6), as is the case with other auditing standards. It also defines the mandatory characteristics of *relevance, completeness, reliability, neutrality* and *understandability* for published information (International Auditing and Assurance Standards Board, 2013, p. 12). Using these characteristics as evaluation criteria, compliance can be more specifically defined. In addition, the ISAE 3000 (Revised) states that inconsistencies indicate material misstatements (International Auditing and Assurance Standards Board, 2013, p. 20), so inconsistent information within sustainability reporting or inconsistencies between financial and non-financial reporting may also indicate a lower level of compliance.

Despite its naming, the ISAE 3000 (Revised) is in need of improvement. When it came into force a decade ago, the area of non-financial reporting covered by it was much smaller and less complex. Furthermore, the importance of this information has increased dramatically over the years. As a result, new auditing standards are being developed, that will eventually replace the ISAE 3000 (Revised). For the German market, the Institute of Public Auditors (Institut der Wirtschaftsprüfer: IDW) has published two drafts for new auditing standards. These drafts address the substantive audit of non-financial reporting with reasonable assurance and limited assurance, respectively.

The drafts are based on the ISAE 3000 (Revised), but are subject to considerable uncertainties of interpretation and therefore may not be used by auditing firms for current audits, also due to their status as drafts and not as finalized auditing standards (IDW Verlag, 2022b, p. 1; IDW Verlag, 2022a, p. 1). The drafts do, however, reveal a certain direction in which the audit procedures for ensuring the quality of sustainability reporting are being intensified and on what they are based. For example, the drafts IDW EPS 990 and IDW EPS 991 refer to the requirements of the EU taxonomy in many places, starting with the scope of application of the future standards to companies included within the EU taxonomy (IDW Verlag, 2022a, p. 4) to the performance of audit procedures according to the information categories of the EU taxonomy (IDW Verlag, 2022a, pp. 26, 29). Furthermore, the drafts explicitly state that the absence of information required by the EU taxonomy is generally to be considered as a

[14] In Germany there are 13 G250 companies. There are also 13 G250 companies in France, although France has not exercised the option for EU member states to be exempt from the audit, and therefore sustainability reports of companies that meet the size criteria are required to undergo an external audit (Reuters, 2021, p. 2).

material misstatement (IDW Verlag, 2022b, p. 31; IDW Verlag, 2022a, p. 30). Other sections focus on the assessment of the process for the identification of taxonomy-eligible economic activities.

Measuring regulatory compliance is challenging, as there are many requirements from different regulatory bodies. The requirements of the NFRD are currently in force but are almost obsolete. The CSRD, which is supposed to replace the NFRD, has not yet come into force. GRI standards exist in parallel and the IFRS Foundation is working on separate new standards. As far as auditing standards are concerned, companies mostly refer to ISAE 3000 (Revised), which is effective but also somewhat outdated. New auditing standards are still being implemented. The EU taxonomy, on the other hand, differs from other regulatory requirements. Its formally adoption in 2021 is relatively recent while it is also already effective for reporting of the recent financial year, 2022 (European Union (EU), 2020, p. 18). The frequent reference of the IDW in new auditing standards underlines the relevance. Due to these factors, this thesis employs the EU taxonomy requirements as a benchmark for regulatory compliance in general.

The EU taxonomy has been applied on a mandatory basis for the second time in the last fiscal year of 2022. In this year, non-financial companies were required to report on eligibility and alignment of their activities for the first two of the six taxonomy objectives. At the same time, financial companies were only required to report on the eligibility of their activities, but not on their alignment. The reporting requirements will gradually increase until the financial year 2025, at which point companies will be expected to report fully on all six environmental targets. This reporting includes the identification of eligible activities, an assessment of whether these activities contribute to at least one of the six objectives while not harming any other objective, and the compliance with the minimum safeguards set of the taxonomy. This ensures a consistent identification of activities to be considered sustainable for the purpose of determining the relevant indicators (PricewaterhouseCoopers, 2023, p. 10).

The EU taxonomy demands, on the one hand, information on the proportion of a company's turnover as well as its investment and operating expenditure, which can be classified as sustainable according to the taxonomy (PricewaterhouseCoopers, 2023, p. 23; European Union (EU), 2020, p. 17). Disclosing these metrics provides insight into the current contribution to environmental goals as well as projecting future contributions. On the other hand, qualitative information must also be provided explicitly. Both the computation logic and the key elements of the indicators need to be disclosed. This qualitative information highlights the transition process from taxonomy-eligible activities to taxonomy-aligned activities (European Commission, 2022, pp. 7–8, European Commission, 2021, p. 4).

The definition of regulatory compliance proves to be difficult due to the many different regulatory bodies involved, although the requirements of the EU taxonomy were identified as a suitable quality characteristic as they are in use today and not going to be superseded by new regulations in the near future. The mandatory requirements for the first two objectives of the taxonomy, *climate change mitigation* and *climate change adaptation*, are appropriate for the analysis of this thesis, since in addition to the key figures, qualitative information is explicitly required, which can be evaluated with the help of textual analysis methods. However, it should be noted that these do not represent an exhaustive quality feature of sustainability reporting. Other approaches to measuring the quality of sustainability reporting are also conceivable.

In order to analyze H1, it is crucial to identify not only the contents to be considered, but also the method most suitable. The detailed presentation of known methods in Chapter 4 serves this purpose. These methods can be divided into three main categories: *readability, sentiment analysis* and *disclosure quantity and similarity*. Within these main categories, the compatible individual methods can then be determined on the basis of the correspondence between the objectives of the method and the research question, as well as on the basis of restrictions, e.g. due to lack of time, computing power or other limited resources.

The measure used to assess the quality of non-financial reporting is the extent to which the relevant sections of the reporting comply with the requirements of the EU taxonomy. To assess those extracts, not only their content but also their characteristics are evaluated, more precisely the characteristics that are also listed in the currently relevant auditing standard ISAE 3000 (Revised) and which are also relevant for various other contents in financial and non-financial reporting: *relevance, completeness, reliability, neutrality* and *understandability* (International Auditing and Assurance Standards Board, 2013, p. 12). The fulfillment of these characteristics indicates reporting quality. Trying to link the characteristics with textual analysis methods (overview in *Table 2*), *understandability* can intuitively be covered by readability measures. A text that is easily readable may not always be understandable. Still, high readability facilitates the reader's comprehension, while poor readability makes reporting more difficult to understand. Next, *neutrality* can be assessed by various methods of sentiment analysis by examining whether the extracts show certain sentiments, such as positively formulated language, which indicates a lack of neutrality. Thus, *understandability* can be analyzed quite well with readability measures and *neutrality* can be analyzed with sentiment measures. *Relevance, completeness* and *reliability* tend to be less intuitive. It is reasonable to argue for analysis methods from the group of disclosure quantity and similarity to compare the extracts with the requirements of the EU taxonomy, but such an approach is likely to be less precise than the assessment of the characteristics of *understandability* and *neutrality*, where the methods correspond to the research problem more well.

To overcome this problem, the analysis in this thesis is carried out using an exhaustive method through the application of an LLM. In *Table 2*, it can be seen that LLMs are among the state-of-the-art methods covering all methodological areas. The developments in NLP and machine learning make

LLMs appealing for research. Especially text-generating LLMs have become very popular due to their low barriers to application and their text comprehension ability (de Kok, 2023, p. 2; Kim et al., 2023, p. 6). Chapter 4.3 describes the most relevant types of models, namely GPT models and BERT models. Both types are alike in that their neural networks consist of a large number of parameters, which allows them to have a human-like *world understanding*. Moreover, both model types operate in the same way, taking natural language as an input and delivering their output in natural language as well (de Kok, 2023, p. 5). A major difference, however, is that BERT models require fine-tuning via training data for most applications, which is not necessarily the case for GPT models. The time constraints of this thesis are the reason why the analysis is performed with a GPT model rather than a BERT model. Another factor is the context window, which for BERT is only 512 tokens. The size of the sustainability reports requires much larger context windows. Current GPT models can process the context of 4,000 to more than 32,000 tokens at once[15], making them more suitable for the analysis of sustainability reports, although as the number of tokens increases, the processing requirements and time required increases correspondingly (de Kok, 2023, pp. 7–8; Kim et al., 2023, p. 7).

