Technische Universität München

Department of Mathematics

Inference Analysis of Existing ESG Scoring Methodologies and an NLP-based Approach to Daily Updates

Master Thesis

by

Vanessa Theel

Supervisor:	Prof. Dr. Rudi Zagst, Prof. Dr. Luis Seco
Advisor:	Jonathan Mostovoy
Submission Date:	31.05.21

Throughout the writing of this thesis I have received a great deal of support and assistance.

First, I would like to thank my fellow researcher in this project, Flora Geske, for the fruitful and exciting teamwork in sharpening our research stream, in developing the framework, and in performing our analyses and evaluations. Your expertise and ideas have greatly enriched our discussions and have brought our work to a much higher level.

Furthermore, I would like to acknowledge my supervisors, Prof. Dr. Rudi Zagst and Prof. Dr. Seco, for their ongoing support and enthusiasm for our project and our results. In addition, I would like to thank my advisors, Jonathan Mostovoy and Alik Sokolov, for providing ongoing technical and methodological feedback and for guiding our research in the right direction.

Last but not least, I want to thank *Archive.org* for providing the Twitter data that enabled us to build and evaluate the proof-of-concept of our framework and the Technical University of Munich for providing access to *Thomson Reuters Datastream.*

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, 31.05.21

Abstract

This thesis investigates the relationship between *Refinitiv's* Environmental, Social and Governance (ESG) Scores and financial performance and introduces a new approach for updating ESG Scores more frequently using stakeholder data and leveraging state-of-the-art natural language processing (NLP) models. Firstly, we find evidence that ESG Scores based on company reportings have a negative effect on how a company is valued in the market, whereas ESG Scores based on global media reports have a considerable positive effect. Exploring the connection between ESG Scores and financial perfomance through a portfolio-based approach, we find that the market allocates a higher risk-premium to companies with low sustainability activities and that this phenomenon is concentrated on those firms that also have a rather positive sentiment portrayed in society by global media. By constructing an ESG Factor, we see a significant outperformance when creating a negative spread on sustainability activities as reported by the company and a positive spread on sustainability activities as reported by the media. Building on these results and leveraging the ever growing body of stakeholder data, we propose an ESG scoring methodology based on Twitter discourse and powered by a pipeline of state-of-the-art NLP models, which dynamically and automatically reflects ESG-relevant events in near real-time. Here, we find that our framework successfully detects material events in a Twitter sample and adjusts the ESG score accordingly, providing an efficient, objective and independent assessment of firms.

Abstract

Diese Masterarbeit untersucht die Beziehung zwischen *Refinitiv's* Environmental, Social and Governance (ESG) Scores und finanzieller Performance und stellt einen neuen Ansatz zur häufigeren Aktualisierung von ESG Scores unter Verwendung von Stakeholder-Daten und der Nutzung modernster Natural Language Processing (NLP) Modelle vor. In einer ersten Untersuchung finden wir Belege dafür, dass ESG Scores, die auf Unternehmensberichten basieren, einen negativen Effekt auf die Bewertung eines Unternehmens am Markt haben, während ESG Scores, die auf globalen Medienberichten basieren, einen erheblichen positiven Effekt haben. Bei der Untersuchung des Zusammenhangs zwischen ESG Scores und finanzieller Performance durch einen portfoliobasierten Ansatz stellen wir fest, dass der Markt Unternehmen mit geringen Nachhaltigkeitsaktivitäten eine höhere Risikoprämie zuweist und dass sich dieses Phänomen auf solche Firmen konzentriert, die zusätzlich in den globalen Medien ein eher positives gesellschaftliches Stimmungsbild aufweisen. Durch die Konstruktion eines ESG-Faktors sehen wir eine signifikante Outperformance, wenn ein negativer Spread auf Nachhaltigkeitsaktivitäten, wie vom Unternehmen berichtet, und ein positiver Spread auf Nachhaltigkeitsaktivitäten, wie von den Medien berichtet, gebildet wird. Aufbauend auf diesen Ergebnissen und unter Nutzung der ständig wachsenden Menge an Stakeholder-Daten schlagen wir eine ESG-Scoring-Methode vor, die auf Twitter-Diskursen basiert und durch eine Pipeline von hochmodernen NLP-Modellen angetrieben wird, die dynamisch und automatisch ESG-relevante Ereignisse nahezu in Echtzeit reflektiert. Wir stellen fest, dass unser Framework erfolgreich wesentliche Ereignisse in einer Twitter-Stichprobe erkennt und den ESG Score entsprechend anpasst, was eine effiziente, objektive und unabhängige Bewertung von Unternehmen ermöglicht.

Contents

Chapter 1

Introduction

> *Today, market participants, policy-makers and governments' highest priority risks — both in terms of probability and magnitude — include extreme weather events, natural disasters, water-related crises, and the failure of climate change mitigation and adaptation. In addition, investors are increasingly incorporating sustainability considerations into their investment decisions and using these indicators to screen potential investments in order to allocate capital to the most efficient users of that capital. ESG data embed a set of criteria that one can use to evaluate a company's adherence to its contract with society.*
>
> **Madelyn Antoncic [Ant20] (p. 107)**

For a very long time, businesses and corporations were fixed on the objective of maximizing shareholder wealth. However, this view has experienced a drastic hit in recent years due to a growing awareness of social responsibility. One prominent example for this change is the latest *Statement of Purpose of a Corporation* of the *Business Roundtable*, a non-profit association based in Washington, D.C. whose members are CEOs of nearly 200 major U.S. companies. In contrast to their periodically issued statements on the principles of corporate governance, which have endorsed shareholder primacy since the late 1970s, this new document outlines a modern standard for corporate responsibility where the focus is a commitment to all stakeholders [Bus20].

In addition to the business side, governments have also long realized the pressing need for a change in perspective. As one result of the *United Nations Conference on Sustainable Development* in Rio de Janeiro in 2012, all UN member states agreed on the necessity of producing a "set of universal goals that meet the urgent environmental, political and economic challenges facing our world" [Uni12]. The 17 goals were adopted by all UN member states in 2015 as part of the 2030 agenda for sustainable development which set out a 15-year plan to achieve the goals. Furthermore, the Principles for Responsible Investment (PRI), an initiative supported by the United Nations, has the goal of creating a set of principals to provide possible actions to an international network of signatories for incorporating environmental, social, governance (ESG) criteria into their investment and ownership decisions. Since the launch in April 2006, more than 3000 investment managers, asset owners and service providers have signed [Pri]. One third of the total US assets under management now use sustainable investing strategies, with a sharp incline from $12 trillion at the start of 2018 to over $17 trillion at the start of 2020. Globally, this value has almost doubled over four years, reaching over $40 trillion in 2020 [Pen20]. This growth can be attributed to rising social, governmental, and consumer attention on the companies' adherence to sustainable norms, as well as to investors and executives who realize the positive correlation between strong ESG performance and reduced risk while safeguarding a company's long-term economic success [HKN19][FBB15].

The growing focus of corporate social responsibility (CSR), socially responsible investment (SRI) or ESG criteria gave rise to a new agent: ESG rating agencies. A recent survey shows that there exist over 600 ESG ratings and rankings globally and there is no consent on a go-to rating between the users [Sus20]. What's more, studies show that the different scores and ratings lack convergence, differing in both distribution and risk [DHN15].

For all that, both investors who aim to incorporate ESG criteria into their investments and researchers who investigate the relationship between economic factors and ESG scores, are limited in gaining insight and economic success from ESG-score data by the precision and speed at which ESG scores reflect ESG-relevant events. For example, not only focusing on level and change of ESG ratings but also on their fluctuations over time is currently de facto impossible due to the static and vague character of scores in the past [DHN15].

The goal of this master thesis is to investigate the relationship between *Refinitiv's* (formerly *Thomson Reuters Asset4*) ESG Scores and financial performance and to introduce a new approach for updating ESG scores more frequently using stakeholder data and leveraging state-of-the-art natural language processing (NLP) models. In detail, we firstly compare the explanatory power of ESG scores based on company reports and global media, respectively, and show that sustainability activities as reported by the media have a significant effect. Building on this result and leveraging the ever growing body of stakeholder data (e.g. news articles or social media posts), we propose a new ESG scoring methodology and full framework for a Tweet-based scoring method. The structure is as follows: chapter 2 gives an overview of the existing research, chapter 3 introduces methods and measures used in inference detection and the NLP-framework set-up. Next, chapter

4 shows the analytical results of the inference analysis for the annual ESG scores and in chapter 5 we propose the NLP-based approach to daily ESG score updates and evaluate it in case studies. Finally, chapter 6 gives a conclusion.

Chapter 2

Motivation and Previous Research

2.1 Sustainability Activities

Sustainability awareness, both on the consumer and the business side, has grown substantially in recent years. For example, a study of *The Business Resarch Company* [The20] acknowledges the fast growing market for ethical fashion, which is expected to grow from \$6.35 billion in 2019 to \$8.25 billion in 2023, and attributes it to the growing awareness about using ethical fashion for sustainability. On the health and nutrition front, the growing awareness about animal health and animal cruelty has been encouraging people to shift from animal-based to plant-based food products. Here, a study by *The Vegan Society* showed that the meat-free food demand grew by 987% from 2012 to 2017 and the size of the plant-based meat market is expected to grow from \$4.3 billion in 2020 to a value of \$8.3 billion by 2025 [Mar20]. To paint a broader picture, a survey by *Deloitte* revealed that even, or rather especially, after the Covid-19 pandemic, millennials and GenZs remain focused on large societal issues. They continue to campaign for a world where both businesses and governments follow them in the commitment to society, emphazising the changing focus of putting people ahead of profits and prioritizing environmental sustainability [Del20]. Furthermore, consumers in general link shopping decisions to corporate values: "87% will purchase a product because a company advocated for an issue they cared about and 76% will refuse to purchase a company's products or services upon learning it supported an issue contrary to their beliefs"[Con17].

As an answer to these growing demands, there have been numerous corporate sustainability activities with the goal of reducing negative and increasing positive impacts in order to achieve greater sustainability performance [ZB17]. Examples include *Adidas*, who committed to using only recycled polyester in all their shoes and clothing within the next six years in order to increase the sustainability of their supply chain and make the business more environmentally friendly [Fin18], and *Apple*, who announced in April 2018 that they achieved 100% renewable electricity, powering its global facilities across 43 countries [RE1]. As of 2019, there are no more all-male boards in the *S&P500* companies and women held 26% of board seats, up from 15.7% in 2010 [Spe19].

More and more investors, both retail and institutional, take sustainability aspects into account when making investment decisions. A 2019 study showed that 85% of the general population and 95% of Millennials are interested in sustainable investing [Mor19]. For institutional investors, Figure 2.1 indicates this growth by depicting the signatories of the *UN Principles for Responsible Investment* (PRI) and their collective assets under management (AUM).

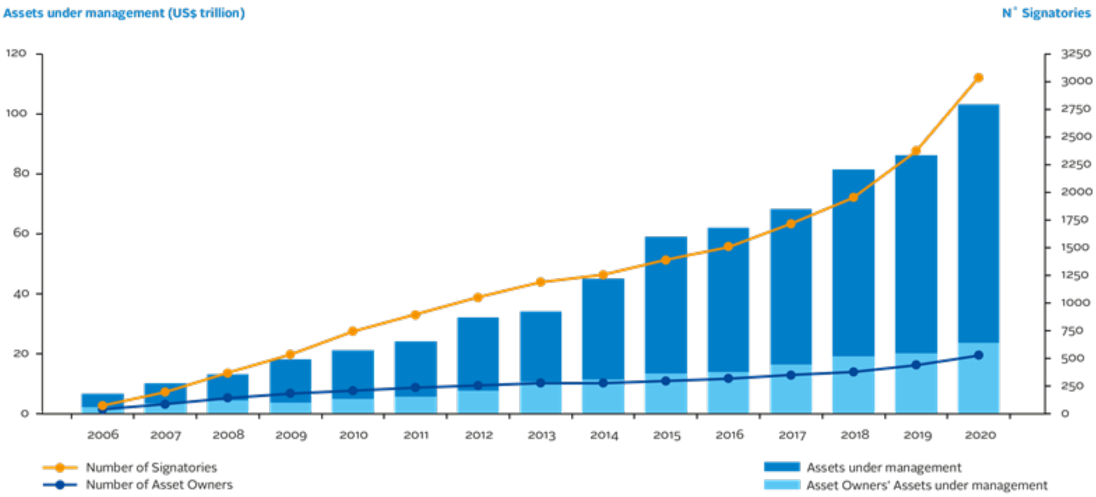

Figure 2.1: Principles for Responsible Investment - Growth [Pri]

2.2 ESG Ratings

This development sparked the need for reporting and measuring a company's sustainability performance: ESG refers to the three pillars used to evaluate a company's operations with respect to their fulfillment of and sustainable contribution to the societal contract. Environmental criteria examine a company's behavior towards nature and risk induced by the environment. Social criteria consider how it manages its business and stakeholder relationships. Governance covers the scope of a company's leadership, management and supervision. In the past decades, we have seen exponential growth in the number of companies measuring and reporting ESG data. Since the early 1990s, the number of companies issuing sustainability reports grew from around 20 to nearly $9,000$ by 2016 [AS18]. In the attempt of providing a comparable, single source for aggregated information about companies' ESG performances, multiple rating agencies started their own indices and came up with their own methodologies for quantification. Some of the largest, most established providers are *Refinitiv* (*Asset4*), *S&P Global* (*RobeccoSAM*), *KLD*, *Sustainalytics* and *MSCI* [BKR19]. Their methodologies mostly focus on expert views, checklists and company disclosures, with a growing focus on monitoring media sources as well [Susb][SP a][MSC]. However, contrary to credit ratings, ESG ratings differ substantially between providers. There is no consent or agreed-on framework, which leads to diverging and, sometimes, contradictory results. [DHN15] and [EK19] show that not only do the distributions and correlations vary, but different measure compositions and weights lead to significant distinctions in the final ratings.

2.3 ESG and Financial Performance

Owing to the extreme increase in attention, ESG ratings are not only incoporated by investors, but are also utilized for a large number of empirical studies. [FBB15] gather data and results from more than 2000 studies on the relationship between ESG and corporate financial performance from the 1970s until 2015 and find that roughly 90% report it to be non-negative and, more importantly, a majority reports it to be positive. [Alb13] compare 52 studies over a 35-year period with differing results for the relationship between environmental and financial performance, and show that the inconsistency of the studies' findings can be explained by the measures used, the regional differences and the duration of the studies. In a more recent paper [PSK19] argue that "there has never been conclusive evidence that socially responsible screens or company positions [...] deliver alpha" (p. 2).

Let us look at some of these studies in more detail. Over 25 years ago, [CFN95] already pointed out the lack of consensus in prior research and proposed an objective way of measuring environmental performance based on government records rather than company disclosures. Based on this, in a portfolio-based analysis on the *S&P500* stocks, they find "the 'low pollution' portfolio does as well as - and often better than - the 'high pollution' group" (p. 4). Revolutionizing the view of which ESG issues to consider, [KSY16] show that there is evidence of a positive relationship between performance on *material* ESG issues and financial performance, whereas there is no significant relationship on *immaterial* ESG issues, by evaluating portfolios based on *KLD* ESG data. Using *Bloomberg* data, [Hen+19] find that companies with a better ESG score have higher valuations, however, there is no statistically significant difference in returns between well and poor performing ESG companies. Using *Asset4* data, like we will be using in this thesis, [RB10] find that the ESG scores provided have significant value as stock selection factors, especially for a long-term investment horizon. Furthermore, [Vel17] evaluate the relationship between *Asset4* ESG data and accounting and market-based measures of financial performance on a sample of companies listed on the German Prime Standard, where they find a positive impact on the Return-on-Assets but not on Tobin's Q. Leveraging the social performance score from *KLD* and with the goal of wanting to overcome the present causality issue in previous studies on the relation between corporate social responsibility (CSR) and firm value, [DKL13] investigate mergers as largely unanticipated events and find that high CSR acquirers are more likely to complete the deal, realize higher merger announcement returns and have larger increases in post-merger long-term operating performance compared to low CSR acquirers.

2.4 ESG Sentiment

Previous literature and research discuss a connection between investor sentiment and financial performance in an overall market context. [BW06] present evidence, as one of the first, that investor sentiment may have significant effects on the cross-section of stock prices and, more importantly, that several firm characteristics only show their predictive power after conditioning on sentiment. Following this, [SYY12] find that a set of market anomalies are more pronounced, i.e. their long-short strategies are more profitable, after a period of high investor sentiment.

In the setting of ESG, [KÇ18] find that companies with both higher environment-focused CSR activities and social performance exhibit lower returns following high sentiment periods, where investor sentiment is proxied for the overall market. To the best of our knowledge, [Ser20] is the first and only one to evaluate the influence of *investor ESG sentiment* (as provided by *TruValue Labs*) combined with corporate ESG data (as provided by *MSCI*) on financial valuation and performance. They find that public sentiment about a company's sustainability activities influences its valuation and returns of portfolios which consider ESG data. In more detail, they present evidence that in the presence of negative public ESG sentiment, the company's sustainability activities are valued less and are related to positive abnormal returns in the future. Inspired by this paper, in the first part of this thesis we will investigate the influences on corporate valuation and portfolio returns using ESG scores and ESG sentiment as provided by *Refinitiv* before proposing our own methodology to measure ESG sentiment as portrayed by stakeholders in social media to update ESG scores on a daily basis in the second part.

2.5 Machine Learning in ESG Ratings

Machine learning (ML) allows us to collect, process and analyze vast amounts of unstructured data to gather more information than ever before when evaluating a company's environmental, social, and governance risks and opportunities. Tasks like manually going through checklists or reading company disclosures can now be complemented or even replaced by ML algorithms that automatically evaluate the tone and content of news articles, NGO reports and social media posts. As the demand for ESG data grows, this gives way to reducing the manual effort while providing accurate, real-time and consistent responses to ESG issues.

In the industry, ESG rating providers are increasingly adopting ML models and other technologies to benefit from these advantages [MSC; Ref; SP b]. Some providers, like *TruValue Labs*, *Covalence* or *Sensefolio*, even offer real-time sustainability data analytics by analyzing publicly available texts about companies (including articles, blog posts, and Twitter) relying (almost) completely on artificial intelligence. However, these models are a complete black-box for users - there is no clear framework or agreed upon data basis on which to evaluate these models in the industry.

In research, there have been many advancements in NLP models for text classification and sentiment analysis, but only few explore the ability of using it in an ESG context. For example, [Nem+19] introduce a framework that detects and monitors controversial events by clustering Tweets using feature vectors. [RBN20] show how state-of-the-art NLP models can be leveraged to detect historical trends in ESG discussions by analyzing the transcripts of corporate earning calls. This thesis builds on the work of [Sok+20] who propose an approach to constructing automatic ESG indices using deep learning techniques for NLP and evaluating social media data.

Chapter 3

Preliminaries

The goal of this master thesis is two-fold: (1) we want to analyze the relation between existing, annual ESG scores and financial valuation, (2) we introduce a novel approach to using Twitter data for daily updated ESG scores. The following sections aim to present the methods, algorithms and metrics used in this thesis to reach these goals. For the first part, we will use *multiple linear regression* models (Section 3.1) to detect inference between the sustainability scores and company valuation in the form of the *Price-to-Book* multiple, and build score-based *portfolios* (Section 3.2) to investigate the financial performance. In a more in-depth analysis, we will further split our investigation into turbulent and calm market phases using *Markov Switching Models* (Section 3.3) to analytically detect times of crises. For the second part, we will leverage a state-of-the-art NLP model (Section 3.4) to build our framework for daily ESG score updates.

3.1 Regression

The first part of this master thesis focuses on providing extensive insight into the relationship between ESG Scores and financial market valuation in the form of a multiple. We investigate the explanatory power through model estimation. The following sections are an introduction into the statistical methods used for inference detection.

Univariate regression analysis is an approach to modelling the expectation and dependence of quantity $\mathbf{Y}$, referred to as *response, regressand* or *dependent variable*, on quantities $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_k}$ referred to as *explanatory variables, regressors, features* or *independent variables* [HKR15]. The relationship between the *dependent variable* and the *independent variable* is estimated on the base of n observations $(x_1, y_1) \ldots (x_n, y_n)$, where each $\mathbf{x_i} = (x_{i1}, \ldots, x_{ik})^T$ is a vector of feature measurements for the ith realisation of $\mathbf{Y}$, generated by some population, for which insights into important characteristics can be derived from the resulting statistical model [Oli03]. *Linear regression* is a simple and widely used and understood method, which often provides an adequate and interpretable description of how the inputs affect the output [FHT+01]. In the following, we refer to [Oli03] and [Gre00] for the definitions and theorems of linear regression estimation and inference.

3.1.1 Model Formulation

We start by introducing the general model definition.

Definition 3.1.1 (Multiple Linear Regression Model)**.** The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad \text{for } i = 1, \ldots, n \tag{3.1}$$

or in matrix notation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3.2}$$

is called a *multiple linear regression* model with sample size n on the response y with k regressors $x_1, \ldots, x_k$. It is called linear, since the equation 3.1 is a linear function of unknown parameters $\beta_0, \ldots, \beta_k$ called the regression coefficients.

The following assumptions are needed for inference on linear regression models.

Assumption A (Linearity)**.** The model 3.2 is linear in the parameters $\boldsymbol{\beta}$.

Assumption B (Full Rank)**.** $\boldsymbol{X}$ is an $n \times k$ matrix with rank k.

Assumption C (Exogeneity of the independent variables)**.** The *independent variables* will not carry useful information for the prediction of $\boldsymbol{\epsilon}$, i.e.

$$\mathbb{E}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{0} \tag{3.3}$$

Assumption D (Homoscedasticity and non-autocorrelation)**.** The disturbance in the regression has conditional variance

$$\boldsymbol{Var}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \boldsymbol{Var}[\boldsymbol{\epsilon}] = \boldsymbol{\sigma^2} \tag{3.4}$$

With reference to a time-series setting, this will be labeled nonautocorrelation

$$\mathbb{E}[\boldsymbol{\epsilon\epsilon^T}|\boldsymbol{X}] = \boldsymbol{\sigma^2 I} \tag{3.5}$$

Assumption E (Data generation)**.** $\boldsymbol{X}$ may be fixed or random.

Assumption F (Normal distribution)**.** The errors are normally distributed

$$\boldsymbol{\epsilon}|\boldsymbol{X} \sim \boldsymbol{\mathcal{N}(0, \sigma^2 I)} \tag{3.6}$$

3.1.2 Model Estimation

Having introduced the general form and assumptions of the *linear regression* model, we now establish its estimation theory. First, we define the estimated model with its fitted values and residuals. Then, we propose the *Ordinary Least Squares* (OLS) estimator of the regression coefficients $\boldsymbol{\beta}$, a fundamental result regarding its properties under the model assumptions and further model estimates. Here, it is important to distinguish between unobserved population quantities such as $\boldsymbol{\beta}$, $\boldsymbol{\epsilon}$ and σ^2 and their respective sample estimates $\boldsymbol{b}$, $\boldsymbol{e}$ and s^2.

Definition 3.1.2. Given the estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{y}} \equiv \widehat{\boldsymbol{y}}(\boldsymbol{b}) = \boldsymbol{Xb}$. Thus, the ith fitted value is given as

$$\hat{y}_i \equiv \hat{y}_i(\mathbf{b}) = \mathbf{x_i^T b} = b_0 + x_{i1}b_1 + \cdots + x_{ik}b_k \tag{3.7}$$

where $\mathbf{x_i^T}$ corresponds to the ith row of $\boldsymbol{X}$.
The vector of residuals is $\mathbf{e} \equiv \mathbf{e}(\mathbf{b}) = \mathbf{Y} - \hat{\mathbf{Y}}(\mathbf{b})$.

Theorem 3.1.3 (Ordinary Least Squares Estimate)**.** The *least squares coefficient vector* $\boldsymbol{b}^* = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$ minimizes the sum of squared residuals, i.e.

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{i=1}^{n} e_i(b)^2 = \arg\min_{\mathbf{b}} \mathbf{e}(\mathbf{b})^{\mathbf{T}}\mathbf{e}(\mathbf{b}) = \arg\min_{\mathbf{b}} (\mathbf{y} - \mathbf{Xb})^{\mathbf{T}}(\mathbf{y} - \mathbf{Xb}) \tag{3.8}$$

Proof. See [Gre00]. □

Theorem 3.1.4 (Gauß-Markov Theorem)**.** In the *linear regression* model 3.1 under Assumptions A-D,

- the *least squares estimator* $\boldsymbol{b}^*$ is the *minimum variance unbiased estimator* of $\boldsymbol{\beta}$
- for any vector of constants $\boldsymbol{w}$, the *minimum variance linear unbiased estimator* of $\boldsymbol{w^T\beta}$ is $\boldsymbol{w^T b}^*$

Proof. See [Gre00]. □

Theorem 3.1.5. An unbiased estimator of the population parameter σ^2 is given by

$$s^2 = \frac{\mathbf{e^T e}}{n-k} \tag{3.9}$$

Proof. See [Gre00]. □

Corollary 3.1.6. An estimator for the conditional covariance matrix of the *least squares estimator* $Var[\boldsymbol{b}|\boldsymbol{X}] = \sigma^2(\boldsymbol{X^T X})^{-1}$ is given by

$$\widehat{Var}[\boldsymbol{b}|\boldsymbol{X}] = s^2(\boldsymbol{X^T X})^{-1} \tag{3.10}$$

Proof. See [Gre00]. □

3.1.3 Model Inference

Now that we know the functional form of a *multiple linear regression* model, have learned how to estimate model parameters using the OLS method and derived properties of the estimates, the next important questions we would like to answer are: (1) How powerful is our model of regressors with respect to being able to explain the regressand? (2) Which regressors are important? For this, we firstly introduce the *residual maker* as a helping concept in order to define a measure for the goodness of fit for the linear regression model.

