Norwegian School of Economics
Bergen, Fall 2022

Is there an ESG anomaly? Using textual analysis to detect market mispricing

HENRIK FIRING MARTINSEN and BJØRN-TERJE KONGSNES ANDREASSEN

SUPERVISOR: CARSTEN BIENZ

Master Thesis, Economics and Business Administration
Major: Financial Economics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

This thesis investigates the relationship between stock returns and ESG scores on primary listed firms on the three major Scandinavian stock exchanges between 2017 and 2021. In conducting the analysis, we have collected ESG scores from Refinitiv and designed two scores based on textual analysis of the companies' annual reports from 2015 to 2019. A two-year lag is applied between the ESG scores and the financial data to ensure that ESG information is known before the returns are produced, and the connection can be explained. We analyze the difference in stock returns on firms with high and low ESG score by constructing long-short portfolios. Then, we apply the Fama-French three-factor and five-factor models to control for different risk exposures. The ESG-based portfolio shows no abnormal return using either ESG scores, which suggests that the ESG risk is priced in the market.

Keywords: ESG, sustainability, asset pricing, textual analysis, annual reports, Scandinavia

Acknowledgement

This thesis is part of the master's degree in *Economics and Business Administration* at the Norwegian School of Economics.

Our shared interest in sustainability and finance led us to choose this topic. It therefore felt like the ideal opportunity to investigate the possibilities in a field that is fascinating, and determine how we could contribute to the existing literature. Writing this thesis has been challenging and educational. The fact that the topic is very current and relevant today, has been a great motivation.

We would like to thank our supervisor Carsten Bienz for advice and valuable feedback, during this process. Furthermore, we would also like to thank family, colleagues, and acquaintances for inspiring and valuable conversations on sustainable investing.

Norwegian School of Economics

Bergen, December 2022

Henrik F. Martinsen

Bjørn-Terje K. Andreassen

Table of Contents

List of Figures

List of Tables

1. Introduction

"Humanity is on a highway to climate hell" as UN's *General Secretary*, António Guterres, phrased the climate reality during the COP27 meeting in Egypt earlier this year (Harvey, 2022). According to the UN is it more important than ever to focus on sustainability (2022). Society and politics have become more influenced by the ESG trend. The terms *environmental*, *social*, and *governance* have been mentioned to exhaustion during the last few years. CEOs of the S&P 500 companies now mention ESG nine times a quarter in earning calls, on average (Economist, 2022). Compare this to 2017, where investors could hear the term once if they were lucky. The rapidly increased focus on firms' ESG position and performance has laid a foundation for a relatively new market, ESG-score providers. Firms that analyze and sell ESG scorings on companies has increased in numbers over the last years (Tayan, 2022).

One of the tremendous problems with the ESG score providers today is the lack of transparency and divergence between the different scores. Another problem is that there are a limited number of companies that receive these scores, mostly the largest listed companies. Berg et al. (2022) investigates divergence between the ESG scores and calls for more transparency from the providers to lower the uncertainty of the ratings, as risk is linked to the precision of the scores. The lack of transparency related to the actual data they use in their analysis is a significant problem because one cannot follow the calculations to see what the firm weighs the most. Uncertainty increases an investor's demand for return, which drives up the cost of capital and ultimately results in less investment into *greener* firms than it could be. This is a substantial problem for society in general and the environment because it does not reallocate capital towards greener firms in the most efficiently manner.

Textual analysis has existed for years and been a valuable tool to apply to different types of analysis to manage large data, for instance how negative publication in newspapers affect different firms' stock prices. In this thesis we will investigate if textual analysis could be applied to expand the ESG score universe. We test to see if textual analysis is superior, complementary or superfluous compared to the ESG scores provided by Refinitiv. To test for this, long-short portfolios are created for both Refinitiv and the textual scores.

1.1. Procedure

The first step in the process of testing the applicability of textual analysis on the ESG scores, is to create our sample. In this instance the geographical area is restricted to Scandinavia, and their respective stock exchanges. An advantage with the textual approach compared to Refinitiv's, is that one can simply download as many annual reports as one wants, and therefore create more ESG scores than Refinitiv can produce. In this thesis the sample is limited to the same Scandinavian firms that Refinitiv has assessed to determine whether the textual approach is better, compatible with, or worse than the current method of computing the scores.

To evaluate a company's ESG performance, three scores are used: The first is based on Refinitiv's ESG scores, and the other two are produced using ESG term weights. Term weights measure the frequency of ESG words in annual reports. Companies with a high ESG score have many ESG words in their annual reports. All the available annual reports of the relevant companies are manually obtained from their official websites and organized into folders for each country to calculate the two textual ratings. The downloaded annual reports cover the years 2015 through 2019, ensuring that the sample contains up to five years' worth of annual reports for each company.

The ESG dictionary created by Baier et al. (2020) is used to determine which words are considered ESG words. Term weights are most frequently employed in textual analysis, are also applied in this thesis. This includes term frequency (tf), which Loughran & McDonald (2011) refer to as proportional weights, and term frequency - inverse document frequency (tf.idf). Formulas for the term weights vary, hence the ones used in this thesis are explained more in detail later.

Financial data of firms on the Scandinavian stock exchanges are downloaded and later merged with the Refinitiv and the two textual ESG scores. The timeframe applied is 2015 to 2019 for the annual reports and 2017 to 2021 for the stock returns. A two-year lag is incorporated because of the assumption that investors incorporate ESG information from two years before in their assessment of stocks.

We analyze the ESG implications by creating a long-short ESG portfolio similar to Lioui & Tarelli (2022), where we go long in top ESG firms and short in bottom ESG firms. To create the long-short portfolios, the *time*-series approach developed by Fama & French (1993) is used. It consists of sorting portfolios by a sorting-variable into different breakpoints (Bali, Engle, & Murray, 2016). We sort Scandinavian stocks into quintiles, because it leaves each portfolio with a reasonable number of shares. If the sample is sorted into deciles instead, for instance, it would have been too few observations in each portfolio. The three separate ESG scores are applied as the sorting variables. We weight companies equally and according to their market capitalization when calculating portfolio returns. Each portfolio is regressed on the Fama-French three-factor model (1993) and the Fama-French five-factor model (2015). If there are abnormal returns to be made, there should be a significant alpha in one of the regressions.

We discover that neither regressions exhibit an ESG anomaly and have significant alpha. This means that it is not possible to trade on ESG information, since it is already priced in the Scandinavian markets. Although the findings were underwhelming, this thesis offers an intriguing approach to apply textual analysis in the context of ESG investing. It demonstrates that constructing ESG scores based on textual analysis is a viable alternative to using ESG scores from a renowned ESG score supplier such as Refinitiv. The advantage with the textual approach is that one can expand the ESG universe, but is implementing and adopting textual analysis worth the effort?

1.2 Related literature

Textual analysis has been applied to many different papers before and is a well renowned method to manage and analyze large amounts of data consisting of words for instance annual reports or 10-K filings, press releases or news. One of the few publications that use textual analysis in the context of ESG investing is Engle et al. (2020) and Ardia et al. (2022). They study US financial markets and use textual analysis on media's attention to climate change. Our thesis uses an approach that is closer to Loughran & McDonald (2011), who analyze the textual context of 10-Ks using a dictionary specific to the business domain. The difference is that we use a dictionary that contains ESG words rather than business semantics.

McDonald et al. (2009) is also among the most relevant textual analysis papers. In this paper the authors find a connection between the firms' number of ethically charged words in their 10-K filings - and their chances of getting sued. They find that the companies with the largest chance of getting sued, *bad firms,* often tend to apply more ethical words in their 10-K filings than the *good firms* – which can suggest that the *bad firms* try to "greenwash" their reputation. This is a possibility one must consider when analyzing the annual reports for EGS terms. In an effort to circumvent this issue, we include *tf.idf* weights as it rewards uniqueness of the ESG words.

This thesis contributes to the existing ESG literature. Pástor, Stambaugh & Taylor (2022) construct a green factor that goes long green and short brown stocks, where stocks are weighted by their greenness. Their focus is on the "E" in ESG, basing the proxy on MSCI's environmental score. They find a significant and positive alpha with respect to the Fama-French three-factor-model. The same results emerge from the green-minus-brown portfolio, which is value-weighted instead of greenness-weighted. Green assets strongly outperform brown assets, even when controlling for the five Fama-French factors (2015), Carhart's momentum factor (1997), the traded liquidity factor produced by Pástor & Stambaugh (2003) and the factors of Hou et al. (2015; 2021). In the regressions, Pástor et al. (2022) find that GMB tilts toward large stocks, consistent with our results.

The positive GMB alpha in Pástor et al. (2022) appears to contradict with our results. However, one has to bear in mind that they use a different timeframe (2012-2020), a different cross-section of securities (they analyze US stocks) and a different proxy for ESG (the MSCI's environmental score). Moreover, Pástor et al. (2022) find that the green outperformance disappears after controlling for climate related shocks. Green stocks typically outperform brown when climate concerns increase. This result is similar to the findings of Choi et al. (2020), Engle et al. (2020), and Ardia et al. (2022). However, when climate shocks are accounted for, in addition to earnings shocks, the green premium becomes slightly negative. Pástor et al. (2022) also analyze green-minus-brown ex ante and find that the green premium is consistently negative throughout the sample period.

Avramov, et al. (2022) get the same empirical results as this thesis. They find that ESG ratings are negatively associated with future performance. However, this is only the case when there is low ESG rating uncertainty. They use dependent portfolio sorts; stocks are first sorted into quintiles according to rating uncertainty, and then within each group further sorted into quintiles according to their ESG rating. The predictability of the ESG rating is weak when there is ESG rating uncertainty. Univariate high-minus-low portfolios of ESG rating uncertainty also show negative and significant CAPM alpha.

Pedersen, Fitzgibbons and Pomorski (2021) analyze the cross-sectional relation between returns and each ESG pillar as well as the complete ESG score (MSCI). They sort stocks into quintiles for each month, and then form a portfolio that goes long in top ESG stocks and short in bottom ESG stocks. Only the portfolio based on *G* has abnormal returns after controlling for the Fama-French (2015) factors augmented with momentum. The alpha is consistent for both value-weighted and equal-weighted portfolios. Pedersen et al. (2021) find little or weak evidence for outperformance in the *E* and overall *ESG* score. The *S* produces some abnormal returns, but only for value-weighted portfolios regressed on CAPM and the Fama-French three-factor model.

Bolton and Kacperczyk (2021; 2022) study whether there is a carbon premium in the US financial markets between 2005 and 2017. They find that there is significant alpha after adjusting for several standard risk factors including *MKTRF*, *HML*, *SMB*, *MOM*, *CMA*, *BAB*[1], *LIQ*, *NET ISSUANCE*[2] and *IDIO VOL*[3] in the case of total emissions and year-on-year growth rates in total emissions. Their results show that the carbon premium contains independent information about the cross-section of average returns that cannot be explained by known risk factors. In Bolton and Kacperczyk (2022), they find that the premium, in relation to emissions growth, is higher in countries with lower economic development, larger energy sectors, and less inclusive political systems. Premia related to total emissions are higher in countries with stricter domestic climate policies. Bolton and Kacperczyk (2021) do not find any carbon premium for emission intensity. They apply the Fama-MacBeth approach, and regress returns on carbon emissions in the cross-section before they run time-series regressions of the carbon premium on the standard risk-factors.

Hsu, Li, and Tsou (2022) research abnormalities in the long-short portfolio constructed from firms with high versus low toxic emission intensity. They find that there are abnormal returns to be made within a given industry, which cannot be explained by several explanations, including existing systematic risks, investors' preference, market sentiment, political connections, and corporate governance (Hsu, Li, & Chi-Yang, 2022). However, when they

[1] BAB is the return of portfolio that is long on low-beta stocks and short on high-beta stocks
[2] NET ISSUANCE is the return of a portfolio that is long on high-net-issuance stocks and short on low-net-issuance stocks; net issuance is the change in split-adjusted shares outstanding from one year to another.
[3] IDIO VOL is the return of a portfolio that is long on low idiosyncratic volatility stocks and short on high idiosyncratic volatility stocks.

control for environmental litigation penalties to measure regime change risk, they find that it helps price the cross-section of emission portfolios' returns. This suggests that investors require a pollution premium in order to be exposed to systematic risk related to environmental policy uncertainty.

Lins, Servaes and Tamayo (2017) illustrate that firms that score high on CSR earn higher returns and are less risky than other firms, because trust and social capital helps them survive a downturn such as a global financial crisis. They use CSR data from MSCI ESG STATS database, and analyze US returns between August 2008 and March 2009. Several variables are controlled for, including financial health, firm characteristics and corporate governance. The CSR-outperformance is robust after the adjustments, although weaker. The model of Lins et al. (2017) is also re-estimated after splitting CSR into quartiles, and it is found that returns increase the most when moving from the lowest to the second lowest quartile. This indicates that investors are most concerned when trust and social capital is very low.

Existing literature suggest that Scandinavian investors might be less sensitive to climate related news. We do not control for climate concerns, but still get a negative return spread, although there is no significant alpha. This could suggest that the predictions of Pástor et al. (2021) has already happened in Scandinavia; the demand for green assets has increased, and the cost of capital has gone down because of investors' preferences towards green assets and their climate hedging capabilities. Another explanation could be that the ESG rating uncertainty is low, which causes the ESG premium to disappear. Nevertheless, similar to Pedersen et al. (2021) we find that the *G* in ESG has the highest return predictability; when applying *tf.idf* weights on ESG words, the governance words dominate the textual ESG score. Simultaneously, environmental words are more prominent when applying the ESG score with tf weights. The ESG *tf.idf* score is better described by the Fama-French models than the ESG tf score.

The thesis proceeds as follows: the first part addresses the data, describing how this was collected as well as how it was cleaned. In the second part we address the ESG scores we have used: the Refinitiv scores as well as the scores based on textual analysis. This is followed by the analysis of the ESG long-short portfolio returns. In the fourth part, we discuss the results and disclose what we have discovered compared to existing literature, as well as addressing the limitations of our model. Lastly, we conclude on our findings and stress why it is beneficial to use textual analysis when constructing a portfolio based on ESG rather than using one of the existing ESG providers' scores.

2. Data

The focus of this thesis will rely on returns from 2017 and 2021 and the respective annual reports published between 2015 and 2019, allowing for a two-year lag between the release of an annual report and the realization of returns. This is because Refinitiv has produced more ESG scores in the most recent years. Companies listed on the Norwegian, Swedish, and Danish stock exchanges with an ESG score in the Refinitiv database are included in the research. The purpose of this is to determine whether a company's performance and ESG focus may be more accurately estimated by textual analysis of annual reports.

Table 1: Sample Creation

This table shows how the annual reports sample and the securities sample was created. The data filters are described on the left-hand side and the sample sizes are shown in the middle. The number of observations removed is on the right-hand side.

Source	*Sample Size*	*Observations Removed*
Text Data		
Annual reports 2015-2019 downloaded complete sample	1,289	
After tokenization	1,288	1
Remove observations that do not match COMPUSTAT data	1,052	236
Filter out reports with less than 2,000 words	970	82
Number of words in tidy text data	5,325,620	
Remove stop words	4,796,190	529,430
Merge with LM11 master dictionary	3,558,837	1,237,353
Remove observations that do not match COMPUSTAT data	2,917,981	640,856
Filter out reports with less than 2,000 words	2,796,890	121,091
Number of firms represented in the final sample	224	
COMPUSTAT Data		

Source	*Sample Size*	*Observations Removed*
Monthly observations for Norway, Sweden, and Denmark	112,258	
Remove duplicates and arrange trade dates	74,531	37,727
Missing returns or market capitalization	72,816	1,715
Listed on Oslo, Copenhagen, or Stockholm exchange	62,306	10,510
Remove financial firms	55,191	7,115
Number of securities in the complete sample	1,832	
Remove duplicates and arrange trade dates	1,711	121
Missing returns or market capitalization	1,670	41
Listed on Oslo, Copenhagen, or Stockholm exchange	1,432	238
Remove financial firms	1,283	149

2.1 Textual Data Wrangling

1,289 annual reports are downloaded manually directly from the official company-websites. To ensure the two-year lag between the returns data and the ESG data, we download every report from 2015 to 2019. This approach ensures that ESG metrics are known before the returns they are used to explain (Fama & French, 2012). To clarify, annual reports for all fiscal years $t-2$ are matched with returns for January of year t to December of year t. This thesis does not follow Fama & French's 6-month approach, as several companies in the data sample have fiscal years between July 1 of year $t-2$ and June 30 of year $t-1$. This ensure that every type of investor is then able to evaluate the ESG performance of a company before investing in it.