Even though GPT models are relatively new, they have already had some use in accounting research. As an example, Bai et al. use ChatGPT to examine the information content of earning calls. They compare responses given by company representatives to responses given by LLMs, including ChatGPT, to the identical questions. The LLMs were given information about industry and company conditions. The answers given by company representatives could then be compared to those of the LLMs using semantic similarity or cosine similarity. These methods were explained in Chapter 4.4.1. and 4.2.3. With the help of such comparisons, investors can recognize when company representatives are actually revealing new information and when answers are merely conveying already known information (Bai et al., 2023, pp. 3–4, 16–18, 26). Kim et al. have also used ChatGPT to identify and quantify risk exposures from earning call transcript information (Kim et al., 2023, pp. 4, 9–12).

GPT models offer some advantages over BERT models. Due to the absence of fine-tuning and a different mode of operation that does not analyze text from both directions, the results of GPT models are sometimes not quite at the level of BERT models. Nevertheless, the models are able to achieve high quality results at a comparable degree (Hu et al., 2023, pp. 12–13, 17), which makes them vulnerable for some use cases like the one of this thesis.

6.3. Processing of dependent variables

6.3.1. Determination of reporting compliance

The previous chapter concluded that sustainability reporting quality can be evaluated by comparing the consistency of the qualitative reporting required by the EU taxonomy with the International Auditing and Assurance Standards Board characteristics that are critical to non-financial reporting. Furthermore, the utilization of an LLM, specifically a GPT model, proves to be the most appropriate for the purposes of this thesis, as it enables a holistic assessment within the given time and resource constraints. As already noted, BERT models are limited to a context of 512 tokens for their analysis while GPT models can capture much larger contexts. However, sustainability reports are still so extensive that they surpass the capabilities of the models completely. Many sustainability reports contain over 100,000 tokens and over half a million characters. In order to evaluate the reports, the context per report has to be greatly reduced. Therefore, the sustainability reports were scanned for keywords indicating relevant passages. All passages were classified as relevant if they contained the term *EU taxonomy* or a related term and additionally at least one term related to the two goals of the EU taxonomy nearby. The list of terms and the code used for this purpose is presented in *Appendix F*. The length of each section is limited to 400 words (about 500 tokens), which on the one hand ensures that the sum of all sections of a report does not become too large for processing it with a GPT model. It also allows for the potential implementation of BERT models in later studies analyzing the relevant sections.

After this selection, a text file containing the relevant sections is obtained for each sustainability report. The sections are self-contained, although this procedure cannot completely prevent the loss of context between individual sections. Nevertheless, this method ensures a drastic reduction in the total amount of information, with as little loss of relevant information as possible. Of the original 90 companies in the sample and 89 sustainability reports that could be analyzed[16], no text sections were identified for twelve reports that did not meet the filter criteria. A manual evaluation indicated that the companies concerned had provided little or no relevant information on the objectives of the taxonomy to be considered. This may be because no information was actually provided or, more likely, because the information was provided outside the sustainability report in the annual statements (PricewaterhouseCoopers, 2023, p. 17). As this thesis explicitly examines sustainability reports, the twelve companies from the sample were excluded for most of the analyses, leaving 77 reports to be further examined[17].

The GPT 3.5 model, which was available free of charge at the time of the study (November 2023), captures a context of just over 8,000 tokens. While there were other fee required models that are not only more powerful and capa-

[15] For an recent overview see: platform.openai.com/docs/models/overview

[16] The company HELLA GmbH & Co. KGaA did not publish a sustainability report in fiscal year 2022 due to a previous acquisition.

[17] Five of the twelve samples were audited by an external auditor with limited assurance, while 7 samples were not externally audited. Accordingly, for the sub-sample there is no indication that the disclosure of EU taxonomy related information outside the sustainability reporting is related to the presence or absence of an external audit of the sustainability reporting.

ble of capturing more context, but also promise significantly less effort due to their combinability with programming interfaces, their use is not feasible for this study. As their costs are calculated not only on the basis of output tokens, but also on the basis of input tokens, their use would become very costly when analyzing long sustainability reports or excerpts from sustainability reports. However, manual use of the free version without a programming interface allows each text file to be analyzed individually. The only restriction is the total number of tokens of an analysis, including input and output, must not exceed 8,000 tokens, otherwise, the context will surpass the model's limit.

Of the 76 text files, 70 could be analyzed directly. The remaining six files contained too many relevant text segments and had to be reduced by up to 2,000 tokens. This was performed by manually discarding apparently irrelevant or recurring elements of the sustainability reports.

However, the text alone is not sufficient. To analyze text files as desired, the model requires a *task*. This task is reflected in the prompt, which represents the input parameter for the LLM. The prompt has a direct and substantial influence on the output received. It is designed to communicate with the LLM and provide all the necessary information needed to run the analysis. Prompts usually consist of three or four components: The actual instructions to the LLM (*What to do?*), the relevant context (*What to involve?*) and the actual data (*What to apply to?*) are essential. Depending on the approach, the prompt may also include one or more examples as a fourth component (de Kok, 2023, p. 14).

The prompt has a very strong impact on the output and thus a correspondingly strong impact on the research results. At the same time is relatively complex in structure and offers the researchers a lot of design flexibility. As a result, a whole field of research has been devoted to the design of prompts - the so-called *prompt engineering*. The goal of prompt engineering is to produce a prompt that is as precise as possible while achieving the best possible results (de Kok, 2023, pp. 15–16, 44). In most cases, the prompts are reduced to the most essential content and the instructions are formulated in an unambiguous manner. *Figure 3* shows the prompt used for testing H1 in its unmodified form as well as clustered by components.

The first subsentence provides context to the LLM by assigning a role to the model. This assignment ensures that the model will respond based on not only all available information, but particularly the information in the domains of reporting and auditing. This is followed by a precise instruction to the model to give an assessment in the form of a rating. The third subsentence defines the data to be evaluated. The fourth subsentence again specifies the context against which the rating should be conducted. The following instructional subsentences ensure that the output is as accurate as possible and in a homogeneous format.

The evaluation of H1 will be based on quantitative ratings only. Therefore, no further qualitative information in the form of text is needed. Models such as GPT 3.5 are generative LLMs that are specifically designed to generate qualitative output. Therefore, in order to actively counteract this behavior, multiple instructions can be necessary in such cases where only one specific information is demanded. Finally, the ninth part serves as a placeholder for the text files to be analyzed. The foregoing prompt, without the text file, consists of only 82 tokens.

The specific wording of the prompt allows for a precise evaluation of the model. In particular, the second and fourth subsentences are geared to the purpose of this thesis and ensure that the assessment is focused on the desired content, which is the evaluation of eligibility and alignment disclosures for the first two taxonomy objectives.

The approach used to test H1 is the so-called *zero-shot* method. Here, the analysis is performed without providing any examples to the model. In contrast, the *few-shot* method would include some previously rated text files that would serve as blueprints for the decision-making process of the model. This can improve the quality of the rating, but can also cause the scoring to go in the wrong direction if the example files are of poor quality or the examples do not scale well to the full sample. Finally, there is the *fine tuning* approach, in which the LLM is adapted to the task in advance using the known procedures. The choice of method is always a trade-off between accuracy and effort (de Kok, 2023, p. 13).

Especially when employing zero-shot approaches, the prompt optimization often occurs through trial and error (Kim et al., 2023, p. 25; A. Huang et al., 2022, p. 3). This particular prompt was also optimized in this manner. For example, using the prompt without the instruction in the sixth subsentence resulted in the model not providing a score, with the reasoning that it could not determine the output with complete reliability. However, this behavior of the model is unfavorable since it is intended to derive a rating from the given context and potentially substitute this context for missing information. It also appeared that GPT 3.5 produced more consistent outputs when the initial prompt was placed at the end as well as at the beginning of the text file rather than at the beginning only. This may reflect the fact that the prompt together with the corresponding text files consists of 8,000 tokens, and that the actual task is being forgotten in the overall context of the model, since it represents only about 1 % of the overall prompt. After successful prompt engineering, GPT 3.5 was able to generate a rating for each sustainability report based on the text file, consisting of a number of relevant sections of that report. This rating is supposed to represent the quality of the reporting in the context of compliance with the requirements of the EU taxonomy. Thus, the data for the relevant dependent variable for H1 could be successfully collected. The evaluation of the results is presented in Chapter 7.1.