Definition 3.1.7 (Residual Maker)**.** Let the $n \times k$ full column rank matrix $\boldsymbol{X}$ be composed of columns $(\boldsymbol{x_1, x_2, \ldots, x_k})$, and let $\boldsymbol{y}$ be an $n \times 1$ column vector. The matrix

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T} \tag{3.11}$$

is a *residual maker* in that when $\boldsymbol{M}$ premultiplies a vector $\boldsymbol{y}$, the resulting $\boldsymbol{My}$ is the column vector of residuals in the *least squares regression* of $\boldsymbol{y}$ on $\boldsymbol{X}$.

As a first idea to measure the fit of the regression line to the data, the fitting criterion used in the OLS method, the sum of squared residuals, seems natural. However, as can easily be verified, the sum of squared residuals is highly dependent on the scale of $\boldsymbol{y}$. Because the *fitted values* of the regression are based on the values of $\boldsymbol{X}$, we might ask instead, whether variation in $\boldsymbol{X}$ is a good predictor of variation in $\boldsymbol{y}$. Hence, we propose the *coefficient of determination* R^2 as a ratio to measure the goodness of fit.

Definition 3.1.8 (R-squared)**.** The total variation in y is the sum of squared deviations, i.e. the total sum of squares

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \boldsymbol{y^T M^0 y} && (3.12)\\ &= \boldsymbol{b^T X^T M^0 X b} + \boldsymbol{e^T e} && (3.13)\\ &= \text{regression sum of squares (SSR)} + \text{error sum of squares (SSE)} && (3.14) \end{aligned}$$

where $\boldsymbol{M^0}$ is the *residual maker* for $\boldsymbol{X} = \boldsymbol{i}$, i.e. the $n \times n$ idempotent matrix that transforms observations into deviations from sample means. From this we can obtain a measure of how well the regression line fits the data by using the *coefficient of determination* R^2

$$R^2 = \frac{SSR}{SST} = \frac{\boldsymbol{b^T X^T M^0 X b}}{\boldsymbol{y^T M^0 y}} = 1 - \frac{\boldsymbol{e^T e}}{\boldsymbol{y^T M^0 y}} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.15}$$

However, this measure for the goodness of fit has a few drawbacks. The biggest is that the R^2 will never decrease when another variable is added to a regression equation [Gre00]. In order to control for the degrees of freedom, the *adjusted* R^2 ($\bar{R}^2$) incorporates a penalty for the number of variables k.

Definition 3.1.9 (Adjusted R-squared)**.** The *adjusted* R^2 is given by

$$\bar{R}^2 = 1 - \frac{\boldsymbol{e^T e}/(n-k)}{\boldsymbol{y^T M^0 y}/(n-1)} \tag{3.16}$$

and may decline when a variable is added to the set of independent variables. It depends on whether the contribtion of the new variable to the fit of the regression more than offsets the correction for the loss of an additional degree of freedom.

The Gauß-Markov Theorem 3.1.4 assures us that the OLS estimate of the regression coefficients has the smallest variance among all unbiased estimators. However, this result gives no actual indication of the absolute magnitude of the variance. If we look at the variance of the estimator in Corollary 3.1.6, we see that it will blow up, if the matrix $\boldsymbol{X^T X}$ is close to being singular. This means that if an *explanatory variable* can almost be explained by a linear combination of other *explanatory variables*, the variance of the *least squares estimator* will become huge.

Definition 3.1.10 (Variance Inflation Factor)**.** The *variance inflation factor* (VIF) for a variable shows the increase in $Var[b_j]$ that can be attributed to the fact that this variable is not orthogonal to the other variables in the model. It is computed by regressing all but the jth *explanatory variables* on the jth *explanatory variable*, yielding an *R-squared* R_j^2:

$$VIF = \frac{1}{1 - R_j^2} \tag{3.17}$$

It follows that the more highly correlated a variable is with the other variables in the model, the greater the VIF will be. As a rule of thumb, values greater than 10 should be further investigated, because they could indicate multicollinearity [FG67].

Hence, before we are able to plug in all of our available *features* to find out, whether and how they influence the *regressand*, we have to investigate how they influence each other. With this result, we come back to methods for model inference. So far, the presented measures have focused on answering question (1) How powerful is our model of *regressors* with respect to being able to explain the *regressand*? Now, we want to deepen the analysis towards answering (2) Which *regressors* are important? To answer this question, a natural approach would be to fit the regression model with our data sample and examine the estimated *regression coefficients*. If they are zero, we would argue that the respective *explanatory variable* is not actually explanatory. However, since the *least squares coefficient* is a random variable which will almost surely never be exactly zero, even if its estimate is, we will instead check whether the sample estimate seems to be close enough to zero for us to conclude that its population counterpart is actually zero.

The following approach is to formulate a hypothesis about the *regression coefficients* as a restriction on the model. In detail, the null hypothesis is defined as the statement that limits the model and the alternative hypothesis is defined as the broader one.

Definition 3.1.11 (General Linear Hypothesis)**.** The *general linear hypothesis* is a set of J restrictions on the *linear regression* model 3.2. The restrictions are written

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1k}\beta_k &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2k}\beta_k &= q_2 \\ &\cdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{Jk}\beta_k &= q_J \end{aligned}$$

and can be written in matrix form

$$\boldsymbol{R\beta} = \boldsymbol{q} \tag{3.18}$$

The hypothesis implied by the restrictions, which are representated by each row of $\boldsymbol{R}$, is written

$$H_0 : \boldsymbol{R\beta} - \boldsymbol{q} = \boldsymbol{0}, H_1 : \boldsymbol{R\beta} - \boldsymbol{q} \neq \boldsymbol{0}$$

The matrix $\boldsymbol{R}$ has k columns to be conformable with $\boldsymbol{\beta}$, J rows for a total of J restrictions, and full row rank, i.e. $J < k$.

We will mainly focus on the following two cases:

1. One of the coefficients is zero, i.e. $\beta_j = 0$,

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}; q = 0$$

2. All of the coefficients in the model except the constant term are zero,

$$\boldsymbol{R} = \begin{pmatrix} 0 & | & \mathbf{I}_{k-1} \end{pmatrix}; q = 0$$

Definition 3.1.12 (Wald test for single coefficient)**.** The *Wald test*, often called a *significance test*, is the most commonly used procedure for testing the variables' significance. Here, the regression is fit without imposing the restrictions in order to assess whether the results agree with the hypothesis.

The *Wald distance* of a *coefficient estimate* from a hypothesized value is the distance measured in standard deviation units. In the single coefficient case, this means the distance of b_j from β_j^0 would be

$$W_j = \frac{b_j - \beta_j^0}{\sqrt{\sigma^2 S^{jj}}} \tag{3.19}$$

where S^{jj} denotes the jth diagonal element of $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}$. W_j has a standard normal distribution assuming that $\mathbb{E}[b_j] = \beta_j$.
Since σ^2 is unknown, we compute W_j using the sample estimate of σ^2 and obtain

$$t_j = \frac{b_j - \beta_j^0}{\sqrt{s^2 S^{jj}}} \tag{3.20}$$

Under the assumption of the null hypothesis that β_j equals β_j^0, t_j has a t distribution with $n-k$ degrees of freedom. This yields the confidence interval

$$Prob\{-t^*_{(1-\alpha/2),[n-k]} < t_j < t^*_{(1-\alpha/2),[n-k]}\} \tag{3.21}$$

where α is the desired level of confidence and $t^*_{(1-\alpha/2),[n-k]}$ the appropriate critical value from the t table.

This test statistic constructed for the hypothesis $\beta_j^0 = 0$ is commonly used and the statistic is usually labeled *t-ratio* for the estimator b_j, which will appear in regression results in the following chapters.

Next, we introduce a test statistic to evaluate the fit of the overall model in order to be able to answer the question whether the regression equation as a whole is significant.

Definition 3.1.13 (F-statistic for complete model)**.** Let $\boldsymbol{b}_*$ be the *least squares estimator* for the model 3.2 under the restrictions $\boldsymbol{Rb}_* = \boldsymbol{q}$. Furthermore, let R^2 and R^2_* denote the goodness of fit measures for the unrestricted and restricted model, respectively. Then we obtain the *F-statistic*

$$F[J; n-k] = \frac{(R^2 - R^2_*)/J}{(1-R^2)/(n-k)} \tag{3.22}$$

where the joint test that all the coefficients except the constant term are zero translates to $R^2_* = 0$.
Under the assumption of the null hypothesis that $\beta_1, \ldots, \beta_k = 0$ and normally distributed disturbances, this statistic has an F distribution with $k-1$ and $n-k$ degrees of freedom. Large values of F are induced by large values of R^2 and lead to a rejection of the hypothesis.

Lastly, we will look at what inference we can draw from the estimated coefficient values. Here, let us consider the following cases:

1. **General Form:**
   If neither the *regressand* nor any of the *regressors* are transformed, i.e. the model is given in form of equation 3.1, the *coefficient estimate* b_j for the jth *explanatory*

variable can be interpreted as the change in y_i for one unit change in x_{ij}, if all other variables remain constant.

$$\begin{aligned}
&\mathbb{E}[y_i|x_{i1}, x_{i2}, \dots, x_{ij}+1, \dots, x_{ik}] - \mathbb{E}[y_i|x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ik}] \\
&= (b_0 + b_1 x_{i1} + \dots + b_j(x_{ij}+1) + \dots + b_k x_{ik}) - (b_0 + b_1 x_{i1} + \dots + b_j x_{ij} + \dots + b_k x_{ik}) \\
&= b_j
\end{aligned}$$

2. **Log-Transformed Regressand:**
If simply the *regressand* is log-transformed but none of the *regressors* are transformed, i.e. the model is given in form of equation

$$ln(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i \quad \text{for } i = 1, \dots, n$$

the *coefficient estimate* b_j for the jth *explanatory variable* can be interpreted as the percentage change in y_i for one unit change in x_{ij}, if all other variables remain constant.

$$\begin{aligned}
\frac{\partial \mathbb{E}[ln(y_i)|x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ik}]}{\partial x_{ij}} &= b_j \\
\frac{\partial \mathbb{E}[ln(y_i)|x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ik}]}{\partial y_i} &= \frac{1}{y_i} \\
\iff \partial \mathbb{E}[ln(y_i)|x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ik}] &= \frac{\partial y_i}{y_i} \\
\Rightarrow b_j = \frac{\partial y_i / y_i}{\partial x_{ij}}
\end{aligned}$$

Having introduced the main aspects, estimators and statistics needed for studying a *multiple linear regression* model, we will now focus on the particularities of this thesis' problem. The data which we will investigate in the following chapters are annual observations from a set of companies, namely the current *S&P500* constituents, over the timeframe 2004-2020. This means, we will be working with panel data, a combination of time series and cross sections. An important characteristic to notice and consider before building models on panel data are *group-* or *time-specific effects* inherent to it. Not taking these into account could lead to heteroscedasticity (compare Assumptions D). The following definition introduces these *fixed effects* into the model structure and gives a result for estimation with observed effects.

Definition 3.1.14 (Fixed Time and Group Effects Model)**.** When analyzing panel data one faces *fixed effects* on group- and time-specific variables, i.e. the regression model 3.1 has the form

$$y_{it} = \mathbf{x_{it}}^T \boldsymbol{\beta} + \boldsymbol{z_i}^T \boldsymbol{\alpha} + \boldsymbol{w_t}^T \boldsymbol{\gamma} + \epsilon_{it} \tag{3.23}$$

If $\boldsymbol{z_i}, \boldsymbol{w_t}$ are observed for all individuals, then the entire model can be treated as an *ordinary linear* model and fit by *least squares.*

Since we have *sector fixed effects* and *yearly fixed effects* in our data, which are observable and part of our sample, all previously stated methods and results are still valid.

3.2 Portfolio Statistics

After firstly investigating the relationship between ESG Scores and financial market valuation on a company-level, the second part of this master thesis focuses on providing extensive insight into the relationship between ESG Scores and return on a portfolio-level. The following section gives are an introduction into the methods used in the approach and for inference detection.

When constructing and managing a portfolio one wants to be aware of all the risks associated with it and monitor its performance. Since our goal in section 4.2 is to evaluate and compare portfolios constructed using different sustainability scores in order to draw conclusions from them, we now discuss techniques and measures to help understand and track both the performance and the risk of a portfolio.

Definition 3.2.1 (Portfolio)**.** A *portfolio* is a collection of financial assets like stocks, bonds, commodities, or cash.

Definition 3.2.2 (Mean Return)**.** Looking back at how a *portfolio* has performed one aims to capture its *trend.* In general, this is done by calculating the average of past returns from time points $t = 1, \dots, T$ with $r = (r_t)_{t=1,\dots,T}$ denoting the time series of returns.

$$\hat{\mu} = \widehat{\mathbb{E}[r]} = \bar{r} = \frac{1}{T}\sum_{t=1}^{T} r_t$$

which can be used as an estimate for the *expected return.*
However, there are different ways of defining a *portfolio return.*

- **Discrete Return**

$$P_{t-1} \cdot (1 + r_t^d) = P_t \iff r_t^d = \frac{P_t - P_{t-1}}{P_{t-1}}$$

- **Continuous Return**

$$P_{t-1} \cdot e^{r_t^c} = P_t \iff r_t^c = \log \frac{P_t}{P_{t-1}} = \log(1 + r_t^d)$$

In the following, we will use *discrete returns*. Furthermore, we have monthly returns $r_t^{p.m.}$ in the analysis in this thesis. In order to annualize the measure of *mean return* we perform the following computations based on the assumption that the monthly returns are independent and identically distributed (i.i.d.):

$$\begin{aligned}\mu_{p.a.} &= \mathbb{E}[r_{p.a.}] = \mathbb{E}[\prod_{t=1}^{12}(1+r_t^{p.m.})-1] \\ &= \prod_{t=1}^{12}\mathbb{E}[(1+r_t^{p.m.})]-1 = (1+\mu_{p.m.})^{12}-1\end{aligned}$$

Hence, we define an estimator of the annual mean return as

$$\hat{\mu}_{p.a.} = (1+\hat{\mu}_{p.m.})^{12}-1$$

Having talked about the first order moment as an estimate for the *portfolio trend*, we now discuss the explanatory power of higher order moments.

Definition 3.2.3 (Standard Deviation - Single Asset)**.** One commonly used and easily computable indicator for the riskiness of an asset is the *standard deviation* of its returns. It represents how large an asset's prices swing around the mean price. Hence, it is a statistical measure of its dispersion of returns.
One looks at past returns to calculate an estimate of the *standard deviation* as a measure for return volatility:

$$\hat{\sigma} = \sqrt{\widehat{Var}(r)} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(r_t-\bar{r})^2}$$

When using this measure of volatility in a portfolio context, one needs to take the *covariance/correlation* between assets in the portfolio into account.

Definition 3.2.4 (Covariance and Correlation)**.** Both terms are statistical tools that are used to determine the relationship between two variables, here the movement of two asset prices. *Covariance* indicates the direction of the linear relationship between asset returns. When two stocks tend to move together, they are seen as having a positive *covariance*; when they move inversely, the *covariance* is negative. An estimate $\hat{\sigma_{1,2}}$ of the *covariance* is obtained from the following, where we consider two assets $i = 1, 2$ with returns $r_i = (r_{i,t})_{t=1,\dots,T}$:

$$\hat{\sigma}_{1,2} = \widehat{Cov}(r_1, r_2) = \frac{1}{T-1}\sum_{t=1}^{T}(r_{1,t}-\bar{r_1})(r_{2,t}-\bar{r_2})$$

Correlation, on the other hand, measures both the strength and direction of the linear relationship and is a function of the *covariance*. The difference is that *correlation* values are standardized. Its values are $\in (-1, 1)$ with -1 meaning they move perfectly opposite

and 1 meaning they move exactly the same.
An estimate $\hat{\rho}$ of the *correlation* is obtained from the following, where we consider two assets $i = 1, 2$ with returns $r_i = (r_{i,t})_{t=1,\dots,T}$:

$$\hat{\rho}(r_1, r_2) = \frac{\hat{\sigma}_{1,2}}{\hat{\sigma}_1 \cdot \hat{\sigma}_2}$$

Now we can look at *standard deviation* as a measure of portfolio volatility.

Definition 3.2.5 (Standard Deviation - Portfolio)**.** Having learned about estimates for both the *covariance* and *correlation* of assets, we can compute an estimate $\hat{\sigma}_\pi$ of the *standard deviation* for a portfolio Π (consisting of assets $i = 1, \dots, N$ with portfolio weights w_i) from the individual asset returns.

$$\hat{\sigma}_\pi = \sqrt{\widehat{Var}(r_\pi)} = \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \hat{\sigma}_{i,j}} = \sqrt{w \hat{\Sigma} w}$$

Again, the analyses in this thesis are based on monthly returns $r_t^{p.m.}$. In order to annualize the measure of *standard deviation* we perform the following computations based on the assumption that the monthly returns are i.i.d. and using the fact that $Var(X + b) = Var(X)$ for a random variable X and scalar b:

$$\begin{aligned}
\sigma_{p.a.} = \sqrt{Var(r_{p.a.})} &= \sqrt{Var(\prod_{t=1}^{12}(1 + r_t^{p.m.}) - 1)} \\
&= \sqrt{\mathbb{E}[\left(\prod_{t=1}^{12}(1 + r_t^{p.m.})\right)^2] - (\mathbb{E}[\prod_{t=1}^{12}(1 + r_t^{p.m.})])^2} \\
&= \sqrt{\prod_{t=1}^{12} \mathbb{E}[(1 + r_t^{p.m.})^2] - ((1 + \mu_{p.m.})^{12})^2} \\
&= \sqrt{\prod_{t=1}^{12}(1 + 2\mathbb{E}[r_t^{p.m.}] + \mathbb{E}[(r_t^{p.m.})^2]) - (1 + \mu_{p.m.})^{24}} \\
&= 2\sqrt{((1 + \mu_{p.m.})^2 + \sigma_{p.m.}^2)^{12} - (1 + \mu_{p.m.})^{24}}
\end{aligned}$$

Hence, we define an estimator of the annual standard deviation as

$$\hat{\sigma}_{p.a.} = \sqrt{((1 + \hat{\mu}_{p.m.})^2 + \hat{\sigma}_{p.m.}^2)^{12} - (1 + \hat{\mu}_{p.m.})^{24}}$$

After computing the *average return* and *standard deviation* of the portfolio of interest, another measure, which combines the view on return and volatility by computing the risk-adjusted excess return over a benchmark, is given by the *Sharpe Ratio*.

Definition 3.2.6 (Sharpe Ratio)**.** The *Sharpe ratio* [Sha94] can help explain whether a portfolio's excess returns are due to good investment decisions or a result of too much risk (under the assumption that risk equals volatility). Even though a portfolio can generate higher returns than others, it is only a good investment if those higher returns do not come with an excess of additional risk.
Hence, we compute the *Sharpe Ratio* by substracting a benchmark return r_B (often set as the risk-free rate r_f) from the portfolio return $\bar{r}_\pi$ and dividing by the portfolio volatility σ_π.

$$SR_\pi = \frac{\bar{r}_\pi - r_B}{\sigma_\pi}$$

An extension of this measure allowing for time-varying benchmark return is the *expected Sharpe Ratio* [Zag+10].

$$\mathbb{E}[SR_\pi] = \frac{\mathbb{E}[r_\pi - r_B]}{\sqrt{Var[r_\pi - r_B]}}$$

A drawback of the *(expected) Sharpe Ratio* measure is the fact that it assumes gaussian returns and it loses explanatory power when the (expected) overperformance (i.e. the numerator) becomes negative, since then a higher (i.e. less negative) Sharpe ratio implies higher volatility given the same negative overperformance.

Having discussed the first order moment as a measure for the *return trend*, the second order moment as a statistical measure for the *volatility of asset/portfolio returns* and a combination of these two given by the *Sharpe Ratio*, we will also shortly take a closer look at the third and fourth central moments: *skewness* and *kurtosis*. They offer additional explanatory power when the underlying distribution of the returns is not gaussian.

Definition 3.2.7 (Skewness and Kurtosis)**.** Both *skewness* and *kurtosis* refer to distortion or asymmetry in relation to a gaussian distribution.
If the curve is shifted to the left or to the right, it is said to be *skewed*. The mean of positively skewed data will be greater than the median, hinting at rather small losses and big wins. In a distribution that is negatively skewed, the exact opposite is the case.
Whereas *skewness* differentiates extreme values in one versus the other tail, *kurtosis* measures extreme values in either tail. It is a measure that describes the shape of a distribution's tails in relation to its overall shape. For investors, high kurtosis (higher than 3, which is the *kurtosis* for normally distributed returns) of the return distribution hints at occasional extreme returns, both positive and negative, more extreme than predicted by the normal distribution of returns. Often, kurtosis is given in the form of *excess kurtosis*, where 3 is substracted to make it more easily comparable to a normal distribution.
However, one has to be very careful when using these two statistics to judge an investment. When the sample estimators are used, extreme outliers have a drastic effect, because their

difference to the mean weighs into the calculation with the power of 3, respectively, 4.

$$\widehat{Skew} = \frac{T}{(T-1)(T-2)} \sum_{t=1}^{T} \frac{(r_t - \bar{r})^3}{\hat{\sigma}^3}$$

$$\widehat{Kurt} = \frac{T(T+1)}{(T-1)(T-2)(T-3)} \sum_{t=1}^{T} \frac{(r_t - \bar{r})^4}{\hat{\sigma}^4} - \frac{3(T-1)^2}{(T-2)(T-3)}$$

When looking at possible investments, investors want to be aware of downside risk, i.e. the amount they would lose in a worst-case scenario. The main problem with volatility, however, is that it doesn't take direction of an investment's movement into account: a stock can be volatile because it suddenly jumps higher (which is a good thing). Therefore, an additional method is needed.

Definition 3.2.8 (Value-at-Risk)**.** One way of measuring and quantifying the downside risk is the *Value at risk (VaR)* [Jor07]. The *VaR* determines the potential loss in the investment under a given probability of occurrence for a given time frame. For example, when calculating the $95\% VaR$, one calculates the highest amount of invested money lost in the next year (here: month) where the probability of losing more is 5%.

$$VaR_\alpha : \quad \mathbb{P}(r \leqslant VaR_\alpha) = 1 - \alpha$$

The historical estimation method calculates the *VaR* as the α-quantile of the past returns. However, if the history of past returns is very small or you want to use a more sophisticated method, a possible way of computing this is to generate a set of return scenarios for the time-frame under assessment with distribution parameters estimated from past returns and further market information.

Definition 3.2.9 (Conditional Value-at-Risk)**.** The *Conditional Value-at-Risk (CVaR)*, also known as *Expected Shortfall*, is the expected portfolio return in the worst $\alpha\%$ cases [RU+00]. It is more sensitive to the shape of the tail of the loss distribution than the *VaR*.

$$CVaR_\alpha : \quad \mathbb{E}[r | r \leqslant VaR_\alpha]$$

Again, the historical estimation method calculates the *CVaR* as the mean of the returns which are lower than the α-quantile of the past returns. However, if the history of past returns is very small or you want to use a more sophisticated method, a possible way of computing this is to generate a set of return scenarios for the time-frame under assessment with distribution parameters estimated from past returns and further market information.

3.3 Markov Switching Model

Aiming for a more in-depth view about the relationship between ESG scores and return on a portfolio-level, we split our later analysis into turbulent and calm market phases. This section will introduce the concept we use for detecting times of crises.

The *Markov switching methodology*, first introduced in the seminal paper [Ham89], aims to analyze econometric time series with the particular interest to model unobservable states (or regimes) on which the distribution of an observable time series is dependent on. Prominent examples for these states (or regimes) are economic recessions and expansions, manic and depressive periods in bipolar disorders or, like in this thesis, calm and turbulent phases in the stock market. In more detail, our *Markov Switching Model* (MSM) with E regimes is established by an observable time series of monthly *S&P500 returns* $(r_t)_{t=1,\dots,T}$, an unobservable state process $(Z_t)_{t=1,\dots,T}$ with $Z_t \in \{0,\dots,E\}$ denoting the market phase and market-phase-dependent parameters $\mu(Z_t)$ and $\sigma(Z_t)$ denoting the *drift* and *volatility* of the return distribution, respectively. The problem at hand is to estimate the timepoints of regime changes and the values of the parameters associated with each regime. In the following, we refer to [Pri13], [Zag+10], [Ham89] and [Ham10] for the definitions and method of estimation of Markov switching models.

We begin by introducing the concepts of a *Markov Chain.*

Definition 3.3.1 (Time Homogeneous Markov Chain)**.** Let $(Z_t)_{t=0,\dots,T}$ be a discrete-time stochastic process taking values in a discrete state space $\mathbb{S}$, here $\mathbb{S} = \{0,\dots,E\}$. The $\mathbb{S}$-valued process $(Z_t)_{t=0,\dots,T}$ is said to be *Markov*, if for all t the probability distribution of Z_{t+1} is determined by the state Z_t of the process at time t, and does not depend on the past values Z_k for $k < t$.
In other words, for all $t \geqslant 1$ and all $i_0, i_1, \dots, i_{t-1}, i, j \in \mathbb{S}$ we have

$$\mathbb{P}[Z_{t+1} = j | Z_t = i, Z_{t-1} = i_{t-1}, \dots, Z_0 = i_0] = \mathbb{P}[Z_{t+1} = j | Z_t = i]$$

Using induction, one can show that the law of the process is given by

$$\begin{aligned}&\mathbb{P}[Z_t = i, Z_t = i_t, Z_{t-1} = i_{t-1}, \dots, Z_0 = i_0]\\&= \mathbb{P}[Z_t = i | Z_{t-1} = i_{t-1}] \cdots \mathbb{P}[Z_1 = j | Z_0 = i_0] \mathbb{P}[Z_0 = i_0]\end{aligned}$$

Furthermore, given by the law of total probability we get the following equation

$$\mathbb{P}[Z_1 = i] = \sum_{j \in \mathbb{S}} \mathbb{P}[Z_1 = i, Z_0 = j] = \sum_{j \in \mathbb{S}} \mathbb{P}[Z_1 = i | Z_0 = j] \mathbb{P}[Z_0 = j] \tag{3.24}$$

for all $i \in \mathbb{S}$.
A *Markov chain* is called *time homogeneous* when the probability $P_{i,j} := \mathbb{P}[Z_{t+1} = j | Z_t = i]$ is independent of t.