Older annual reports for certain companies are not accessible on their investor relations page, particularly among those listed on the Swedish stock exchange. As a consequence of this, we can only download the most current annual reports from these businesses. The format of some of the annual reports also has flaws. Several of the 2015 and 2016 annual reports were scanned documents and could not be included in the textual analysis. Additionally, certain businesses do not provide an English version of their annual report, especially for the earlier years within the relevant timespan. The ESG dictionary that we will apply for the textual analysis is written in English, hence this presents an issue.

Making a dictionary in Norwegian, Swedish and Danish is possible, but could lead to other problems. When documents are written in one of three different languages, the same ESG words will not appear across different countries. Since, the number of firms in each industry vary by country, ESG materiality will also vary, making it difficult to compare ESG scores. This is the case regardless of whether one uses parallel corpora or a bilingual dictionary. One example is *hydro*, which is material in Denmark, but not in Norway. A workaround could be to translate every annual report into one language, for instance Norwegian. However, it would be time-consuming to do this properly.

Not all companies have English annual reports, but this is fine. Ignoring a few of the companies' annual reports reduces the data foundation, but it must be emphasized that the selection in this dataset is quite large. After removing firms that are not primary listed in Oslo, Stockholm, or Copenhagen, and that have no returns data, there are *970* annual reports left. Hence, it should not have a damaging effect on our results. Companies that are not primary-listed on one of the three major exchanges are removed because reporting standards are either different or disclosure requirements are less stringent. Annual reports with less than *2,000* words are excluded, following Loughran & McDonald (2011).

Before the appropriate data is filtered, it is converted into a tidy text format. The tidy text format is a table with a one-token-per-row structure (Silge & Robinson, 2022). The token is a meaningful unit of text and in this case it is a single word, but it could also be a sentence. Nonetheless, we want a table that displays the name of the company and the year of the annual report in the left column, and the word that appears in the annual report in the right column. This is done through tokenization which is essentially splitting text into tokens. This is easily done with an R function developed by Silge & Robinson (2022) in the *tidytext* package. By default, this function strips punctuation and converts tokens into lower case.

Numbers are removed to only keep the textual data. This operation is done with regular expressions using the *stringr* package developed by Wickham (2022). To complete the tidy text transformation, the words appearing in the annual reports are counted by applying a simple function in dplyr, *count*. After it is filtered, the text is ready to be visualized. The figure below summarizes each step in the tidy text transformation (**Figure 1**).

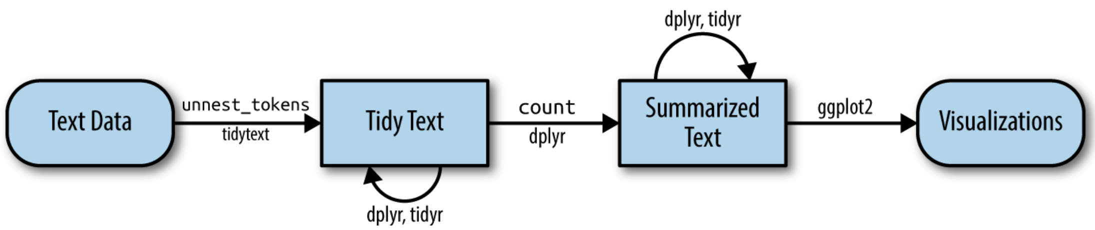

Figure 1: The tidy text flowchart retrieved from Silge & Robinson (2022).

One problem with the *tokenization* function is that some companies are incorrectly tokenized. This became apparent after a manual control was conducted by scrolling through the words. In order to analyze words in a more sophisticated way, we apply a R package called *hunspell*, developed by Ooms (2022). This package allows to extract all tokens that were misspelled or not found in an *American English lexicon*. A few of the misspelled tokens were words written in *British English*. Other tokens were words that made no sense or several words that were joined together with no space or hyphen in-between. Some tokens were also abbreviations, and multiple words were names of places or people.

There were around *80,000* observations of these "misspellings". It would be too time-consuming to manually go through each of these observations. To solve this issue, the textual dataset was merged with an American English lexicon instead. More specifically, it was merged with the Loughran & McDonald (2011) master dictionary (LM11) as this contains a business-related jargon. Every word that did not match this dictionary was stripped. An advantage with this approach is that it was timesaving, and we did not need to remove stop words as they were not included in the master dictionary. The disadvantage, however, was that British words were removed, and some companies may prefer to write in British English.

2.2 Securities Sample

First step is importing the *Compustat Global Securities Data* from Wharton Research Data Services. The variables selected are the stock's international security *ID* (i.e. ISIN), the trade date, the company name, the *GIC* industry that the company belongs to, the country of incorporation, the number of shares outstanding, the daily closing price, the end of month indicator, the stock exchange code, the adjustment factor, and the daily total return factor.

Trade dates between 2017 and 2021 are selected. Swedish, Norwegian, and Danish stocks are filtered using the country variable. The stock exchange code is used to identify the name of the stock exchanges. This code is important, as we exclude companies that are not primary-listed on the three major exchanges: Nasdaq Stockholm, Euronext Oslo and Nasdaq Copenhagen. The month-end indicator is used to calculate monthly returns and the adjustment factor adjusts the stock price for stock splits. By dividing the closing price by the adjustment factor, one can determine the adjusted closing price. The monthly returns also include dividends and cash equivalent distributions through the daily total return factor.

In order to get comparable market capitalization numbers, foreign exchange rates from the European Central Bank Statistical Data Warehouse are imported. This is done with the *priceR* package in R, made by Condylios (2022). It allows for converting all market capitalization numbers into *euros*. Market capitalization is calculated by taking the product of the monthly closing price and the number of shares outstanding. Fama and French (1992) suggest excluding financial firms from the factor regressions, as they often differ in terms of leverage and capital structure. Hence, this is also done in this paper. After the removal of the missing values and duplicates from the dataset, we are left with 55,191 observations.

Both the Fama-French three-factor model and the Fama-French five-factor model for the European markets are applied. The R package developed by Areal (2021) allows for reading all financial data from Kenneth French's data library. The returns are calculated as specified by Fama & French (2012). Similar to the Compustat dataset, the dates fall between 2017 and 2021.

2.3 ESG Dictionary

The dictionary that is used to classify words as ESG related or not, is Baier, Berninger & Kiesel (2020). It consists mostly of unigrams and separates words into each ESG pillar: *environmental*, *social* and *governance*. Within each pillar, words are assigned to a specific category and subcategory. This allows for a comprehensive analysis of ESG related words. The dictionary is formed based on 10-K reports and proxy statements of the 25 largest firms in the US in a four-year period (Baier, Berninger, & Kiesel, 2020). Personal judgment is used to extract words that appear in an ESG related context in the majority of its occurrences. A word must appear in at least 5% of all annual reports to be considered. The ESG dictionary is downloaded from Florian Kiesel's personal Google site. It is merged with the tidy text dataset to create a data frame that only shows the count of ESG related words in annual reports. This makes it possible to make effective visualizations and informative tables.

Figure 2: Most frequently mentioned ESG words in companies' annual reports.

This figure shows the ESG words that are most frequently mentioned in companies' annual reports. The different colors represent the three ESG pillars. Green is *environmental*, grey is *social,* and orange is *governance*.

As presented in the figure above (**Figure 2**), the most frequently mentioned words in annual reports are *audit, pension, remuneration, control, governance*, and *sustainability*. Most of these words belong to the governance pillar. Governance related words often occur in the annual reports of the Norwegian, Swedish and Danish firms because they are required by law to disclose their corporate governance practices. The dominance of governance related words is also highlighted in Baier et al. (2020), which states that governance has been on the companies' agenda for a long time. *Sustainability* does not belong to the governance pillar though; it belongs to the environmental pillar. Despite this, it is conceivable that it has a high word count as more companies report on their sustainability practices. Although there is no legal requirement for this, stakeholders have for a long time expected firms to report on climate related risks. However, with the extension of the current EU directive on the disclosure of non-financial information, firms are likely to face stricter requirements (Stehl, 2022).

Table 2: Most frequent ESG words in annual reports by pillar and category.

This table shows the most frequent ESG words in companies' annual reports by pillar and category. *Count* is the word frequency in each pillar and category across all companies and years. Percent is the share of ESG words in a specific pillar or category on total ESG words.

	Count	*Percent*
ESG	1,333,947	100.00%
Governance	863,011	64.70%
Corporate Governance	713,306	53.47%
Business Ethics	34,070	2.55%
Sustainability Management and Reporting	89,186	6.69%
Social	244,785	18.35%
Public Health	38,062	2.85%
Human Rights	14,938	1.12%
Labor Standards	74,881	5.61%
Society	80,582	6.04%
Environmental	226,151	16.95%
Climate Change	65,045	4.88%
Ecosystem Service	14,288	1.07%
Environmental Management	41,302	3.10%

Table *2* above confirms that the ESG words that occur most often are words categorized as governance. Almost 65% of all ESG words are governance related. The share of words that belong to the corporate governance category is 53.47%, which is remarkably high. It signalizes that corporate governance accounts for more than 80% of governance related words. Remarkably few words relate to business ethics. The rest of governance related words are mostly about sustainability management and reporting, which is no huge surprise. Stakeholders are increasingly more concerned about sustainability disclosure.

When it comes to *social*, words about *public health* and *human rights* are less frequently mentioned. *Public health* concerns itself with epidemics, pandemics, illness, and childbirth, whilst *human rights* cover topics including discrimination, quality, and freedom (Baier, Berninger, & Kiesel, 2020). *Social* words are more frequently used in the context of labor standards and philanthropy. This is logical, as trade unions are strong in Scandinavia and corporate philanthropy is a well-established phenomenon. *Environmental* words, on the other hand, are mentioned less frequently than *social* words. This is quite unexpected because of the media coverage and focus on environmental challenges. Climate change is the most prevalent category in the environmental pillar. Less words occur in relation to ecosystem service and environmental management.

Table 3: The distribution of ESG words across industries.

This table presents the distribution of ESG words across industries. Both the share of total ESG words and the share of total words in each pillar is presented.

Industry	*ESG*	*Environmental*	*Social*	*Governance*
Communication Services	4.18%	0.49%	0.80%	2.89%
Consumer Discretionary	12.46%	1.77%	2.35%	8.34%
Consumer Staples	9.65%	1.66%	2.05%	5.95%
Energy	3.70%	0.36%	0.51%	2.83%
Health Care	10.70%	0.73%	2.54%	7.44%
Industrials	33.72%	5.96%	5.83%	21.92%
Information Technology	5.57%	0.55%	0.88%	4.14%
Materials	10.73%	3.11%	1.95%	5.67%
Real Estate	7.10%	1.57%	1.21%	4.32%
Utilities	2.19%	0.75%	0.24%	1.19%

The table above (**Table 3**) shows that ESG related words are most frequently reported by businesses in the industrial sector. This is no coincidence, as the securities sample is dominated by firms operating in *industrials*, *information technology* and *health care*. The *information technology* sector reports ESG words rarely compared to the other sectors though. Only 5.57% of all ESG words are mentioned by firms operating in the IT sector. One reason could be that stakeholders perceive IT firms as more sustainable. They often monetize software, which is not very capital-intensive. Nevertheless, it is simple to overlook that what happens in "the cloud" actually takes place in a physical data center (Haga, 2022). The emissions that these centers produce is currently on par with air traffic, which is about two percent of global greenhouse gas emissions.

Health care has a high proportion of all ESG related words mentioned in annual reports. Interestingly, most of this is due to governance related words. The same applies to consumer staples and consumer discretionary. An interesting remark is the fact that consumer staples accounts for a small number of firms in the data sample. Consumer staples constitutes less than half as many as the number of firms in consumer discretionary. Despite that, both sectors mention ESG words with nearly equal frequencies. Consumer staples also reports substantially on social, second only to industrials. Materials is another sector that frequently mention social words, and it constitutes about the same number of firms as consumer staples in the data sample.

Table 4: The distribution of ESG words by *pillar* and category from 2015 to 2019.

This table shows the distribution of ESG words by pillar and category from 2015 to 2019. The percentages are the number of ESG words in one pillar or category divided by the total ESG words in the sample.

	2015	*2016*	*2017*	*2018*	*2019*
ESG	13.78%	17.09%	20.03%	22.66%	26.44%
Governance	9.28%	11.56%	13.14%	14.45%	16.26%
Corporate Governance	7.74%	9.56%	10.83%	11.93%	13.42%
Business Ethics	0.31%	0.43%	0.54%	0.59%	0.68%
Sustainability Management and Reporting	0.94%	1.22%	1.38%	1.49%	1.66%
Social	2.52%	3.04%	3.67%	4.20%	4.93%
Public Health	0.39%	0.46%	0.55%	0.66%	0.79%
Human Rights	0.14%	0.17%	0.24%	0.25%	0.31%
Labor Standards	0.73%	0.90%	1.14%	1.31%	1.54%
Society	0.85%	1.03%	1.19%	1.37%	1.61%
Environmental	1.99%	2.49%	3.22%	4.01%	5.25%
Climate Change	0.55%	0.68%	0.90%	1.15%	1.60%
Ecosystem Service	0.14%	0.17%	0.22%	0.24%	0.29%
Environmental Management	0.41%	0.49%	0.60%	0.71%	0.88%

As presented in **Table 4**, the share of ESG words on total words in annual reports has increased from 2015 to 2019. This is due to even more governance related words, particularly corporate governance. There could be several explanations for why this is the case. For example, annual reports could be getting longer or there could be more companies that are being listed on the three stock exchanges. If the former is the case, the quality of corporate governance reporting may be improving. It is possible that companies are choosing higher transparency rather than pure compliance. There is a noticeable increase in the share of social and environmental words.

However, *social,* and *environmental* words still comprise a small percentage of total words. This could be due to less regulatory scrutiny. The findings suggest a large increase in words related to tackling climate change amongst the Scandinavian listed companies, with an increase of almost 200% from 2015 to 2019.

3. Method

3.1 Refinitiv's ESG Score

The ESG score provided by Refinitiv is highly used by finance practitioners and by finance academicians alike. It is often used to measure ESG risk and the correlation with stock returns. The ESG score is based on over 630 measures on ESG performance which are grouped into 10 categories. These 10 categories are bundled into the three ESG pillar scores: *environmental*, *social* and *governance*. The ESG score is a result of a weighted sum of the *E*, *S* and *G* scores. In the *governance* pillar, the weights are constant, but depending on the industry, they can change in the *environmental* and *social pillars*. Refinitiv's approach to ESG performance is illustrated in the **Figure 3** below.

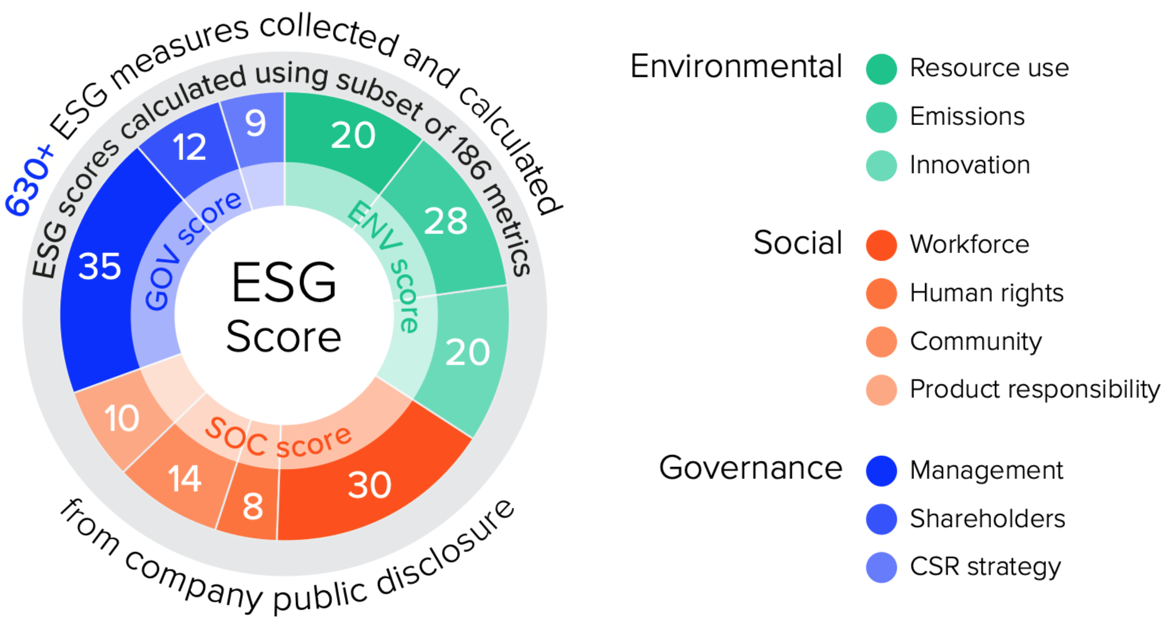

Figure 3: This figure shows the ESG categories that is used by Refinitiv to calculate ESG scores.