6.3.2. Determination of reporting quantity and the level of assurance

The research design and data collection for H1 was complex, as it was necessary to specify the content to be targeted and the textual analysis procedure to be applied. The

Original:

As a sustainability reporting and auditing expert, please rate the eligibility and alignment disclosures of this sustainability report abstract with environmental objectives related to climate change mitigation and adaptation. Provide a numerical rating from 0 to 100. Give a rating even if you are unsure. Respond in the format: Score: [number]. Do not provide further information. [Individual text file]

Clustered:

As a sustainability reporting and auditing expert (1), please rate the eligibility and alignment disclosures (2) of this sustainability report abstract (3) with environmental objectives related to climate change mitigation and adaptation (4). Provide a numerical rating from 0 to 100 (5). Give a rating even if you are unsure (6). Respond in the format: Score: [number] (7). Do not provide further information (8). [Individual text file] (9)

Instructions
Context
Data

Figure 3: Prompt applied for testing H1

operationalizing and data collection for H2 and H3 is more straightforward. The two related hypotheses state that sustainable companies disclose more information (H2) and that they are more likely to seek voluntary assurance (H3). One central point of these hypotheses is the classification of which companies can be considered sustainable. A simple and effective way of achieving such a classification is to rely on the assessments of rating agencies. Their ratings allow for a transparent classification without having to carry out a costly assessment independently (Drempetic et al., 2020, p. 354). There are many different rating agencies and ratings. This thesis relies on Bloomberg ratings to classify whether companies are sustainable. Bloomberg ratings are considered to be one of the most important rating frameworks, along with ratings from other agencies such as Thomson Reuters and MSCI (Sorrosal-Forradellas et al., 2023, p. 2). In addition to their popularity, Bloomberg ratings are particularly suitable for the purposes of this thesis because they explicitly incorporate the requirements of the EU taxonomy and other regulations into the evaluation process (Bloomberg, 2023). In addition to the overall score, Bloomberg also includes the individual scores for each of the *environmental, social* and *governance* components. The summary statistics for the scores of the sample are presented in Panel A of *Appendix C*.

Besides determining the degree of sustainability of a company, H2 also involves determining the quantity of information disclosed. Generally, publications such as annual reports or non-financial reports contain information. One could argue that within these reports, the quantity of information can be influenced by whether the authors present the information clearly and to the point, or whether they talk around the subject and stuff the reporting with meaningless sentences without actually revealing any relevant information. Since this problem is a distinct area of research, for the purpose of measuring the quantity of information, it is assumed that all text contains information and that an increase in text length is associated with an increase in information quantity.

To measure the information quantity of sustainability reports, it would be sufficient to count the sentences, words or letters in these reports. Alternatively, this thesis utilizes a measure that has been applied by Loughran and McDonald (2014) in a similar form to measure readability - the file size. File size refers to the size of the reporting document in kilobytes or megabytes. Loughran and McDonald measure this size for files that companies report to the SEC (Loughran & McDonald, 2014, p. 1667). For this thesis, the file size is measured for sustainability reports published on company websites. However, not the Portable Document Format (PDF) file is used, but a converted version that extracts the text from the PDF files.[18] This ensures that the size is only driven by actual text and not by other factors such as formatting or the use of graphics and images, which account for most of the PDF file size. Additionally, there is a distinction between standalone non-financial reporting and integrated reporting. Since the integrated reporting files contain further reporting components, these files are reduced to approximate the portion of non-financial reporting in the total file.[19] Based on the ESG scores and file size values, H2 can be targeted. The evaluation of the results is presented in Chapter 7.2.

H3 suggests that sustainable companies are more likely to undergo a voluntary audit of their non-financial reporting than less sustainable companies. This hypothesis is based on the theoretical foundation of Chapter 2.3 of this thesis. *Table 1* shows that the disclosure of favorable information should be covered by audit assurance in particular in order to achieve an appropriate level of reliability. However, information that does not present a company as sustainable does not require audit assurance as a company would have no interest in publishing incorrect information that presents itself in a poor manner. Conversely, information about sustainable activities requires assurance from external auditors in order to

[18] The code used for the conversion is available in *Appendix F*.

[19] The approximation is made using a correction factor of 0.3312, which was determined by dividing the average file size of the non-financial reports in the sample by the average file size of the integrated reports in the sample.

be considered reliable by external parties. According to this theory, companies with better ESG ratings should be more likely to have their non-financial reporting verified by external auditors. The results of H3 are presented in Chapter 7.3.

7. Main results

7.1. Regression results for H1

Following the procedure described in Chapter 6, the results of the analyses are presented below. In addition, the summary statistics and the corresponding correlation matrices are provided in *Appendix C*.

H1 states that audit assurance for sustainability reports increases their compliance with regulatory requirements. To test this hypothesis, ratings are gathered using the procedure described in Chapter 6.3.1. *Figure 4* shows the correlation between the variables *GPT_Rating* and *Assurance_LVL*.

Assurance_LVL 1 represents no assurance from an external auditor, *Assurance_LVL 2* represents a limited level of assurance, and *Assurance_LVL 3* represents a reasonable level of audit assurance. Remarkably, the trendline shows a negative correlation between the rating assigned by the LLM (*GPT_Rating*) and *Assurance_LVL*. This would indicate, ceteris paribus, that an increase in *Assurance_LVL* results in a decrease in the *GPT_Rating* and thus a decrease in the compliance of the reports with the requirements of the EU taxonomy.

The relationship is tested in a regression analysis with *GPT_Rating* as the dependent variable and *Assurance_LVL* as the independent variable. In addition to *Assurance_LVL*, twelve other independent variables are included in the regression. These are *ESG_Score, ENV_Score, SOC_Score, GOV_Score, Market_CAP, Total_ASS, File_Size, File_Size_ABS, Neutrality_Score_ABS, Fog_Index_ABS, Neutrality_Score* and *Fog_Index*. The control variables include diverse Bloomberg ratings regarding ESG disclosures, company data including market capitalization and total assets, as well as significant metrics from traditional textual analysis methods from the areas of readability and sentiment analysis. All variables are defined in *Appendix A*. The full regression results are presented in *Appendix D*. The regression analysis, as well as all subsequent analyses, are conducted on a 95 % confidence level.

Omitting any clearly insignificant variables results in a multiple linear regression with three independent variables and the following regression of *Equation 6*:

$$GPT\_Rating_{it} = \alpha - \beta_1 Assurance\_LVL_{it} + \beta_2 ESG\_Score_{it} + \beta_3 ENV\_Score_{it} + \varepsilon_{it} \quad (6)$$

i indexes the respective company from the sample and t denotes the company's financial year. ε represents an error term. The independent variable *Assurance_LVL* is the variable of interest for H1. However, it shows to be not significant with a t-statistic of -1.49. In addition, *ESG_Score* and *ENV_Score* determine *GPT_Rating*. Interestingly, *ENV_Score* is negatively correlated with *GPT_Rating* and *ESG_Score* is positively correlated, with *ENV_Score* being a subset of *ESG_Score*. However, *ENV_Score* is not significant (t-statistic: -1.54) in this regression analysis, leaving *ESG_Score* as the only significant variable (t-statistic: 2.06).

The regression in *Equation 6* thus shows that *ESG_Score* is positively correlated with *GPT_Rating* and therefore, companies with a higher ESG scoring are more likely to meet the requirements of the EU taxonomy. This correlation is in line with reasonable expectations. At the same time, *Assurance_LVL* has no significant impact on *GPT_Rating*, hence H1 cannot be confirmed. The level of assurance does therefore not appear to have a significant impact on the compliance of the disclosures with the requirements of the EU taxonomy.

In principle, the regression results should be corrected for the companies' industry. Some industries are naturally more concerned with the issues required by the EU taxonomy. It was examined whether more text in the disclosures has a fundamentally positive effect on *GPT_Rating*. This would have the consequence that companies from industries that naturally have to deal with these issues in more detail would receive unjustified higher or lower ratings. However, no significant correlation has been found (*Appendix D*). Therefore, industry was not controlled for.