As seen above, the random evolution of a *time homogeneous Markov chain* is determined by the probabilities $P_{i,j}$.

Definition 3.3.2 (Transition Matrix)**.** The transition matrix of a *time homogeneous Markov chain* $(Z_t)_{t=0,\dots,T}$ is given by $(P_{i,j})_{i,j\in\mathbb{S}} = \mathbb{P}[Z_1 = j|Z_0 = i]_{i,j\in\mathbb{S}}$, also written as

$$(P_{i,j})_{i,j\in\mathbb{S}} = \begin{pmatrix} P_{0,0} & \cdots & P_{0,E} \\ \vdots & \ddots & \vdots \\ P_{E,0} & \cdots & P_{E,E} \end{pmatrix}$$

where the initial state i is given by the row number, while the final state j corresponds to the column number.

In the following, we assume that the return process is normally distributed in each state with

$$r_t = \mu(Z_t) + \sigma(Z_t)\epsilon_t \quad \forall t \in \{0, \dots, T\}$$

where $\epsilon_t \sim \mathcal{N}(0,1)$ and $(\epsilon_t)_{t=0,\dots,T}$ i.i.d. Furthermore, we only consider a *time homogeneous Markov chain* with 2 regimes, i.e. $\mathbb{S} = \{0,1\}$, and transition matrix

$$(P_{i,j})_{i,j\in\mathbb{S}} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

with $p = \mathbb{P}[Z_{t+1} = 1|Z_t = 0]$ and $q = \mathbb{P}[Z_{t+1} = 0|Z_t = 1]$. Let $(\delta, 1-\delta)$ with $\delta := \mathbb{P}[Z_0 = 0] \in [0,1]$ denote the initial distribution of the *Markov Chain* $(Z_t)_{0=1,\dots,T}$, then the MSM is completely determined by the *parameter vector*

$$\theta = (p, q, \mu(0), \mu(1), \sigma(0), \sigma(1), \delta)$$

The conditional densities of $r = (r_t)_{t=0,\dots,T}$ which differ between states are then given by

$$f_{i,t} = f(r_t|z_t = i) \sim \mathcal{N}(\mu(Z_t = i), \sigma(Z_t = i)^2) \tag{3.25}$$

Since the regime Z_t is unobservable, the true state at any point in time is never known. Hence, past and current observations are used to make an inference on the probability of a regime.

Definition 3.3.3 (Hamilton Filter)**.** The *Hamilton filter* gives the conditional distribution of the state Z_t on the data up to time t, i.e.

$$p_t^i = \mathbb{P}[Z_t = i|r_t, r_{t-1}, \dots, r_1; \hat{\theta}_t]$$

for $i \in \mathbb{S}$, where $\hat{\theta}_t$ denotes the estimate of the *parameter vector* derived from observations $\boldsymbol{r_t}= (r_t, r_{t-1}, \dots, r_0)$.

We use it to compute the probability density of r_t for the next period in the following formula, where we leverage the Markov porperty of $(Z_t)_{t=0,\dots,T}$, the law of total probability and the definition of conditional probability:

$$\begin{aligned}
\mathbb{P}[r_{t+1}|r_t, r_{t-1}, \dots, r_0] &= \sum_{i\in\mathbb{S}} \mathbb{P}[r_{t+1}, Z_{t+1} = i|\boldsymbol{r_t}] \\
&= \sum_{i\in\mathbb{S}} \mathbb{P}[r_{t+1}|Z_{t+1} = i, \boldsymbol{r_t}]\mathbb{P}[Z_{t+1} = i|\boldsymbol{r_t}] \\
&= \sum_{i\in\mathbb{S}} (\mathbb{P}[r_{t+1}|Z_{t+1} = i, \boldsymbol{r_t}] \sum_{j\in\mathbb{S}} \mathbb{P}[Z_{t+1} = i, Z_t = j|\boldsymbol{r_t}]) \\
&= \sum_{i\in\mathbb{S}} (\mathbb{P}[r_{t+1}|Z_{t+1} = i, \boldsymbol{r_t}] \sum_{j\in\mathbb{S}} \mathbb{P}[Z_{t+1} = i|Z_t = j, \boldsymbol{r_t}]\mathbb{P}[Z_t = j|\boldsymbol{r_t}]) \\
&= \sum_{i\in\mathbb{S}} (\mathbb{P}[r_{t+1}|Z_{t+1} = i, \boldsymbol{r_t}] \sum_{j\in\mathbb{S}} \mathbb{P}[Z_{t+1} = i|Z_t = j]\mathbb{P}[Z_t = j|\boldsymbol{r_t}]) \\
&= \sum_{i\in\mathbb{S}} \sum_{j\in\mathbb{S}} \mathbb{P}[r_{t+1}|Z_{t+1} = i, \boldsymbol{r_t}]\mathbb{P}[Z_{t+1} = i|Z_t = j]\mathbb{P}[Z_t = j|\boldsymbol{r_t}]
\end{aligned}$$

where $\mathbb{P}[Z_{t+1} = i|Z_t = j]$ is the transition probability $P_{j,i}$ and $\mathbb{P}[r_{t+1}|Z_{t+1} = i, \boldsymbol{r_t}]$ is given through Equation 3.25. The filtered probability of Z_t is easy to compute when the filtered probability of Z_{t-1} is known:

$$\begin{aligned}
\mathbb{P}[Z_t = j|\boldsymbol{r_t}] &= \frac{\mathbb{P}[Z_t = j, r_t|\boldsymbol{r_{t-1}}]}{\mathbb{P}[r_t|\boldsymbol{r_{t-1}}]} \\
&= \frac{\mathbb{P}[r_t|Z_t = j, \boldsymbol{r_{t-1}}]\mathbb{P}[Z_t = j|\boldsymbol{r_{t-1}}]}{\mathbb{P}[r_t|\boldsymbol{r_{t-1}}]} \\
&= \frac{\mathbb{P}[r_t|Z_t = j, \boldsymbol{r_{t-1}}] \sum_{i\in\mathbb{S}} \mathbb{P}[Z_t = j|Z_{t-1} = i, \boldsymbol{r_{t-1}}]\mathbb{P}[Z_{t-1} = i|\boldsymbol{r_{t-1}}]}{\mathbb{P}[r_t|\boldsymbol{r_{t-1}}]}
\end{aligned}$$

Hence, given the initial distribution of the first-period state, the filtered probability of Z_t and the associated one-period ahead probability density can be calculated iteratively.

With this result, we can now construct the *likelihood function* and estimation algorithm, which conclude the chapter about the fundamentals of MSMs.

Definition 3.3.4 (Estimation)**.** The *likelihood function* is constructed as the product of the conditional probability densities:

$$\mathbb{P}[r|\theta] = \prod_{t=0}^{T} \mathbb{P}[r_t|\boldsymbol{r_{t-1}}, \theta]$$

With this, the maximum likelihood estimation is based on an iterative expectation-maximization (EM) algorithm (see [Ham89]).

3.4 Natural Language Processing

In the final part of this master thesis we introduce a methodology for updating ESG scores on a daily basis using a state-of-the-art Natural Language Processing model to evaluate stakeholder discourse as portrayed on Twitter. In order to provide a deeper understanding of the proposed methodology, this section will give a short overview of the history, most important architectures and terminology needed to understand the model.

In machine learning, NLP represents the set of theories and technologies that concern the intersection of computers and human language, i.e. how computers can be used to understand and manipulate natural langauge [Cho03]. [Lid01] offers the definition:

> *Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.*

The idea of viewing language as a ”system” already emerged in the early 1900s [De 89] and made its way into the world of computer science in the late 1950s [Cho20]. Even though the concept was picked up so early, no true advancements were made until the 1980s, when the steady increase in computational power and the shift towards using statistical methods enabled researchers to replace complex, rule-based approaches with machine learning algorithms.
The development of *Word2Vec* in 2013 [Mik+13a] is a milestone in modern NLP. It leverages the idea that the meaning of a word can be inferred by the company it keeps, by training words against other words that neighbor them in the input corpus to generate a vector representation of each word in a two-layer neural net. This way, instead of having a one-to-one mapping like a one-hot vector, the representation of a word is spread across all the elements in the vector and each element in the vector contributes to the definition of many words. An example of the power of this model is given by the fact that, when trained on a huge *Google News* corpus, the result of the vector calculation *vec(“Madrid”) - vec(“Spain”) + vec(“France”)* is closer to *vec(“Paris”)* than to any other word vector [Mik+13b]. These word embeddings gave rise to the usage of various algorithms for NLP related tasks, including but not limited to question answering, sentiment analysis and machine translation [You+18].

However, a major drawback of this approach to translate words into machine-readable vectors is that the vector representation of a word is fixed and only extracted by the context given in the input corpus. This means NLP related tasks will struggle with ambiguous words. Take, for example, the word *play*: if the training corpus only consists of sentences where it is used in a sports context, the representation completely fails to express other meanings like referring to a dramatic work intended for performance by actors on a stage. In the following years, various different machine learning architectures were exploited to attempt modelling the context into vector representations. The majority are based on *neural networks*, which are a system of layers with interconnected *nodes/neurons* inspired by biological neural networks like the brain. Here, instead of calculating a linear transformation of the input like in the *multiple linear regression*, each *neuron* calculates the weighted sum of its inputs, which is then passed through a typically non-linear so-called *activation function* as new input to all connected *neurons* in other layers until the calculation reaches the output layer.

Definition 3.4.1 (Recurrent Neural Network)**.** A *recurrent neural network* (RNN) is a neural network that is specialized for processing a sequence of values $\boldsymbol{x} = \boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n_x)}$ by computing *hidden state vectors* $\boldsymbol{h} = \boldsymbol{h}^{(0)}, \dots, \boldsymbol{h}^{(n_x)}$ [Goo+16]. In contrast to *vanilla neural networks*, they do not only take input data from the present time stamp into account, but rather also take a *hidden state vector* representing the context based on prior inputs into account.

$$\boldsymbol{h}^{(t)} = \boldsymbol{f}(\boldsymbol{h}^{(t-1)}, \boldsymbol{x}^{(t)}) \tag{3.26}$$

Hence, the same input could produce a different output depending on previous inputs of the sequence.

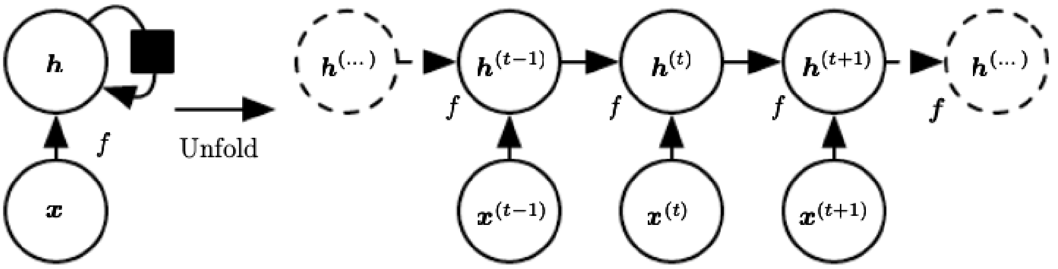

Figure 3.1: This RNN processes information from the input $\boldsymbol{x}$ by incorporating it into the state $\boldsymbol{h}$ that is passed forward through time [Goo+16]. A loop allows information to be passed from one step of the network to the next.

Definition 3.4.2 (Long Short-Term Memory)**.** A *Long Short Term Memory Network* (LSTM) is a modification of an RNN that allows for a longer retention time. They are based on the idea of creating paths through time that have derivatives that neither vanish nor explode, by using connection weights that may change at each time step.

Instead of a unit that simply applies an element-wise nonlinearity to the affine transformation of inputs and recurrent units (see Equation 3.26), LSTM recurrent networks have

LSTM cells that have an internal recurrence (a self-loop), in addition to the outer recurrence of the RNN. Each cell has the same inputs and outputs as an ordinary recurrent network, but has more parameters and a system of gating units that controls the flow of information [Goo+16].

Definition 3.4.3 (Encoder-Decoder or Sequence-to-Sequence Architecture)**.** Let the context C be a vector or sequence of vectors that summarize the input sequence $\boldsymbol{x} = \boldsymbol{x^{(1)}, \ldots, x^{(n_x)}}$. The idea of an encoder-decoder architecture is as follows:

1. An *encoder* (e.g. RNNs, LSTMs) processes the input sequence.
2. The *encoder* emits the context C, usually as a simple function of its final hidden state.
3. A *decoder* (e.g. RNNs, LSTMs) is conditioned on that fixed-length context vector to generate the output sequence $\boldsymbol{y} = \boldsymbol{y^{(1)}, \ldots, y^{(n_y)}}$.

At each step the model is auto-regressive, i.e. it consumes the previously generated *hidden state vector* as additional input when generating the next [Gra13].

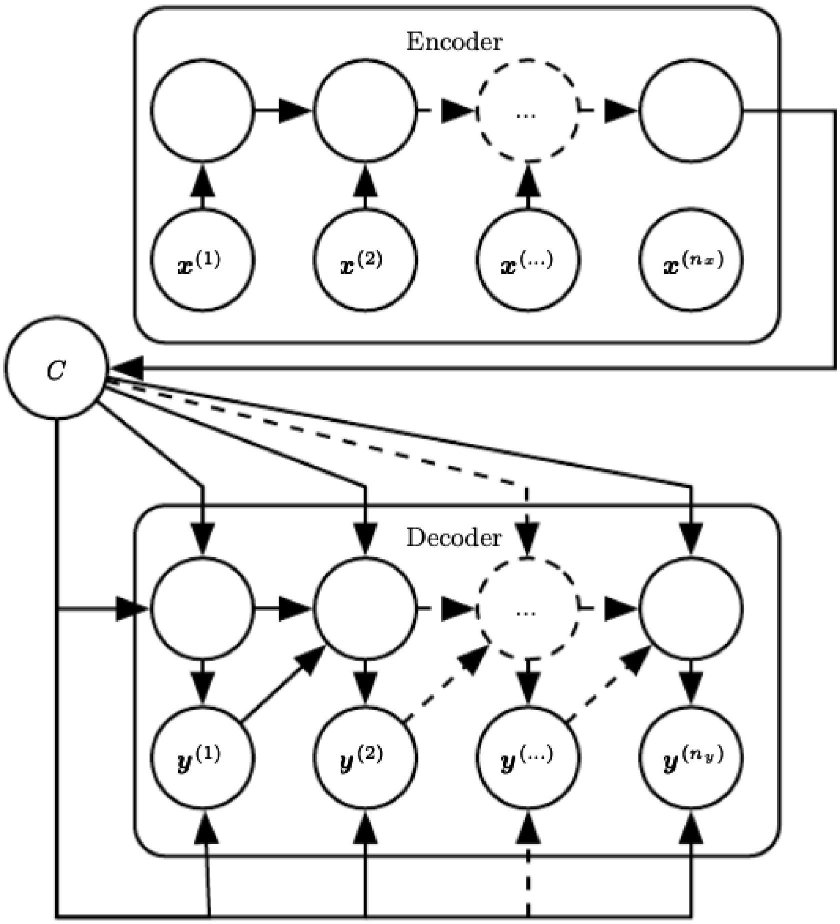

Figure 3.2: Example of an *encoder-decoder* or *sequence-to-sequence RNN* architecture, for learning to generate an output sequence $\boldsymbol{y} = \boldsymbol{y^{(1)}, \dots, y^{(n_y)}}$ given an input sequence $\boldsymbol{x} = \boldsymbol{x^{(1)}, \dots, x^{(n_x)}}$ [Goo+16].

Using multiple instances of RNNs, LSTMs or even *Gated Recurrent Units* (GRUs) in a sequence based model with an *encoder-decoder* architecture allows information to persist within the network by successfully learning on data with long range temporal dependencies [SVL14] and capturing the context in the so-called *hidden state vector*. Even though these architectures performed very well in, for example, translation tasks for sentences, there was still the problem of retaining meaningful long term information consistently.

A solution called *Attention* was introduced by [LPM15] and [BCB14] which allows the model to focus on, or ”pay attention” to, different parts of the input sequence at every stage of the output sequence, allowing the context to be preserved from beginning to end. Simply put, since the issue with previous algorithms was that passing a single *hidden state vector* to the *decoder* was not enough, now as many *hidden state vectors* as the number of instances in the input sequence are generated and combined differently for each instance in the output sequence. However, the attention mechanism was still used in combination with RNNs, LSTMs and GRUs.

Definition 3.4.4 (Attention)**.** In order to predict the next position $\boldsymbol{y^{(t)}}$ in the output sequence $\boldsymbol{y} = \boldsymbol{y^{(1)}, \ldots, y^{(n_y)}}$, the *attention mechanism* searches for a set of positions in the input sequence $\boldsymbol{x} = \boldsymbol{x^{(1)}, \ldots, x^{(n_x)}}$ where the most relevant information for $\boldsymbol{y^{(t)}}$ is concentrated. Instead of simply receiving the fixed *context vector* $\boldsymbol{c}$ as additional input, it receives a *position-dependent context vector* $\boldsymbol{c^{(t)}}$ which is an *attention*-weighted average of the *encoder hidden state vectors* $\boldsymbol{h} = \boldsymbol{h^{(1)}, \ldots, h^{(n_x)}}$.
Here, one way of determining the *attention weights* is to simply calculate the dot product of each *encoder hidden state vector* with the current *decoder hidden state vector* and scale them with a *softmax* function to compute the relative importance - if an *encoder hidden state vector* is similar to the current *decoder hidden state vector*, the respective *encoder hidden state vector* will get a high weight. Hence, this position in the input sequence will have a big influence on the prediction of $\boldsymbol{y^{(t)}}$.

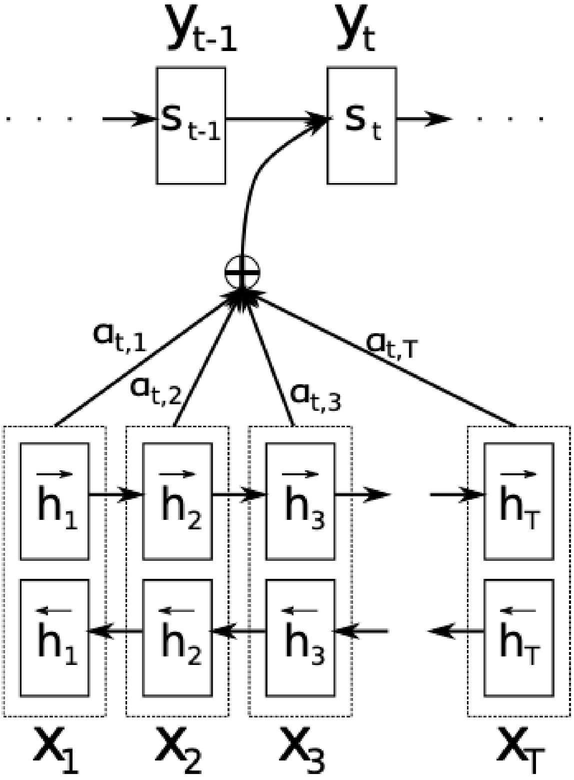

Figure 3.3: Example of using *attention* in an *encoder-decoder* RNN architecture, where the *encoder* consists of a bidirectional RNN and the *decoder* is a simple RNN. Pictured above is the step to generate the tth output $\boldsymbol{y_t}$ given an input sequence $\boldsymbol{x} = \boldsymbol{x_1}, \ldots, \boldsymbol{x_T}$, *encoder hidden state vectors* $\boldsymbol{h} = \boldsymbol{h_1}, \ldots, \boldsymbol{h_T}$, *attention weights* $\boldsymbol{a_{t,1}}, \ldots, \boldsymbol{a_{t,T}}$ to generate a *position-dependent context vector* $\boldsymbol{c_t}$ as the weighted sum of the *encoder hidden state vectors* and a *decoder hidden state vector* $\boldsymbol{s_{t-1}}$ [BCB14]. Instead of using the same *context vector* $\boldsymbol{c_t}$ for every decoder step as seen in Figure 3.2, the *attention mechanism* generates *position-dependent context vectors* $\boldsymbol{c_t}$ for each $t \in \{1, \ldots, n_y\}$.

It wasn't until 2017, when the seminal paper "Attention is all you need" [Vas+17] revolutionized research in NLP by completely dispensing with recurrence and, instead, entirely relying on the *Attention* mechanism in the so called *Transformer* architecture.

Definition 3.4.5 (Transformer)**.** Similar to *sequence-to-sequence* models, the *Transformer* architecture encompasses two main building blocks: an *encoder unit* and a *decoder unit*. Here, they are each stacks of identically built blocks, which share the same inherent structure, but do not share parameters.
In contrast, however, the commonly used *recurrent layers* in *sequence-to-sequence encoders* or *decoders* are substituted by *self-attention layers*, which evaluate the similarity between the positions in the input sequence. Instead of requiring $O(n)$ sequential operations, a

self-attention layer connects all positions with a constant number of sequentially executed operations, which reduces the total computational complexity per layer [Vas+17]. For each input, the *attention* mechanism in each *encoder* and *decoder layer* weighs the relevance of every other input and draws information from them in order to generate the output, with the *decoder layer* also having an additional *attention* mechanism to draw information from the outputs of previous decoders.

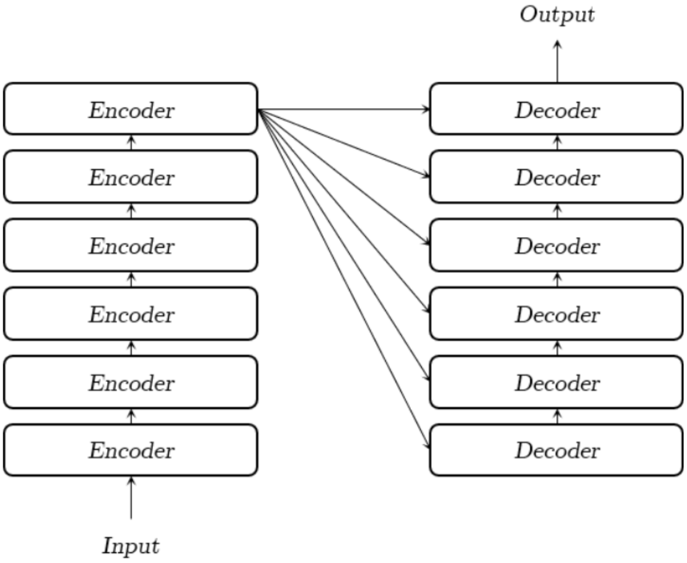

Figure 3.4: A high-level presentation of the transformer model architecture [Ges21].

Having introduced the main underlying principles and ideas, we will now give a short introduction into the state-of-the-art NLP model, a variant (*BERTweet*) of which we use in this thesis: *BERT*.

Definition 3.4.6 (Bidirectional Encoder Representations from Transformers (BERT))**.** *BERT* is an NLP model comprising stacked layers of *Transformer encoders*, which was proposed by researchers at *Google AI Language* [Dev+19]. Its key technical innovation is the application of bidirectional training to the *Transformer* architecture. For training, instead of only looking at a sequence from left-to-right in a typical *next sequence prediction* (NSP), *BERT* additionally uses a *masked language model* (MLM) to overcome the challenge of limited context learning. The MLM randomly "masks" some positions from the input, and the model is trained with the objective to predict the original values of the masked input positions based only on their context.
Being pre-trained on all of English Wikipedia and BooksCorpus in the two tasks of MLM and NSP allows *BERT* to achieve state-of-the-art accuracy in various NLP tasks, such as question answering and language inference.

Numerous models building on the innovation of *BERT* have been proposed since its publication, aiming to either improve it with respect to size or speed, catering to other languages or specializing on specific training data. One of these, namely *BERTweet* [NVN20], is the model used in this thesis. It has the same architecture as *BERT*, but leverages 80GB of uncompressed texts containing 850M English Tweets for pre-training. Since Tweets tend to be rather different from the data used to pre-train the classical *BERT*, with their typically short length of no more than 140 characters, frequent use of informal language, typographical errors and hashtags [HCB13][Eis13], this model is the ideal choice for our use case.

To conclude this section, we will give a short introduction into some important hyperparameters and methods for model training and present three measures for model performance. First, we will introduce algorithms used for model parameter training.