Refinitiv has received widespread criticism due to the divergence in ratings compared to other providers. Yet, this is not a firm-specific challenge. All rating providers use different methodologies, causing ESG ratings to greatly diverge. Researchers such as Berg, Kölbel & Rigobon (2022) call for greater transparency on how ESG ratings are calculated. One problem with the existing ESG ratings, and Refinitiv's score in particular, is the fact that it tries to account for a large company bias. Refinitiv argues that large-cap companies suffer from higher media attention than smaller-cap companies (Refinitiv, 2022). What they fail to address however, as noted by the European Securities and Market Authority, is the fact that large companies have more resources to respond to provider surveys and to generally address ESG perceptions among stakeholders (Halper, Grieve, Bussiere, & Shriver, 2022).

Other problems highlighted by Halper et al. (2022), is that there is potentially a geography specific bias in the large ESG providers' scores. European companies tend to score higher in terms of ESG performance, while companies in emerging markets generally receive lower ESG ratings. It also seems to be great tension between the different ESG pillars. One company may score high on E and bad on S and G, for example the car manufacturing company Tesla. Tesla scores high on environmental challenges but seems to ignore issues related to executive compensation and shareholder rights (Bloomberg Intelligence, 2022). They have a staggered bord, whereby they only elect one third of board members each term. Tesla falls short in terms of gender diversity as well, with only 25% of its board being female.

3.2 Textual Analysis

Textual data analysis can be used to retrieve information about ESG performance. The advantage of using the textual ESG approach is that it is highly objective. ESG word counts in annual reports cannot be altered or favored by the rating providers; only the companies themselves have an influence. The only part that is affected by the author's personal opinions and preferences is the classification of words as whether they are ESG related or not. This is not a problem if context checks of ESG related words in annual reports are put in place. For example, one must ensure that words are used in relation to an ESG situation and not as a name for a company or an asset.

The fact that textual ESG analysis offers more observations is another argument in favor of employing it. We can see in the table (**Table 5**) that the textual analysis provides more scores than Refinitiv for the total securities sample. Theoretically, by using the textual technique, it is possible to obtain an ESG score for every publicly listed firm. **Table 5** below shows the number of ESG observations in the Scandinavian markets for Refinitiv and for the textual approach, respectively.

Table 5: Number of ESG Scores from Refinitiv and Textual Analysis

This table shows the total number of companies that have received an ESG score from Refinitiv and from the textual approach in each year. The number of observations in each stock exchange per year is also shown for both methodologies.

	2015	*2016*	*2017*	*2018*	*2019*
Refinitiv	84	90	100	173	236
Euronext Oslo	16	16	19	37	48
Nasdaq Copenhagen	20	21	24	35	38
Nasdaq Stockholm	48	53	57	101	150
Textual	160	181	197	211	221
Euronext Oslo	35	40	41	43	46
Nasdaq Copenhagen	23	28	33	34	34
Nasdaq Stockholm	102	113	123	134	141

As shown, when it comes to the overall number of observations over the course of the sample period, the textual ESG approach is superior. Thus, one would prefer to use the textual ESG score rather than the ESG score provided by Refinitiv. The only exception is 2019, where Refinitiv has 25 more observations. The text sample that year initially included the same number of firms, but due to several data filters applied during the sample construction process, the number has decreased marginally. For example, one must take into consideration that only English annual reports were downloaded. Yet, this does not reduce the value of the textual ESG analysis. If one would like to have more observations than Refinitiv that particular year, one can simply download more annual reports.

The sentiment analysis method used in computational linguistics, serves as the foundation for the textual ESG analysis. Despite the fact that we are not analyzing sentiment, it is similar since it also make use of a vocabulary or lexicon. As illustrated in the tidy text flowchart below (***Figure 4***), a dictionary is merged with the textual dataset. This is the ESG dictionary created by Baier et al. (2020) in our case. The difference between our approach and the sentiment analysis, is the fact that we do not assign each word a “positive” or “negative”, but rather an “ESG” or “Not ESG”.

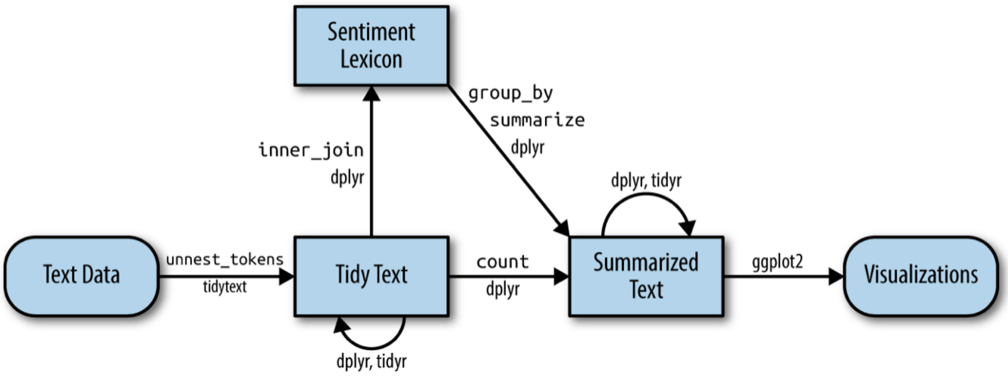

Figure 4: This figure shows the flow chart for tidy sentiment analysis

There are two main approaches in sentiment analysis: the lexicon-based approach and the machine learning approach (Taboada, 2011). This thesis applies the former, because with the machine learning approach one must assign each document with a predefined ESG score. ESG scores cannot be predefined in our case because we want to determine the companies' ESG scores ourselves. The lexicon-based method is applied because we are more interested in forecasting firms' stock returns using ESG ratings as predictors than forecasting firms' ESG performance.

When applying the machine learning approach, one would train an algorithm to recognize the patterns in a text that relates to high ESG performance. Our approach is much simpler, as lexicons can be created manually. After the ESG dictionary is merged with the textual data, each ESG word is given a numerical statistic that reflects how often it appears or how unique it is in the corpus. These numerics are also referred to as weighting factors or weighting schemes. After the numerics are calculated for each word, they are aggregated in order to achieve a single score for every document. These single scores, which are the ESG scores, will be analyzed later.

Even though it could be interesting to apply machine learning in this paper, Toboada et al. (2011) points out that, when tested in other domains than initially trained, the machine learning approach is worse than the lexicon-based approach. For instance, an algorithm trained on movie reviews would have little accuracy if tested on annual reports. Machine learning is also worse if implemented in a linguistic context because it is a much more complex task to teach the algorithm valence shifters. Valence shifters are essentially words that can change the sentiment of the polarized words. They include negation words such as 'not' and 'never', as well as intensification words such ‘very’ and ‘highly’. Ultimately, the most important reason for applying the lexicon-based approach, is the fact that two of the most renowned names in textual analysis apply this approach in their own papers, both Baier et al. (2020) and Loughran & McDonald (2011).

Baier et al. (2020) follow Loughran & McDonald (2011) and only include words that appear in at least 5% of all documents in its dictionary. They also analyze unigrams instead of bigrams, arguing that the meaning of compound ESG words can typically be covered by single ESG words. According to Baier et al. (2020), the noise of double counting is stronger than the additional signal of context. While Loughran & McDonald (2011) builds its analysis on 10-Ks only, Baier et al. (2020) go further in their approach. They download proxy statements to get more terms related to governance. Stop words such as “and”, “the” and “of” are removed as they occur so often and do not carry any important meaning. Similar to Loughran & McDonald (2011), Baier et al. (2020) include inflections of root words. “Greener” is an inflection of “green” for instance. Inflections are included if they also appear in an ESG context.

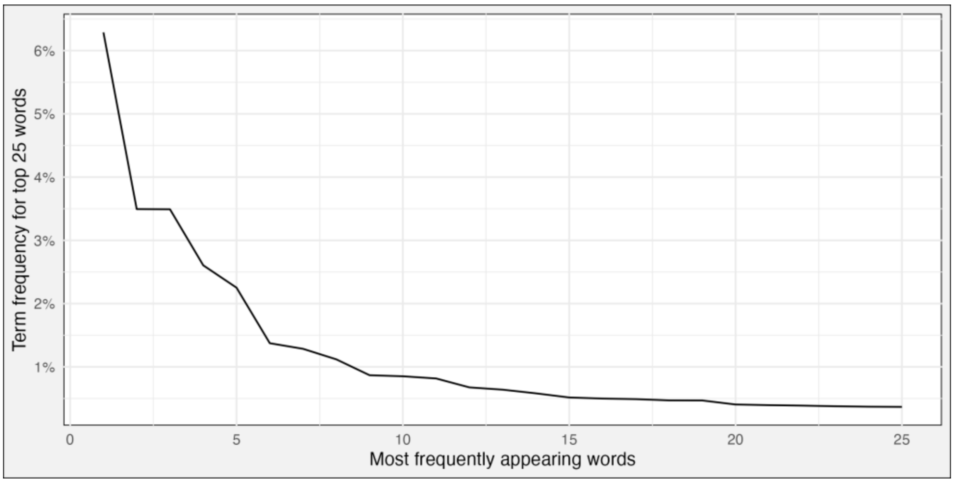

Figure 5: This figure shows the 25 most frequently appearing words and their term frequencies.

The plot above, **Figure 5**, shows why it is so important to remove stop words. These frequently appearing uninformative words account for the majority of word occurrences. This is known as *Zipf’s law*. Our textual data is merged with Loughran & McDonald’s master dictionary, as mentioned previously in this thesis. The Loughran & McDonald’s master dictionary does not include stop words, which means we do not actually need to remove stop words ourselves. However, it is a simple task to do by using the *tidytext* package. This R package has a premade stop-words dictionary that can be applied to filter out stop words from the textual dataset.

3.3 Weighting Schemes

We apply the "bag of words" method that summarizes each document D in a vector of term weights: $D = \{w_1, w_2, \dots, w_{|C|}\}$, where $|C|$ is the size of the entire corpus, i.e. the vector space, and $w_i, i = 1, \dots, |C|$ is the weight of term i in document D (Paltoglou & Thelwall, 2010). By representing documents as vectors, we can compute a score for each document, also known as vector space scoring (Manning, Raghavan, & Schütze, 2009). We select two different weighting schemes instead of raw word counts as they are better measures of a word's information content (Loughran & McDonald, 2011). The objective of these weighting schemes is to address the importance of a word, both in the relevant document and the entire corpus. There is often some sort of normalization that takes the length of each document into account.

In the information retrieval (IR) discipline, there are several weighting schemes to choose from, for example the Boolean (Manning, Raghavan, & Schütze, 2009; Paltoglou & Thelwall, 2010). This thesis focuses on configuring the measures term frequency (*tf*) and term frequency inverse document frequency (*tf.idf*), because they are widely used in finance literature. Both these weighted schemes have their strengths and weaknesses, that is why both are applied. ESG *tf* is very easy to interpret; It tells which company has the highest percent of ESG words in its annual report. ESG *tf.idf*, on the other hand, captures the importance and uniqueness of each ESG word. Hence, one could argue that it takes ESG materiality into account. Nevertheless, this would highly depend on whether the ESG word that is unique to a company is material to it. This could be determined by comparing it to other companies that operate in the same industry.

How the *tf* score and the *tf.idf* score is calculated in the IR discipline often varies. The following formulas applied to calculate these scores, are provided underneath. *tf* is defined as the number of times a word occurs in a document compared to the total number of word occurrences in that document:

$$w_{i,j}^{tf} = \frac{f_{i,j}}{\sum_{i \in D} f_{i,j}}$$

Here $f_{i,j}$ is the raw word count of term j in document i. The denominator is the sum of all word counts in document D. The aggregated ESG *tf* score is calculated by taking the sum of all *tf* weights of ESG related words:

$$Score_i^{tf} = \sum_{j=1}^{J} w_{i,j}^{tf}$$

Here, the total number of words in the ESG dictionary is represented by J. The *tf* score reflects the ratio of ESG word counts to total word counts within each document i. This score is easy to interpret; a firm with a high score in a given year has a high proportion of ESG related words in its annual report.

Another way to account for document length is to use log normalization of term frequencies. This is done in the *tf.idf* approach described by Jegadeesh & Wu (2013). The difference between the *tf.idf* compared to *tf* is that each word is also weighted inversely proportional to document frequency, that is the number of documents that a word appears at least once in. Each word j is assigned the following weight:

$$w_j^{idf} = \log \frac{N}{df_j}$$

Here, N is the total number of documents in the corpus and df_j is the document frequency. As one can see from the equation, the *idf* of a rare term is high, whereas the *idf* of a frequent term is likely to be low (Manning, Raghavan, & Schütze, 2009). This will help us filter out immaterial ESG words. The TF-IDF weight is defined as follows:

$$w_{i,j}^{tf.idf} = \begin{cases} 1 + \log f_{i,j} \cdot w_j^{idf}, & f_{i,j} > 0 \\ 0, & otherwise \end{cases}$$

Following Manning et al. (2009), *tf.idf* is highest when j occurs many times within a small number of documents and lower when j occurs in many documents. The weight is also lower when it occurs fewer times in a document and lowest when it occurs in almost every document.

To achieve a score at the document level, *tf.idf* is aggregated using the formula mentioned by Jegadeesh & Wu (2013):

$$Score_i^{tf.idf} = \frac{1}{(1 + \log a_i)} \cdot \sum_{j=1}^{J} w_{i,j}^{tf.idf}$$

Here, a_i is the total number of words in document i, and J is the total number of words in the ESG dictionary. The *tf.idf* score is a normalized sum of *tf.idf* weights that takes logged document lengths into consideration. Loughran & McDonald (2011) uses a similar approach.

In order to compare the two textual ESG scores with Refinitiv's ESG scores, the scores are normalized to be between 0 and 100, as this is the interval Refinitiv applies when measuring ESG performance. Both the textual term weights and Refinitiv' ESG scores are normalized. The formula for the normalization is provided below:

$$NS_{i,j} = \frac{(S_{i,j} - \min S_{i,j})}{(\max S_{i,j} - \min S_{i,j})} \cdot 100$$

Here, $NS_{i,j}$ is the normalized ESG score for company i in year j and $S_{i,j}$ is the ESG score for company i in year j prior to normalization.

3.4 Fama-French Models

In this thesis we have applied the Fama-French three-factor (1993) and Fama-French five-factor model (2015). The models were developed by Eugene Fama and Kenneth French to describe stock returns. All the factors in both the three-factor and five-factor model act like independent variables in regressions with returns as dependent variables. The Fama-French factors are easily accessible at Kenneth French's web site (2022). We aim to see if there are any abnormal returns from ESG portfolios that cannot be described by the factors.

We apply univariate portfolio sorting as described by Bali, Engle & Murray (2016), with ESG score as the sorting variable. Firms are divided into quintiles each month in the time-period between 1st of January 2017 and the 31st of December 2021, and the first and fifth quintile firms are classified as low ESG performers and high ESG performers respectively. The portfolio returns of the low and high performers are calculated with equal weights and weighted by market capitalization. Afterwards, the long-short portfolio is created by taking the difference in returns between the top ESG portfolio and the bottom ESG portfolio. Thereafter, we apply both the three-factor and the five-factor model and run the regressions.