Generally, the results of the analysis should be considered in the context of the overall environment and the data used. Although the overall regression is statistically significant, it also has a relatively low coefficient of determination and, most importantly, a small sample size of only 73 observations. The two independent variables, *ESG_Score* and *ENV_Score*, share a correlation coefficient of 0.71 and are therefore highly interdependent. This is a strong signal for multicollinearity. Multicollinearity can result in imprecise regression analysis outcomes. The strong correlation between two independent variables may affect the predictive accuracy of the model as it becomes difficult to determine what proportion of the variation is explained by a particular variable (Kutner et al., 2004, p. 283).

An examination using the widely accepted Variance inflation factor (VIF) has revealed the presence of multicollinearity in this regression (*ESG_Score* VIF: 18.13; *Environmetnal_Score* VIF: 5.62)[20]. In further regression analyzes where one of the two variables was removed in order to deal with the multicollinearity problem, the statistical significance of the regression decreased (*Appendix D*).

Furthermore, the dependent variable *GPT_Rating* is prone to errors. The procedure described in Chapter 6.3.1 ensures that the ratings are of the highest possible quality, but they still contain deficiencies. Although LLMs have been used in previous research to provide ratings on the content of textual data based on defined criteria, the outcome is, to some extent, the product of a *black box*. Traditional textual analysis methods offer more transparency and allow better traceability of results.

[20] Multicollinearity is expected when the VIF exceeds ten, thereby strongly impacting the analysis results (Kutner et al., 2004, pp. 408–409).

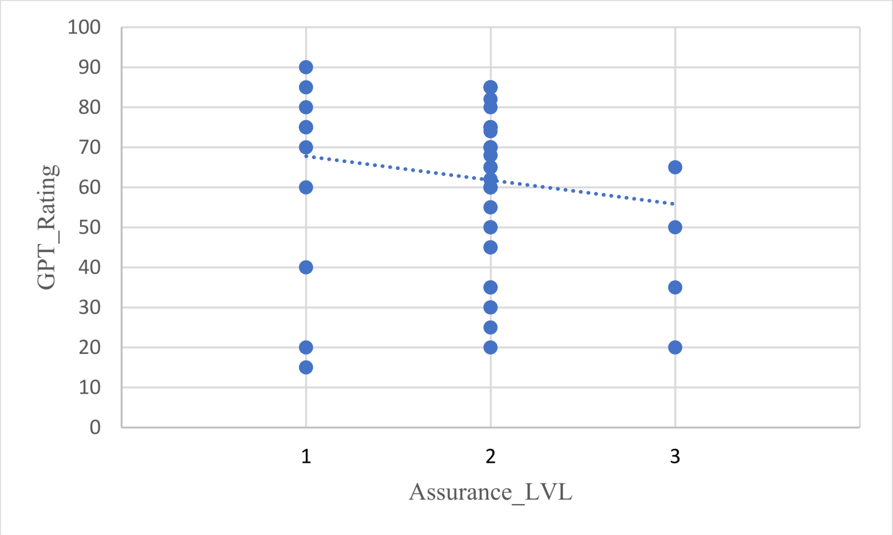

Figure 4: Relationship between Assurance_LVL and GPT_Rating

The length and related context of the text passages to be evaluated pushed the GPT 3.5 model to its limits. In addition, the evaluation criteria and the subject matter of the reports are more complex than those of many other texts. The combination of all these factors results in less reliable ratings. This has even led to the same text file receiving different ratings from GPT 3.5 when evaluated multiple times. The ratings assigned are based on probabilistic methods and are therefore influenced by statistical probabilities and judgments based on the text and information available to the model. In particular, longer texts increase the likelihood that the model will not provide an identical response or rating. In addition, GPT 3.5 is designed to produce outputs in text form. The model calculates each word within the answer based on the prompt itself, as well as on the previously given words of its own output. Thus, the model utilizes the previously produced part of its own output to generate the remaining part of the output, word by word. This generally improves the quality of the output (de Kok, 2023, p. 44). In the application case of this thesis, only the rating as a two-digit number is assigned as an output. Therefore, this effect is absent at the expense of output consistency.

The standard deviation of the ratings is 12.27 % (*Appendix E*). Thus, although the ratings are not randomly determined by the model, they are still subject to a certain degree of uncertainty. In order to increase the reliability of the *GPT_Rating* variable for possible future research, it is strongly recommended to verify it against valid ratings or the results of other textual analysis methods, like the control variables *Neutrality_Score* and *Fog_Index*.

7.2. Regression results for H2

H2 states that sustainable companies disclose more information. This hypothesis builds on the voluntary disclosure theory, which suggests that companies disclose information only when it is beneficial to them, and on the assumption that sustainable companies derive more of these benefits from disclosure and therefore provide more information in form of higher reporting quantity. The direct relationship between *ESG_Score* and reporting *File_Size* is presented in *Figure 5*.

In order to gain confidence in the statistical significance of this relationship, some regression analyzes were performed. In those regressions, *File_Size* represented the dependent variable. A first analysis included eleven independent variables besides *ESG_Score*. These control variables are identical to those for H1, except for the absence of *File_Size_ABS*, as the extracted text sections cannot have any effect on the dependent variable File_Size, since they are themselves taken from the originating report. Among these control variables, *Market_CAP* shows a significant positive correlation (t-statistic 2.66) and *Neutrality_Score* has a significant negative correlation (t-statistic -3.72). *Neutrality_Score* reflects the ratio of positive and negative words to the total amount of words, whereas a higher ratio of positive and negative words indicates a less neutral writing style (*Appendix A*).

A second analysis using only *ESG_Score* and the two significant control variables shows a slightly higher adjusted coefficient of determination and a generally higher significance for the overall regression analysis. The dependent variable *File_Size* and the control variable *Neutrality_Score* have a high negative correlation coefficient of 0.8590. *Neutrality_Score* differs greatly from the other significant control variables in that it is a measure derived directly from the reporting being analyzed. To prioritize relevant indicators that are not directly related to reporting, one further regression is performed, in which the dependent variable is only compared with *ESG_Score* and the control variable *Market_CAP*. *Equation 7* expresses the relationship between these two variables and the dependent variable as follows:

$$File\_Size_{it} = \alpha + \beta_1 ESG\_Score_{it} + \beta_2 Market\_CAP_{it} + \varepsilon_{it} \qquad (7)$$

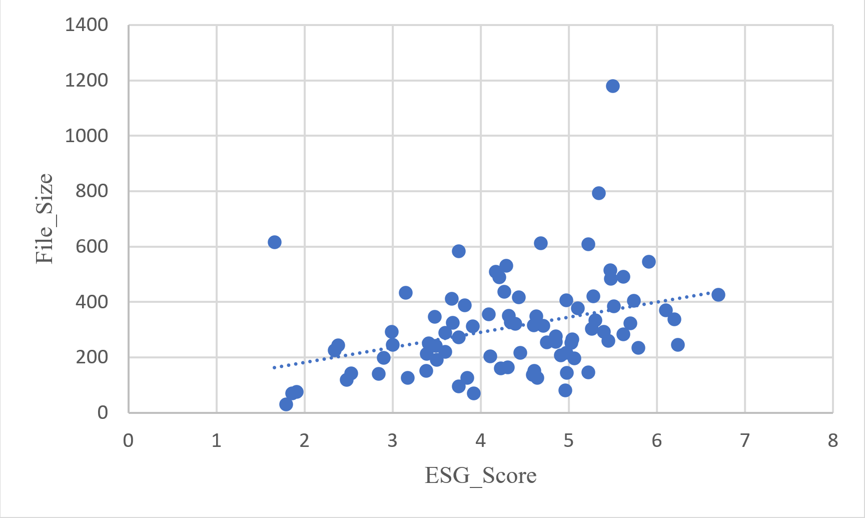

Figure 5: Relationship between ESG_Score and File_Size

Within this regression model, *ESG_Score* has a significant t-statistic of 2.33, while *Market_CAP* is also significant with a t-statistic of 4.23 (at the 95% confidence level). When the regression is not adjusted for *Market_CAP*, the t-statistic of *ESG_Score* increases to 3.35, but the overall regression generally becomes less significant and its coefficient of determination decreases.