Definition 3.4.7 (Stochastic/Mini-Batch Gradient Descent)**.** The goal of training an NLP model is finding model parameters, i.e. weights $\boldsymbol{w}$ for the network, that, together, minimize the *loss function* $\boldsymbol{J}(\boldsymbol{y}, \hat{\boldsymbol{y}})$ (e.g. *cross-entropy loss*). This is usually done by using a variant of the *stochastic gradient descent.*
In each step, a prediction $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ is made on a training sample $\boldsymbol{x}_{\boldsymbol{i}}$ using the model with the current set of parameters $\boldsymbol{w}$ and compared to the real outcome $\boldsymbol{y}_{\boldsymbol{i}}$. Then, the error (loss) is calculated and used to update the model parameters.
In detail, an initial weight vector $\boldsymbol{w}$ is (randomly) chosen and updated with every training sample i

$$\boldsymbol{w} := \boldsymbol{w} - \eta * \boldsymbol{\nabla J}(\boldsymbol{y}_{\boldsymbol{i}}, \hat{\boldsymbol{y}}_{\boldsymbol{i}})$$

where η is the learning rate and $\boldsymbol{J}(\boldsymbol{y}_{\boldsymbol{i}}, \hat{\boldsymbol{y}}_{\boldsymbol{i}})$ is the loss function evaluated for the ith training sample [Rud17].
When training deep learning models, optimizers often use so called *mini-batch gradient descent.* This means, instead of updating the model in every single training step, a batch of samples is propagated through the model and then backward propagated to calculate gradients for every sample. The respective gradients are averaged (or summed up) and will then be used as an input to the updating rule of the chosen optimizer.
This has been shown to perform significantly better than "true" stochastic gradient descent described, because the code can make use of vectorization libraries rather than computing each step separately. It may also result in smoother convergence, as the gradient computed at each step is averaged over more training examples.

Definition 3.4.8 (Adaptive Moment Estimation (Adam))**.** *Adam* is a learning algorithm based on *Stochastic Gradient Descent* used to update the model parameters $\boldsymbol{w}$ which adapts the learning rate to the parameters [KB14]. In addition to using the gradient, it also stores an exponentially decaying average of *past gradients* $\boldsymbol{m_i}$ and the decaying average of *past squared gradients* $\boldsymbol{v_i}$ as estimates of the first and second order moment of the gradient, respectively:

$$\begin{aligned}\boldsymbol{m_i} &= \beta_1 \boldsymbol{m_{i-1}} + (1-\beta_1)\boldsymbol{\nabla J}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i}) \\ \boldsymbol{v_i} &= \beta_2 \boldsymbol{v_{i-1}} + (1-\beta_2)\boldsymbol{\nabla J}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i})^{\mathbf{2}}\end{aligned}$$

It then computes bias-corrected estimates

$$\begin{aligned}\boldsymbol{\hat{m}_i} &= \frac{\boldsymbol{m_i}}{1-\beta_1^i} \\ \boldsymbol{\hat{v}_i} &= \frac{\boldsymbol{v_i}}{1-\beta_2^i}\end{aligned}$$

to finally update the model parameters $\boldsymbol{w}$ with the rule:

$$\boldsymbol{w} := \boldsymbol{w} - \eta \frac{\boldsymbol{\hat{m}_i}}{\sqrt{\boldsymbol{\hat{v}_i}} + \epsilon} * \boldsymbol{\nabla J}(\boldsymbol{y_i}, \boldsymbol{\hat{y}_i})$$

The authors propose default values of 0.9 for β_1, 0.999 for β_2, and 10^{-8} for ϵ.

There are numerous other variations of *SGD* which can be used for model training, but we refrain from introducing them, since the model used in this thesis is pre-trained with the *Adam* optimizer.

Next, we give an overview of some of the most important hyperparameters that can be set for a learning algorithm.

Definition 3.4.9 (Batch Size)**.** The *batch size* is a hyperparameter of the learning algorithm (e.g. *Adam*). It controls the number of training samples to work through before a parameter update is performed.

The *batch size* hyperparameter is limited by memory constraints on the training device, since the more samples are being propagated through the model the larger the intermediate calculations are. However, the smaller the batch, the less information is contained and, hence, the gradient estimates resulting from each batch will be less accurate [Goo+16]. This can sometimes make it difficult for the model to converge. One way of tackling this drawback is *Gradient Accumulation.*

Definition 3.4.10 (Gradient Accumulation)**.** *Gradient accumulation* makes use of the fact that the *loss* in *mini-batch optimization* is usually calculated by taking the average loss of each sample in the batch. This means, they do not necessarily need to be calculated in parallel, but can actually be calculated sequentially and then averaged afterwards.
In detail, we run n smaller batches through the model without updating the parameters,

accumulate their total losses and average over all n batches before backward-propagating to calculate gradients for the bigger batch consisting of the n smaller batches and, finally, updating the model parameters. The amount n of batches accumulated is called *gradient accumulation steps.*

Definition 3.4.11 (Epochs)**.** *Epochs* is a hyperparameter in model training and sets the number of times the learning algorithm processes the entire training data set. When increasing the number of *epochs*, there is a trade-off between better finding the model parameters that lead to the minimal *loss* and overfitting on the training examples, meaning the model will most likely suffer from poor generalizability.

Definition 3.4.12 (Maximum Sequence Length)**.** Generally speaking, NLP models break down language into shorter, more basic pieces, called tokens, in order to have a mathematical representation before processing the input. The *maximum sequence length* is the maximum number of tokens which can be handled as an input sequence for the model.
This parameter is set in advance for pre-training and, theoretically, has no limit. However, longer sequences require more memory, and memory usage scales quadratically with sequence length. Hence, most publically available pre-trained NLP models have a *maximum sequence length* of 128 or 512.

In order to find the "best" combination of hyperparameters for model training, we first need to define how to measure a model's performance. In the following, we introduce three commonly used metrics in classification tasks [SL09]. As a basis, let us define class-specific *True Positives*, *False Positives*, *True Negatives* and *False Negatives* in the following way:

Predicted/True	i	**not** i
i	True Positive (TP_i)	False Positive (FP_i)
not i	False Negative (FN_i)	True Negative (TN_i)

Table 3.1: Confusion Matrix for Classification

Before creating a confusion matrix (Table 3.1), one has to choose a cut-off value on which to perform the classification.

Definition 3.4.13 (Accuracy)**.** *Accuracy* is a well known metric for evaluating the performance of classification models. It is calculated as the fraction of samples the model classified correctly:

$$Accuracy = \frac{\#\text{correct predictions}}{\#\text{all predictions}}$$

In multi-class classification with $n \geqslant 3$ class labels one can also compute a *(weighted) average accuracy* based on the class-specific accuracies $Accuracy_i$ for $i \in \{1, \dots, n\}$:

$$Accuracy_i = \frac{TP_i + TN_i}{TP_i + FP_i + FN_i + TN_i}$$

[SL09]

Even though this metric is widely used and easily understood, it poorly characterizes the performance of a model with highly unbalanced classes. Take, for example, a binary classifier on a dataset with only 0.1% *Positives* - if we always predict *Negative*, we automatically have an *accuracy* of 99.9%. In order to tackle this drawback, let us introduce the next metrics [HM13].

Definition 3.4.14 (Precision-Recall Curve)**.** The *precision* of a classifier for each class i is the number of correctly classified class examples divided by the number of examples labeled by the model as belonging to class i:

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

The *recall/true positive rate* of a classifier for each class i is the number of correctly classified class examples divided by the number of class examples in the data set:

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

A *precision-recall curve* shows the relationship between *precision* and *recall* for every possible cut-off, i.e. it summarizes the trade-off between the *true positive rate* and the *positive predictive value* for a model using different probability thresholds [SL09].
A helpful tool for comparing these curves is by calculating the *area under the curve* (AUC) to summarize the curve with a range of threshold values as a single score. The score is a value between 0 and 1 for a perfect classifier.

Definition 3.4.15 (Receiver Operating Characteristic (ROC) Curve)**.** The *false positive rate (FPR)* of a classifier for each class i is the number of falsely classified class examples divided by the number of examples labeled by the model as not belonging to class i:

$$FPR_i = \frac{FP_i}{FP_i + TN_i}$$

A *ROC curve* shows the relationship between the *recall/true positive rate* and the *false positive rate* for every possible cut-off [Bra97].
Again, a helpful tool for comparing these curves is given by calculating the *area under the curve* (AUC) to give a single score for a classification model across all threshold values. The score is a value between 0 and 1 for a perfect classifier.

Having introduced measures to judge the performance of a model, there are a few common tuning methods to find the "best" combination: manual, grid search, random search, bayesian hyperparameter optimization and many more. The goal here is to find the hyperparameters of a given algorithm that return the best performing model as measured on a validation set. The biggest problem concerning hyperparameter optimization is that with a large number of hyperparameters and complex models, testing every combination quickly becomes intractable: each time we try different hyperparameters, we have to train the model on the training data, make predictions on the validation data, and calculate the performance metric. In this thesis, we use *bayesian hyperparamter optimization*, because it has shown promising results with respect to speed, computational cost and performance [Ber+11][DMC16].

Definition 3.4.16 (Bayesian Hyperparameter Optimization)**.** The search for good hyperparameters is cast as an optimization problem, where the decision variables x are the hyperparameters of the model that can take any value in the set $\mathcal{X}$:

$$x^* = \arg\min_{x \in \mathcal{X}} f(x)$$

Commonly, the function f to be optimized is the error on the validation set.
In general, *bayesian hyperparameter optimization* keeps track of past evaluation results and, based on these, uses a Bayesian approach to estimate a probability model of the objective function:

$$\hat{\mathbb{P}}[f|x]$$

This is updated in each step and leveraged to propose the next, most promising set of hyperparameters for training [Goo+16][DMC16].

To sum up this section, we have seen algorithms and models used in *Natural Language Processing*, how to train and how to evaluate them. Building on this knowledge, we will present a complete framework for updating ESG Scores on a daily basis using Tweets as a proxy for stakeholder discourse in Chapter 5.

Chapter 4

Inference from Annual Scores

As mentioned in Section 2.3, there is extensive research on the relationship between a company's social and financial performance based on ESG Scores from different rating agencies, in different time periods and in different markets. From more than 2000 studies on the relationship between ESG and corporate financial performance from the 1970s until 2015, roughly 90% find it to be non-negative and, more importantly, a majority reports it to be positive [FBB15]. More concrete, [KSY16] show that there is evidence of a positive relationship between performance on *material* ESG issues and financial performance. In contrast, in a more recent paper [PSK19] argue that throughout the research one cannot find definite confirmation that considering ESG in investment decisions delivers alpha.

This chapter is dedicated to the investigation of the relationship between multiple ESG-related scores provided by *Refinitiv* and the *S&P500* companies' financial performance in the form of a valuation multiple on the company-level and returns on a portfolio-level. We will derive insights into which scores have the biggest, significant explanatory power with respect to corporate valuation principles and whether, how and why score-based portfolios generate outperformance.

4.1 Market Valuation

This section focuses on exploring the connection between ESG scores and financial performance through a *regression model*. In detail, we will regress the natural logarithm of a company's price-to-book ratio on two different sustainability scores while controlling for other market, time and group effects. For this analysis, we will first build two separate models for inference detection and test their robustness by adding other, possibly explanatory variables. Second, we will combine the models in order to also consider the scores' influence on each other and, again, perform a test of robustness. Lastly, we take a more granular look at the relationship by estimating *time-varying coefficients.* A more in-depth look at the data preparation, exploratory data analysis (EDA) and analysis of the linear regression assumptions can be found in Appendix B.

4.1.1 Approach

Inspired by the methodology of [Ser20], we attempt to model the relationship between the (natural logarithm of the) company's price-to-book ratio and its ESG and Controversies Score, respectively, which are presented in detail below, by estimating a *cross-sectional* and *longitudinal regression* model. In order to control for financial market effects, we include the company's return on equity ROE, leverage Lev, size (= ln(market cap)) $Size$, net sales growth $\Delta NetSales$, annual return R as well as industry and yearly fixed effects f_i, f_t. All of this data is collected at year-end. Since adding both *Refinitiv*'s ESG Score and Controversies Score as explanatory variables results in a *variance inflation factor* of around 10 for both of them (compare Appendix B), we decide to first look at their inference in two separate models. We also do a z-score transformation (for each year) for $Size$, i.e.

$$z(Size)|_{year=t} = \frac{Size - \mu(Size)|_{year=t}}{\sigma(Size)|_{year=t}} \tag{4.1}$$

in order to control for multicollinearity, since simply adding it without regularization leads to a *variance inflation factor* of around 15. Hence, we first analyze the following two separate models:

$$\begin{aligned} \ln(PTB)_{i,t} =& b_0 + b_1 ESGScore_{i,t} + b_2 z(Size)_{i,t} + b_3 ROE_{i,t} \\ &+ b_4 \Delta NetSales_{i,t} + b_5 R_{i,t} + b_6 Lev_{i,t} + f_i + f_t + \epsilon_{i,t} \end{aligned} \tag{4.2}$$

$$\begin{aligned} \ln(PTB)_{i,t} =& b_0 + b_1 ControversiesScore_{i,t} + b_2 z(Size)_{i,t} + b_3 ROE_{i,t} \\ &+ b_4 \Delta NetSales_{i,t} + b_5 R_{i,t} + b_6 Lev_{i,t} + f_i + f_t + \epsilon_{i,t} \end{aligned} \tag{4.3}$$

As a second step, we add both scores as input variables in the same model in order to investigate their influence on each other in the context of market valuation models.

$$\begin{aligned} \ln(PTB)_{i,t} =& b_0 + b_1 ESGScore_{i,t} + b_2 ControversiesScore_{i,t} + b_3 z(Size)_{i,t} \\ &+ b_4 ROE_{i,t} + b_5 \Delta NetSales_{i,t} + b_6 R_{i,t} + b_7 Lev_{i,t} + f_i + f_t + \epsilon_{i,t} \end{aligned} \tag{4.4}$$

Furthermore, we estimate the above models for every year in order to investigate whether and how the influence of ESG has changed over time. For that we simply leave out the yearly fixed effects and fix the time t. For tests of robustness, we also add further explanatory variables, namely Capital Expenditures, Price-Earnings Ratio and Dividend Yield.

4.1.2 Data

As a basis for our analysis we look at data of the current (end-of-year 2020) *S&P500* stocks from 2004 until 2020, which we collect from Thomson Reuters (now Refinitiv) Datastream. Here, we access the following sustainability scores for this anaylsis:

- **ESG Score:** *Refinitiv*'s (formerly Asset4) ESG Score provides a measure of a company's relative ESG performance based on verifiable publicly-reported data. It is a combined measure of over 450 company-level measures grouped into three pillars: *Environmental* (resource use, emissions, innovation), *Social* (workforce, human rights, community, product responsibility) and *Governance* (management, shareholders, CSR strategy). The pillar weights are normalized to percentages ranging between 0 (poor relative ESG performance and insufficient degree of transparency in reporting material ESG data publicly) and 100 (excellent relative ESG performance and high degree of transparency in reporting material ESG data publicly). [Ref21]
- **Controversies Score:** *Refinitiv*'s (formerly Asset4) Controversies Score is calculated based on 23 ESG controversy topics and reportings from global media, NGOs, etc.. All recent controversies are counted in the latest closed fiscal year and are benchmarked on industry group. The score calculation addresses the market cap bias, i.e. the fact that larger companies tend to attract more media attention than smaller companies. The score ranges from 0 to 100 (no controversies). [Ref21]

Our data consists of around 300 companies each year, since we exclude those firms which have missing and extreme values for any of the variables of interest (see Appendix B). The exclusion of extreme values is done by computing the z-score grouped by year (as to not introduce a forward-looking bias) and limiting our sample to those observations, that lie within 3.5 standard deviations of the mean observation in each year, i.e. filtering for observations that have a z-score between -3.5 and 3.5. Our sample comprises more than 5000 complete observations.

Table 4.1: Summary Stastics for Market Valuation Analysis Sample

	ESG Score	Controversies Score	Z-norm: Size	ROE	Net Sales Change Pct	Annual Return	Leverage	CapEx	Price-Earnings Ratio	Dividend Yield	Ln(PTB)
count	5958	5953	7536	6811	6625	7101	7000	7579	6919	7459	7276
mean	49.33	82.71	0.00	17.70	0.16	0.13	86.65	899097.73	28.56	1.72	1.07
std	19.91	28.95	0.92	21.13	4.33	0.29	126.73	1882274.00	36.76	1.56	0.71
min	0.63	0.91	-3.04	-271.23	-78.62	-0.84	-996.62	291.00	2.30	0.00	-3.91
25%	33.59	76.47	-0.66	9.33	-0.13	-0.03	28.00	97275.50	15.60	0.18	0.58
50%	49.14	100.00	-0.09	15.23	0.09	0.11	60.60	282560.00	20.90	1.49	1.01
75%	65.58	100.00	0.58	23.16	0.29	0.29	117.59	924429.00	29.75	2.69	1.47
max	95.07	100.00	3.44	383.56	147.79	2.64	992.73	29504000.00	1353.80	9.73	5.12

Please find a detailed description of the variables in the Appendix in Table A.1 and an in-depth explanation of the data preparation as well as a complete EDA in Sections B.1 - B.2 in Appendix B.

To incorporate the industry fixed effects we decide to group the *S&P 500* companies with respect to their *Industry Classification Benchmark* (ICB) which is a widely used standard for the categorization and comparison of companies by industry and sector.

4.1.3 Results

As a first step, we will look at the results generated by the estimation of the panel data regressions in equations 4.2 and 4.3. Table 4.2 and Table 4.3 show the results of the models, respectively.
We can observe for the time period 2004-2020, that the coefficient for the ESG Score has a negative loading of -0.0040, whereas the coefficient for the Controversies Score has a positive loading of 0.0022. Both are highly significant (p-value < 0.001), meaning they explain variation in this corporate valuation multiple. This roughly translates to an approximately 4.0% lower market valuation associated with a 10 point increase in ESG Score and an approximately 2.2% higher market valuation associated with a 10 point increase in Controversies Score. As a test for robustness of our estimation and its interpretation, we estimate the above models without fixed effects as well as with added control variables (Capital Expenditures, Price-Earnings Ratio, Dividend Yield). This analysis yields very similar results (Tables B.5 - B.6 in the Appendix), with the coefficients for ESG and Controversies Score still being highly significant, but slightly less extreme. The bigger model explains more variation in the natural logarithm of the Price-to-Book ratio with an adjusted R^2 of 0.538 vs. 0.5 in the models without the additional explanatory variables. Since adding these variables does not change the principal inferences, we conclude that, overall, the sustainability activities as reported by the companies (i.e. the basis for the ESG Score) have a negative effect on their market valuation, whereas the sustainability activities reported by the media have a positive effect.

	coef	std err	t	P> |t|	[0.025	0.975]
const	0.7428	0.052	14.181	0.000	0.640	0.845
ESG Score	-0.0040	0.000	-8.873	0.000	-0.005	-0.003
Z-norm: Size	0.0798	0.010	8.338	0.000	0.061	0.099
ROE	0.0099	0.000	28.000	0.000	0.009	0.011
Net Sales Change Pct	-0.0018	0.002	-1.081	0.280	-0.005	0.001
Annual Return	-0.2574	0.028	-9.286	0.000	-0.312	-0.203
Leverage	0.0014	6.91e-05	20.332	0.000	0.001	0.002
R-squared	0.506					
Adj. R-squared	0.500					
No. Observations	5435					

Table 4.2: Market Valuation Inference - ESG Score

	coef	std err	t	P> |t|	[0.025	0.975]
const	0.4257	0.056	7.604	0.000	0.316	0.535
Controversies Score	0.0022	0.000	8.299	0.000	0.002	0.003
Z-norm: Size	0.0689	0.009	7.537	0.000	0.051	0.087
ROE	0.0096	0.000	27.008	0.000	0.009	0.010
Net Sales Change Pct	-0.0015	0.002	-0.883	0.377	-0.005	0.002
Annual Return	-0.2387	0.028	-8.658	0.000	-0.293	-0.185
Leverage	0.0014	6.93e-05	20.526	0.000	0.001	0.002
R-squared	0.505					
Adj. R-squared	0.499					
No. Observations	5430					

Table 4.3: Market Valuation Inference - Controversies Score

Please find the detailed regression outputs which include the fixed effects in the Appendix (see Tables B.3, B.4).

Now, we want to look at the results of estimating model 4.4, which can be found in Table 4.4. In this model we can see that the influence (i.e. the coefficient) of both the ESG and Controversies Score is smaller, however, it still explains variation in the Price-to-Book ratio. Again, we test the robustness of these results by adding further, potentially explanatory variables: Capital Expenditures (CapEx), Price-Earnings Ratio (PE), Dividend Yield (DY). The estimation results reported in Table 4.5, again, suggest that the loadings on the sustainability score are still highly significant, but slightly less extreme. Hence, we can also conclude here that the influence of sustainability activities as reported by the companies (i.e. the basis for the ESG Score) on their market valuation is negative, whereas the sustainability activities reported by the media have a significant, positive effect.

	coef	std err	t	P> |t|	[0.025	0.975]
const	0.5551	0.058	9.580	0.000	0.442	0.669
ESG Score	-0.0036	0.000	-8.016	0.000	-0.005	-0.003
Controversies Score	0.0019	0.000	7.399	0.000	0.001	0.002
Z-norm: Size	0.1027	0.010	10.255	0.000	0.083	0.122
ROE	0.0097	0.000	27.478	0.000	0.009	0.010
Net Sales Change Pct	-0.0017	0.002	-1.000	0.317	-0.005	0.002
Annual Return	-0.2670	0.028	-9.662	0.000	-0.321	-0.213
Leverage	0.0014	6.89e-05	20.814	0.000	0.001	0.002
R-squared	0.511					
Adj. R-squared	0.505					
No. Observations	5430					

Table 4.4: Market Valuation Inference - ESG and Controversies Score

	coef	std err	t	P> |t|	[0.025	0.975]
const	0.5732	0.059	9.703	0.000	0.457	0.689
ESG Score	-0.0026	0.000	-5.526	0.000	-0.003	-0.002
Controversies Score	0.0014	0.000	5.358	0.000	0.001	0.002
Z-norm: Size	0.1340	0.011	12.398	0.000	0.113	0.155
ROE	0.0118	0.000	31.282	0.000	0.011	0.013
Net Sales Change Pct	-0.0076	0.002	-3.253	0.001	-0.012	-0.003
Annual Return	-0.2989	0.028	-10.524	0.000	-0.355	-0.243
Leverage	0.0014	7.04e-05	19.457	0.000	0.001	0.002
CapEx	-4.346e-08	5.02e-09	-8.657	0.000	-5.33e-08	-3.36e-08
Price-Earnings Ratio	0.0015	0.000	8.181	0.000	0.001	0.002
Dividend Yield	-0.0220	0.006	-3.703	0.000	-0.034	-0.010
R-squared	0.547					
Adj. R-squared	0.541					
No. Observations	5006					

Table 4.5: Market Valuation Inference - ESG and Controversies Score - added variables

A detailed investigation of the model assumptions is given in Section B.3 in Appendix B.

As a second step, we investigate the change of the relationship between the companies' sustainability activities and their market valuation over the years. In detail, we estimate models 4.2 and 4.3 without the yearly fixed effects for each year between 2004 and 2020. Figure 4.1 graphically depicts the estimation results. Again, as a robustness check, we estimate the full model 4.4 with the added variables Capital Expenditures (CapEx), Price-Earnings Ratio (PE) and Dividend Yield (DY). These results are shown in Figure 4.2. We can see that the ESG Score loading is almost always negative, agreeing with the previous results, and has an overall downward trend. The Controversies Score loading, on the other hand, is almost alway positive and has an overall upwards trend. Furthermore, the picture in the full model looks very similar, hinting at robust results.

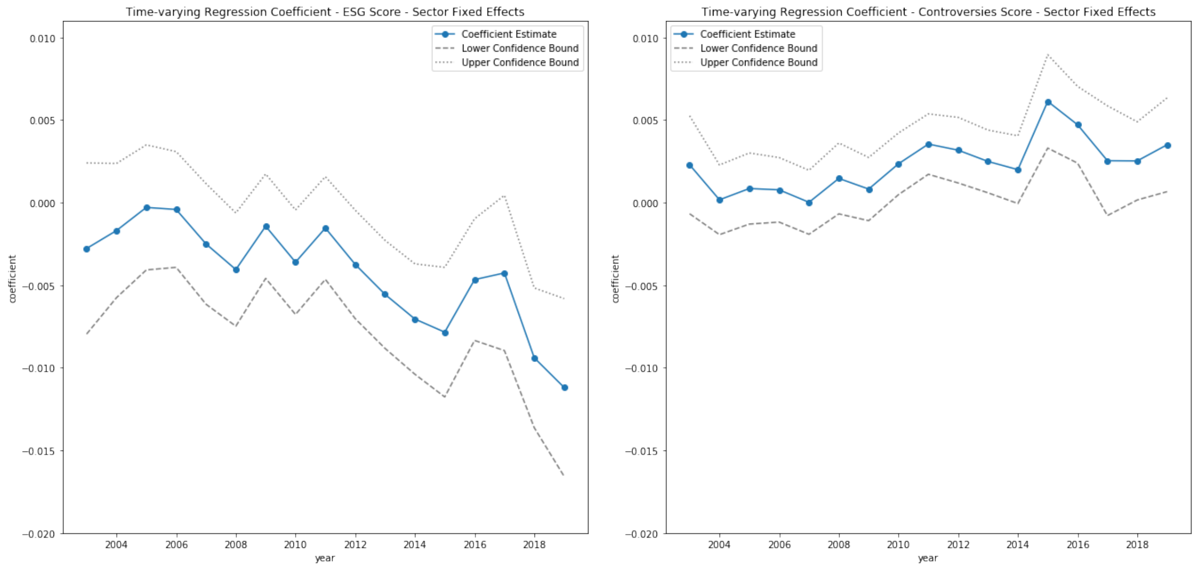

Figure 4.1: Market Valuation Inference - Time-Varying Coefficient

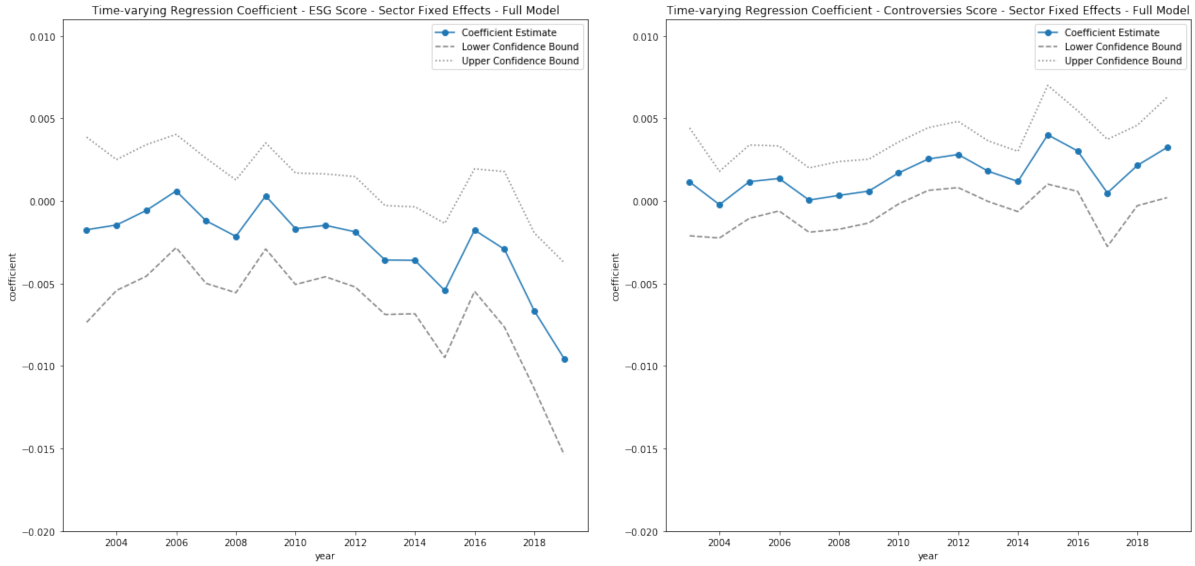

Figure 4.2: Market Valuation Inference - Time-Varying Coefficient - Full Model

Overall, the results presented in this section suggest that public ESG sentiment picked up by the media (as reported in the Controversies Score) has a considerable positive effect on how a company is valued in the market. On the contrary, ESG ratings based on company reportings (as reported by the ESG Score) seem to have a negative effect. Both effects seem to be explainable by other market factors such as Capital Expenditures, Price-Earnings Ratio or Dividend Yield to some degree, since the loadings are less extreme in models with the added variables. However, their influence is still significant and the principal inferences stay the same.