The Fama-French three-factor model expands on the CAPM asset pricing model (Sharpe, 1964; Lintner, 1965; Mossin, 1966) with value risk and size risk as additional factors. By adding these factors, the model considers the outperforming tendencies of small-cap companies and value companies compared to the market.

The formula for the three-factor model is:

$$R_{it} - R_{ft} = \alpha_{it} + \beta_1(R_{Mt} - R_{ft}) + \beta_2 SMB_t + \beta_3 HML_t + \epsilon_{it}$$

The Fama-French five-factor model builds further on the three-factor model, by adding two additional factors, RMW_t and CMA_t. RMW_t is a factor for the most profitable firms minus the least profitable firms. CMA_t are a factor for firms that invest conservatively minus the firms that invests more aggressively.

The formula for the five-factor model is:

$$R_{it} - R_{ft} = \alpha_{it} + \beta_1(R_{Mt} - R_{ft}) + \beta_2 SMB_t + \beta_3 HML_t + \beta_4 RMW_t + \beta_5 CMA_t + \epsilon_{it}$$

4. Results

4.1 Evaluation of ESG term weights

In order to evaluate the performance of the ESG dictionary and the weighting schemes, the ESG words $w_{i,j}, \dots, w_{i,J}$ that have the highest weights are queried from the textual dataset. We try to locate which firms mention these words, and reflect upon whether the relationship makes sense. The top words that occur for each weighting scheme are then verified through a context check in a keyword-in-context table. This table precents the words that most often appear to the left and to the right of the top words.

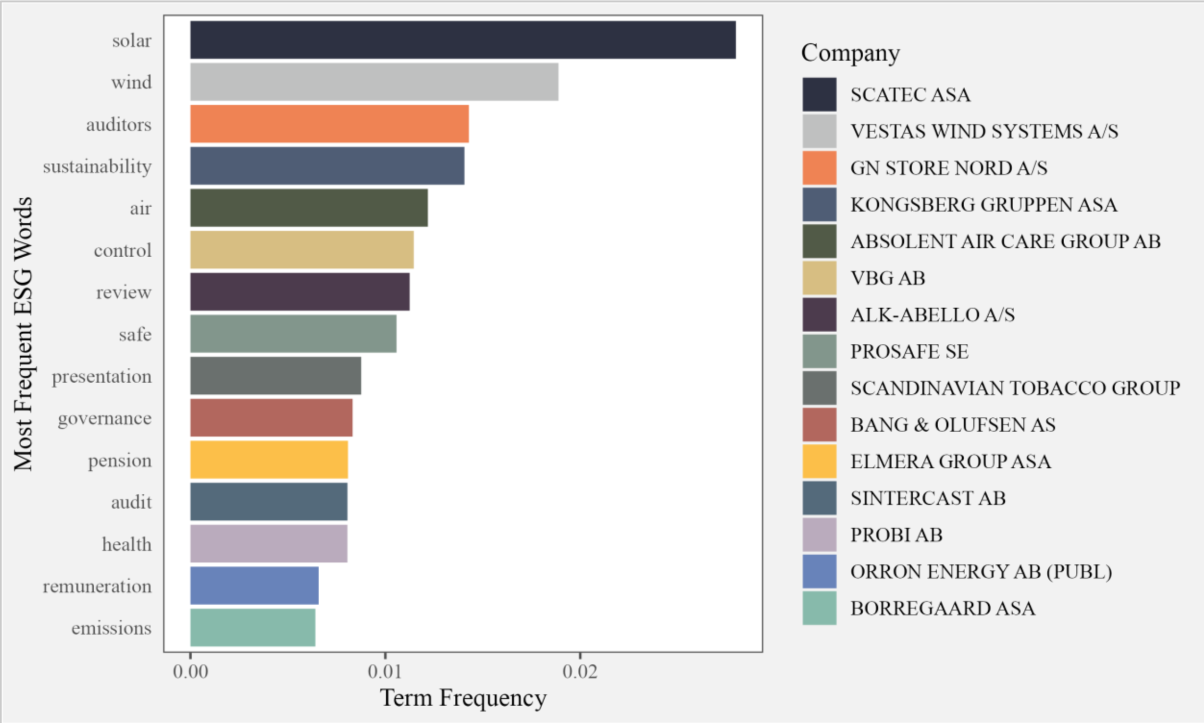

Figure 6: The ESG words with the highest tf weights

This figure shows the ESG words that have the highest term frequency within a document across the entire corpus. These words appear in several documents, but only the document in which the word has the highest *tf* weight is shown. This is because it is more interesting to look at different words and not the same word for different years and companies. The legend bar includes the companies that these documents belong to.

The most noticeable feature in the graph (**Figure *6***) is the dominance of the words *solar*, *wind* and *air*. As illustrated, these words are written by Scatec, Vestas and Absolent respectively. One logical reason for their dominance is the fact that they are part of the company names. Scatec was formerly called as Scatec Solar, but in 2021 the company decided to change its name as a reflection of a broader renewable's strategy (Scatec, 2021). Vestas and Absolent, on the other hand, are still called Vestas Wind Systems and Absolent Air Care. As a result of this, both *solar*, *wind* and *air* are mentioned many times in the respective companies' annual reports. They make up around 3%, 2% and 1% of all terms in the annual report in which they are mentioned the most. This may bias their textual ESG weights and lead to a more inaccurate ESG score overall. Nevertheless, we choose to not exclude these words, because they also appear in an ESG context for other companies. This will be illustrated in the context check further down. Another interesting find in the annual report is that Kongsberg Gruppen mentions *sustainability* a lot in their annual report. One reason could be that they have an exceptional commitment to climate related efforts, or it could be that they mention it several times in a non-ESG related context.

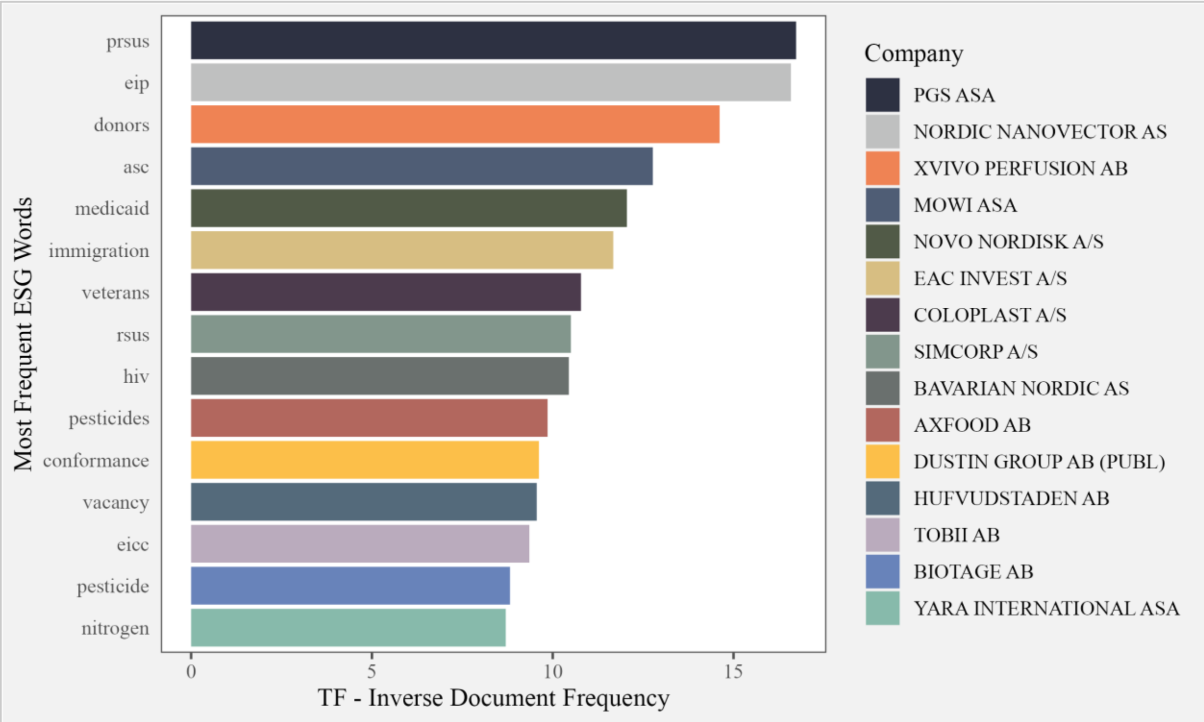

Figure 7: The ESG words with the highest tf.idf weights

This figure shows the ESG words that have the highest term frequency - inverse document frequency within a document across the entire corpus. These words appear in several documents, but only the document in which the word has the highest *tf.idf* weight is shown. This is because it is more interesting to look at different words and not the same word for different years and companies. The legend bar includes the companies that these documents belong to.

The *tf.idf* weights produce an entirely different bar chart, as one can see above in **Figure** *7*. The ESG word with the highest *tf.idf* weight is *PRSUs*, which stands for Performance-based Restricted Stock Units. It is mentioned the most in one of PGS's annual reports. There are not a lot of other companies that mention *PRSUs*. This may be the reason for why it got such a high *tf.idf* weight in PGS's case. Another word that has a very high *tf.idf* weight is *EIP* when it is mentioned by Nordic Nanovector. It is an abbreviation for Equity Incentive Plan, which is a more general term for performance-based compensation that also includes *PSUs* and other incentive plans. *PSUs* are often more related to a company's key performance indicators, while *PRSUs* are more related to how long employees remain with the company (Global Shares, 2022). Both terms are related to corporate governance and mentioned in the corporate governance section of the companies' annual reports. Thus, the ESG words might not be material to the companies, but they are unique in that there are not that many other companies in Scandinavia that mention these remuneration programs.

Table 6: KWIC Table

This table shows the words that most frequently appear before and after *solar* and *prsus*. These words are chosen as keywords due to their high *tf* and *tf.idf* weights. The point of the KWIC table is to highlight the corpus' most important words and the context in which they appear.

solar		*prsus*	
Pre	Post	Pre	Post
scatec	asa	prsus	prsus
wind	power	settlement	granted
rec	panels	pool	will
elkem	energy	plan	subsequent
new	sa	awarded	remaining
construction	grade	units	average
use	group	date	settlement
electricity	pv	dividends	ceo
operate	plants	remaining	rsus
operating	panel	award	lti

To summarize the context check we investigate the KWIC table (**Table *6***). The keywords that are most interesting to look at are the words that score the highest on *tf* and *tf.idf*. One can see that the word that appears most often before *solar* is Scatec. This is not a surprise, because we have discussed earlier why this may be the situation. Several other companies also have *solar* in their names, for example Rec Solar, a former subsidiary of Elkem. Other words that often occur before *solar* are *new* and *construction*, indicating that several companies plan to produce solar energy. Words that appear after *solar* include ASA, SA and Group, which confirms that the word is often used in relation to company names. However, it is also used before power and energy, so it cannot be excluded all together. The occurrence of *plants* and *panels* seems to suggest that the companies are operating power plants or building solar panels for public and residential use. Thus, one could argue that *solar* is a material ESG word for these companies. The word *grade* often occurs next to *solar*. This is because solar grade is a term that describes how pure the silicon must be in order be eligible for solar cell production. PV, which stands for photovoltaics, is the industry that concerns itself with solar cell manufacturing.

The term *PRSUs*, on the other hand, often appear next to *PRSUs*. This may seem strange, but is likely due to the term occurring frequently in table columns. In textual data cleaning, it is often advised to remove text that appears inside table columns. This is pointed out in the internet appendix provided by Loughran & McDonald (2011) for instance. However, this did not damage the validity of our textual analysis considerably. The term *PRSUs* often occurs in regular text next to *settlement*, as the settlement of stock-based compensation is often a topic of discussion. Terms that regularly appear to the left of *PRSUs* are *award* and *awarded*. This is valuable knowledge for investors, because firms that award a lot of *PRSUs* could potentially dilute its existing shareholders. On the other hand, it can be seen as a positive signal to investors for good corporate governance practices. Words that frequently appear after *PRSUs* are *CEO* and *LTI*. *CEO* often appears after *PRSUs* because stock-based compensation is often used on executives. *LTI* stands for Long-Term Incentive plan.

4.2 Descriptive statistics of ESG scores

In this part, descriptive statistics relevant for the analysis will be presented to illustrate what the data look like, as well as the dispersion of the parameters. A combination of statistics and plots are used to illustrate the characteristics of each ESG score. We attempt to visualize the properties of each quintile for every sorting variable.

Table 7: Descriptive statistics of ESG scores

Statistic	*N*	*Mean*	*St. Dev.*	*Min*	*Max*
ESG Refinitiv	8,196	53.504	19.317	1.334	92.002
ESG tf	11,623	0.055	0.013	0.024	0.100
ESG tf.idf	11,623	14.421	6.033	3.493	43.489

As one can see in **Table** *7* above, Refinitiv has fewer observations than the ESG weighting schemes that were created through textual analysis. The difference is more than 3,000 observations. This highlights the advantage of constructing textual ESG scores, as the sample increases considerably. When stocks are sorted into portfolios, more observations are beneficial because the number of firms in each quintile is often below 30, which could lead to statistically misleading results. Textual analysis can mitigate this error by having larger cross-sectional data. More data can ensure statistical significance in the Fama-French regressions in the context of ESG hedging.

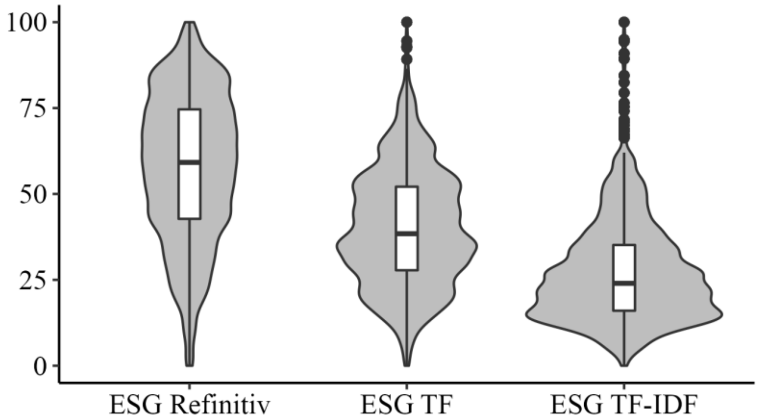

Figure 8: Violin plots with box plots of normalized ESG scores

As illustrated in the violin plots in **Figure *8***, Refinitiv's ESG score has fat tails in both directions. This suggests that a smaller number of firms will get an ESG score centered around the median, the horizontal line seen in the box plot. ESG term frequency on the other hand, has a more bell-shaped distribution, although it has more outliers in the upper tail. The median is lower than Refinitiv, hence fewer companies will get a good ESG score. ESG *tf.idf* is even more extreme, with a lot of outliers on the higher end and most scores centered around the median. The median is even lower than ESG *tf,* scoring around 25 out of 100. This means the majority of companies will get a lower ESG score.

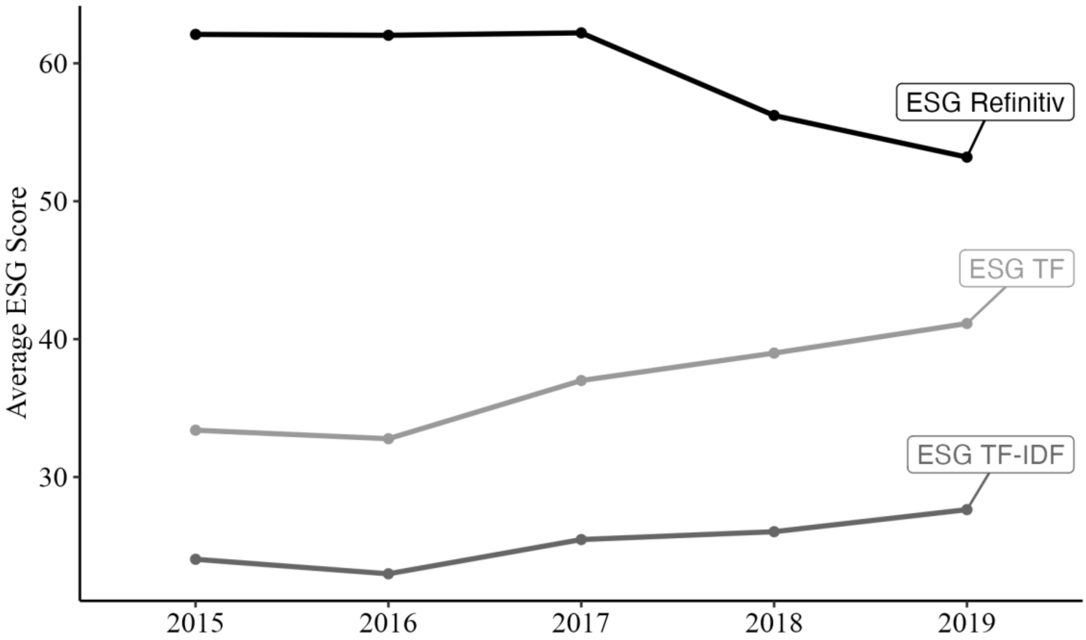

Figure 9: The average of the three ESG scores over time

The figure above, **Figure *9***, shows that the cross-sectional average of Refinitiv's ESG scores declines from 2017 to 2019. This might be surprising, but may be due to Refinitiv assigning more mid-cap and small-cap stocks an ESG score in the Scandinavian region. It was previously highlighted that there are a lot of missing values in Refinitiv's scores, especially in Scandinavian markets. Smaller firms often have lower ESG scores, and this will reduce the cross-sectional average overall. This may be due to the fact that smaller firms have fewer resources available for ESG reporting. Another explanation for the decline could be that Refinitiv is revising the way they measure ESG performance. Research conducted by Berg, Fabisik & Sautner (2020) finds that Refinitiv lowered its ESG scores to improve correlation between ESG performance and future stock returns.