Next to *File_Size*, a number of additional regressions are performed with *File_Size_ABS* as the dependent variable. *File_Size_ABS* differs in that it does not measure the size of the entire report, but only the size of the file containing specific abstracts. These are the abstracts related to the EU taxonomy that were extracted to determine the scores from GPT 3.5. The extraction procedure is described in Chapter 6.3.1 and is further illustrated in *Appendix F*.

When running the regression with the new dependent variable *File_Size_ABS* and *ESG_Score* as well as the same eleven control variables as independent variables, the only control variable to be significant is *Neutrality_Score_ABS*. This variable, like *Neutrality_Score*, measures the neutrality of the writing style, but with respect to the abstracts instead of the full reports. It is reasonable that this control variable would be significant with respect to *File_Size_ABS*, since *Neutrality_Score* was also significant for overall report *File_Size*. The correlation coefficient of -0.7348 is also close to that of *File_Size* and *Neutrality_Score* (-0.8590). When the control variable *Neutrality_Score_ABS* is dropped from the regression model, no other independent variables remain that have a significant effect on the dependent variable *File_Size_ABS*.

In summary, the analysis reveals a significant positive relationship between *ESG_Score* and *File_Size*. Therefore, the analyzes are in support of H2. However, no significant relationship can be found between *ESG_Score* and *File_Size_ABS*. This suggests that sustainable companies generally disclose a higher quantity of information in their non-financial reports. Meanwhile, information based on compliance with the EU taxonomy is not affected. The rationale for this could be that the additional information disclosed by sustainable companies does not concern this particular domain and that the reports are being enriched with redundant information instead or that the focus lies on other domains. At the same time, the reason may also be that the method used to extract the abstracts does not properly capture all relevant parts.

7.3. Regression results for H3

H3 states that sustainable companies are more likely to obtain an external audit of their non-financial reporting. This hypothesis arises from attribution theory, which suggests that the readers of reports are more likely to challenge published information if it presents the company in a favorable way. The relationship between the sustainability of a company and the level of assurance is presented in *Figure 6*.

Similar to *Figure 5*, *ESG_Score* represents the independent variable to determine the dependent variable, in this case *Assurance_LVL*. In addition to *ESG_Score*, eleven other control variables were tested for significance in a regression analysis. These are the same as for H2, with only *File_Size* and *File_Size_ABS* replacing *GPT_Rating* and *Assurance_Level*. In combination, all control variables as well as *ESG_Score* are found to be insignificant, with *Market_CAP* just falling below the threshold with a t-score of 1.93 (at the 95% confidence level). When attempting to determine the dependent variable using only *ESG_Score* or *Market_CAP* seperately, a significant positive correlation is found in each case (t-statistic *ESG_Score*: 2.64; t-statistic *Market_CAP*: 3.00 at the 95% confidence level). To increase the coefficient of determination and the overall power of the regression analysis, the dependent variable in *Equation 8* is determined by the two independent variables *ESG_Score* and *Market_CAP* combined.

$$\begin{aligned} Assurance\_LVL_{it} = {} & \alpha + \beta_1 ESG\_Score_{it} \\ & + \beta_2 Market\_CAP_{it} + \varepsilon_{it} \end{aligned} \tag{8}$$

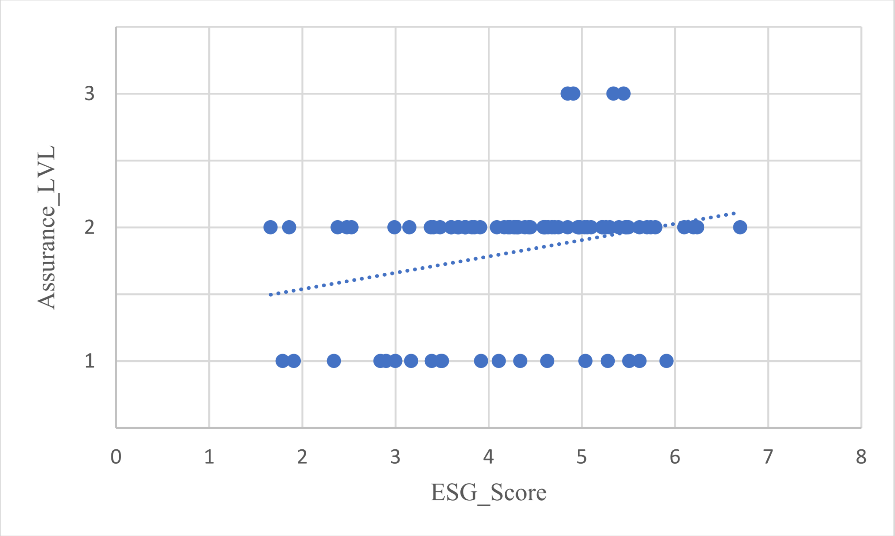

Figure 6: Relationship between ESG_Score and Assurance_LVL

In this analysis, *Market_CAP* shows a significant correlation (t-statistic: 2.38) with *Assurance_LVL*, while *ESG_Score* is just below the significance threshold (t-statistic: 1.93). Based on this result, H3 cannot be confirmed. The correlation between a company's sustainability and level of audit assurance cannot be statistically proven. However, it appears that larger companies with a higher capitalization are more likely to undergo an external audit. This relationship is also shown in *Figure 7*.

19 of the sample companies do not provide an external audit of their non-financial reporting.[21] Only three of those companies exceed $10 billion in market capitalization, while the mean for the entire sample is $21.2 billion and the median is $8.7 billion (*Appendix B*). While many small-cap companies also undergo an audit of their non-financial reporting, at the same time a voluntary audit appears to be inevitable once a company reaches a certain market capitalization. Since the audit of non-financial reporting was not mandatory for the sample of German companies in the most recent fiscal year, implementation at the firm level is a trade-off between costs and benefits (Widmann et al. 2021: 457), similar to the publication of supplementary information according to the voluntary disclosure theory. Audit fees are primarily driven by the size and complexity of firms, but other factors such as a so-called *BIG-4 premium* also play a role, as firms hope to achieve higher audit quality by hiring the large, prestigious audit firms (Widmann et al., 2021, pp. 473–475, 479).

Based on the previous findings, larger firms in particular consider the benefits of an audit to outweigh the associated costs. The reasoning behind this opens up possibilities for further research. At the same time, smaller companies may not be able to afford the voluntary audit, as they do not have the same financial flexibility as larger companies. Nevertheless, even the smallest companies included in the sample are listed on the MDAX, meaning that they are still relatively large compared to other companies in terms of total assets or market capitalization, so that this effect is unlikely to be observed in this case.

The results of *Equation 8* and the results of the examination of the control variable *Market_CAP* should be considered in relation to the conditions of the data set. First, the sample size of 85 is relatively small. Second, the dependent variable *Assurance_LVL* is not a continuous variable but a categorical variable that distinguishes only between *no assurance, limited assurance* and *reasonable assurance*. It is particularly difficult to capture a categorical variable through a linear regression because the outcome of the regression equation provides values that need to be assigned to one of the three assurance levels since there are no intermediate levels.

8. Conclusion

Sustainability reporting has emerged as a major element of corporate reporting in recent years. The challenges arising from climate change as well as increasing stakeholder awareness are some of the key driving factors. While a number of companies have been reporting on these issues voluntarily for some time, today almost all of the major companies are getting on board, partly driven by regulatory requirements. The EU taxonomy represents one of the key regulatory foundations in this field. Not only does it tighten reporting requirements, it also encourages real effects by redirecting capital flows and company activities towards more sustainability and reducing practices such as greenwashing (European Union (EU), 2020, p. 14).