This is in line with research and expert views on ESG investing. Firstly, the awareness on a company's sustainability activities has only grown in recent years [20], meaning that the market view is still in a shifting and adjusting state. Secondly, ESG ratings differ dramatically between providers [Cha+16]. This indicates a discord both between investors and the overall conception of how sustainability is to be measured. Lastly, the influence of investor sentiment on stock prices has been widely acknowledged [SYY12] [BWY12] [BC05]. Only recently, [Ser20] have shown that ESG sentiment as picked up by global media (using *TrueValue Labs* ESG Sentiment Score) influences investor views about the value of corporate sustainability activities.

Now, having examined the relationship between ESG Scores and financial market valuation on a company level, we will take a closer look at the relationship between ESG Scores and returns on a portfolio basis.

4.2 Portfolio Returns

This section focuses on exploring the connection between ESG Scores and financial perfomance through a portfolio-based approach. After having seen how the ESG and Controversies Score influence a company's market valuation in the form of a valuation multiple, we are now interested in the behavior of future returns of portfolios that consider corporate ESG data. The performance of portfolios that take sustainability actions into account might be influenced by different perspectives and actions. Here, possible driving factors could be that investors expect future legal problems or reputational damages for lowly scored companies, an event reported by the media might overthrow the perception of the underlying sustainability activities or that specific sustainability domains (e.g. environment) have bigger awareness in society than others.

For this analysis, we will first build portfolios dependent on differently focused sustainability scores and examine their performance. Second, we will build a factor from the respective portfolios in order to draw conclusions from its behavior. Finally, we will not only look at their performances in an overall time series, but split the analysis for times of crises.

4.2.1 Approach

We follow a standard factor construction method [FF15] to build our different, value-weighted (with respect to market capitalization $MV_{i,t}$) portfolios. We rebalance every month, however, the portfolio constituents only change once a year, since the scores are only updated annually. Furthermore, we allow three months for the data to be published in order to guarantee implementable trading strategies.
At each time point t, i.e. every month-end, and each sustainability score $S_{i,t}$ for company i at time point t we build portfolios through the following steps:

1. Sort companies with respect to their sustainability scores $S_{i,t}$.

2. Group them into the top and bottom tercile (half)

$$\mathcal{S}_{t,top} = \{company_i | S_{i,t} > q_{top}\}$$
$$\mathcal{S}_{t,bottom} = \{company_i | S_{i,t} < q_{bottom}\}$$

   with $\mathbb{P}(S_{.,t} \leqslant q_{top}) \geqslant 2/3$ (resp. 1/2) and $\mathbb{P}(S_{.,t} \leqslant q_{bottom}) \geqslant 1/3$ (resp. 1/2).

3. Split both sets of companies $\mathcal{S}_{t,top}, \mathcal{S}_{t,bottom}$ into three size groups (based on their

market capitalization)

$$\begin{aligned}
\mathcal{S}_{t,top,1} &= \{company_i | S_{i,t} > q_{top} \text{ and } MV_{i,t} \geqslant q^m_{2/3}\} \\
\mathcal{S}_{t,top,2} &= \{company_i | S_{i,t} > q_{top} \text{ and } q^m_{2/3} > MV_{i,t} \geqslant q^m_{1/3}\} \\
\mathcal{S}_{t,top,3} &= \{company_i | S_{i,t} > q_{top} \text{ and } MV_{i,t} < q^m_{1/3}\} \\
\mathcal{S}_{t,bottom,1} &= \{company_i | S_{i,t} < q_{bottom} \text{ and } MV_{i,t} \geqslant q^m_{2/3}\} \\
\mathcal{S}_{t,bottom,2} &= \{company_i | S_{i,t} < q_{bottom} \text{ and } q^m_{2/3} > MV_{i,t} \geqslant q^m_{1/3}\} \\
\mathcal{S}_{t,bottom,3} &= \{company_i | S_{i,t} < q_{bottom} \text{ and } MV_{i,t} < q^m_{1/3}\}
\end{aligned}$$

with $\mathbb{P}(MV_{.,t} \leqslant q^m_{2/3}) \geqslant 2/3$, $\mathbb{P}(MV_{.,t} \leqslant q^m_{1/3}) \geqslant 1/3$.

4. Compute the value-weighted returns of these portfolios for the following month, $r_{t,top,1}, r_{t,top,2}, r_{t,top,3}, r_{t,bottom,1}, r_{t,bottom,2}, r_{t,bottom,3}$

5. Set the average return of these as the portfolio returns, $r_{t,top} = 1/3(r_{t,top,1} + r_{t,top,2} + r_{t,top,3})$ and $r_{t,bottom} = 1/3(r_{t,bottom,1} + r_{t,bottom,2} + r_{t,bottom,3})$ for $\mathcal{S}_{t,top}$ and $\mathcal{S}_{t,bottom}$, respectively.

We follow this approach for the *level* in sustainability score and the *change* in sustainability score in the past year. For the latter, simply substitute $S_{i,t}$ by $S_{i,t} - S_{i,t-1}$. Furthermore, we also use this approach with a combination of two scores $S^1_{i,t}, S^2_{i,t}$ resulting in the combinations "top-top", "top-bottom", "bottom-top", "bottom-bottom":

$$\begin{aligned}
\mathcal{S}_{t,top/top} &= \{company_i | S^1_{i,t} > q^1_{top} \text{ and } S^2_{i,t} > q^2_{top}\} \\
\mathcal{S}_{t,top/bottom} &= \{company_i | S^1_{i,t} > q^1_{top} \text{ and } S^2_{i,t} < q^2_{bottom}\} \\
\mathcal{S}_{t,bottom/top} &= \{company_i | S^1_{i,t} < q^1_{bottom} \text{ and } S^2_{i,t} > q^2_{top}\} \\
\mathcal{S}_{t,bottom/bottom} &= \{company_i | S^1_{i,t} < q^1_{bottom} \text{ and } S^2_{i,t} < q^2_{bottom}\}
\end{aligned}$$

with $\mathbb{P}(S^{1/2}_{.,t} \leqslant q^{1/2}_{top}) \geqslant 2/3$ (resp. 1/2) and $\mathbb{P}(S^{1/2}_{.,t} \leqslant q^{1/2}_{bottom}) \geqslant 1/3$ (resp. 1/2), leading to portfolio returns after computing the average of three size groups:

$$\begin{aligned}
r_{t,top/top} &= 1/3(r_{t,top/top,1} + r_{t,top/top,2} + r_{t,top/top,3}) \\
r_{t,top/bottom} &= 1/3(r_{t,top/bottom,1} + r_{t,top/bottom,2} + r_{t,top/bottom,3}) \\
r_{t,bottom/top} &= 1/3(r_{t,bottom/top,1} + r_{t,bottom/top,2} + r_{t,bottom/top,3}) \\
r_{t,bottom/bottom} &= 1/3(r_{t,bottom/bottom,1} + r_{t,bottom/bottom,2} + r_{t,bottom/bottom,3})
\end{aligned}$$

In order to test the portfolio performances, we regress the monthly portfolio returns on the Fama-French 5 factors. This means, we control for well-known market anomalies such as the size effect, value effect, profitability effect and investment effect [FF15].

- **Size Effect:** The size effect describes the market anomaly that smaller companies tend to outperform larger companies over the long-term. The *Small Minus Big* (SMB) factor is defined as the return on a diversified portfolio of small stocks minus the return on a diversified portfolio of big stocks (average return on the nine small stock portfolios minus the average return on the nine big stock portfolios):

$$\begin{aligned} SMB_{(B/M)} &= 1/3(\mathit{Small\ Value} + \mathit{Small\ Neutral} + \mathit{Small\ Growth}) \\ &\quad - 1/3(\mathit{Big\ Value} + \mathit{Big\ Neutral} + \mathit{Big\ Growth}) \\ SMB_{(OP)} &= 1/3(\mathit{Small\ Robust} + \mathit{Small\ Neutral} + \mathit{Small\ Weak}) \\ &\quad - 1/3(\mathit{Big\ Robust} + \mathit{Big\ Neutral} + \mathit{Big\ Weak}) \\ SMB_{(INV)} &= 1/3(\mathit{Small\ Conservative} + \mathit{Small\ Neutral} + \mathit{Small\ Aggressive}) \\ &\quad - 1/3(\mathit{Big\ Conservative} + \mathit{Big\ Neutral} + \mathit{Big\ Aggressive}) \\ SMB &= 1/3(SMB_{(B/M)} + SMB_{(OP)} + SMB_{(INV)}) \end{aligned}$$

- **Value Effect:** The value effect describes the market anomaly that companies with a high book-to-market ratios (value stocks), outperform those with lower book-to-market ratios (growth stocks). The *High Minus Low* (HML) factor is defined as the difference between the returns on diversified portfolios of high and low B/M stocks (average return on the two value portfolios minus the average return on the two growth portfolios):

$$HML = 1/2(\mathit{Small\ Value} + \mathit{Big\ Value}) - 1/2(\mathit{Small\ Growth} + \mathit{Big\ Growth})$$

- **Profitability Effect:** The profitability effect is similar to the value effect and describes the market anomaly that companies with higher profitability tend to earn better results. The *Robust Minus Weak* (RMW) factor is defined as the difference between the returns on diversified portfolios of stocks with robust and weak profitability (average return on the two robust operating profitability portfolios minus the average return on the two weak operating profitability portfolios):

$$RMW = 1/2(\mathit{Small\ Robust} + \mathit{Big\ Robust}) - 1/2(\mathit{Small\ Weak} + \mathit{Big\ Weak})$$

- **Investment Effect:** The *Conservative Minus Aggressive* (CMA) factor is defined as the difference between the returns on diversified portfolios of the stocks of low and high investment firms, which we call conservative and aggressive (average return on the two conservative investment portfolios minus the average return on the two aggressive investment portfolios):

$$\begin{aligned} CMA &= 1/2(\mathit{Small\ Conservative} + \mathit{Big\ Conservative}) \\ &\quad - 1/2(\mathit{Small\ Aggressive} + \mathit{Big\ Aggressive}) \end{aligned}$$

Controlling for the above listed market anomalies in our analysis for portfolio performance yields the following regression model as a basis for inference detection:

$$r_{t,.} - r_{t,f} = \alpha_{.} + \beta_{1,.}(r_{t,M} - r_{t,f}) + \beta_{2,.}SMB_t + \beta_{3,.}HML_t + \beta_{4,.}RMW_t + \beta_{5,.}CMA_t + e_{t,.} \quad (4.5)$$

where $r_{t,.}$ is a portfolio return, $r_{t,f}$ is the risk free rate, $(r_{t,M} - r_{t,f})$ is the excess market return, SMB_t is the return spread on small cap minus large cap stocks, HML_t is the return spread on value minus growth stocks, RMW_t is the return spread on most profitable minus least profitable stocks and CMA_t is the return spread on stocks that invest conservatively minus aggressively [FF15].

For an in-depth analysis of portfolios constructed using annual sustainability scores, we want to compare portfolio behavior between calm and turbulent market phases. In order to identify these market phases, we estimate Markov Switching Models (MSM) using the *MarkovRegression* class from the Python package *statsmodels* to fit the MSM by maximum likelihood estimation using the Hamilton filter. Here, the monthly returns of the *S&P500* between September 1983 and January 2021 build the basis for identifying times of crises, i.e. turbulent market phases. We estimate a *2-regime* and *3-regime* MSM with regime-dependent drift and volatility. Since both yield very similar results for a turbulent regime (i.e. regime with high volatility and possibly negative drift) but estimating a *3-regime* MSM is generally unstable ([Bol+96], [SLL98]), we decide to apply the *2-regime* MSM in our further analysis. For this model, in contrast to the returns of the index in the whole period, where the Jarque-Bera test rejects normality with a p-value smaller than 10^{-5}, the normality of the returns in each regime cannot be rejected.

Table 4.6 shows that two significantly different market phases could be identified: a calm phase with positive expected return and low volatility and a turbulent phase with negative expected return and high volatility. Furthermore, Table 4.7 depicts the transition probabilities for the two regimes. The probability of remaining in a calm phase from month to month is over 92% and the probability of remaining in a turbulent phase is over 87%, meaning they can be considered persistent.

Mean			Standard Deviation		
Overall	**Calm**	**Turbulent**	**Overall**	**Calm**	**Turbulent**
0.85%	1.42%	-0.12%	4.33%	2.65%	6.08%

Table 4.6: Mean and Standard Deviation of monthly S&P500 returns - Regime-dependent

	Calm	**Turbulent**
Calm	0.927	0.073
Turbulent	0.123	0.877

Table 4.7: Transition Probabilities

Since our goal is not to forecast changes in regimes, but to identify the two market phases from our point of view today, we want to make probability statements about the regimes that incorporate the overall information. A way of doing this is by computing the smoothed marginal probabilities (can be recursively obtained from Definition 3.3.3, see [Kim94]) of the *2-regime Markov Switching Model*. Figure 4.3 depicts both the performance of the *S&P500* in blue and the turbulent market phases in grey blocks. In detail, we define turbulent market phases as those periods, where the smoothed marginal probability of being in the turbulent regime is greater than the smoothed marginal probability of being in the calm regime.

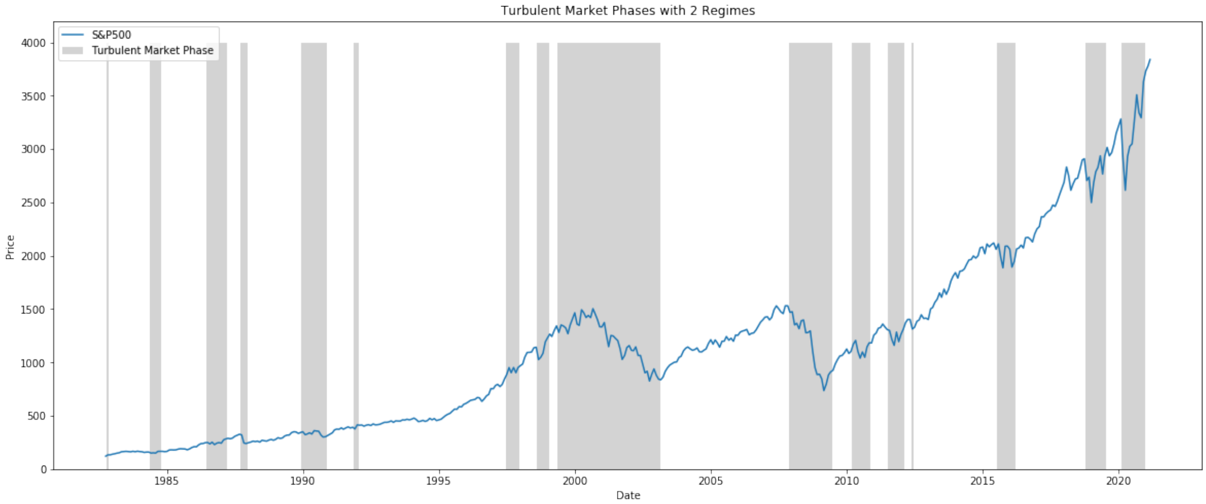

Figure 4.3: S&P 500 - Times of Crises

Hence, in our period of interest from January 2004 until December 2020, we can identify the following times of crises:

- **Nov 2007 - May 2009:** Financial Crisis [Fed09][MA13]
- **Mar 2010 - Oct 2010:** European Debt Crisis [BBC12][BBC10]
- **July 2011 - Jan 2012:** European Debt Crisis [BBC12][The11]
- **May 2012:** European Debt Crisis [BBC12]

- **July 2015 - Feb 2016:** Stock Market Selloff [Reu16][For16]
- **Oct 2018 - June 2019:** China-US Trade War [CNB18][BBC20][LW18]
- **Feb 2020 - Nov 2020:** Covid-19 pandemic [Yil20]

4.2.2 Data

As a basis for our analysis we, again, look at data of the current *S&P500 stocks* from 2004 until 2020, which we collect from *Refinitiv* Datastream. In order to build ESG portfolios, we perform the procedure described in the previous section for the following sustainability scores:

- **ESG Score:** see Section 4.1.2
- **Controversies Score:** see Section 4.1.2
- **E Score:** *Refinitiv*'s Environmental (E) Score is the relative sum of the category weights (which vary per industry) for the measures belonging to environmental topics: *emissions, waste, biodiversity, environmental management systems, product innovation, green revenues, water, energy, sustainable packaging, environmental supply chain.* [Ref21]
- **S Score:** *Refinitiv*'s Social (S) Score is the relative sum of the category weights (which vary per industry) for the measures belonging to social topics: *community, human rights, responsible marketing, product quality, data privacy, diversity and inclusion, career development and training, working conditions, health and safety.* [Ref21]
- **G Score:** *Refinitiv*'s Governance (G) Score is the relative sum of the category weights for the measures belonging to governance topics: *CSR strategy, ESG reporting and transparency, structure, compensation, shareholder rights, takeover defenses.* [Ref21]
- **ESG Combined Score:** *Refinitiv*'s (formerly Asset4) ESG Combined (ESGC) Score is based on the reported information for the three pillars with the ESG Controversies overlay captured from global media sources. It aims to "discount the ESG performance score based on negative media stories" [Ref21] (p. 7). It is calculated as the weighted average of the ESG Score and ESG Controversies Score per fiscal period when companies are involved in ESG controversies, otherwise it is equal to the ESG Score. [Ref21]

Table 4.8 presents summary statistics for the sustainability scores. We see that more than 75% of companies in our sample have a Controversies Score which is higher than 72 and more than half of the companies even have the highest possible score of 100. For the other scores, the values seem to be more evenly distributed.

Table 4.8: Summary Statistics for Portfolio Analysis Sample

	ESG Score	**E Score**	**S Score**	**G Score**	**Controversies Score**	**ESG Combined Score**
count	7167	7161	7161	7167	7161	7167
mean	48.89	37.72	50.97	54.19	81.40	45.29
std	20.11	30.31	22.01	21.82	30.27	18.39
min	0.63	0.00	0.26	0.36	0.62	0.63
25%	32.90	5.51	33.37	37.92	72.92	31.31
50%	48.63	37.14	51.14	56.04	100.00	44.38
75%	65.34	64.73	67.76	71.52	100.00	59.06
max	95.07	98.55	97.75	98.76	100.00	92.53

The data for the monthly returns of the Fama French five factors was obtained from Kenneth R. French's website [Fre]. Since the companies we focus on are the *S&P500* stocks it makes sense to benchmark on the Fama French five factors. They include all NYSE, AMEX, and NASDAQ firms that have available data for computation of the respective factors. Hence, the portfolio companies' stock universe matches the benchmark universe.

4.2.3 Results

First, we investigate the behaviour for portfolios built with the approach explained in section 4.2.1 for each score presented in section 4.2.2. We compare different portfolio metrics and regress their performance on the Fama French 5 factors (see Equation 4.5). Second, we build portfolios based on multiple scores to investigate interactions. For both approaches, we split the analyses with respect to times of crises for a more granular view and investigation of the behaviour. To conclude, using the insights we have gained, we construct a factor from these portfolios and look at its relationship with other market anomalies.

Single-Score Portfolios

In this first analysis, we aim to present and compare the behaviour of portfolios constructed on both the *levels* and *change* in different sustainability scores. In detail, we start by looking at a portfolio strategy using the overall ESG Score which is based on company-reported details. Next, we compare the drivers of the ESG Score by looking at the portfolios based on the E, S and G pillar, respectively. Finally, we also consider a portfolio strategy using the ESG Controversies Score, which is based on media-reported details, and the ESG Combined Score.

Firstly, Tables 4.9 and 4.10 give an overview of the different portfolio constituents' characteristics. We compute the average scores, market capitalization and number of companies for each portfolio and the average standard deviation of the scores and market capitalization inside each portfolio, where the average is taken over the 17 different portfolio constituents from 2003-2020. An interesting finding here is the fact that each portfolio that is constructed on the level of a sustainability score based on company-reportings (i.e. ESG Score, E Score, S Score, G Score) exhibits a higher average Controversies Score for the "bottom" portfolios of the respective score when compared to their "top" counterpart. Actually, we can observe a negative correlation of almost -0.3 of the ESG Score and the Controversies Score in our complete sample. Furthermore, the average market capitalization is significantly higher for all the "top" portfolios, with the exception of the portfolios constructed on the Controversies Score. This finding will be corroborated when we examine the portfolio loadings on the Fama French 5 Factors. The final remark we want to make on Tables 4.9 and 4.10 is that there seems to be a high correlation between the ESG Score and its Pillar Scores: each "bottom" level portfolio exhibits significantly lower values for each of the other scores compared to the "top" level portfolio. Indeed, the correlation between the ESG Score and the Environment/Social Score is around 0.87 and 0.7 between the ESG Score and the Governance Score. Between the Pillar Scores, Environment and Social have a rather high correlation of 0.73, whereas their correlations with the Governance Score are only around 0.4. This indicates that high ratings with respect to environmental activities seem to go hand-in-hand with high ratings with respect to social activities. However, sustainability activities with respect to governance issues seem to be more independent.