Figure *9* also shows that average ESG *tf* and ESG *tf.idf* scores increase over time. The number of observations per year is more constant for the textual weighting schemes. Therefore, we must make a different argument than the one mentioned above concerning Refinitiv. One reason for the increase in average ESG score could be that there has been more focus on sustainability reporting during the last few years. Soon it will be mandatory to disclose sustainable practices for European publicly listed companies (European Commission, 2021). This leads to a larger number of ESG words being written in annual reports. Paradoxically, this could diminish the value of making scores based on annual reports, as it becomes more difficult to differentiate the ESG performance of each company. Nevertheless, term uniqueness could still be an efficient tool to measure ESG materiality, even though one would expect a lot of similarities in terms of ESG materiality within the same industry.

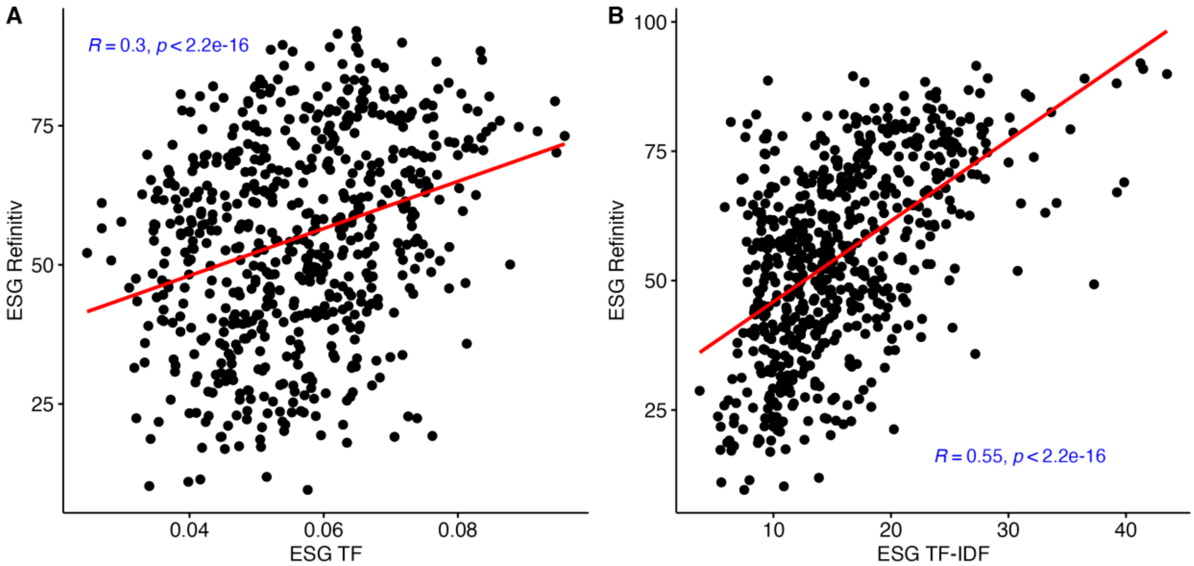

Figure 10: Scatter plots of Refinitiv score and textual scores

This figure shows the Pearson correlation between Refinitiv's ESG score and ESG tf, and between Refinitiv's ESG score and ESG *tf.idf*. The correlation coefficients and the p-values are written in blue. The red lines are the fitted regression lines for the different observations.

As presented in **Figure *10***, the Pearson correlation is relatively high between Refinitiv's ESG score and the textual ESG scores, with 0.30 and 0.55 in correlation respectively. This means that ESG term weightings explain much of the same properties in a firm's ESG performance as Refinitiv. Berg et al. (2022) findings shows that the correlation between the scores of the six large ESG rating providers range between 0.38 and 0.71. This underlines the quality of the textual ESG metric. If Refinitiv's score is considered a good proxy for ESG performance that is. As expected, correlation is the highest between Refinitiv's ESG score and ESG *tf.idf* with a correlation of 0.55 compared to ESG *tf*'s correlation of 0.30. Hence, this suggests that *tf.idf*'s ability to find more material ESG words in annual reports seems to be of importance.

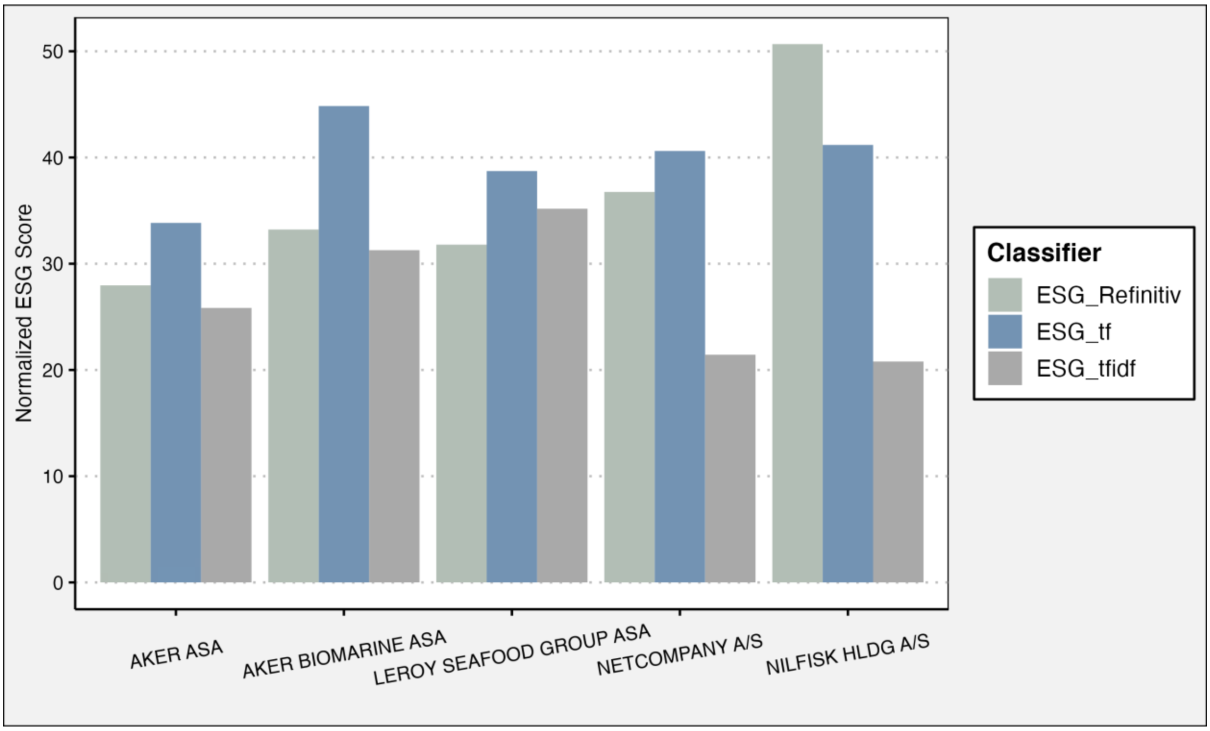

Figure 11: The difference between the three ESG scores; Refinitiv's, ESG tf and ESG tf_idf, of five randomly picked companies from the sample

The last figure of interest, **Figure *11***, presents the ESG scores of five randomly drawn companies from the data sample. This graph aims to illustrate the discrepancy between the Refinitiv ratings and the assigned ESG document weights. This is to ensure that the ESG scores are not identical. It is also a reliable method of verifying and controlling for the quality of textual ESG scores. As one can see, the scores produce almost the same results for Aker, Aker BioMarine and Lerøy. However, for Netcompany and Nilfisk there are large differences in the assigned ESG scores.

4.3 Descriptive statistics of quintile portfolios

We discover that there are several differences after examining the characteristics of the various ESG ratings. For example, ESG *tf.idf* has a much lower median value than ESG *tf* and ESG Refinitiv. Despite this, ESG *tf.idf* seems to have the highest correlation with Refinitiv's ESG score. This section of the thesis focuses on the descriptive statistics of the different quintile portfolios. The performance of the first and fifth quintiles for the various sorting variables will also be examined.

Table 8: Average market capitalization and average monthly returns for equal-weighted portfolios.

This table shows the average market capitalization in millions of euros and average monthly returns for equal-weighted portfolios sorted on ESG score. Each row represents one quintile, and all three proxies for ESG performance are represented.

	Refinitiv ESG		*ESG tf*		*ESG tf.idf*	
Quintile	Market Cap	Return	Market Cap	Return	Market Cap	Return
5	13,883	1.37%	5,807	1.45%	10,459	1.36%
4	9,759	1.30%	5,042	1.53%	4,362	1.49%
3	5,052	1.39%	3,997	1.59%	2,753	1.56%
2	2,774	1.50%	4,571	1.52%	3,333	1.49%
1	2,087	1.42%	3,446	1.78%	1,995	1.96%

From **Table *8*** it is clear that the fifth quintile consists of companies that have a high market capitalization. The average market capitalization is almost seven times that of the first quintile when sorting for Refinitiv's ESG score. Differences are not as large when sorting for ESG term frequency, but for *tf.idf* they are quite substantial. Portfolio returns are also, in the majority of cases, lower for the companies with a high ESG score. However, there are some variations. The second quintile has the highest monthly return for Refinitiv's ESG scores for instance. It is not a perfectly ascending distribution for *tf* and *tf.idf* either. The third quintile has higher average monthly returns than the second quintile.

Table 9: Number of firms with Refinitiv score in each quintile

	Refinitiv ESG scores				
Quintile	2017	2018	2019	2020	2021
1	17	18	20	35	48
2	17	18	20	35	47
3	17	18	20	35	47
4	17	18	20	34	47
5	16	18	20	34	47

As presented in **Table *9*** above, the average number of observations in each quintile is very low in the earlier years for Refinitiv. In the years 2017, 2018 and 2019 the quintile portfolios consist of less than 30 firms. Thus, one cannot be sure whether sample means are representative of the true means when it comes to portfolio returns. The number of observations increases beyond

30 in 2020, which is good. In 2021, Refinitiv has more observations in each quintile than for the textual approach due to previously discussed reasons.

Table 10: Number of firms with textual score in each quintile

Textual ESG scores					
Quintile	2017	2018	2019	2020	2021
1	32	36	40	42	45
2	32	36	39	42	44
3	32	36	39	42	44
4	32	36	39	42	44
5	32	36	39	41	44

Table 10 illustrates the number of observations in each quintile, and that the number of firms in each quintile is more stable for portfolios sorted on ESG *tf* and ESG *tf.idf*. Every year have on average more than 30 observations in each quintile. In the years 2017, 2018 and 2019 the textual scores produce more than double the number of observations compared to Refinitiv. Quintile portfolios are also larger in 2020, but a bit smaller compared to Refinitiv in 2021. However, the difference is only 3 firms, which is fairly insignificant.

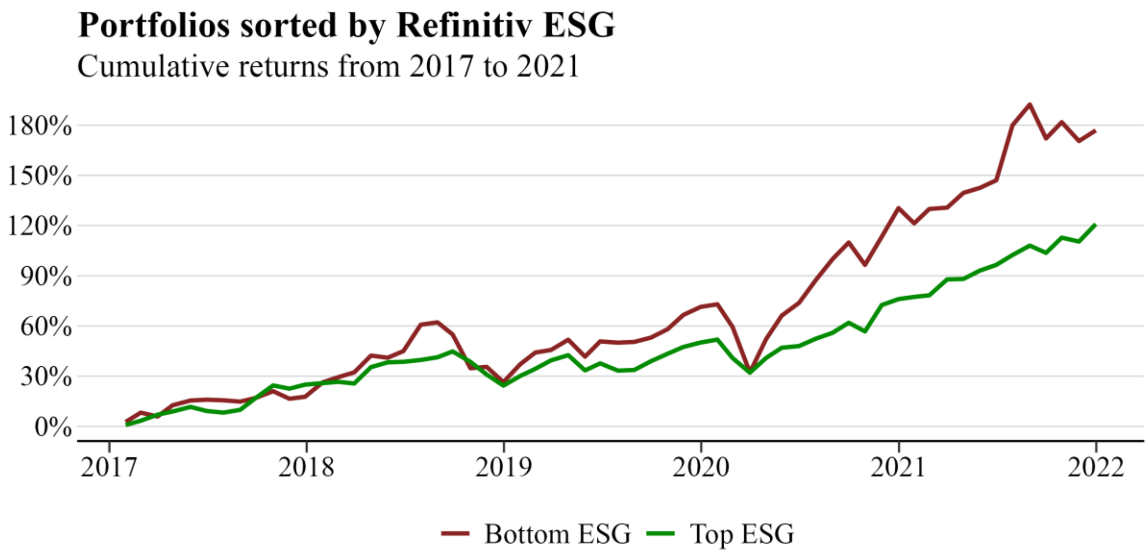

Figure 12: Cumulative returns for top ESG and bottom ESG, Refinitiv's ESG score

The graph above (**Figure** ***12***) shows the cumulative returns for the top ESG and bottom ESG portfolios in terms of Refinitiv's ESG score. It is clear that the top ESG portfolio has average lower returns compared to the low performers. This is especially evident from the years following the covid-19 outbreak in March 2020. The first two years in the five-year period yielded have somewhat similar returns for the two portfolios, however the cumulative returns for the high performer have mostly been higher. Total cumulative returns for the lower-performing ESG portfolio are approximately 180% according to the data, corresponding to an impressive cumulative annual gross return (CAGR) of 22.9%. The top ESG portfolio, on the other hand, has had cumulative returns of 121%, with a corresponding CAGR of 17.2%.

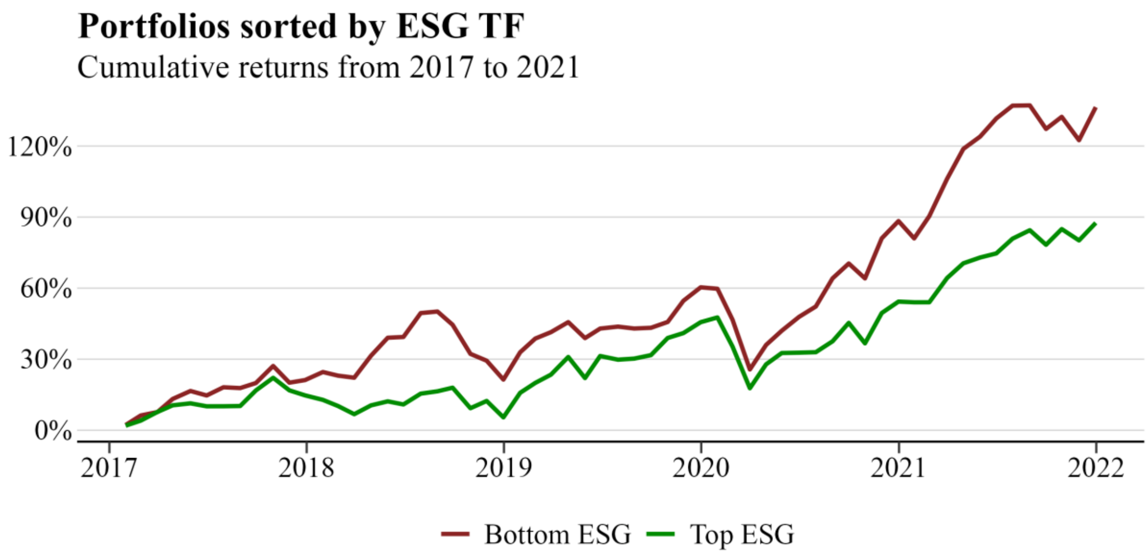

Figure 13: Cumulative returns for top ESG and bottom ESG, ESG tf score

When portfolios are sorted by ESG term frequency and split into "top ESG" and "bottom ESG", bottom ESG outperforms by a great margin. The bottom ESG portfolio has a cumulative return of 136.4% over the five-year period, compared to 87.6% for the top ESG portfolio. This constitutes to CAGRs of 18.8% and 13.4% respectively. Investing in the bottom ESG portfolio therefore looks like the most attractive option, but it is uncertain whether this is due to higher risk or if it is simply an arbitrage opportunity.