This thesis merges the content topic of sustainability reporting with the methodological topic of textual analysis. As sustainability reporting is mainly presented in qualitative text

[21] *One company (Sixt SE) excluded due to incomplete data.*

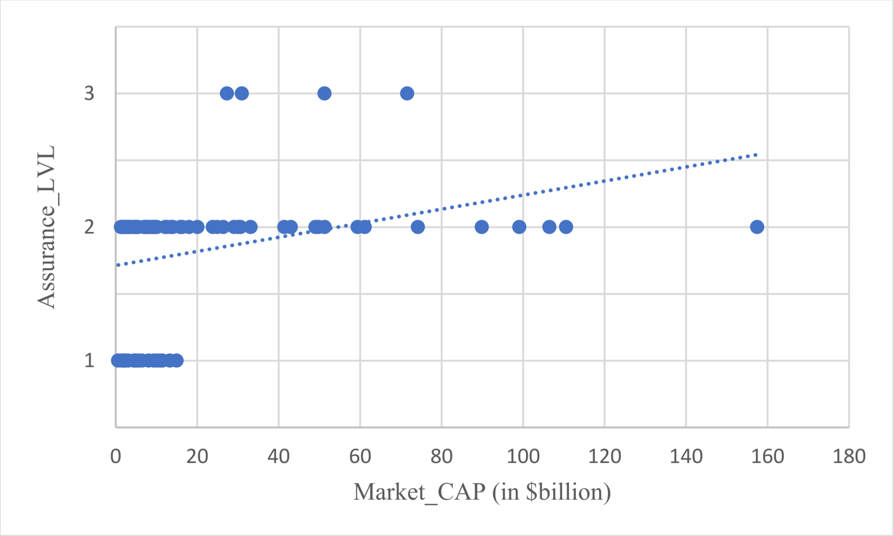

Figure 7: Relationship between Market_CAP and Assurance_LVL

form, this combination fits together quite well. The methodology of textual analysis has become very popular in academic research and among the general public through the introduction of text-generating models such as ChatGPT.

This thesis makes several contributions. It contributes to the literature in the area of auditing, especially the auditing of non-financial reporting, and to the literature in the area of European reporting requirements. The two main areas of focus of this thesis are divided, with the first being a literature review and analysis of various textual analysis methods. The literature review focuses primarily on the finance and accounting domain, detailing the textual analysis techniques utilized in prior research. The methodology overview provides a comprehensive examination of the advantages, disadvantages, and limitations of each method. This review contributes to the reader's understanding of textual analysis capabilities and potential applications, which will help readers identify appropriate methods for their own textual analysis research. A further contribution lies in the results of the investigation of the impact of audit assurance on the quality of sustainability reporting. The results were obtained through a combination of textual analysis methods and multiple linear regressions.

The unique aspect of the methodology in this thesis is that the essential data for the analysis is obtained with the help of GPT 3.5, a freely accessible LLM with text comprehension and generation capabilities. The model is used to analyze a defined problem. Ratings or scores have been used to identify genuine economic connections frequently in prior studies. In this case, the LLM is utilized to generate a rating based solely on the text of the non-financial reports of the sampled companies. By constructing a specific prompt, it is possible to determine exactly which factors should be included to create such a rating. For the purpose of this thesis, the compliance of the reporting with the requirements of the EU taxonomy in relation to two specific taxonomy objectives has been defined as a quality feature that defines the rating. Typically, there are no established ratings for such specific objectives.

LLMs have been utilized in prior studies in a similar manner. For example, Kim et al. used another GPT 3.5 model to summarize components of corporate disclosure, and found that these summaries generated each had a stronger positive or negative sentiment than the original reporting. Accordingly, GPT appears to be able to filter noise from the texts, improve the information content and present more relevant insights than the original reporting (Kim et al., 2023, pp. 1, 2, 5, 15–16, 19–20, 30).

The information content of sustainability reports in this thesis was significantly reduced, with only the *GPT_Rating* remaining. While similar, the approach here is much more drastic, as Kim et al. eliminated about 70% of the original reporting, while here the entire text was eliminated and replaced by the *GPT_Rating* as a single number remaining. This significant reduction may account for why the analyses of this thesis did not yield many significant results.

The regression results from the analysis of H1 show a negative but insignificant correlation between the level of assurance and the reporting compliance, displayed via the *GPT_Rating* variable, contrary to the expectations. Additionally, it was found that companies with a higher ESG score also have a higher *GPT_Rating* and are therefore more likely to meet the requirements of the EU taxonomy. The analysis of H2 indicates a positive correlation between the *ESG_Score* and the *File_Size variable*. Thus, sustainable acting companies disclose more information, which supports H2. Finally, the last regression analyses could not confirm the hypothesis of H3, which states that sustainable companies are more likely to have their non-financial reporting externally audited, as the attribution theory would suggest.

The research results are subject to a number of limitations. First, the sample size which ranges from 73 to 90 companies, depending on the analysis, is relatively small. This is due to the fact that the sustainability reports had to be manually retrieved from company websites and being further processed for the analysis. In addition, the variable *GPT_Rating* is prone to error. On one hand, the analysis of this variable is founded solely on sections extracted from sustainability reports, and not on the entire reports themselves. The extraction procedure was based on an intuitive yet untested approach. The second and more influential source of error is the rating of the sections by ChatGPT itself. Due to time constraints, the rating was done using a zero-shot approach. The model was not fine-tuned and has not been fed with training data in advance. Additionally, the results were not subjected to robustness tests. Such robustness testing could be considered in future research, either by comparing the results with existing, externally validated scores, or by comparing them with a portfolio of results from other textual analysis methods. In this thesis, the results present data that are conceivable but not fully comprehensible.

Changes in the prompt can affect the output, resulting in different rating results for the same report. This even happens if the prompt is not changed (*Appendix E*). The cause may be an overload of the GPT 3.5 model utilized, which is optimized for text output and has limits of approximately 8,000 tokens. OpenAI has announced new models capable of capturing an input context of up to 128,000 tokens. These models are designed to produce reproducible outputs, resulting in lower variances. Such models can enhance the quality of analyses and ensure comprehensive reports. However, at the time of performing the analysis (November 2023), the new models were not yet available.[22]

This thesis has introduced a number of textual analysis methods, ranging from the simplest models with almost no mathematical or technical depth, to models using the latest technical achievements in the field of machine learning. For the analysis of the hypotheses of this thesis, the application of an LLM has proven to be an exhaustive instrument. Despite the emergence and popularity of modern techniques, traditional methods are still commonly utilized in research due to their ease of understanding and intuitive results. However, the future of textual analysis lies in the possibilities presented by the advancing state-of-the-art. Generative LLMs open up numerous possibilities in all fields of research. Textual data is becoming increasingly relevant in the field of accounting. Models like ChatGPT can be particularly helpful in dealing with the growing importance of text in reporting and the associated information overload (Kim et al., 2023, p. 30). Researchers should not overlook these opportunities and utilize the possibilities offered by these and similar models in their research.

References

Aggarwal, C. (2018). *Machine Learning for Text*. Springer.

Agrawal, S., Azar, P., Lo, A., & Singh, T. (2018). Momentum, mean-reversion, and social media: Evidence from StockTwits and Twitter. *Journal of Portfolio Management*, *44*, 85–95.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In M. Berry, A. Mohamed, & B. Yap (Eds.), *Supervised and Unsupervised Learning for Data Science*. Springer Nature.

Amel-Zadeh, A., & Serafeim, G. (2018). Why and how investors use ESG information: Evidence from a global survey. *Financial Analysts Journal*, *74*(3), 87–103.

Antle, R. (1982). The Auditor as an Economic Agent. *Journal of Accounting Research*, *20*(2), 503–527.

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review*, *53*(5), 941–973.

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

Bae, J., Hung, C., & van Lent, L. (2023). Mobilizing Text As Data [forthcoming]. *European Accounting Review*.

Bai, J., Boyson, N., Cao, Y., Liu, M., & Wan, C. (2023). Executives vs. Chatbots: Unmasking Insights through Human-AI Differences in Earnings Conference Q&A. https://ssrn.com/abstract=4480056

Ball, R., & Brown, P. (1968). An Empirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research*, *6*(2), 159–178.

Biddle, G., Hilary, G., & Verdi, R. (2009). How Does Financial Reporting Quality Relate to Investment Efficiency? *Journal of Accounting and Economics*, *48*, 112–131.

Black, F., & Schloes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, *81*(3), 637–654.