			∅ ESG Score	∅ E Score	∅ S Score	∅ G Score	∅ Controversies Score	∅ ESG Combined Score	∅ Market Cap	∅#Companies
ESG Score	**Level**	**Bottom**	27.3	10.9	30.5	36.9	90.8	26.8	14851.3	130.2
		Top	67.5	61.1	69.2	68.6	70.2	59.6	55529.1	130.8
	Change	**Bottom**	44.5	34.5	47.5	48.2	80.7	41.0	32308.5	128.1
		Top	52.0	39.2	53.4	59.0	81.0	48.4	33215.5	129.6
E Score	**Level**	**Bottom**	31.8	6.6	35.5	44.2	89.8	31.2	15528.4	145.9
		Top	64.3	66.0	65.0	62.2	70.8	56.8	56043.2	131.0
	Change	**Bottom**	50.0	39.1	52.1	56.0	78.0	45.5	35887.2	120.7
		Top	53.2	47.3	54.1	56.6	79.9	49.4	35353.0	128.1
S Score	**Level**	**Bottom**	30.3	15.6	27.2	45.5	90.7	29.7	14030.4	130.5
		Top	65.0	56.5	71.9	61.8	70.9	57.5	55007.7	131.2
	Change	**Bottom**	46.0	35.2	46.3	53.4	81.3	42.7	32570.9	127.8
		Top	50.5	37.7	54.5	54.7	81.5	47.1	31856.0	129.3
G Score	**Level**	**Bottom**	33.6	24.8	41.8	28.8	86.2	32.0	24791.9	129.7
		Top	60.6	46.6	56.6	76.9	76.2	54.8	44486.4	131.3
	Change	**Bottom**	45.3	36.2	49.8	46.0	81.0	42.0	32974.9	128.5
		Top	50.3	36.4	50.1	60.4	80.6	46.5	31591.8	129.4
Controversies Score	**Level**	**Bottom**	54.8	45.5	57.2	58.0	51.6	46.0	59658.9	148.8
		Top	42.6	28.9	44.3	50.8	100.0	42.6	16702.5	243.8
	Change	**Bottom**	53.9	44.3	56.3	57.1	48.7	44.5	51296.1	98.6
		Top	54.1	44.2	56.4	58.1	80.9	50.6	47078.8	93.4
ESG Combined Score	**Level**	**Bottom**	29.4	14.7	32.9	37.7	80.2	25.9	27846.4	129.9
		Top	64.5	56.8	65.9	67.0	86.5	62.5	37033.9	131.1
	Change	**Bottom**	46.1	35.8	48.9	50.0	72.2	40.1	34466.2	128.2
		Top	52.5	40.4	54.0	58.9	86.1	50.6	31634.6	129.6

Table 4.9: Statistics of Portfolios - Mean

			∅Std. ESG Score	∅ Std. E Score	∅ Std. S Score	∅ Std. G Score	∅ Std. Controversies Score	∅ Std. ESG Combined Score	∅ Std. Market Cap
ESG Score	**Level**	**Bottom**	7.6	12.2	11.5	17.6	21.4	7.7	21411.1
		Top	8.4	17.4	12.9	16.2	34.4	12.6	75718.6
	Change	**Bottom**	17.7	26.5	20.0	20.7	30.4	16.0	53049.4
		Top	17.0	27.6	19.6	19.6	29.6	15.6	54493.0
E Score	**Level**	**Bottom**	11.8	6.2	15.0	20.3	22.6	11.6	22146.6
		Top	11.6	11.7	15.8	19.5	34.2	14.1	74833.0
	Change	**Bottom**	17.6	25.5	20.1	21.0	31.6	15.9	53954.8
		Top	16.0	23.6	18.8	20.6	30.1	15.1	52878.4
S Score	**Level**	**Bottom**	10.6	17.5	7.8	20.7	21.3	10.5	17623.9
		Top	11.5	21.0	9.9	19.7	34.3	13.7	74988.2
	Change	**Bottom**	17.6	26.7	19.4	21.0	29.7	16.1	55200.1
		Top	17.4	27.1	18.9	21.2	29.5	16.1	51404.6
G Score	**Level**	**Bottom**	14.3	24.4	18.7	10.0	26.0	13.4	38880.7
		Top	14.7	25.8	19.6	7.9	32.2	14.4	75577.3
	Change	**Bottom**	17.6	27.3	20.0	19.3	30.2	16.2	52901.4
		Top	17.5	27.2	20.2	18.4	30.2	16.0	48843.1
Controversies Score	**Level**	**Bottom**	17.5	26.7	19.8	20.9	30.0	15.7	81725.2
		Top	16.7	24.9	18.6	21.0	0.1	16.7	20049.4
	Change	**Bottom**	17.8	27.4	20.0	21.1	29.7	15.3	70223.5
		Top	17.8	27.3	20.5	20.7	28.0	16.7	63338.9
ESG Combined Score	**Level**	**Bottom**	11.4	17.5	14.4	18.5	32.8	6.9	55157.6
		Top	9.6	19.9	13.9	16.6	19.8	8.6	45868.6
	Change	**Bottom**	18.5	27.0	20.4	21.6	35.5	15.2	53747.6
		Top	17.0	27.6	19.6	19.8	22.8	15.9	47691.1

Table 4.10: Statistics of Portfolios - Average Standard Deviation

In Figures 4.4 - 4.9 we present the portfolio performance of the strategies based on the different sustainability scores, where we compare the "top" and "bottom" portfolios with the market (Fama French Market Return) as a benchmark portfolio. We start with a portfolio value of 1. It is important to note that the portfolios depicted can start in different years due to the fact that the data was unavailable before (i.e. if we construct the portfolios based on score changes, we can only start a year after the first observation of a score level).

Throughout, we notice that the market portfolio has the weakest performance in comparison to all portfolios constructed with ESG data. Both, the ”top” and ”bottom” sustainability portfolios have a better traction in bullish times. We can see that they sustain considerable losses in bearish times - during the financial crisis in 2008 all sustainability portfolios drop down to the price of the market portfolio. However, after 2008 they, again, outperform the market, giving them such an advantage that their performance outweighs the losses they sustain in more recent downward facing times. Furthermore, what is interesting to see is the performance and losses during the start of the COVID pandemic in 2020. Almost all of the portfolios constructed on the change of the respective sustainability score in the past year have a substantial drop in value which is not balanced with the outperformance up to this point, leaving them with portfolio values close to overall market performance. For the portfolios based on the level of the respective sustainability score we can also see a significant drop, however, the outperformance of the ”bottom” (”top” for ESG Controversies Score) portfolio in the time prior to the event more than outweighs the losses. In contrast, the ”top” (”bottom” for ESG Controversies Score) portfolios forfeit their gains completely, leaving them with a portfolio value close to the market and even worse than the market during the recovery stage. We will investigate these performances in times of crises in more detail.

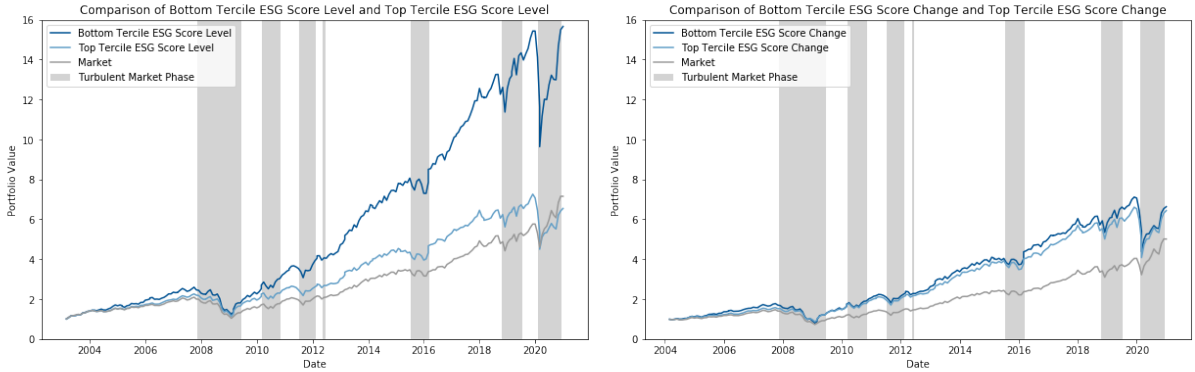

Figure 4.4: Portfolio Performance - ESG Score Level and Change

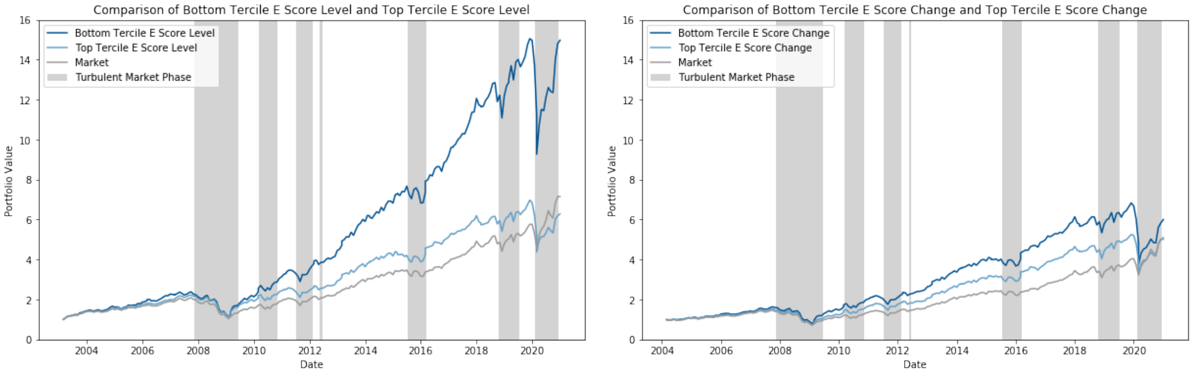

Figure 4.5: Portfolio Performance - E Score Level and Change

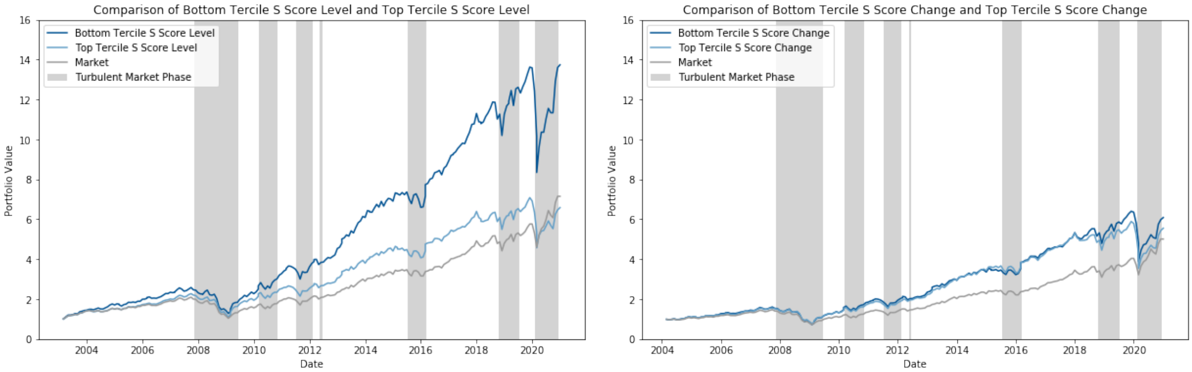

Figure 4.6: Portfolio Performance - S Score Level and Change

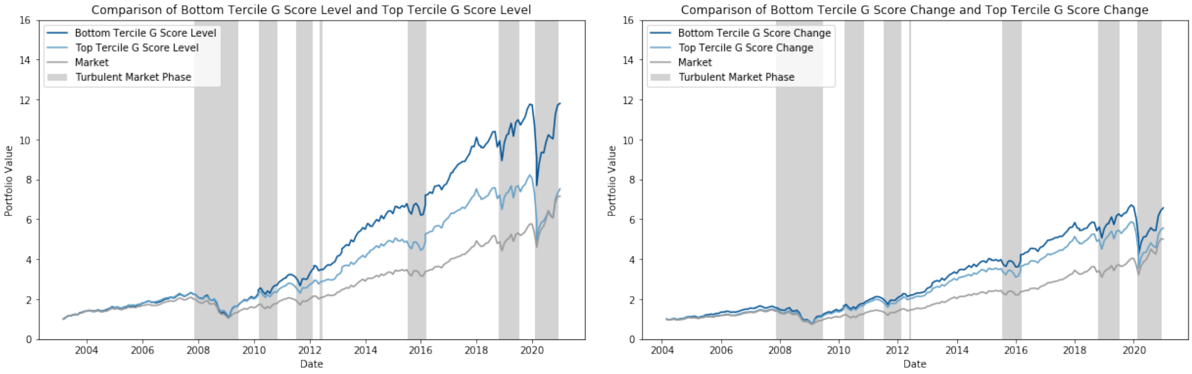

Figure 4.7: Portfolio Performance - G Score Level and Change

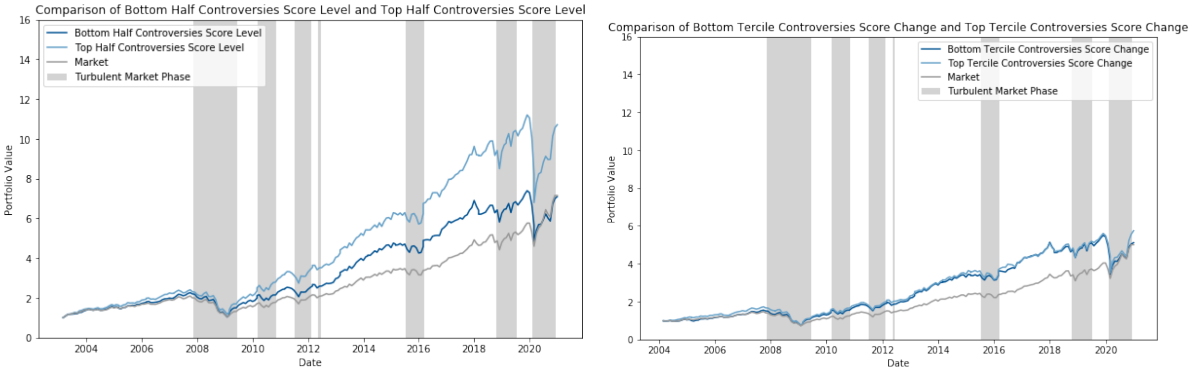

Figure 4.8: Portfolio Performance - ESG Controversies Score Level and Change

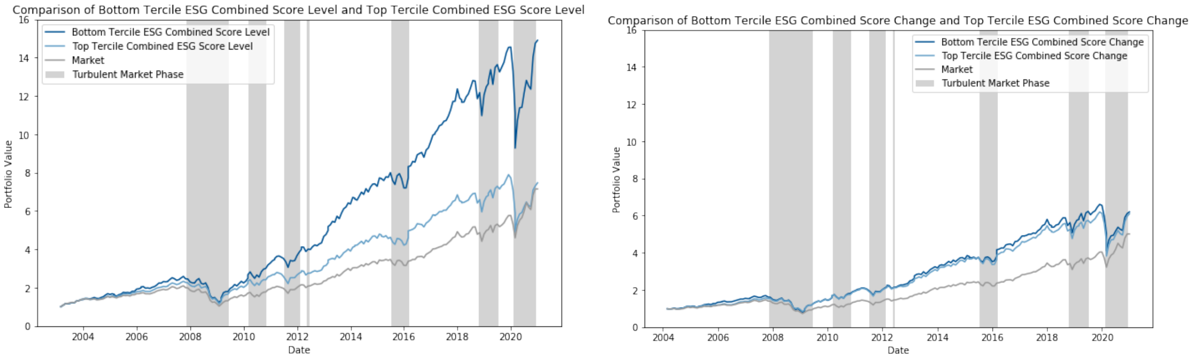

Figure 4.9: Portfolio Performance - ESG Combined Score Level and Change

Looking at the difference between the "top" and "bottom" portfolios, we can clearly see two high-level results: First, the "top" and "bottom" portfolios constructed on the *change* in scores behave very similarly and show no obvious differences, whereas the "top" and "bottom" portfolios constructed on the score *level* present definite dissimilar performance. Second, the "bottom" portfolio outperforms the "top" portfolio for all sustainability scores which are based on company-reported information, while it is the other way around for the ESG Controversies Score (see Figure 4.8). Going into more detail and comparing the portfolios based on the overall ESG Score (Figure 4.4) with the portfolios built on the different pillar scores (Figures 4.5 - 4.7), it seems like the main drivers for the "bottom" level ESG Score portfolio are the Environment and Social Score, since the "bottom" portfolio based on the level of Governance Score exhibits less extreme performance.

The Table 4.11 below summarizes key portfolio metrics, which emphasize the results we have seen so far. Here, kurtosis is given as the Fisher definition, i.e. 3 is substracted from the result to obtain 0 for a normal distribution. For the first two moments we can see that almost all sustainability portfolios, independent of "bottom" or "top", have a higher mean return than the market, but also have a higher volatility associated with it. However, looking at the expected Sharpe ratio, we can clearly see, that all *level* portfolios have a higher expected risk-adjusted return than the market, whereas all of the *change* portfolios have a similar or lower expected risk-adjusted return. Furthermore, we again notice that except for the ESG Controversies Score portfolios, the ”bottom” portfolio has a better (risk-adjusted) expected return than the ”top” portfolio. Overall, the ESG level bottom portfolio has the highest (risk-adjusted) return, closely followed by the ESG Combined level bottom portfolio and the single pillar level bottom portfolios. When looking at the higher moments, we notice slightly better values for the ”bottom” vs. the ”top” portfolios. Finally, comparing the (Conditional) Value-at-Risk values, we can see that the ”top” portfolios have slightly less negative values than their ”bottom” counterparts, indicating a minor advantage in times of crises. Here, the Environment level top portfolio yields the least negative Conditional Value-at-Risk.

This is in line with the results we have seen so far - it harmonizes with the results seen on company-level in Section 4.1.3, where we could already witness a negative influence of the ESG Score and a positive influence of the ESG Controversies Score on company valuation, and the behaviour depicted in Figures 4.4- 4.9.

Overall Market			Mean Return	Mean Return p.a.	Return Std.	Return Std. p.a.	$\mathbb{E}$[SR]	Skewness	Kurtosis	$VaR_{0.95}$	$CVaR_{0.95}$
Market			0.89%	11.24%	4.33%	16.63%	0.18	-0.59	1.87	-7.58%	-9.69%
ESG Score	**Level**	**Bottom**	1.31%	16.92%	4.83%	19.42%	0.2502	-0.85	3.97	-7.07%	-11.45%
		Top	0.91%	11.53%	4.44%	17.09%	0.1827	-0.89	3.44	-6.65%	-10.95%
	Change	**Bottom**	0.99%	12.55%	4.91%	19.08%	0.1806	-0.69	3.65	-7.51%	-12.33%
		Top	0.96%	12.13%	4.54%	17.55%	0.1884	-0.84	3.35	-6.73%	-10.94%
E Score	**Level**	**Bottom**	1.29%	16.62%	4.79%	19.22%	0.2476	-0.77	3.65	-6.75%	-11.46%
		Top	0.89%	11.25%	4.37%	16.79%	0.1808	-0.84	3.31	-6.89%	-10.70%
	Change	**Bottom**	0.95%	12.01%	5.00%	19.35%	0.1692	-0.87	4.75	-7.62%	-12.39%
		Top	0.86%	10.82%	4.71%	18.0%2	0.1607	-0.92	3.86	-8.00%	-11.92%
S Score	**Level**	**Bottom**	1.25%	16.11%	4.79%	19.15%	0.2398	-0.79	3.99	-6.65%	-11.34%
		Top	0.92%	11.61%	4.53%	17.43%	0.1807	-0.70	3.09	-7.34%	-10.85%
	Change	**Bottom**	0.95%	12.02%	4.90%	18.95%	0.1729	-0.81	3.74	-7.86%	-12.44%
		Top	0.90%	11.34%	4.71%	18.11%	0.1690	-0.93	3.66	-7.65%	-11.80%
G Score	**Level**	**Bottom**	1.19%	15.20%	4.79%	19.03%	0.2258	-0.68	3.37	-7.24%	-11.39%
		Top	0.98%	12.42%	4.58%	17.75%	0.1919	-0.91	4.07	-7.12%	-11.19%
	Change	**Bottom**	0.98%	12.45%	4.85%	18.83%	0.1813	-0.69	3.24	-7.90%	-11.98%
		Top	0.90%	11.35%	4.71%	18.09%	0.1692	-0.87	3.50	-7.07%	-11.52%
Controversies Score	**Level**	**Bottom**	0.95%	12.07%	4.56%	17.63%	0.1869	-0.83	3.31	-7.25%	-11.27%
		Top	1.14%	14.59%	4.73%	18.69%	0.2194	-0.84	3.89	-6.96%	-11.47%
	Change	**Bottom**	0.86%	10.78%	4.64%	17.74%	0.1626	-0.76	2.82	-7.69%	-11.33%
		Top	0.92%	11.68%	4.93%	19.02%	0.1667	-0.68	3.96	-7.22%	-12.17%
ESG Combined Score	**Level**	**Bottom**	1.29%	16.60%	4.80%	19.25%	0.2469	-0.77	3.42	-6.92%	-11.29%
		Top	0.98%	12.36%	4.55%	17.63%	0.1921	-0.82	3.53	-7.50%	-11.09%
	Change	**Bottom**	0.95%	12.07%	4.81%	18.61%	0.1769	-0.79	3.44	-7.37%	-11.91%
		Top	0.94%	11.86%	4.63%	17.89%	0.1802	-0.77	3.35	-7.26%	-11.33%

Table 4.11: Properties of Monthly Portfolio Returns

In order to examine the portfolios' behavior more deeply, Tables 4.12 and 4.13 report key portfolio metrics of the portfolios in a turbulent and in a calm market, respectively. A first observation we want to make is about the overall magnitude of skewness and kurtosis: as suggested by our findings for the MSM, the skewness and kurtosis of the returns when split into two regimes are significantly closer to those of a normal distribution than when assuming only one state for the complete time period in Table 4.11. Additionally, all portfolios exhibit significantly lower, negative return, higher volatility and worse (Conditional) Value-at-Risk in the turbulent market phases compared to calm market phases. This confirms the impression we had gained from the performance plots about the fact that the outperformance gained in calm phases is forfeited to an extent in times of crises.

Looking at the sustainability portfolios, we again see that except for the ESG Controversies Score portfolios, the "bottom" portfolio has a better, i.e. less negative in times of crises, (annual) return than the "top" portfolio, but exhibits higher volatility in both market phases. In a calm market, the higher volatility does not offset the higher return, since the expected Sharpe Ratio is always higher for the "bottom" portfolios. However, we cannot draw conclusions from the Sharpe Ratio in turbulent market phases, because a negative Sharpe Ratio does not convey any useful meaning. The previously stated observations are reversed for the ESG Controversies Score.

In turbulent times, the higher order moments give mixed results. The "bottom" portfolios exhibit less negative skewness, but positive kurtosis, hinting at rather big losses and more extreme events than for normally distributed returns, whereas the "top" portfolios exhibit more negative skewness and negative kurtosis, hinting at even more, bigger losses but less extreme events. In calm times, the "top" portfolios display better values for the higher order moments, since they have a more positive skewness and positive kurtosis.
To conclude, the findings and conclusions drawn for the sustainability portfolios in the overall time period (Table 4.11) mainly translate to the trends seen when splitting into calm and turbulent market phases.

Turbulent Market			Mean Return	Mean Return p.a.	Return Std.	Return Std. p.a.	$\mathbb{E}$[SR]	Skewness	Kurtosis	$VaR_{0.95}$	$CVaR_{0.95}$
Market			-0.32%	-3.77%	6.79%	23.00%	-0.06	-0.03	-0.57	-9.35%	-12.47%
ESG Score	Level	Bottom	-0.00%	-0.01%	7.81%	27.51%	-0.01	-0.27	0.14	-10.26%	-16.78%
		Top	-0.40%	-4.66%	7.13%	23.97%	-0.07	-0.28	-0.13	-11.25%	-15.92%
	Change	Bottom	-0.24%	-2.86%	7.86%	26.99%	-0.04	-0.14	0.04	-11.99%	-16.75%
		Top	-0.20%	-2.35%	7.25%	24.93%	-0.04	-0.27	-0.17	-10.03%	-15.75%
E Score	Level	Bottom	0.01%	0.13%	7.73%	27.26%	-0.01	-0.22	0.04	-11.05%	-16.57%
		Top	-0.39%	-4.62%	7.00%	23.54%	-0.07	-0.26	-0.15	-10.37%	-15.38%
	Change	Bottom	-0.25%	-3.00%	8.03%	27.54%	-0.04	-0.31	0.48	-10.49%	-18.19%
		Top	-0.32%	-3.74%	7.61%	25.86%	-0.05	-0.31	-0.07	-12.70%	-17.20%
S Score	Level	Bottom	-0.09%	-1.12%	7.75%	27.01%	-0.02	-0.21	0.16	-10.03%	-16.73%
		Top	-0.37%	-4.39%	7.28%	24.56%	-0.06	-0.14	-0.21	-11.04%	-15.39%
	Change	Bottom	-0.23%	-2.75%	7.88%	27.09%	-0.04	-0.25	-0.02	-12.47%	-17.31%
		Top	-0.29%	-3.47%	7.54%	25.68%	-0.05	-0.34	-0.08	-10.98%	-16.75%
G Score	Level	Bottom	-0.03%	-0.31%	7.79%	27.35%	-0.01	-0.16	-0.12	-10.75%	-15.76%
		Top	-0.44%	-5.16%	7.37%	24.68%	-0.07	-0.27	0.16	-12.09%	-17.04%
	Change	Bottom	-0.19%	-2.24%	7.73%	26.67%	-0.03	-0.17	-0.14	-11.35%	-16.23%
		Top	-0.31%	-3.67%	7.52%	25.58%	-0.05	-0.28	-0.11	-10.47%	-16.60%
Controversies Score	Level	Bottom	-0.49%	-5.70%	7.33%	24.43%	-0.08	-0.20	-0.18	-11.09%	-15.75%
		Top	-0.11%	-1.34%	7.66%	26.64%	-0.02	-0.28	0.08	-10.79%	-16.86%
	Change	Bottom	-0.39%	-4.54%	7.32%	24.67%	-0.06	-0.20	-0.30	-10.71%	-15.65%
		Top	-0.43%	-5.03%	7.89%	26.54%	-0.06	-0.09	0.19	-12.32%	-16.94%
ESG Combined Score	Level	Bottom	-0.03%	-0.39%	7.72%	27.08%	-0.01	-0.21	-0.04	-10.20%	-16.08%
		Top	-0.34%	-4.03%	7.35%	24.90%	-0.06	-0.23	-0.11	-11.31%	-16.22%
	Change	Bottom	-0.21%	-2.53%	7.71%	26.54%	-0.04	-0.23	-0.12	-11.35%	-16.48%
		Top	-0.24%	-2.85%	7.39%	25.31%	-0.04	-0.22	-0.12	-10.71%	-16.11%

Table 4.12: Properties of Monthly Portfolio Returns - Turbulent Market

Calm Market			Mean Return	Mean Return p.a.	Return Std.	Return Std. p.a.	$\mathbb{E}$[SR]	Skewness	Kurtosis	$VaR_{0.95}$	$CVaR_{0.95}$
Market			1.42%	18.50%	2.47%	10.00%	0.52	-0.07	-0.40	-2.71%	-3.40%
ESG Score	Level	Bottom	1.84%	24.52%	2.70%	11.46%	0.64	0.13	-0.29	-2.68%	-3.28%
		Top	1.45%	18.80%	2.52%	10.24%	0.53	0.17	0.17	-2.65%	-3.48%
	Change	Bottom	1.53%	20.03%	2.61%	10.70%	0.54	0.07	-0.06	-2.36%	-3.41%
		Top	1.47%	19.11%	2.42%	9.87%	0.55	0.13	-0.15	-2.39%	-3.20%
E Score	Level	Bottom	1.81%	24.00%	2.70%	11.43%	0.62	0.14	-0.35	-2.63%	-3.22%
		Top	1.41%	18.36%	2.50%	10.12%	0.52	0.22	0.18	-2.52%	-3.35%
	Change	Bottom	1.48%	19.28%	2.64%	10.79%	0.51	0.15	-0.12	-2.54%	-3.63%
		Top	1.38%	17.85%	2.41%	9.71%	0.52	0.16	0.00	-2.65%	-3.25%
S Score	Level	Bottom	1.80%	23.86%	2.67%	11.27%	0.63	0.22	0.05	-2.71%	-3.34%
		Top	1.44%	18.78%	2.56%	10.40%	0.52	0.21	0.07	-2.67%	-3.43%
	Change	Bottom	1.47%	19.16%	2.56%	10.45%	0.53	0.14	-0.24	-2.60%	-3.30%
		Top	1.43%	18.51%	2.50%	10.15%	0.52	0.15	0.03	-2.96%	-3.39%
G Score	Level	Bottom	1.68%	22.10%	2.67%	11.13%	0.58	0.18	-0.16	-2.80%	-3.40%
		Top	1.56%	20.38%	2.55%	10.48%	0.56	0.24	0.05	-2.66%	-3.19%
	Change	Bottom	1.50%	19.55%	2.64%	10.77%	0.52	0.14	0.24	-2.77%	-3.71%
		Top	1.43%	18.62%	2.51%	10.18%	0.52	0.07	-0.14	-2.70%	-3.44%
Controversies Score	Level	Bottom	1.54%	20.13%	2.54%	10.41%	0.56	0.21	-0.12	-2.53%	-3.20%
		Top	1.65%	21.70%	2.65%	11.01%	0.58	0.21	0.03	-2.73%	-3.37%
	Change	Bottom	1.41%	18.23%	2.57%	10.41%	0.50	0.12	-0.09	-2.74%	-3.54%
		Top	1.52%	19.87%	2.58%	10.57%	0.54	0.12	0.02	-2.35%	-3.41%
ESG Combined Score	Level	Bottom	1.82%	24.23%	2.73%	11.55%	0.62	0.09	-0.41	-2.73%	-3.38%
		Top	1.51%	19.72%	2.52%	10.33%	0.55	0.28	0.19	-2.51%	-3.14%
	Change	Bottom	1.47%	19.12%	2.55%	10.40%	0.53	0.04	-0.15	-2.39%	-3.44 %
		Top	1.46%	18.97%	2.50%	10.16%	0.53	0.19	0.04	-2.44%	-3.23%

Table 4.13: Properties of Monthly Portfolio Returns - Calm Market

In order to clean the performances, i.e. the monthly portfolio returns, for known market anomalies to pave the way for assessing the inherent influence of the companies' sustainability activities in the respective portfolios, we now perform a regression on the Fama

French 5 factors (see equation 4.5). Table 4.14 below lists the loadings with their respective significance level (detailed regression statistics can be found in *appendix.pdf* in the public Git repository https://github.com/VanessaTheel/master_thesis.git.