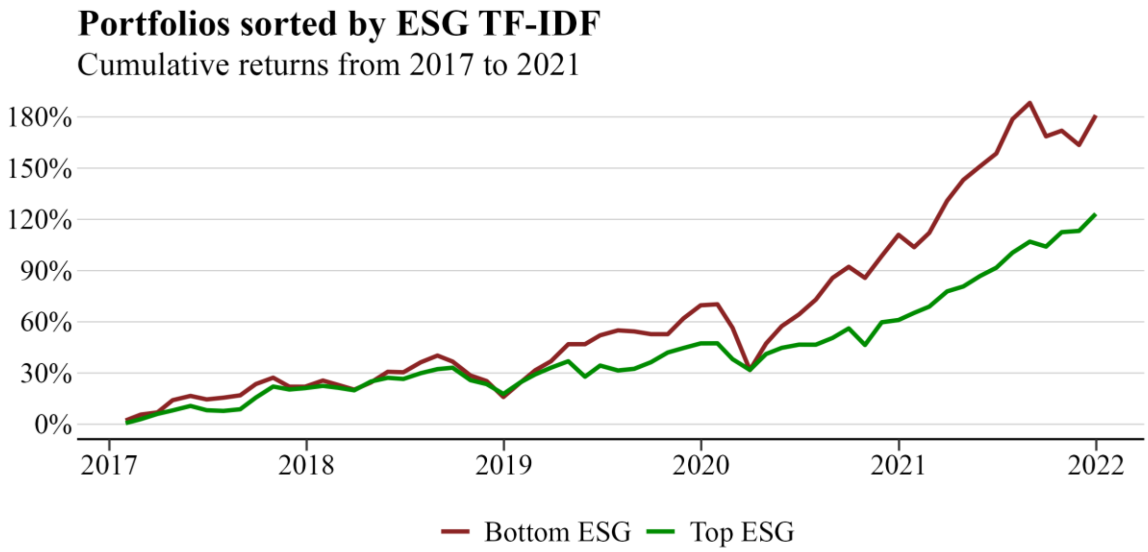

Figure 14: Cumulative returns for top ESG and bottom ESG, ESG tf score

As illustrated in **Figure *14***, the difference in returns is almost as high as when the portfolios are sorted by Refinitiv's ESG score. The bottom ESG portfolio has a cumulative return of 181.0%, which is substantially higher than the top ESG portfolio's return of 123.2%. 181.0%, corresponds to an annual return of 23.0%, while 123.2% corresponds to about 17.4%.

4.4 Fama-French Regressions

Three long-short portfolios are regressed on the Fama-French factors. Each portfolio is sorted on a different ESG metric, either Refinitiv's ESG score, ESG *tf* or ESG *tf.idf*. The best ESG-performer is the portfolio of companies with the 20% highest score, and the worst ESG-performer is the portfolio of companies with the 20% lowest score. Average portfolio returns are calculated monthly for the best and worst. The long-short portfolio is formed by taking the difference in returns between the two. As the companies are given only one ESG score per year, the portfolios are rebalanced annually. The main focus in this section is to investigate whether a long-short trading strategy can generate alpha, that is a risk-less return, or not.

Table 11: Regression with the ESG Refinitiv score

This table shows the regression of long-short ESG portfolios on the Fama-French three- and five-factor models. Portfolio returns are calculated both using equal weights and by weighting each firm by its market capitalization. The sorting variable used for this long-short portfolio is Refinitiv's ESG score. Standard deviations are shown in parenthesis. ***, ** and * denotes significance at the 1%, 5% and 10% level.

	Refinitiv ESG			
	Fama-French Three-Factor and Five-Factor Model			
	EW	VW	EW	VW
Mkt-Rf	-0.199^{**}	-0.418^{***}	-0.228^{**}	-0.405^{***}
	(0.086)	(0.110)	(0.096)	(0.100)
SMB	-0.817^{***}	-0.727^{**}	-0.751^{***}	-0.554
	(0.220)	(0.281)	(0.253)	(0.410)
HML	0.316^{**}	0.398^{**}	0.508^{*}	0.318
	(0.132)	(0.168)	(0.264)	(0.359)
RMW			0.470	0.333
			(0.389)	(0.582)
CMA			0.147	0.640
			(0.490)	(0.638)
Alpha	0.005	0.002	0.005	0.004
	(0.004)	(0.005)	(0.004)	(0.004)
Observations	60	60	60	60
R^2	0.376	0.383	0.388	0.391
Adjusted R^2	0.342	0.349	0.331	0.334
Note:				*p<0.1; **p<0.05; ***p<0.01

There are no significant alphas in the regressions above, **Table *11***. Therefore, it is not possible to earn positive and abnormal risk-adjusted returns with a long-short trading strategy. The implications are that ESG is already sufficiently priced in the market, either through the model factors or some other unknown risk-premia. This is consistent with *the efficient market hypothesis* outlined by Fama (1970). Nevertheless, the three-factor model presented above appears to describe the variation in monthly returns quite well. All the independent variables are significant at a 5% confidence level, both in the equal-weighted and value-weighted portfolio regression. It is therefore feasible to explain why the best-in-class ESG portfolio has lower returns than the worst-in-class ESG portfolio.

The portfolio of companies with high ESG scores has a smaller beta compared to the portfolio of bottom ESG firms. This suggest that the top ESG firms have less systematic exposure to the market, signalizing that they are less volatile than the companies with a low ESG score. Much of the discrepancy in portfolio returns can also be explained by the market size of the portfolio stocks. The high performing portfolio tend to consist of stocks that have a high market capitalization. This is confirmed by the negative sign and statistical significance of the SMB factor, which is significant at the 5% confidence level. The only exception is found in the value-weighted five-factor model, in which only the excess market return has explanatory power. Furthermore, the HML factor is positive and significant in several cases. Stocks with high ESG scores tend to have high book-to-market ratios. The value premium contributes to higher monthly returns for the top ESG firms and reduces the gap in returns between top ESG and bottom ESG portfolios.

A general insinuation is that the five-factor model is not particularly good at explaining the difference in monthly returns between the high performing and low performing portfolio. Neither the equal-weighted, nor the value-weighted portfolio have statistically significant RMW- and CMA-coefficients. This means that there is no tendency of top ESG firms to have higher operating profitability or to invest more aggressively than bottom ESG firms. By including more risk factors in the regression, the existing independent variables SMB and HML even loses statistical significance in the value-weighted case. This suggests that there are too many factors in the model, causing the standard deviations of the factors to rise. Thus, it is difficult to make statistical inferences on the model. The adjusted R-squared reflects this problem well. The explanatory power is higher in the three-factor model than in the five-factor model adjusted for the number of independent variables.

Table 12: Regression with ESG tf weights.

This table shows the regression of long short ESG portfolios on the Fama-French three- and five-factor model. Portfolio returns are calculated both using equal weights and by weighting each firm by its market capitalization. The sorting variable used for this long-short portfolio is ESG *tf* weight. Standard deviations are shown in parenthesis. ***, ** and * denotes significance at the 1%, 5% and 10% level.

	Textual Data Analysis ESG (tf)			
	Fama-French Three-Factor and Five-Factor Model			
	EW	VW	EW	VW

Mkt-Rf	-0.022	-0.028	-0.082	-0.075
	(0.094)	(0.081)	(0.100)	(0.089)
SMB	-0.457*	-0.345	-0.293	-0.266
	(0.240)	(0.209)	(0.263)	(0.236)
HML	-0.364**	-0.065	-0.076	0.175
	(0.144)	(0.125)	(0.275)	(0.247)
RMW			1.030**	0.618*
			(0.405)	(0.363)
CMA			0.390	0.090
			(0.510)	(0.457)
Alpha	-0.004	-0.003	-0.004	-0.004
	(0.004)	(0.003)	(0.004)	(0.004)
Observations	60	60	60	60
R^2	0.185	0.070	0.272	0.115
Adjusted R^2	0.142	0.021	0.205	0.034
Note:				*p<0.1; **p<0.05; ***p<0.01

After sorting the stocks by term frequency of ESG words, the constructed portfolios paint a different picture on the ESG anomaly. Alphas are still insignificant; however, the market excess return factor has lost its explanatory power. This suggest that the stocks included in the top ESG portfolio are no less volatile than the stocks included in the bottom ESG portfolio. This stands in stark contrast to the findings in the Refinitiv regressions, where market excess return is the only statistically significant independent variable across both the three-factor and five-factor model. Now the HML factor has a negative coefficient in all regressions, which implies that companies with a high ESG score have low book-to-market ratios, suggesting that they are growth stocks. This contradicts the fact that high performing ESG companies are usually largely capitalized with high book-to-market ratios. HML is only significant in the equal-weighted three-factor regression though.

In the equal-weighted five-factor model, RMW is the only statistically significant explanatory variable. This is a new finding. The coefficient is positive, indicating that ESG firms have higher operating profitability than others. In the value-weighted five-factor model, RMW is also the only statistically significant factor at the 10% level. The lack of statistically significant variables in the Fama-French models suggest that term frequency is either a bad proxy for ESG performance or that the asset pricing model is not a good fit for the data. After calculating the model's variance-inflation factors, it appears that there is some collinearity between HML and the other risk factors. This could be the reason why there is noise in the coefficient estimates.

Table 13: Regression with ESG tf.idf weights.

This table shows the regression of long short ESG portfolios on the Fama-French three- and five-factor models. Portfolio returns are calculated both using equal weights and by weighting each firm by its market capitalization. The sorting variable used for this long-short portfolio is ESG *tf.idf* weight. Standard deviations are shown in parenthesis. ***, ** and * denotes significance at the 1%, 5% and 10% level.

	Textual Data Analysis ESG (tf.idf)			
	Fama-French Three-Factor and Five-Factor Model			
	EW	VW	EW	VW
Mkt-Rf	-0.213^{***}	-0.229^{**}	-0.224^{**}	-0.250^{***}
	(0.078)	(0.093)	(0.088)	(0.079)
SMB	-0.677^{***}	-0.806^{***}	-0.698^{***}	-0.874^{***}
	(0.201)	(0.248)	(0.232)	(0.325)
HML	0.065	0.184	0.186	0.386
	(0.120)	(0.167)	(0.242)	(0.304)
RMW			0.127	0.064
			(0.357)	(0.433)
CMA			-0.042	-0.269
			(0.449)	(0.417)
Alpha	-0.002	0.0001	-0.002	-0.001
	(0.003)	(0.004)	(0.004)	(0.004)
Observations	60	60	60	60
R^2	0.352	0.370	0.358	0.376
Adjusted R^2	0.317	0.337	0.298	0.318
Note:				*p<0.1; **p<0.05; ***p<0.01

Table ***13*** applies *tf.idf* as a proxy for ESG performance. Alphas are still statistically insignificant, hence there is no market anomaly. Both the market factor and the SMB are statistically significant at the 5% level across all regressions. Thus, it could be stated, with confidence, that the top ESG portfolio consists of mostly large companies with lower market volatility compared to others. Furthermore, HML has no statistical significance in any of the regressions. RMW and CMA also fail to add any value in explaining the difference in monthly portfolio returns. The results are disappointing; the long-short strategy on top ESG firms and bottom ESG firms does not fit according to this model either. Yet the Fama-French factors struggle to explain why.

To summarize the results, there is little consistency in the coefficient estimates when controlling for the different ESG proxies. Market excess return, SMB and HML are all instrumental in explaining returns for the Refinitiv ESG regressions. However, the value premium disappears for *tf.idf* portfolio sorts, and market volatility becomes insignificant for ESG *tf.* The size premium and market volatility may be the most reliable risk factors, as they are significant in both the Refinitiv and *tf.idf* models. This complements the descriptive statistics presented earlier. In that section, it was discovered that the average market capitalization was higher for the fifth quintile compared to the first quintile.

4.5 Regression validation

In order for the statistical inferences to be valid, there must be no heteroscedasticity or serial correlation in the error term. Heteroscedasticity is examined through Breusch-Pagan tests and scatter plots of residuals versus predictors. Autocorrelation is examined through Durbin-Watson tests and lag plots. Heteroscedasticity is found in the value-weighted five-factor model for portfolios sorted on Refinitiv's ESG score, and it is also found in the equal-weighted factor models for portfolios sorted on *tf.idf*. White's heteroscedasticity-consistent standard errors are therefore used for these regressions. There seems to be no autocorrelation in any of the regressions, although the t-statistic for first-order autocorrelation is somewhat low for the value-weighted *tf.idf* models. In other words, there is no need to use Newey-West robust standard errors which is common when managing financial data.

4.6 The Market Size Bias

Even though there are conflicting results in our regression analysis, we find that there are significant SMB coefficients for two portfolio sorts. In other words, the long-short portfolio creates an increased exposure to firms with a high market capitalization. The exposure is correlated with lower expected stock returns. There are several implications: 1) Investors might favor large market capitalization stocks, and less capital is therefore allocated to smaller firms; 2) Investing in high ESG performers may also reduce the effect of diversification, leading to worse portfolio returns; 3) Lastly, the way ESG scores are calculated may give an advantage to larger firms with more resources, while not providing socially responsible investors with the information needed to make decisions based on their beliefs (Drempetic, Klein, & Zwergel, 2020).

4.7 ESG Market Volatility

It seems that the best-in-class ESG portfolio is less volatile than the bottom ESG portfolio. The betas in our regressions are negative and statistically significant for portfolios sorted on Refinitiv's ESG score and ESG *tf.idf*. This means that high ESG performers have lower exposure to the market index than low ESG performers, hence high performers get lower stock returns. Our findings coincide with that of Ouchen (2022), who finds that stock returns of high ESG performers are relatively less turbulent compared to the market index. Ouchen (2022) studies daily return of "MSCI USA ESG Select" and the "S&P 500" using Markov-switching GARCH models. Investing in top ESG firms may therefore yield more stable but lower stock returns compared to bottom ESG firms and the financial market.

4.8 The Long-Short ESG Portfolio

Teti, Dallocchio & L'Erario (2022) findings suggest that a long-short portfolio of top ESG firms and bottom ESG firms generates positive – though not statistically significant – alpha. The relationship holds when regressed on both the Fama-French three-factor and the Fama-French five-factor model, suggesting that ESG could have a positive impact on stock performance. Lioui & Tarelli (2022) also find that highly rated firms outperform low-rated firms, however alphas feature strong time variation during the sampling period. They measure the return spread between responsible and irresponsible portfolios using both the *time-series* approach in Fama & French (1993), as it is named by Fama & French (2020), and the *cross-sectional* approach which builds on Fama & MacBeth (1973).

This contradicts Hartzmark & Sussman (2019), who do not find that sustainability outperforms non-sustainability. On the contrary, their evidence suggests the opposite. Furthermore, Auer & Schumacher (2016) find that investors in Europe tend to pay a premium for socially responsible investment. In this thesis we find that high ESG performance correlates with lower stock returns, but for the Scandinavian stock markets specifically. Thus, it is hard to tell which results to trust. The conflicting empirical results may be due to the dispersion of ESG ratings used in asset pricing models, as noted by Berg et al. (2022), or due to ESG rating uncertainty as Avramov et al. (2022) refers to it.