Blackwell, D., Noland, T., & Winters, D. (1998). The Value of Auditor Assurance: Evidence from Loan Pricing. *Journal of Accounting Research*, *36*(1), 57–70.

Bloomberg. (2021). Applying the EU taxonomy to your investments, how to start? https://www.bloomberg.com/professional/blog/applying-the-eu-taxonomy-to-your-investments-how-to-start/

Bloomberg. (2023). ESG Data. https://www.bloomberg.com/professional/product/esg-data/

Bochkay, K., Brown, S., Leone, A., & Tucker, J. (2023). Textual analysis in accounting: What's next? *Contemporary Accounting Research*, *40*, 765–805.

Bonsall, S., Leone, A., Miller, B., & Rennekamp, K. (2017). A Plain English Measure of Financial Reporting Readability. *Journal of Accounting and Economics*, *62*(2), 329–357.

BookCorpus. (2023). BookCorpus dataset. https://paperswithcode.com/dataset/bookcorpus

Bradbury, M. (1990). The Incentives for Voluntary Audit Committee Formation. *Journal of Accounting and Public Policy*, *9*, 19–36.

Breijer, R., & Orij, R. (2022). The Comparability of Non-Financial Information: An Exploration of the Impact of the Non-Financial Reporting Directive (NFRD, 2014/95/EU). *Accounting in Europe*, *19*(2), 332–361.

Brown, N., Crowley, R., & Elliott, W. (2020). What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, *58*(1), 237–291.

Brown, S., & Knechel, W. (2016). Auditor-client compatibility and audit firm selection. *Journal of Accounting Research*, *54*(5), 1331–1364.

Brown, S., Tian, X., & Tucker, J. (2018). The spillover effect of SEC comment letters on qualitative corporate disclosure: Evidence from the risk factor disclosure. *Contemporary Accounting Research*, *35*(2), 622–656.

Brown, S., & Tucker, J. (2011). Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research*, *49*(2), 309–346.

Busco, C., D'Eri, A., & Novembre, V. (2022). *Corporate Disclosure* (N. Linciano, P. Soccorso, & C. Guagliano, Eds.). Springer Nature.

Chakrabarty, B., Seetharaman, A., Swanson, Z., & Wang, X. (2018). Management risk incentives and the readability of corporate disclosures. *Financial Management*, *47*, 583–616.

[22] For an recent overview see: platform.openai.com/docs/models/overview

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling.

Chen, H., De, P., Hu, Y., & Hwang, B. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*(5), 1983–2020.

Chen, W., & Srinivasan, S. (2023). Going digital: implications for firm value and performance [forthcoming]. *Review of Accounting Studies*. https://ssrn.com/abstract=4177947

Chevalier, J. A., & Mayzlin, D. (2006). The effect of Word of Mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354.

Christensen, B., Glover, S., Omer, T., & Shelley, M. (2016). Understanding Audit Quality: Insights from Audit Professionals and Investors. *Contemporary Accounting Research*, *33*(4), 1648–1684.

Christensen, H., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: economic analysis and literature review. *Review of Accounting Studies*, *26*, 1176–1248.

Christofi, A., Christofi, C., & Sisaye, S. (2012). Corporate sustainability: historical development and reporting practices. *Management Research Review*, *35*(2), 157–172.

Chychyla, R., Leone, A., & Minutti-Meza, M. (2019). Complexity of financial reporting standards and accounting expertise. *Journal of Accounting and Economics*, *67*, 226–253.

Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, *110*, 103539.

Coram, P., Monroe, G., & Woodliff, D. (2009). The Value of Assurance on Voluntary Nonfinancial Disclosure: An Experimental Evaluation. *Auditing: A Journal of Practice & Theory*, *28*(1), 137–151.

de Kok, T. (2023). Generative LLMs and Textual Analysis in Accounting: (Chat)GPT as Research Assistant? https://ssrn.com/abstract=4429658

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

Dissanayake, D., Tilt, C., & Qian, W. (2019). Factors influencing sustainability reporting by Sri Lankan companies. *Pacific Accounting Review*, *31*(1), 84–109.

Dorfleitner, G., Hornuf, L., & Kreppmeier, J. (2023). Promise not fulfilled: FinTech, data privacy, and the GDPR. *Electronic Markets*, *33*(33), 1–29.

Drempetic, S., Klein, C., & Zwergel, B. (2020). The Influence of Firm Size on the ESG Score: Corporate Sustainability Ratings Under Review. *Journal of Business Ethics*, *167*, 333–360.

Dumrose, M., Rink, S., & Eckert, J. (2022). Why Do Firms in Emerging Markets Report? A Stakeholder Theory Approach to Study the Determinants of Non-Financial Disclosure in Latin America. *Sustainability*, *10*(9), 1–20.

Duran, I., & Rodrigo, P. (2018). Disaggregating confusion? The EU Taxonomy and its relation to ESG rating. *Finance Research Letters*, *48*, 1–8.

Dusík, J., & Bond, A. (2022). Environmental assessments and sustainable finance frameworks: will the EU Taxonomy change the mindset over the contribution of EIA to sustainable development. *Impact Assessment and Project Appraisal*, *40*(2), 90–98.

Efretuei, E., & Hussainey, K. (2023). The fog index in accounting research: contributions and challenges. *Journal of Applied Accounting Research*, *24*(2), 318–343.

European Commission. (2019). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. The European Green Deal. https://eur-lex.europa.eu/resource.html?uri=cellar:b828d165-1c22-11ea-8c1f-01aa75ed71a1.0002.02/DOC_1&format=PDF

European Commission. (2020). Taxonomy: Final report of the Technical Expert Group on Sustainable Finance. https://finance.ec.europa.eu/system/files/2020-03/200309-sustainable-finance-teg-final-report-taxonomy_en.pdf

European Commission. (2021). FAQ: What is the EU Taxonomy Article 8 delegated act and how will it work in practice? https://finance.ec.europa.eu/system/files/2021-07/sustainable-finance-taxonomy-article-8-faq_en.pdf

European Commission. (2022). FAQs: How should financial and non-financial undertakings report Taxonomy-eligible economic activities and assets in accordance with the Taxonomy Regulation Article 8 Disclosures Delegated Act? https://finance.ec.europa.eu/system/files/2022-01/sustainable-finance-taxonomy-article-8-report-eligible-activities-assets-faq_en.pdf

European Union (EU). (2014). Directive 2014/95/EU. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014L0095

European Union (EU). (2016). Paris Agreement. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:22016A1019(01)

European Union (EU). (2020). Regulation 2020/852. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32020R0852

European Union (EU). (2022). Directive 2022/2464/EU. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2464+

Filighera, A., Steuer, T., & Rensing, C. (2019). Automatic text difficulty estimation using embeddings and neural networks. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming Learning with Meaningful Technologies*. Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-34971-2_17

Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science*, *68*(7), 5514–5532.

Frankel, R., Mayew, W., & Sun, Y. (2010). Do pennies matter? Investor relations consequences of small negative earnings surprises. *Review of Accounting Studies*, *15*, 220–242.

Gaulin, M., & Peng, X. (2022). Semantic vs. literal disclosure similarity. https://ssrn.com/abstract=3971286

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Guay, W., Samuels, D., & Taylor, D. (2016). Guiding through the Fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics*, *62*(2-3), 234–269.

Guidry, R., & Patten, D. (2012). Voluntary disclosure theory and financial control variables: An assessment of recent environmental disclosure research. *Accounting Forum*, *36*, 81–90.

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, *40*(1), 75–87.

Hoberg, G., & Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Review of Financial Studies*, *23*(10), 3773–3811.

Hu, N., Liang, P., & Yang, X. (2023). Whetting All Your Appetites for Financial Tasks with One Meal from GPT? A Comparison of GPT, FinBERT, and Dictionaries in Evaluating Sentiment Analysis. https://ssrn.com/abstract=4426455

Huang, A., Lehavy, R., Zang, A., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, *64*(6), 2833–2855.

Huang, A., Wang, H., & Yang, Y. (2022). FinBERT—A large language model for extracting textual information from financial text [forthcoming]. *Contemporary Accounting Research*.