We see that the level "bottom" portfolios based on the ESG Score, Environment Score, Social Score and ESG Combined Score produce significant positive alpha. Overall, except for the Controversies Score, again, the "top" portfolios have less alpha than the "bottom" portfolios. This can be interpreted as portfolios consisting of companies with worse sustainability scores, which are based on company-reported information, have a higher abnormal return that cannot be explained by well-known market factors in contrast to portfolios consisting of companies with better sustainability scores. Additionally, the loading on the market factor *Mkt-RF* is always significant and higher for the "bottom" portfolios. Since the market factor is almost always bigger than 1, this means that the sustainability portfolios move overproportionately with the market. Furthermore, we can again observe that this behavior is switched when it comes to media-reported sustainability information: the "top" portfolios of the Controversies Score perform better than their "bottom" counterparts and are more driven by the market.

Looking at the other factors, we observe the trend that "bottom" portfolios have a significant and more positive loading on the size factor *SMB*, which can mean that these portfolios tend rather to consist of small-cap stocks. Again, this is the other way around for the Controversies Score. This is in line with what we have seen in Table 4.9. All portfolios have positive loadings on the value factor *HML*, the majority being highly significant. There is no clear indication, whether the "top" or "bottom" portfolios have a bigger trend towards value-stocks. However, for portfolios constructed on score levels, the majority of the "bottom" portfolios have a more positive loading on the value factor than their "top" counterpart, whereas it is vice versa for the portfolios based on score changes. The profitability factor *RMW* and the investment factor *CMA* are only sparsely significant. A clear trend, however, can be seen for the "bottom" portfolios constructed using the level of company-reported sustainability scores: they have a significantly negative loading on the investment factor, giving rise to the interpretation that they are biased towards high investment stocks.

Overall Market			Alpha	Alpha p.a.	Mkt-RF	SMB	HML	RMW	CMA
ESG Score	Level	Bottom	***0.25%	***3.08%	***1.02	***0.22	0.04	0.05	***-0.22
		Top	-0.05%	-0.54%	***0.99	-0.03	***0.14	0.03	0.06
	Change	Bottom	0.09%	1.08%	***1.02	***0.11	***0.18	0.05	*-0.13
		Top	0.07%	0.81%	***0.99	***0.09	*0.06	0.03	-0.04
E Score	Level	Bottom	***0.30%	***3.66%	***0.99	***0.19	***0.13	-0.04	***-0.22
		Top	-0.08%	-1.00%	***0.98	0.00	***0.11	**0.10	0.04
	Change	Bottom	0.02%	0.29%	***1.04	***0.13	***0.22	***0.16	-0.07
		Top	-0.03%	-0.42%	***1.01	*0.08	**0.09	0.01	-0.03
S Score	Level	Bottom	**0.19%	**2.36%	***1.01	***0.24	0.05	0.07	**-0.14
		Top	-0.05%	-0.64%	***1.00	-0.01	***0.13	-0.00	0.03
	Change	Bottom	0.03%	0.33%	***1.04	***0.11	***0.15	0.08	-0.07
		Top	0.01%	0.17%	***1.00	**0.09	***0.13	0.03	-0.07
G Score	Level	Bottom	0.14%	1.71%	***1.03	***0.13	0.04	-0.01	**-0.16
		Top	-0.01%	-0.10%	***0.99	**0.08	***0.17	** 0.11	0.04
	Change	Bottom	0.07%	0.85%	***1.02	***0.13	***0.12	0.04	-0.07
		Top	-0.01%	-0.11%	***1.02	***0.09	***0.09	0.04	-0.01
Controversies Score	Level	Bottom	-0.02%	-0.18%	***1.01	-0.01	***0.14	-0.02	-0.03
		Top	0.11%	1.33%	***1.01	***0.17	***0.09	0.05	-0.07
	Change	Bottom	-0.00%	-0.06%	***1.00	0.02	***0.13	-0.03	-0.03
		Top	0.00%	0.03%	***1.06	0.05	***0.12	0.02	0.06
ESG Combined Score	Level	Bottom	**0.22%	**2.64%	***1.03	***0.19	0.03	0.07	***-0.17
		Top	-0.01%	-0.13%	***1.00	0.04	***0.12	0.03	0.03
	Change	Bottom	0.06%	0.75%	***1.02	**0.08	***0.15	0.02	*-0.11
		Top	0.03%	0.40%	***1.01	***0.09	** 0.06	0.03	-0.01

The above table shows results for (annualized) alphas and factor loadings and the corresponding significance level ($* \leqslant 0.05, ** \leqslant 0.01, *** \leqslant 0.001$) when monthly returns of the denoted portfolios are cleaned for Fama-French 5 Factors.

Table 4.14: Regression Results - Monthly Returns on Fama French 5 Factors

Below, Table 4.15 and 4.16 list the loadings with their respective significance level for the more granular analysis in turbulent and calm market phases, respectively (again, detailed regression statistics can be found in *appendix.pdf* in the public Git repository https://github.com/VanessaTheel/master_thesis.git). We see that, in contrast to the regression results of the complete time period's returns, in times of crises only the Environment Score "bottom" level portfolio and the Governance Score "bottom" level/change portfolios exhibit significant positive alpha. Here, especially the former yields overperformance which cannot be explained by other market factors. In calm market phases, however, the picture is similar to the one obtained from looking at the whole time frame and our previous analyses. Here, all "bottom" level portfolios of sustainability scores based on company-reported information produce significant positive alpha, whereas it is the "top" portfolios for the Controversies Score.

The market factor is highly significant throughout, where the loadings are mostly around 1 in turbulent market phases and smaller than 1 in calm phases. Again, we can observe that it is mostly higher for the "bottom" portfolios. Looking at the other factors, we notice that the trend we saw for the size factor can also be observed when splitting the analysis, but is more clear in times of crises. The loadings on the value factor become more extreme in turbulent times, with no clear indication, whether the "top" or "bottom" portfolios have a bigger trend towards value-stocks, but are only sparsely significant in calm market phases with lower, sometimes negative (especially for the "top" portfolios) loadings. This could indicate that in normal market circumstances the "top" sustainability level portfolios except for the Governance Score behave more like growth stocks. The profitability factor is only sparsely significant in both phases while the investment factor exhibits noticeably different results when comparing the two states: in calm times we see that the "top" portfolios have a significant positive loading giving rise to the interpretation that they are biased towards conservative investment stocks. This does not contradict what we see in Table 4.14, but extends our insight.

Turbulent Market			Alpha	Alpha p.a.	Mkt-RF	SMB	HML	RMW	CMA
ESG Score	Level	Bottom	0.39%	4.82%	***1.01	**0.27	0.08	0.02	-0.20
		Top	0.13%	1.52%	***0.97	0.02	***0.20	-0.08	-0.02
	Change	Bottom	0.36%	4.35%	***1.00	*0.20	***0.23	-0.07	-0.11
		Top	0.09%	1.09%	***0.99	*0.16	0.05	0.09	-0.09
E Score	Level	Bottom	**0.58%	**7.25%	***0.97	**0.23	***0.23	-0.05	-0.25
		Top	0.03%	0.35%	***0.97	0.02	***0.15	-0.02	-0.00
	Change	Bottom	0.27%	3.34%	***1.02	**0.22	***0.29	0.15	-0.13
		Top	0.21%	2.52%	***1.03	0.09	**0.14	-0.13	-0.01
S Score	Level	Bottom	0.22%	2.66%	***0.99	***0.35	0.06	0.12	-0.18
		Top	0.18%	2.20%	***0.99	0.05	***0.17	*-0.16	0.00
	Change	Bottom	0.33%	3.99%	***1.03	*0.18	***0.22	-0.02	-0.04
		Top	0.23%	2.75%	***1.01	0.11	***0.18	-0.05	-0.06
G Score	Level	Bottom	*0.44%	*5.39%	***1.03	*0.18	0.08	-0.11	-0.13
		Top	-0.04%	-0.53%	***0.98	*0.17	***0.18	0.13	0.00
	Change	Bottom	*0.36%	*4.37%	***1.01	*0.17	***0.17	-0.09	-0.08
		Top	0.09%	1.07%	***1.00	**0.20	*0.10	0.02	-0.07
Controversies Score	Level	Bottom	0.06%	0.75%	***1.00	0.03	***0.16	*-0.16	-0.01
		Top	0.33%	4.01%	***1.00	***0.24	*0.14	0.02	-0.09
	Change	Bottom	0.15%	1.76%	***0.98	0.06	***0.17	-0.12	-0.11
		Top	0.13%	1.62%	***1.05	0.16	*0.12	-0.17	0.13
ESG Combined Score	Level	Bottom	0.31%	3.75%	***1.04	**0.20	0.06	0.06	-0.12
		Top	0.19%	2.30%	***0.98	*0.13	***0.16	-0.11	0.02
	Change	Bottom	0.36%	4.35%	***1.01	0.13	***0.21	-0.08	-0.12
		Top	0.11%	1.33%	***1.01	*0.14	0.08	0.03	-0.05

The above table shows results for (annualized) alphas and factor loadings and the corresponding significance level ($* \leqslant 0.05, ** \leqslant 0.01, *** \leqslant 0.001$) when monthly returns of the denoted portfolios are cleaned for Fama-French 5 Factors.

Table 4.15: Regression Results - Monthly Returns on Fama French 5 Factors - Turbulent Market

Calm Market			Alpha	Alpha p.a.	Mkt-RF	SMB	HML	RMW	CMA
ESG Score	Level	Bottom	***0.30%	***3.62%	***0.99	***0.18	-0.05	0.03	-0.11
		Top	0.02%	0.25%	***0.95	*-0.06	0.00	*0.09	***0.27
	Change	Bottom	0.12%	1.46%	***0.98	0.06	0.04	0.08	0.04
		Top	*0.15%	*1.86%	***0.92	0.05	0.01	-0.02	0.07
E Score	Level	Bottom	***0.27%	***3.33%	***0.98	***0.14	-0.02	-0.07	-0.08
		Top	-0.04%	-0.43%	***0.96	0.00	0.02	***0.16	***0.18
	Change	Bottom	0.08%	0.91%	***0.98	0.06	0.06	*0.11	*0.15
		Top	0.10%	1.20%	***0.89	*0.07	-0.04	0.06	**0.18
S Score	Level	Bottom	***0.27%	***3.29%	***0.97	***0.17	-0.03	0.03	-0.03
		Top	-0.02%	-0.22%	***0.97	*-0.05	0.03	0.06	***0.21
	Change	Bottom	0.08%	0.91%	***0.98	*0.06	-0.00	*0.09	0.11
		Top	0.07%	0.84%	***0.94	*0.07	0.03	0.05	0.07
G Score	Level	Bottom	*0.15%	*1.85%	***0.99	**0.09	-0.06	-0.01	-0.04
		Top	0.10%	1.24%	***0.95	0.02	**0.08	*0.08	***0.17
	Change	Bottom	0.08%	0.95%	***0.99	**0.10	0.01	0.08	0.08
		Top	0.07%	0.89%	***0.96	0.03	-0.01	0.02	***0.16
Controversies Score	Level	Bottom	0.08%	0.98%	***0.97	-0.03	*0.06	0.04	*0.11
		Top	*0.13%	*1.61%	***0.97	***0.12	-0.01	0.02	0.07
	Change	Bottom	0.04%	0.43%	***0.97	0.00	0.03	0.02	*0.15
		Top	*0.16%	*1.61%	***0.97	-0.00	-0.00	0.09	***0.24
ESG Combined Score	Level	Bottom	***0.25%	***3.07%	***1.00	***0.18	-0.03	0.04	*-0.13
		Top	0.08%	1.00%	***0.94	-0.02	-0.02	0.08	***0.25
	Change	Bottom	0.09%	1.05%	***0.97	0.04	0.03	0.05	0.06
		Top	0.12%	1.44%	***0.94	0.05	-0.02	0.00	*0.13

The above table shows results for (annualized) alphas and factor loadings and the corresponding significance level ($* \leqslant 0.05, ** \leqslant 0.01, *** \leqslant 0.001$) when monthly returns of the denoted portfolios are cleaned for Fama-French 5 Factors.

Table 4.16: Regression Results - Monthly Returns on Fama French 5 Factors - Calm Market

Overall, these results suggest that portfolios leveraging high public ESG sentiment picked up by the media (as reported in the Controversies Score) only yield significant excess return in calm market phases. On the contrary, portfolios built on low sustainability ratings based on company reportings (as reported by the ESG Score and its Pillar Scores) seem to yield a significant overperformance, both over the market and portfolios built on high sustainability ratings in calm market phases and overall. All portfolio returns can be explained very well with known market factors, since the regressions have an adjusted R^2 of over 0.9 each. However, the before mentioned overperformance is not accounted for by any factors.

In conclusion, what we have seen in this section could mean that the market allocates a higher risk-premium for companies with low sustainability activities in order to compensate for the hazards these companies face and inherently possess. In the past, they succeeded, hence providing high excess returns. However, with the ever-growing awareness on sustainability in society and the expanding threat of climate change, this might shift in the future. Having seen that the Controversies Score has a significant positive influence on market valuation on a company-level in the previous section, which cannot be confirmed in this single-score portfolio analysis, we will look at interactions in the following section and build multiple-score portfolios.

Multiple-Score Portfolios

In this second analysis, we aim to present and compare the interaction of the ESG Score and Controversies Score by constructing portfolios based on their combinations. In detail, we look at a portfolio created by clustering companies with respect to the level of overall ESG Score, which is based on company-reported details, and the ESG Controversies Score, which is based on media-reported details. Here, we cluster companies with a good (bad) ESG Score and a good (bad) Controversies Score in the ”top-top” (”bottom-bottom”) portfolio. With this approach we want to further break down the drivers of the previously seen portfolios and pave the way for creating an ESG Factor in the following section.

Similar to the previous section, Tables 4.17 and 4.18 give an overview of the different portfolio constituents’ characteristics. We can see that the majority of companies in the ESG Score level ”bottom” portfolio have a Controversies Score of 100 (i.e. the highest possible score translating to no media controversies about the company) and the average number of companies in the ”bottom-bottom” portfolio is significantly lower than the number of companies in the ”bottom-top” portfolio. This ratio seems more even for the ESG Score level top portfolio. Furthermore, we can again observe that the ”top-bottom” portfolio consists of the companies with the highest market capitalization. We will also see this in our return analysis when we examine the loadings on the Fama French 5 Factors.

		∅ ESG Score	∅ E Score	∅ S Score	∅ G Score	∅ Controversies Score	∅ ESG Combined Score	∅ Market Cap	∅#Companies
ESG Score Bottom	**Controversies Score Bottom**	28.3	12.6	31.8	36.0	56.0	26.4	29497.4	27.5
	Controversies Score Top	27.0	10.4	30.2	37.0	100.0	27.0	11197.7	102.4
ESG Score Top	**Controversies Score Bottom**	68.8	62.7	71.0	68.8	47.6	55.1	79093.7	74.1
	Controversies Score Top	65.9	59.1	66.8	68.3	99.9	65.9	26641.7	56.7

Table 4.17: Statistics of Portfolios - Mean

		∅Std. ESG Score	∅ Std. E Score	∅ Std. S Score	∅ Std. G Score	∅ Std. Controversies Score	∅ Std. ESG Combined Score	∅ Std. Market Cap
ESG Score Bottom	**Controversies Score Bottom**	7.7	12.9	11.9	17.1	26.9	7.8	39803.9
	Controversies Score Top	7.6	11.8	11.4	17.6	0.1	7.6	10252.6
ESG Score Top	**Controversies Score Bottom**	8.9	17.4	13.0	16.6	30.8	13.5	91467.8
	Controversies Score Top	7.5	17.2	12.4	15.6	0.2	7.5	27572.8

Table 4.18: Statistics of Portfolios - Average Standard Deviation

Figure 4.10 shows the portfolio performance of the strategies based on the different sustainability scores, where we compare the ”bottom-bottom” and ”bottom-top” portfolios with the market (Fama French Market Return) as a benchmark portfolio on the left and the ”top-bottom” and ”top-top” portfolios with the market (Fama French Market Return) as a benchmark portfolio on the right.

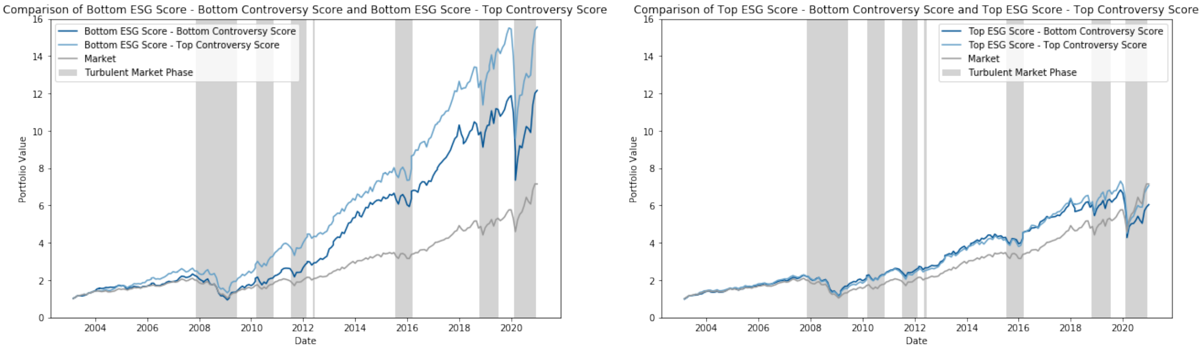

Figure 4.10: Portfolio Performance - ESG Score Level and Controversies Score Level

The first observation we want to make is that splitting the portfolio consisting of companies with a low sustainability score originating from company-reports into two portfolios by further distinguishing between those that have a relatively bad vs. good sustainability score originating from media-reports (i.e. the left part in Figure 4.10) yields two rather different performances. Comparing the picture drawn here with the portfolio simply based on ”bad” ESG Score levels, one notices that it is the companies with a rather ”good” Controversies Score that generate the overperformance. This finding is corroborated by the portfolio metrics of the overall time series depicted in Table 4.19 and of the split into turbulent and calm market phases in Tables 4.12, 4.13. Comparing the statistics of the ”bottom-bottom” and ”bottom-top” portfolios, we see a clear winner: the mean return of the ”bottom-top” portfolio is comparable or even higher while also having a lower volatility, resulting in a significantly higher expected Sharpe ratio. This is not only true in calm market phases, but especially in times of crises. Furthermore, it has a lower (Conditional) Value-at-Risk. Overall, the ”bottom-top” portfolio yields very similar and even slightly better results than the overall ESG Score level ”bottom” portfolio (compare Tables 4.11 - 4.13), making it the best performing sustainability portfolio with respect to expected (risk-adjusted) return.

Overall Market		Mean Return	Mean Return p.a.	Return Std.	Return Std. p.a.	$\mathbb{E}$[SR]	Skewness	Kurtosis	$VaR_{0.95}$	$CVaR_{0.95}$
Market		0.89%	11.24%	4.33%	16.63%	0.18	-0.59	1.87	-7.58%	-9.69%
ESG Score Bottom	Controversies Score Bottom	1.22%	15.65%	5.19%	20.68%	0.22	-0.66	2.62	-7.83%	-12.42%
	Controversies Score Top	1.31%	16.86%	4.79%	19.24%	0.25	-0.88	4.32	-6.72%	-11.36%
ESG Score Top	Controversies Score Bottom	0.88%	11.05%	4.40%	16.86%	0.18	-0.83	3.29	-6.63%	-10.78%
	Controversies Score Top	0.96%	12.08%	4.61%	17.83%	0.19	-0.97	3.73	-7.11%	-11.41%

Table 4.19: Properties of Monthly Portfolio Returns

Turbulent Market		Mean Return	Mean Return p.a.	Return Std.	Return Std. p.a.	$\mathbb{E}$[SR]	Skewness	Kurtosis	$VaR_{0.95}$	$CVaR_{0.95}$
Market		-0.32%	-3.77%	6.79%	23.00%	-0.06	-0.03	-0.57	-9.35%	-12.47%
ESG Score Bottom	Controversies Score Bottom	-0.40%	-4.73%	8.21%	27.73%	-0.06	-0.08	-0.31	-12.54%	-17.10%
	Controversies Score Top	0.10%	1.18%	7.72%	27.48%	0.00	-0.34	0.34	-10.10%	-16.84%
ESG Score Top	Controversies Score Bottom	-0.45%	-5.32%	6.99%	23.35%	-0.08	-0.24	-0.08	-10.94%	-15.54%
	Controversies Score Top	-0.34%	-3.97%	7.43%	25.19%	-0.06	-0.37	-0.05	-12.97%	-17.04%

Table 4.20: Properties of Monthly Portfolio Returns - Turbulent Market

Calm Market		Mean Return	Mean Return p.a.	Return Std.	Return Std. p.a.	$\mathbb{E}$[SR]	Skewness	Kurtosis	$VaR_{0.95}$	$CVaR_{0.95}$
Market		1.42%	18.50%	2.47%	10.00%	0.52	-0.07	-0.40	-2.71%	-3.40%
ESG Score Bottom	Controversies Score Bottom	1.88%	25.01%	3.04%	12.95%	0.58	0.11	0.19	-2.94%	-4.20%
	Controversies Score Top	1.80%	23.84%	2.73%	11.51%	0.61	0.16	-0.23	-2.78%	-3.36%
ESG Score Top	Controversies Score Bottom	1.42%	18.40%	2.56%	10.37%	0.51	0.13	0.05	-2.89%	-3.65%
	Controversies Score Top	1.48%	19.28%	2.60%	10.59%	0.52	0.17	0.03	-2.85%	-3.46%

Table 4.21: Properties of Monthly Portfolio Returns - Calm Market

Again, let us clean the performances, i.e. the monthly portfolio returns, for known market anomalies to pave the way for assessing the inherent influence of the companies' sustainability activities in the respective portfolios (see equation 4.5). Table 4.22 below lists the loadings with their respective significance levels for the overall market and Tables 4.23, 4.24 list the loadings for the turbulent and calm market, respectively (detailed regression statistics can be found in *appendix.pdf* in the public Git repository https://github.com/VanessaTheel/master_thesis.git).

We see that the "bottom-top" portfolio is the only one to produce significant positive alpha in the overall market, times of crises and calm market phases. Especially in times of crises, this portfolio exhibits very high alpha. This can be interpreted as a sign that it is actually the companies with a relatively bad ESG Score but good ESG media reports which have a high abnormal return that cannot be explained by well-known market factors and perform well in turbulent times. Again, the loading on the market factor *Mkt-RF* is always significant but there is no clear direction to its magnitude. Since the market factor for the "bottom-top" portfolio is approximately equal to 1, this portfolio moves with the market. Looking at the other factors, we observe the trend that "bottom" Controversies Score portfolios have a significantly lower loading on the size factor *SMB*. All portfolios have positive loadings on the value factor *HML* when regressing on the whole return time series and in times of crises, but they are only significant for the ESG Score ”top” portfolios. Furthermore, there is an indication that the ”bottom” Controversies Score portfolios have a bigger trend towards value-stocks. In calm market phases, the loadings on the value factor are mostly negative, a phenomenon which could also be seen in the previous section. The profitability factor *RMW* is not significant for any of the portfolios under none of the analyzed circumstances. However, the loading on the investment factor *CMA* again becomes highly significant in a calm market for the ESG Score ”top” portfolios, but no clear trend can be observed.