The large selection of ESG measures is not only produced by the major ESG providers; researchers apply other measures as well (Bolton & Kacperczyk, 2022; Hsu, Li, & Chi-Yang, 2022). Bolton & Kacperczyk (2022) uses carbon data provided by Trucost for instance. They find that stocks of firms with higher total carbon emissions earn higher risk-adjusted returns. However, when they change the way that carbon emissions are measured, from total carbon emissions to carbon intensity, emissions per unit of sales, the effect vanishes. This highlights the confusion of how to measure ESG and how to integrate it into portfolios.

Some research take a more innovative approach, arguing that ESG outperformance is not a result of higher expected returns, but rather climate related shocks (Pástor, Stambaugh, & Taylor, 2022; Choi, Gao, & Jiang, 2020; Engle, Giglio, Kelly, Lee, & Stroebel, 2020). Pástor et al. (2022) construct a green-minus-brown portfolio from US stock data and find that the difference in returns between green and brown disappears after controlling for increases in climate related concerns. Engle et al. (2020) do something similar, but they apply textual analysis of climate news from *The Wall Street Journal* instead of the media index provided by Ardia et al. (2022). They implement it as a way to hedge against climate risk by going long in the winners and short in the losers.

The main takeaway is that there are many alternative ways to measure ESG and sustainability, and the method one chooses will affect the outcomes. We have demonstrated how to create ESG scores using textual analysis of annual reports. This is advantageous in terms of sample size when compared to scores made by the large ESG providers. The score, however, only evaluates how well corporations disclose ESG, not if they have a real social impact. So, if investors are looking for doing socially good, they may look elsewhere. Most existing research are also done on US stocks, while we use stocks from the Scandinavian stock markets only. In addition, different periods of time are studied in the existing studies. This thesis focuses on the last five years, from 2017 to 2021. Most research goes way longer back in time, and this can impact the regression results.

5. Textual ESG Scores

One question that we want to answer with this thesis, is whether the textual ESG scores are representative of Scandinavian firms' ESG performance. As one can observe from the Fama-French regressions, both the market premium and SMB factor are significant on the equal-weighted and value-weighted long-short portfolios of Refinitiv ESG and ESG *tf.idf*. This implies that the ESG *tf.idf* score is at least as good as the Refinitiv score in explaining the cross-sectional relation between returns and ESG performance. The correlation between the ESG *tf.idf* and the Refinitiv score is also relatively high at 0.55, falling in the range that Berg et al. (2022) find for the different ESG rating providers (0.38-0.71). This indicates that ESG tf.idf is a good proxy for ESG performance.

In contrast to the ESG *tf.idf* score, the ESG *tf* score has much lower correlation with the Refinitiv score, at only 0.30. The Fama-French regressions also explain less of the variation in the return differential between top ESG firms and bottom ESG firms. Most of the risk factors have insignificant coefficients, both the market premium and the SMB factor are insignificant, and the ones that are significant are usually not significant in the Refinitiv regressions, referring to the robust-minus-weak factor. In other words, the ESG *tf* score does not explain ESG performance the same way that Refinitiv's score does. It is unclear whether this is positive or negative. However, if Refinitiv's score is used as a quality benchmark, ESG *tf* may be less of a good proxy for ESG performance. We conclude that ESG *tf.idf* is a better measure for ESG performance, as it captures the materiality of ESG words and contains more similar attributes to Refinitiv's score.

6. Limitations

Our ambiguous regression results may be due to the Fama-French not being a good model for Scandinavian stock data. See Appendix for regressions of the whole sample on the Fama-French three factor and five factor models. The value-weighted portfolio tends to yield few significant variables. To mitigate the lack-of-model-fit one could expand the sample to European stock data, or one could construct the Fama-French factors oneself using the Scandinavian securities sample. If the model is still inaccurate, one could look to other asset pricing models, for example the q-factor model (Hou, Xue, & Zhang, Digesting Anomalies: An Investment Approach, 2015). Admittedly, the regressions should not give abnormal returns. This would violate the efficient-market hypothesis and make room for a trading opportunity.

Even Loughran & McDonald (2011) find that there is no arbitrage opportunity using US stock data, for which the Fama-French models arguably fits best.

There are also a few potential issues and challenges related to the method of textual analysis on annual reports. In most cases, companies incorporate their ESG report in their annual report, also known as *Integrated Reporting* (IR), which is overseen by the Value Reporting Foundation. However, some companies choose to include sustainability matters in a separate report. We did not retrieve these separate ESG reports. ESG scores could be negatively impacted by the fact that companies in our sample have different sustainability reporting practices. Nevertheless, in recent years IR has become much more common. Additionally, the Corporate Sustainability Reporting Directive, which would make inclusion mandatory, was adopted by the EU. As a result, the issue of several reporting frameworks will no longer exist in the coming years.

Another issue that one could potentially encounter by using textual analysis, is the possibility of 'brown' companies using their annual report as a means to *greenwash* themselves to get "rewarded" with a higher ESG score. It is not necessarily a compliance between what the company reports and what it practices. Loughran et al. (2009) learn in their analysis that the companies with the most frequent count of ethics-related words in 10-K filings are companies that are significantly more likely to be sued or that are "bad" in terms of ethics. Nonetheless, the study is done in a different market, and one could argue that it does not apply, at least not to the same extent, to Scandinavian listed companies.

In the process of cleaning the data, we could have followed a more stringent set of rules. For example, we could have followed the parsing procedure described by the Appendix in Loughran & McDonald (2011). Their procedure consists of removing graphics, headers, exhibits, tables, and HTML code. This could be done in R with regex, loops, and functions to ensure more a more comprehensive text parsing. We chose to merge our textual dataset with the Loughran & McDonald Master-Dictionary. Consequently, we could have extracted more words than we actually did. As mentioned earlier, several un-processed words were removed including British words and misspellings.

Furthermore, we could have downloaded more annual reports, for all the firms on the stock exchange and not only the ones that Refinitiv has given a score. The reasoning for not doing this was that we wanted to compare our score to Refinitiv's score. The size of the dataset was also quite large, and it took some time to download all the reports and read them in R. To get even more precise data, we could have downloaded all the annual reports in their original language, and not only the English ones, as this restricted the number of observations. Yet, then one would face problems with the ESG dictionary written in English.

7. Recommendations

Instead of aggregating the term weights in each document, one could have found the cosine similarity between each document and a benchmark document. The benchmark document would be a document that is best-in-class in terms of ESG reporting. The cosine similarity could also serve as a good addition to the other ESG scores. It is used in several financial journal articles in the context of textual analysis, for example Hanley & Hoberg (2010), Cohen, Malloy, & Nguyen (2020) and Girardi et al. (2021). The difficulty with the approach though, is to determine which document is best-in-class. For instance, one could look at the company with the best ESG score or pick the annual report in the sample that has the highest ESG term frequency.

Another recommendation is to supplement the Fama-French regressions with other regressions. For example, one could do an event study of the market's reaction to the release of an annual report and analyze cumulative abnormal returns (CAR) for top ESG firms and bottom ESG firms. Loughran & McDonald (2011) analyze the CAR on the four days around the filing date. They regress the day [0, 3] filing period buy-and-hold excess return on the proportion of negative words in the 10-K reports as well as on several control variables. Using both *tf* and *tf.idf* weights, one might carry out the same procedure but instead focus on the proportion of ESG words. One could also use filing period abnormal volume and post-event return volatility as dependent variables.

Lastly, one could make a different wordlist than the one provided by Baier et al. (2020). For instance, one could analyze the General Reporting Initiative (GRI) standards, which are available to download online. GRI is the most widely used reporting framework for sustainability, with 82% of the world's largest 250 companies using it. It is therefore a good basis for ESG dictionary creation, as the words are present in many annual reports. In the wordlist, one could opt for bigrams instead of unigrams to capture compound ESG words.

Since some words derive their meaning from their collocation with other words, they could be beneficial to include (Loughran & McDonald, 2015). An idea is to create two dictionaries: one that applies to all firms and one for each industry. The GRI Sector Standards and the GRI Universal Standards could be used to calculate term weights to produce two distinct scores: one that is universal and one that is material to the company.

8. Conclusion

Our main purpose in this thesis is to study the Scandinavian stock exchanges and investigate whether ESG is priced into the financial markets, or if there are abnormal returns to be made, which will suggest high ESG companies outperform low ESG companies. We also test to see if the textual scores are superior, complimentary or superfluous compared to the Refinitiv's scores. Refinitiv's ESG score and two scores that we have developed through textual analysis, are applied to all companies in the sample. Textual analysis is performed on annual reports in the period between 2015 and 2019. A two-year lag is used to ensure that investors have time to react after the publishing date and before the returns are realized.

There are no significant abnormal returns according to our results, and it seems like the market have priced-in ESG risk, for all scores, in the timeframe between 2017 and 2021. It seems like more awareness related to ESG reporting, as a result of demand from owners, consumers, and society in general, has led to correct pricing with no abnormal returns present. This hypothesis is therefore rejected, as there is no conclusive evidence of a relationship between the companies' stock performance and ESG scores. Hence, there are no trading opportunities to exploit. The Fama-French regressions and the Pearson correlation suggest that the ESG *tf.idf* score is complimentary to the Refinitiv score, but that the ESG *tf* may be superfluous.

In further research, we suggest using a larger corpus and securities sample to get more precise regression results. Additionally, more textual pre-processing is required to get more unbiased ESG scores. Going forward, it could also be interesting to apply the q-factor model to see if regression results will differ. Perhaps using a shorter lag than two years is wise, even so if annual reports are published sooner in a given year and investors interpret ESG matters faster. Not least, it could be interesting to apply textual analysis in other contexts, for example in an event study to see how fast share prices adjust to bad ESG publicity.

References

Ardia, D., Bluteau, K., Boudt, K., & Inghelbrecht, K. (2022). Climate Change Concerns and the Performance of Green Versus Brown Stocks. *SSRN Working Paper*, 68 pages.

Areal, N. (2021, September 10.). *frenchdata: Download Data Sets from Kenneth's French Finance Data Library Site.* Retrieved from frenchdata: Download Data Sets from Kenneth's French Finance Data Library Site: https://cran.r-project.org/web/packages/frenchdata/index.html

Auer, B. R., & Schuhmacher, F. (2016). Do socially (ir)responsible investments pay? New evidence from international ESG data. *The Quarterly Review of Economics and Finance, 59*, 51-62.

Avramov, D., Cheng, S., Lioui, A., & Tarelli, A. (2022, August). *Journal of Financial Economics, Volume 145, Issue 2, Part B, Pages 642-664.* Retrieved from Sustainable investing with ESG rating uncertainty: https://doi.org/10.1016/j.jfineco.2021.09.009

Baier, P., Berninger, M., & Kiesel, F. (2020). *Financial Markets, Inst & Inst, 29: 93-118.* Retrieved from Environmental, social and governance reporting in annual reports: A textual analysis: https://doi.org/10.1111/fmii.12132

Bali, T. G., Engle, R. F., & Murray, S. (2016). *Empirical asset pricing: the cross section of stock returns.* Hoboken, New Jersey, United States: John Wiley & Sons.

Berg, F., Fabisik, K., & Sautner, Z. (2020, November 05.). *European Corporate Governance Institute – Finance Working Paper 708/2020.* Retrieved from Is History Repeating Itself? The (Un)Predictable Past of ESG Ratings: https://dx.doi.org/10.2139/ssrn.3722087

Berg, F., Kölbel, J., & Rigobon, R. (2022, November). *Review of Finance, Volume 26, Issue 6, Pages 1315–1344.* Retrieved from Aggregate Confusion: The Divergence of ESG Ratings: https://doi.org/10.1093/rof/rfac033

Bloomberg Intelligence. (2022, May 26.). *Bloomberg Professional Service*. Retrieved from Is Tesla ESG? Many funds own it. We say it's not for everyone:

https://www.bloomberg.com/professional/blog/is-tesla-esg-many-funds-own-it-we-say-its-not-for-everyone/

Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics Volume 142, Issue 2*, Pages 517-549.

Bolton, P., & Kacperczyk, M. T. (2022). Global Pricing of Carbon-Transition Risk. *Journal of Finance, Forthcoming*, 78 pages.

Carhart, M. M. (1997, March). *Journal of Finance, Volume 52, Issue 1, Page 57-82.* Retrieved from On Persistence in Mutual Fund Performance: https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Choi, D., Gao, Z., & Jiang, W. (2020). Attention to Global Warming. *The Review of Financial Studies, 33*(3), 1112-1145.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy Prices. *The Journal of Finance, 75*(3), 1371-1415.

Condylios, S. (2022). *priceR: Economics and Pricing Tools.* Retrieved from priceR: Economics and Pricing Tools: https://cran.r-project.org/package=priceR

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, https://doi.org/10.1007/s10551-019-04164-1.

Economist. (2022, September 29.). *Economist.* Retrieved from The fundamental contradiction of ESG is being laid bare: https://www.economist.com/leaders/2022/09/29/the-fundamental-contradiction-of-esg-is-being-laid-bare

Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020, March). *The Review of Financial Studies, Volume 33, Issue 3, Pages 1184–1216.* Retrieved from Hedging Climate Change News: https://doi.org/10.1093/rfs/hhz072

European Commission. (2021, April 21.). *EUR-Lex.* Retrieved from Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Directive 2013/34/EU, Directive 2004/109/EC, Directive 2006/43/EC and Regulation (EU) No 537/2014, as regards corporate sustainability reporting: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0189

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance, 25*(2), 383-417.

Fama, E. F., & French, K. (1992, June). *Journal of Finance, Volume 42, Issue 2, Pages 427-465.* Retrieved from The Cross-Section of Expected Stock Returns: https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fama, E. F., & French, K. (2012, September). *Journal of Financial Economics, Volume 105, Issue 3, Pages 457-472.* Retrieved from Size, value, and momentum in international stock returns: https://doi.org/10.1016/j.jfineco.2012.05.011

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*(1), 3-56.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116*(1), 1-22.

Fama, E. F., & French, K. R. (2020). Comparing Cross-Section and Time-Series Factor Models. *The Review of Financial Studies, 33*(5), 1891-1926.

Fama, E. F., & MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. *The Journal of Political Economy, 81*(3), 607-636.

French, K. R. (2022). *Data Library*. Retrieved from Kenneth R. French: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Girardi, G., Hanley, K. W., Nikolova, S., Pelizzon, L., & Sherman, M. G. (2021). Portfolio similarity and asset liquidation in the insurance industry. *Journal of Financial Economics, 142*(1), 69-96.

Global Shares. (2022, August 3.). *Global Shares.* Retrieved from PSUs vs RSUs – What you need to know: https://www.globalshares.com/academy/rsus-vs-psus/

Haga, A. (2022). *Statskraft*. Retrieved from Data centres: The new power-intensive industry: https://www.statkraft.com/newsroom/news-and-stories/archive/2018/data-centres-the-new-power-intensive-industry/

Halper, J. M., Grieve, D., Bussiere, S., & Shriver, T. (2022, October 24.). *Lexology*. Retrieved from ESG Ratings: A Call for Greater Transparency and Precision:

https://www.lexology.com/library/detail.aspx?g=3cd1276b-eb1e-4c53-8912-2f3dd256fd62

Hanley, K. W., & Hoberg, G. (2010). The Information Content of IPO Prospectuses. *The Review of Financial Studies, 23*(7), 2821-2864.

Hartzmark, S. M., & Sussman, A. B. (2019). Do Investors Value Sustainability? A Natural Experiment Examining Ranking and Fund Flows. *The Journal of Finance, 74*(6), 2789-2837.

Harvey, F. (2022, November 07.). *The Guardian*. Retrieved from World is on 'highway to climate hell', UN chief warns at Cop27 summit: https://www.theguardian.com/environment/2022/nov/07/cop27-climate-summit-un-secretary-general-antonio-guterres

Hou, K., Mo, H., Xue, C., & Zhang, L. (2021, February). *Review of Finance, Volume 25, Issue 1, Page 1-41.* Retrieved from An Augmented q-Factor Model with Expected Growth: https://doi.org/10.1093/rof/rfaa004

Hou, K., Xue, C., & Zhang, L. (2015). Digesting Anomalies: An Investment Approach. *The Review of Financial Studies, 28*(3), 650-705.