Huang, X., S., T., & Zhang, Y. (2014). Tone Management. *The Accounting Review*, *89*(3), 1083–1113.

IDW Verlag. (2022a). Entwurf eines IDW Prüfungsstandards: Inhaltliche Prüfung mit begrenzter Sicherheit der nichtfinanziellen (Konzern-)Berichterstattung außerhalb der Abschlussprüfung (IDW EPS 991 (11.2022)). https://www.idw.de/IDW/IDW-Verlautbarungen/IDW-PS/EPS-991-11-2022.pdf

IDW Verlag. (2022b). Entwurf eines IDW Prüfungsstandards: Inhaltliche Prüfung mit hinreichender Sicherheit der nichtfinanziellen (Konzern-)Berichterstattung außerhalb der Abschlussprüfung (IDW EPS 990 (11.2022)). https://www.idw.de/IDW/IDW-Verlautbarungen/IDW-PS/EPS-990-11-2022.pdf

International Auditing and Assurance Standards Board. (2013). ISAE 3000 (Revised), Assurance Engagements Other than Audits or Reviews of Historical Financial Information. International Framework for

Assurance Engagements and Related Conforming Amendments. https://www.ifac.org/_flysystem/azure-private/publications/files/ISAE%203000%20Revised%20-%20for%20IAASB.pdf

International Sustainability Standards Board. (2022a). [Draft] IFRS S1 General Requirements for Disclosure of Sustainability-related Financial Information. https://www.ifrs.org/content/dam/ifrs/project/general-sustainability-related-disclosures/exposure-draft-ifrs-s1-general-requirements-for-disclosure-of-sustainability-related-financial-information.pdf

International Sustainability Standards Board. (2022b). [Draft] IFRS S2 Climate-related Disclosures. https://www.ifrs.org/content/dam/ifrs/project/climate-related-disclosures/issb-exposure-draft-2022-2-climate-related-disclosures.pdf

Kim, A., Muhn, M., & Nikolaev, V. (2023). Bloated Disclosures: Can ChatGPT Help Investors Process Information? https://ssrn.com/abstract=4425527

King, R. (2007). *The Regulatory State in an Age of Governance*. Palgrave Macmillan.

Kothari, S., Li, X., & Short, J. (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, *84*(5), 1639–1670.

KPMG. (2022). Big shifts, small steps. Survey of Sustainability Reporting 2022. https://assets.kpmg.com/content/dam/kpmg/sg/pdf/2022/10/ssr-small-steps-big-shifts.pdf

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied Linear Statistic Models* (Fifth). McGraw-Hill Irwin.

Lang, M., & Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, *60*, 110–135.

Lawrence, A. (2013). Individual Investors and Financial Disclosure. *Journal of Accounting and Economics*, *56*, 130–147.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lepore, L., & Pisano, S. (2023). *Environmental disclosure. Critical issues and new trends*. Routledge.

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

Long, C. S., Lucey, B., Xie, Y., & Yarovaya, L. (2023). “I just like the stock”: The role of Reddit sentiment in the GameStop share rally. *Financial Review*, *58*, 19–37.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, *66*, 35–65.

Loughran, T., & McDonald, B. (2014). Measuring Readability in Financial Disclosures. *The Journal of Finance*, *69*, 1643–1671.

Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*, 1187–1230.

Loughran, T., & McDonald, B. (2020). Textual Analysis in Finance. http://dx.doi.org/10.2139/ssrn.3470272

Lucarelli, C., Mazzoli, C., Rancan, M., & Severini, S. (2020). Classification of sustainable activities: EU taxonomy and scientific literature. *Sustainability*, *12*(16), 6460.

Madrazo, A., & Pera, M. (2020). Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, *71*(6), 644–656.

Martinc, M., Pollak, S., & Robnik-Sikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, *47*(1), 141–179.

McGurk, Z., Nowak, A., & Hall, J. (2020). Stock returns and investor sentiment: textual analysis and social media. *Journal of Economics and Finance*, *44*, 458–485.

Miller, B. (2010). The Effects of Reporting Complexity on Small and Large Investor Trading. *The Accounting Review*, *85*, 2107–2143.

Nagy, W., & Anderson, R. (1984). How Many Words Are There in Printed School English? *Reading Research Quarterly*, *19*(3), 304–330.

Porter, M. (1980). An Algorithm for Suffix Stripping. *Programm*, *14*(3), 130–137.

PricewaterhouseCoopers. (2023). EU Taxonomy Reporting 2023. Analysis of the financial and non-financial sector. https://www.pwc.de/de/content/20e6bff9-ea5a-4d03-b375-a6a58f7b8b46/pwc-eu-taxonomy-reporting-2023.pdf

Purda, L., & Skillicorn, D. (2015). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research*, *32*, 1193–1223.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding with unsupervised learning* (tech. rep.). OpenAI.

Reuters. (2021). Regulatory Intelligence. Country Update-France: ESG Reporting. https://www.gide.com/sites/default/files/2021_06_17_france_esg_guide.pdf

Reuters. (2023). ChatGPT’s explosive growth shows first decline in traffic since launch. https://www.reuters.com/technology/booming-traffic-openais-chatgpt-posts-first-ever-monthly-dip-june-similarweb-2023-07-05/

Sautner, Z., van Lent, L., Vilkov, G., & Zhang, R. (2023). Firm-Level Climate Change Exposure. *Journal of Finance*, *78*, 1449–1498.

Siano, F., & Wysocki, P. (2021). Transfer learning and textual analysis of accounting disclosures: Applying big data methods to small(er) data sets. *Accounting Horizons*, *35*(3), 217–244.

Sorrosal-Forradellas, M., Barbera-Marine, M., Fabregat-Aibar, L., & Li, X. (2023). A new rating of sustainability based on the Morningstar Sustainability Rating. *European Research on Management and Business Economics*, *29*, 1–6.

Technical Readiness Working Group (IFRS Foundation). (2021). General Requirements for Disclosure of Sustainability-related Financial Information Prototype. https://www.ifrs.org/content/dam/ifrs/groups/trwg/trwg-general-requirements-prototype.pdf

Tetlock, P., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms’ fundamentals. *Journal of Finance*, *63*, 1437–1467.

Tworzydło, D., Gawroński, S., Opolska-Bielańska, A., & Lach, M. (2022). Changes in the demand for CSR activities and stakeholder engagement based on research conducted among public relations specialists in Poland, with consideration of the SARS-COV-2 pandemic. *Corporate Social Responsibility and Environmental Management*, *29*(1), 135–145.

United Nations. (2022). Global Issues. Climate Change. https://www.un.org/en/global-issues/climate-change

United States Securities and Exchange Commission (SEC). (1998). A Plain English Handbook. How to create clear SEC disclosure documents. https://www.sec.gov/pdf/handbook.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*.

Velte, P. (2023). Prüfung von Nachhaltigkeitsberichten nach der Corporate Sustainability Reporting Directive (CSRD) durch den Wirtschaftsprüfer – Fluch oder Segen? *Schmalenbach Impulse*, *3*(1), 1–13.

Verrecchia, R. (1983). Discretionary Disclosure. *Journal of Accounting and Economics*, *5*, 179–194.

Wagenhofer, A., & Ewert, R. (2015). *Externe Unternehmensrechnung* (Third). Springer Gabler.

Walker, C. (2014). Organizational Learning: The Role of Third Party Auditors in Building Compliance and Enforcement Capability. *International Journal of Auditing*, *18*, 213–222.

Widmann, M., Follert, F., & Wolz, M. (2021). What is it going to cost? Empirical evidence from a systematic literature review of audit fee determinants. *Management Review Quarterly*, *71*, 455–489.

Zaman, M., Hudaib, M., & Haniffa, R. (2011). Corporate Governance Quality, Audit Fees and Non-Audit Services Fees. *Journal of Business Finance & Accounting*, *38*, 165–197.

Zamil, I., Ramakrishnan, S., Jamal, N., Hatif, M., & Khatib, S. (2023). Drivers of corporate voluntary disclosure: a systematic review. *Journal of Financial Reporting and Accounting*, *21*(2), 232–267.

Zhou, Z. (2021). Introduction. In *Machine Learning*. Springer.