Hsu, P.-H., Li, K., & Chi-Yang, T. (2022). The Pollution Premium. *Journal of Finance, Forthcoming*, 157 pages.

Jegadeesh, N., & Wu, D. (2013, December). *Journal of Financial Economics, Volume 110, Issue 3, Pages 712-729.* Retrieved from Word power: A new approach for content analysis: https://doi.org/10.1016/j.jfineco.2013.08.018

Lins, K. V., Servaes, H., & Tamayo, A. (2017, August). *Journal of Finance, Volume 72, Issue 4, Page 1785-1824.* Retrieved from Social Capital, Trust, and Firm Performance: The Value of Corporate Social Responsibility during the Financial Crisis: https://doi.org/10.1111/jofi.12505

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 13-37.

Lioui, A., & Tarelli, A. (2022). Chasing the ESG factor. *Journal of Banking & Finance, 139*, 106498.

Loughran, T., & McDonald, B. (2011, February). *The Journal of Finance, Volume 66, Issue 1, Page 35-65.* Retrieved from When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks: https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2015). The Use of Word Lists in Textual Analysis. *Journal of Behavioral Finance, 16*(1), 1-11.

Loughran, T., McDonald, B., & Yun, H. (2009, May). *Journal of Business Ethics, 89 (Suppl 1), Pages 39-49.* Retrieved from A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports: https://doi.org/10.1007/s10551-008-9910-1

Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Cambridge University Press*. Retrieved from Introduction to Information Retrieval: https://nlp.stanford.edu/IR-book/

Mossin, J. (1966). Equilibrium in a Capital Asset Market. *Econometrica*, 768-783.

Ooms, J. (2022, April 04.). *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker.* Retrieved from hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker: https://cran.r-project.org/web/packages/hunspell/index.html

Ouchen, A. (2022). Is the ESG portfolio less turbulent than a market benchmark portfolio? *Risk Management, 24*, 1-33.

Paltoglou, G., & Thelwall, M. (2010, July). *Association for Computational Linguistics*. Retrieved from A Study of Information Retrieval Weighting Schemes for Sentiment Analysis. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1386–1395 : https://aclanthology.org/P10-1141/

Pástor, Ľ., & Stambaugh, R. (2003, June). *Journal of Political Economy, Volume 111, Number 3.* Retrieved from Liquidity Risk and Expected Stock Returns: https://doi.org/10.1086/374184

Pástor, Ľ., Stambaugh, R. F., & Taylor, L. A. (2022). Dissecting green returns. *Journal of Financial Economics, 146*(2), 403-424.

Pástor, Ľ., Stambaugh, R., & Taylor, L. (2021, November). *Journal of Financial Economics, Volume 142, Issue 2, Pages 550-571.* Retrieved from Sustainable investing in equilibrium: https://doi.org/10.1016/j.jfineco.2020.12.011

Pedersen, L. H., Fitzgibbons, S., & Pomorski, L. (2021, November). *Journal of Financial Economics, Volume 142, Issue 2, Pages 572-597.* Retrieved from Responsible investing: The ESG-efficient frontier: https://doi.org/10.1016/j.jfineco.2020.11.001

Refinitiv. (2022, May). *Refinitiv*. Retrieved from Environmental, Social and Corporate Governance - ESG: https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf

Scatec. (2021, January 5.). *GlobalNewswire.* Retrieved from Scatec launches new logo and profile and changes ticker to SCATC: https://www.globenewswire.com/news-release/2021/01/05/2153051/0/en/Scatec-launches-new-logo-and-profile-and-changes-ticker-to-SCATC.html

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 425-442.

Silge, J., & Robinson, D. (2022, November 02.). *Text Mining with R: A Tidy Approach.* Retrieved from Welcome to Text Mining with R: https://www.tidytextmining.com/

Stehl, K. (2022, November 27.). *Harward Law School Forum on Corporate Governance.* Retrieved from Linking Executive Compensation to ESG Performance: https://corpgov.law.harvard.edu/2022/08/23/eu-corporate-sustainability-reporting-directive-what-do-companies-need-to-know/

Taboada, M. (2011, June 01.). *MIT Press Direct*. Retrieved from Lexicon-Based Methods for Sentiment Analysis: https://direct.mit.edu/coli/article/37/2/267/2105/Lexicon-Based-Methods-for-Sentiment-Analysis

Tayan, B. (2022, August 24.). *Harvard Law School Forum on Corporate Governance.* Retrieved from ESG Ratings: A Compass without Direction: https://corpgov.law.harvard.edu/2022/08/24/esg-ratings-a-compass-without-direction/

Teti, E., Dallocchio, M., & L'Erario, G. (2022). The impact of ESG tilting on the performance of stock portfolios in times of crisis. *Finance Research Letters, 52*.

UN. (2022, April 04.). *UN News*. Retrieved from UN climate report: It's 'now or never' to limit global warming to 1.5 degrees: https://news.un.org/en/story/2022/04/1115452

Wickham, H. (2022). *stringr.* Retrieved from stringr: Simple, Consistent Wrappers for Common String Operations: https://stringr.tidyverse.org/

Appendix

A.1 Diagnostic Plots

A.1.1 Refinitiv ESG regressions

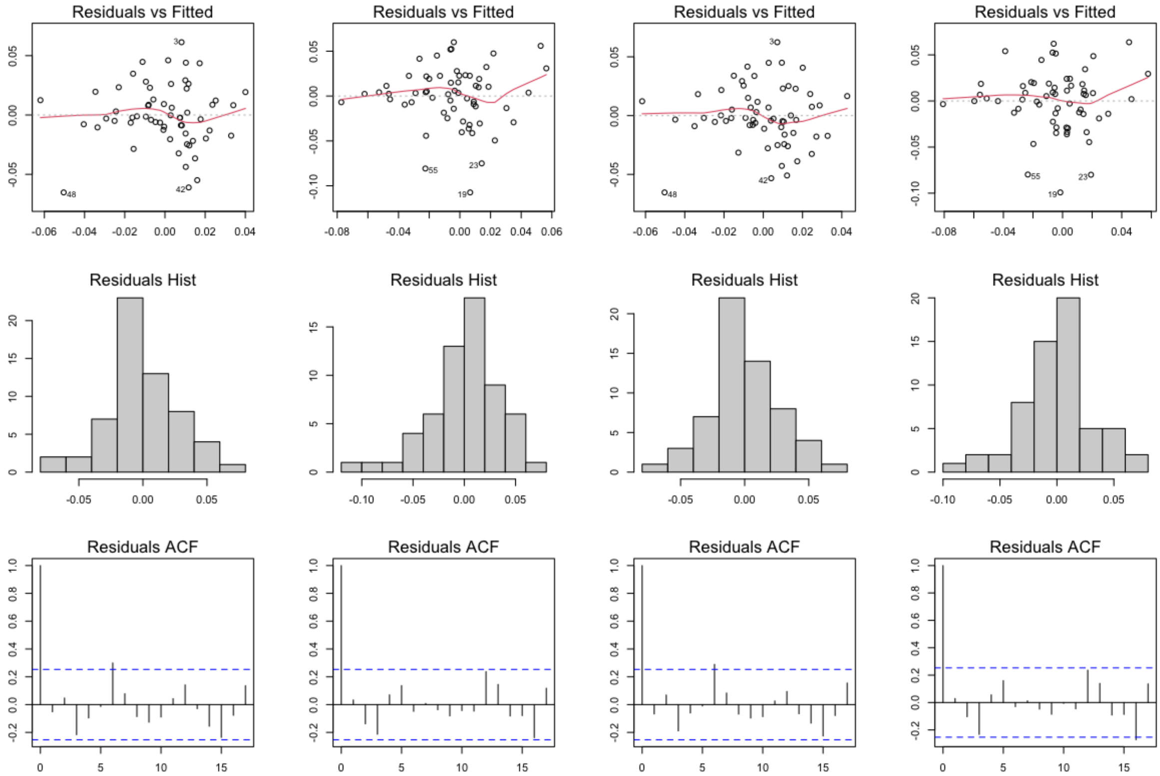

Figure 15: Diagnostic plots for the Refinitiv ESG regressions

This figure shows the residual plots, the residual histograms and the residual autocorrelation plots for the Refinitiv ESG regressions. From left to right is the equal-weighted Fama-French three-factor regression, the value-weighted Fama-French three-factor regression, the equal-weighted Fama-French five-factor regression and the value-weighted Fama-French five-factor regression.

A.1.2 ESG tf regressions

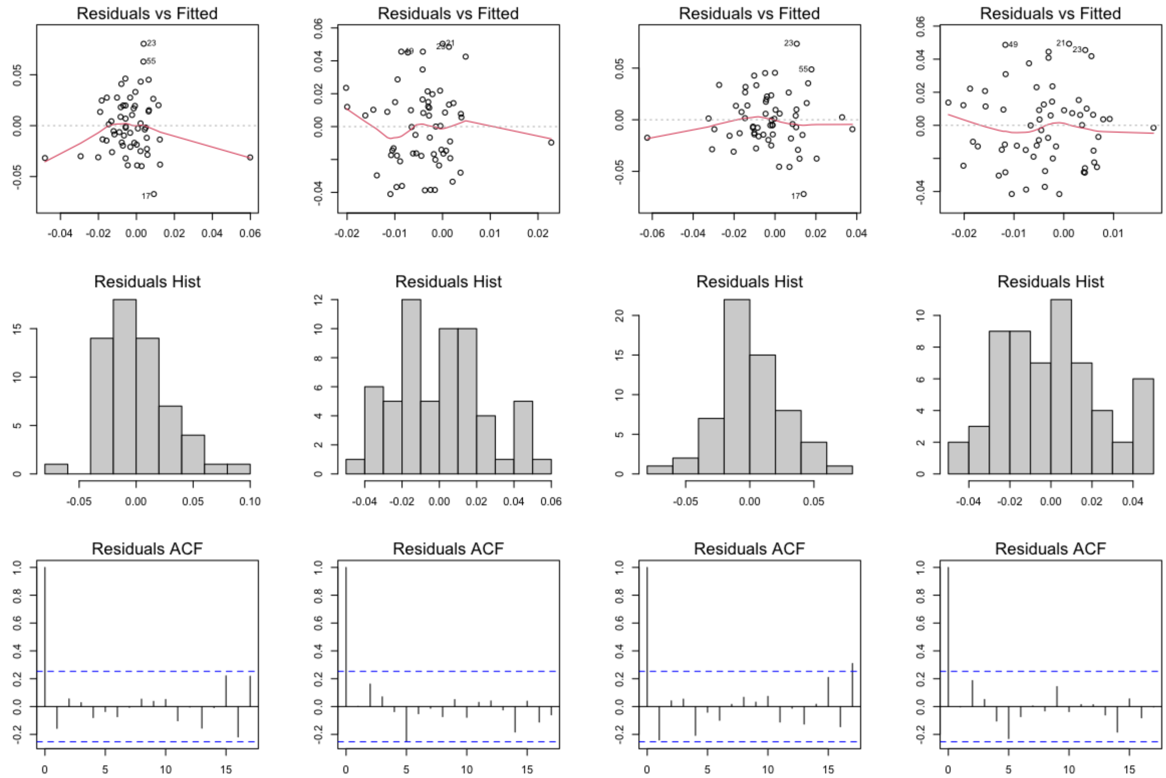

Figure 16: Diagnostic plots for the ESG tf regressions

This figure shows the residual plots, the residual histograms and the residual autocorrelation plots for the ESG *tf* regressions. From left to right is the equal-weighted Fama-French three-factor regression, the value-weighted Fama-French three-factor regression, the equal-weighted Fama-French five-factor regression and the value-weighted Fama-French five-factor regression.

A.1.3 ESG tf_idf regressions

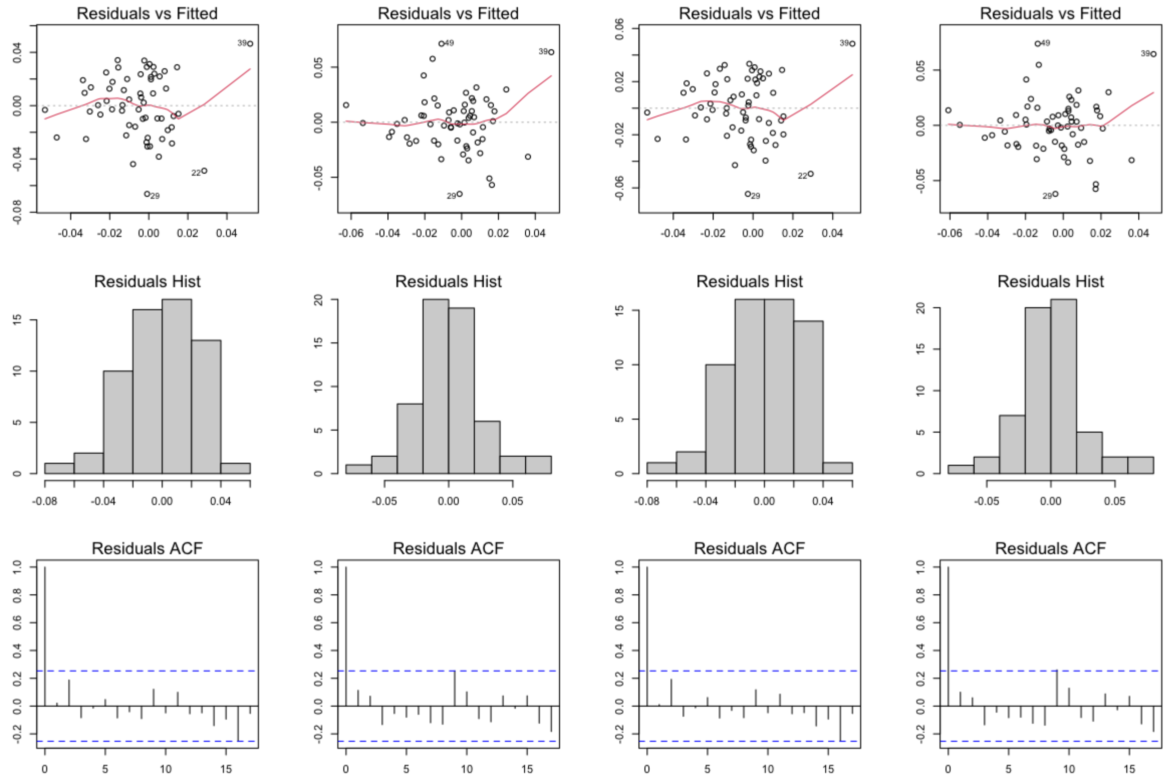

Figure 17: Diagnostic plots for the ESG tf regressions

This figure shows the residual plots, the residual histograms and the residual autocorrelation plots for the ESG *tf.idf* regressions. From left to right is the equal-weighted Fama-French three-factor regression, the value-weighted Fama-French three-factor regression, the equal-weighted Fama-French five-factor regression and the value-weighted Fama-French five-factor regression.

A.2 Fama-French Model Fit

Table 14: Regressions with the whole COMPUSTAT sample.

This table shows the regression of long short ESG portfolios on the Fama-French three- and five-factor models, portfolio returns are calculated both using equal weights and by weighting each firm by its market capitalization. The sorting variable used for this long-short portfolio is the whole COMPUSTAT sample. Standard deviations are shown in parenthesis. ***, ** and * denotes significance at the 1%, 5% and 10% level.

	The Whole COMPUSTAT Sample			
	Fama-French Three-Factor and Five-Factor Model			
	EW	VW	EW	VW
Mkt-Rf	0.785^{***}	0.762^{***}	0.721^{***}	0.687^{***}
	(0.087)	(0.056)	(0.087)	(0.058)
SMB	0.783^{***}	-0.080	0.539^{*}	-0.085
	(0.222)	(0.144)	(0.276)	(0.154)
HML	-0.211	-0.325^{***}	0.153	0.074
	(0.133)	(0.086)	(0.250)	(0.161)
RMW			-0.193	0.590^{**}
			(0.397)	(0.237)
CMA			-1.175^{**}	-0.355
			(0.461)	(0.299)
Alpha	0.006^{*}	0.005^{**}	0.003	0.003
	(0.004)	(0.002)	(0.003)	(0.002)
Observations	60	60	60	60
R^2	0.722	0.794	0.750	0.824
Adjusted R^2	0.707	0.783	0.727	0.808
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		