Overall Market		Alpha	Alpha p.a.	Mkt-RF	SMB	HML	RMW	CMA
ESG Score Bottom	**Controversies Score Bottom**	0.14%	1.73%	***1.08	0.10	0.07	0.00	-0.15
	Controversies Score Top	***0.25%	***3.09%	***1.00	***0.25	0.03	0.07	***-0.21
ESG Score Top	**Controversies Score Bottom**	-0.05%	-0.57%	***0.97	**-0.07	***0.18	0.02	0.06
	Controversies Score Top	-0.04%	-0.51%	***1.00	0.05	*0.08	0.01	0.04

The above table shows results for (annualized) alphas and factor loadings and the corresponding significance level ($* \leqslant 0.05, ** \leqslant 0.01, *** \leqslant 0.001$) when monthly returns of the denoted portfolios are cleaned for Fama-French 5 Factors.

Table 4.22: Regression Results - Monthly Returns on Fama French 5 Factors

Turbulent Market		Alpha	Alpha p.a.	Mkt-RF	SMB	HML	RMW	CMA
ESG Score Bottom	**Controversies Score Bottom**	0.15%	1.76%	***1.09	0.11	0.16	-0.10	-0.03
	Controversies Score Top	*0.45%	*5.59%	***0.99	***0.30	0.07	0.06	-0.21
ESG Score Top	**Controversies Score Bottom**	0.08%	0.96%	***0.96	-0.06	***0.24	-0.06	0.00
	Controversies Score Top	0.22%	2.62%	***1.00	0.07	*0.14	-0.19	-0.02

The above table shows results for (annualized) alphas and factor loadings and the corresponding significance level ($* \leqslant 0.05, ** \leqslant 0.01, *** \leqslant 0.001$) when monthly returns of the denoted portfolios are cleaned for Fama-French 5 Factors.

Table 4.23: Regression Results - Monthly Returns on Fama French 5 Factors - Turbulent Market

Calm Market		Alpha	Alpha p.a.	Mkt-RF	SMB	HML	RMW	CMA
ESG Score Bottom	**Controversies Score Bottom**	*0.35%	*4.32%	***1.00	0.07	-0.11	0.02	0.01
	Controversies Score Top	**0.25%	**3.04%	***0.98	***0.21	-0.05	0.05	-0.12
ESG Score Top	**Controversies Score Bottom**	0.00%	0.04%	***0.95	**-0.09	0.06	0.05	***0.22
	Controversies Score Top	0.05%	0.64%	***0.93	0.03	-0.06	0.11	***0.29

The above table shows results for (annualized) alphas and factor loadings and the corresponding significance level ($* \leqslant 0.05, ** \leqslant 0.01, *** \leqslant 0.001$) when monthly returns of the denoted portfolios are cleaned for Fama-French 5 Factors.

Table 4.24: Regression Results - Monthly Returns on Fama French 5 Factors - Calm Market

In conclusion, these results suggest that portfolios built on companies with a **high public ESG sentiment** picked up by the media (as reported in the Controversies Score) and a **low ESG Score**, yield significant excess return not accounted for by other well-known market anomalies. Hence, what we have seen in this section further confirms and expands what we have seen in the previous sections: on the one hand, the market allocates a higher risk-premium for companies with low sustainability activities, possibly due to hazards these companies face and inherently possess. On the other hand, this phenomenon is concentrated on those firms which also have a rather positive sentiment portrayed in society. Now, building on these findings, we present an ESG Factor constructed as a long-short portfolio and evaluate its relationship with the Fama French 5 Factors.

ESG Factor

In this section, we want to assess whether sustainability activities account for yet unexplained effects by leveraging the insight we have gained from the previous analyses. Specifically, we construct a long-short portfolio to create a negative spread on the ESG Score and positive spread on the Controversies Score:

$$r_{t,ESG} = r_{t,bottom/top} - r_{t,top/bottom}$$

i.e. we go long in a portfolio consisting of companies with a relatively low ESG Score and high Controversies Score and go short in a portfolio consisting of companies with a relatively high ESG Score and low Controversies Score.

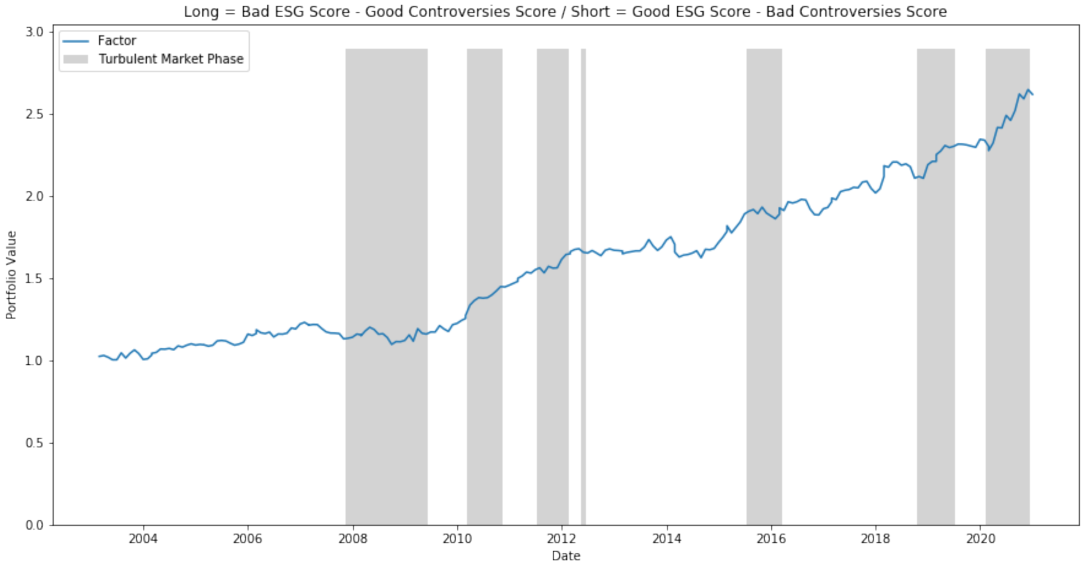

Figure 4.11: Portfolio Performance - ESG Factor

	Alpha	Alpha p.a.	Mkt-RF	SMB	HML	RMW	CMA
ESG Factor: Bottom/Top - Top/Bottom	*0.2%	*2.38%	0.03	***0.33	***-0.15	0.06	***-0.26

The above table shows results for (annualized) alphas and factor loadings and the corresponding significance level ($* \leqslant 0.05, ** \leqslant 0.01, *** \leqslant 0.001$) when monthly returns of the denoted portfolios are cleaned for Fama-French 5 Factors.

Table 4.25: Regression Results - Monthly ESG Factor Returns on Fama French 5 Factors

Above, we present the factor performance and the results when regressing the constructed ESG Factor returns on the Fama French 5 Factors (Figure 4.11, Table 4.25). We observe a significant (with $p \leqslant 0.05$) alpha of 2.38% annually, but also highly significant loadings on the size factor, value factor and investment factor. This means that the ESG factor is correlated with known market factors, but still generates a significant, unexplained overperformance. The complete regression results can be found in Table **??**.

To conclude what we have seen in this section, we will give a short summary of the main findings. Firstly, we have seen that the market allocates a higher risk-premium to companies with low sustainability activities, possibly due to hazards these companies face and cause. Secondly, in a more granular analysis, we discovered that this phenomenon is concentrated on exactly those firms which also have a rather positive sentiment portrayed in society by global media. Thirdly, by constructing an ESG Factor and studying its performance and relationship with the Fama French 5 Factors, we saw a significant outperformance when creating a negative spread on sustainability activities as reported by the company and a positive spread on sustainability activities as reported by the media. This could mean that the market overvalues bad sustainability activities in the presence of positive sentiment.

Building on our findings and the gained knowledge from investigating the annual sustainability scores provided by *Refinitiv*, we will now present and evaluate a way of leveraging state-of-the-art Natural Language Processing models to generate daily score updates for a sustainablity score based on stakeholder views as portrayed on social-media.

Chapter 5

Social Media based ESG Scoring Methodology

With an ever growing body of stakeholder data (e.g. news articles or social media posts) and NLP technology now available, there is tremendous potential to design ESG scoring methodologies which dynamically and automatically represent ESG-relevant events in near real-time. Furthermore, generating ESG-score data without manual interference is an important step towards an objective and independent assessment of firms. Treating information then follows consistent principles and becomes dramatically more efficient in terms of scoring costs. Particularly the latter aspect democratizes ESG scores beyond the current custom of a firm commissioning the scoring body to evaluate it. This chapter builds upon first attempts in research to incorporate NLP technology in ESG scoring methodology and proposes a full framework for a Tweet-based scoring method. Beyond that, we investigate the performance of the methodology in a case study and show the potential of it in terms of a dynamic, efficient, objective and independent score calculation. It is important to note that this research has been done in cooperation with Flora Geske. Hence, we refer to [Ges21] for the following chapter.

5.1 Approach

Originated in accounting, the concept of materiality was defined by the Financial Accounting Standards Board (FASB) as follows:

> *The omission or misstatement of an item in a financial report is material if, in the light of surrounding circumstances, the magnitude of the item is such that it is probable that the judgment of a reasonable person relying upon the*

> *report would have been changed or influenced by the inclusion or correction of the item. [US 99]*

Even though it began as a guiding principle for reporting [Jeb19], it has long been dominated by an internal perspective on company issues and risks. In the realm of ESG, this concept has only gained attention in the last few years. A Harvard Business School study from 2015 [KSY16] put this topic on the spot. They found that, in contrast to ESG Scores calculated based on *immaterial* issues, companies with a good ESG Score based on *material* issues significantly outperform companies with a bad ESG Score. This suggests that not all ESG issues matter equally and scores based on financially material issues are better predictors of investment return, making materiality a desirable distinction. For example, the ESG issue *Data Security* is not material for companies in the *metals & mining* industry, whereas it is highly material for companies in the *internet media & services* industry [Sus21].

With increasing influence from society represented by stakeholders, which are able to connect and voice their observations and opinions about companies online to a giant audience, the assessment of the materiality of an issue has shifted from internal to external. Stakeholders, in addition to the classical internal information given by the company itself, have become a relevant source of information for investors when evaluating an investment opportunity. Particularly with controversial issues, such as issues with relevance to ESG topics, investors can not rely solely on the company disclosing the information itself.

Based on the terminology developed by [CEG20], who link ESG performance on material issues to financial performance, we introduce the terms of **material relevance** and **material intensity**:

- **Material relevance** refers to the potential of an ESG issue to impact the ESG score.
- **Material intensity** refers to the magnitude of change of the ESG score induced by an issue.

With this, we propose the following framework for extracting changes in ESG scores from unstructured public data in the form of Tweets:

(i) **Classify** each Tweet with respect to *company*, *ESG issue* and *polarity*

(ii) **Group** by *company* and *ESG issue*

(iii) **Assign** *material relevance* and **compute** *material intensity* by aggregating the classified Tweets

5.1.1 Definitions and Notation

Definition 5.1.1. We define the set of documents (e.g. tweets) with respect to a given company f at instance t as $\mathcal{D}_{f,t} := \{d_1, \ldots, d_{N_{f,t}}\}$. In the following we consider the set of daily generated posts.

Definition 5.1.2. Let the predicted probabilities of an ESG-dimension classifier of a document d_i belonging to ESG-dimension k be given by $p_k(d_i)$, where $k = 0, 1, \ldots, N_{\text{ESG dimensions}}$ and $k = 0$ corresponds to non-ESG related.
We define the ESG-dimension of each document d_i through an indicator function that specifies where an ESG-dimension classifier predicts the greatest probability $p_k(d_i)$.

$$c_k(d_i) := \begin{cases} 1 & \text{if } p_k(d_i) > p_j(d_i) \quad \forall j = 0, \ldots, N_{\text{ESG dimensions}} \backslash k \\ 0 & \text{else} \end{cases} \tag{5.1}$$

for $i = 1, \ldots, N_{f,t}$.
In case $p_k(d_i) = p_l(d_i) > p_j(d_i)\ \forall j = 1, \ldots, N_{\text{ESG dimensions}} \backslash \{k, l\}$ and $k < l$, $c_k(d_i) = 1$ and $c_l(d_i) = 0$.

Definition 5.1.3. Let the predicted probabilities of a Polarity classifier of a document d_i being positive or negative be given by $p_{pos}(d_i)$ and $p_{neg}(d_i)$, respectively.
We define the polarity $polarity(d_i)$ of each document d_i as an indicator function given through the detected probabilities $p_{pos}(d_i), p_{neg}(d_i)$ of a polarity-classifier.

$$polarity(d_i) := \begin{cases} 1 & \text{if } p_{pos}(d_i) > p_{neg}(d_i) \\ 0 & \text{else} \end{cases} \tag{5.2}$$

for $i = 1, \ldots, N_{f,t}$.

Definition 5.1.4. We define the set of documents for a company f on a given day t regarding ESG-dimension k as $\mathcal{M}_{f,t,k}$. Furthermore, we define the set of positive and negative documents, respectively, as $\mathcal{M}^+_{f,t,k}$ and $\mathcal{M}^-_{f,t,k}$.

$$\mathcal{M}^+_{f,t,k} := \{d_i \in \mathcal{D}_{f,t} : c_k(d_i) = 1 \text{ and } polarity(d_i) = 1\} \tag{5.3}$$

$$\mathcal{M}^-_{f,t,k} := \{d_i \in \mathcal{D}_{f,t} : c_k(d_i) = 1 \text{ and } polarity(d_i) = 0\} \tag{5.4}$$

$$\mathcal{M}_{f,t,k} := \mathcal{M}^-_{f,t,k} \cup \mathcal{M}^+_{f,t,k} \tag{5.5}$$

5.1.2 Material Relevance Detection

As a first step, we discuss different approaches of defining *material relevance* of a set of documents $\mathcal{M}_{f,t,k}$. In order to detect material relevance, two approaches can be found in existing literature with the first one being based on the *SASB Materiality Map* [Sus21], which identifies industry-specific material issues and is largely incorporated in practice and academia [CEG20]. Here, a set of documents $\mathcal{M}_{f,t,k}$ would be flagged as *materially relevant* if the ESG-dimension k is considered material by the *SASB Materiality Map* for the industry to which company f belongs. For this, we define material relevance $m_{f,t}(k)$ as an indicator function for company f on a given day t regarding ESG-dimension k as the following:

$$m_{f,t}(k) = \begin{cases} 1 & \text{if company } f\text{'s industry is flagged as material for ESG-dimension } k \text{ by SASB} \\ 0 & \text{else} \end{cases} \tag{5.6}$$

However, the static fashion of the *SASB Materiality Map* and its subjective basis on industry experts calls for a more dynamic approach.

Looking for such an approach, [Tho+20] found that the majority of ESG topics discussed by the public align with the materiality issues defined by SASB, regardless of industry. Therefore, they suggest that material issues are very well reflected in the stakeholder discourse online and can be captured by merely monitoring the discourse and filtering for company and ESG topics. This yields a dynamic approach to detecting material relevance, which accounts for natural changes in material issues due to changing stakeholders' ideals and beliefs over time, in contrast to evaluating predefined, static material issues for industries, e.g. by SASB. This understanding of ”dynamic materiality” resonates well with the general notion of a stakeholder view, as stakeholders' ideals and beliefs about companies and their impact on society can guide ESG-wary investors in their decision to invest in a company, either because they fear a decrease in purchases or because of their own beliefs. Hence, a set of documents $\mathcal{M}_{f,t,k}$ could be flagged as *materially relevant* if the volume of documents regarding the ESG-dimension exceeds a company-specific, dynamic threshold $\theta_{f,t}$. For this, we define material relevance $m_{f,t}(k)$ as an indicator function for company f on a given day t regarding ESG-dimension k as the following:

$$m_{f,t}(k) = \begin{cases} 1 & \text{if } |\mathcal{M}_{f,t,k}| > \theta_{f,t} \\ 0 & \text{else} \end{cases} \tag{5.7}$$

In order to define a dynamic threshold $\theta_{f,t}$ we propose to compute the historical distribution of a company's non-ESG media coverage as a rolling window of x days backward-looking. This has the advantage of including the changing nature of stakeholders' ideals, the ever-increasing number of documents generated daily and possible shifts in focus on companies. The reason why we decide to only look at the **non-ESG** media coverage for material relevance detection is because otherwise big ESG-events skew the distribution and this might lead to non-detection of following smaller ESG-events. Following this, we propose to define the dynamic threshold as an $n - sigma$ event:

$$\theta_{f,t} = \mu(\{|\mathcal{M}_{f,t-x,0}|, \dots, |\mathcal{M}_{f,t-1,0}|\}) + n * \sigma(\{|\mathcal{M}_{f,t-x,0}|, \dots, |\mathcal{M}_{f,t-1,0}|\}) \tag{5.8}$$

where μ, σ denote the mean and standard deviation, respectively.

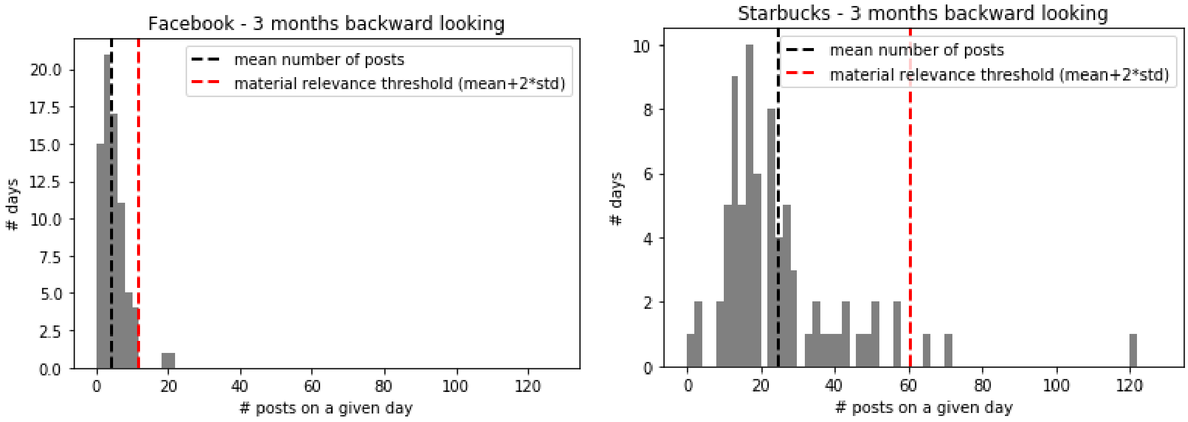

Figure 5.1: Dynamic Threshold for Material Relevance Detection - Snapshot 31.12.18

However, measuring material relevance based on frequency of stakeholder discourse makes it vulnerable to the flaws of monitoring stakeholder discourse in reality. These flaws include dealing with large amount of noisy data, misuse of social media platforms and media channels, and a general imbalance of representation of stakeholder groups on social media platforms and in the media.

A more holistic approach would incorporate a variety of additional features into the material relevance detection, which would lead to a non-binary but continuous measure for materiality linked to actual economic impact of past ESG issues in order to project, whether a current issue is materially relevant. These additional features could include information about business models and major areas of business activities of individual companies in order to detect whether a particular ESG issue hurts a companies' major profitability lever. However, since such features are not yet accessible in a well-structured manner, we propose the holistic approach for future research and adopt a dynamic understanding of material relevance detection based on document frequency.

5.1.3 Material Intensity Detection

As a next step, after classifying a set of documents $\mathcal{M}_{f,t,k}$ as *materially relevant*, we discuss different approaches of computing its *material intensity*. A first option, built on the static

SASB Materiality, is given by [CEG20], who use the number of material issues per category and industry in the map to calculate an issue- and industry-specific materiality index. However, the drawback of implementing a static materiality concept is again substantial.

As a second option, we propose to use the stakeholder discourse to measure material intensity. The intuition behind this approach is the following: The larger the volume of stakeholder discourse, the bigger the impact of an ESG issue on society. This principle holds for negative as well as positive and controversial issues. A first, naive approach would be to simply subtract the number of negative documents from the number of positive documents:

$$I_{f,t,k} = |\mathcal{M}^{+}_{f,t,k}| - |\mathcal{M}^{-}_{f,t,k}| \tag{5.9}$$

This approach could be further refined when considering the phenomenon that people tend to talk more about negative experiences and events than positive ones [RR01], by adding a *negativity shrinkage factor* λ:

$$I_{f,t,k} = |\mathcal{M}^{+}_{f,t,k}| - |\mathcal{M}^{-}_{f,t,k}| * (1 - \lambda) \tag{5.10}$$

This is a simple way of using frequency of stakeholder discourse as a proxy for material intensity while accounting for a *negativity bias*. However, this aggregation could lead to skewed effects, assuming more popular companies tend to have more media coverage (e.g. because of their products or company size). As a way of overcoming this *popularity bias*, we propose to scale the aggregated polarities by normalizing with respect to a *popularity factor* $\gamma_{f,t,k}$. This approach could define the material intensity $I_{f,t,k}$ of a materially relevant (i.e. if $m_{f,t}(c_k) = 1$) company-issue pair f, k on a given day t as:

$$I_{f,t,k} = \frac{|\mathcal{M}^{+}_{f,t,k}| - |\mathcal{M}^{-}_{f,t,k}| * (1 - \lambda)}{\gamma_{f,t}} \tag{5.11}$$

where λ = negativity shrinkage factor. One example for this *popularity factor* could be the total number of documents generated with respect to the company of interest on the given day. However, this could potentially cause unwanted effects, since the material intensity is highly dependent on the number of documents generated on that day (e.g. if another big, ESG-unrelated event is happening at the same time, the material intensity would be forced to be small). Hence, we propose to use some kind of average of the past distribution. One possibility could be the *median* of the number of documents of the past x days: $\gamma_{f,t} = median(\{|\mathcal{D}_{f,t-x}|, \ldots, |\mathcal{D}_{f,t-1}|\})$. It has the advantage of being more robust to extreme values than the mean, is dynamic and evolves over time.

The computed *material intensity* of a set of *materially relevant* documents regarding ESG dimension k for company f on day t can now be used as a proxy for the change in a (social-)media-based sustainability score for dimension k, since it represents the relative change in polarity of the stakeholders.

5.1.4 Training Data

In order to build our pipeline with the NLP model, we decide to use Twitter data, as it is readily available, plays a vital role in sharing content and public opinion among the communities and state-of-the-art, transformer-based NLP models have already been pre-trained on this type of text data. Since we want to classify tweets with respect to both *ESG dimension* and *polarity*, two models and, hence, two sets of labelled data for model training are needed. Our approach for generating labelled data as ground truth is as follows:

- We align our ESG dimensions to SASB Dimensions [Susa], namely *Environment, Social Capital, Human Capital, Business Model & Innovation* and *Leadership & Governance.*
- Our data was obtained through *Archive Team's Twitter* "Spritzer" stream, a random sample of 1% of all tweets. We use tweets from January 2019, February 2019 and January 2020, which we pre-filter with a set of keywords (see Appendix C) and *S&P500* company names. These tweets are then manually labelled and reviewed. For non-ESG relevant tweets, we randomly sample tweets from the above mentioned months and review them on exemplary basis.
- ESG relevant tweets are further manually categorized as *positive* or *negative*, depending on the writer's polarity towards the subject.
- In order to adhere to the data conventions used to train *BERTweet* [NVN20], we delete the mentioning of retweets (*RT @USER:*). Furthermore, quoted tweets are translated into <*ANSWER TWEET*> + *QUOTED:* + <*ORIGINAL TWEET*>.
- Our final data sets for the ESG-classification task and the polarity-classification task comprise 5,223 tweets and 1,610 tweets, respectively, with the following distributions:

Category	#Tweets
Environment	380
Social Capital	772
Human Capital	491
Business Model & Innovation	80
Leadership & Governance	724
None	2,776
Total	5,223

Table 5.1: ESG-Classification Task

Category	#Tweets
Negative	1,279
Positive	331
Total	1,610

Table 5.2: Polarity-Classification Task

5.1.5 Model Training

We utilize the pre-trained *BERTweet* model [NVN20] as provided by *SimpleTransformers* which is based on the Transformers library of *HuggingFace* [Wol+20] and pre-trained on a large text body of Tweets. We then fine-tune this pre-trained model on both of our tasks, resulting in two models. We optimize our models with respect to accuracy with *Bayesian Hyperparameter Optimizatio*n and fixed test sets, which are randomly drawn 30% samples from our full data sets. In detail, this yields the following hyperparameters for model training:

- **ESG-Classification Model**: We use a max sequence length of 128, a batch size of 4 with a gradient-accumulation of 20 and a learning rate of $5e^{-5}$ for 20 epochs.

- **Polarity-Classification Model**: We use a max sequence length of 128, a batch size of 2 with a gradient-accumulation of 14 and a learning rate of $2e^{-5}$ for 20 epochs.

We evaluate the performance of our final models based on the area under the ROC curve (ROC-AUC) as well as under the Precision-Recall curve (PR-AUC) to assess the model performance for each task. The table below gives an overview of the models' performance on the test set.

Category	ROC	PR
Environment	0.99	0.92
Social Capital	0.97	0.86
Human Capital	0.99	0.91
Business Model & Innovation	0.94	0.54
Leadership & Governance	0.97	0.89
None	0.99	0.99

Table 5.3: AUC: ESG-Classification

Category	ROC	PR
Negative	0.93	0.97
Positive	0.93	0.84

Table 5.4: AUC: Polarity-Classification

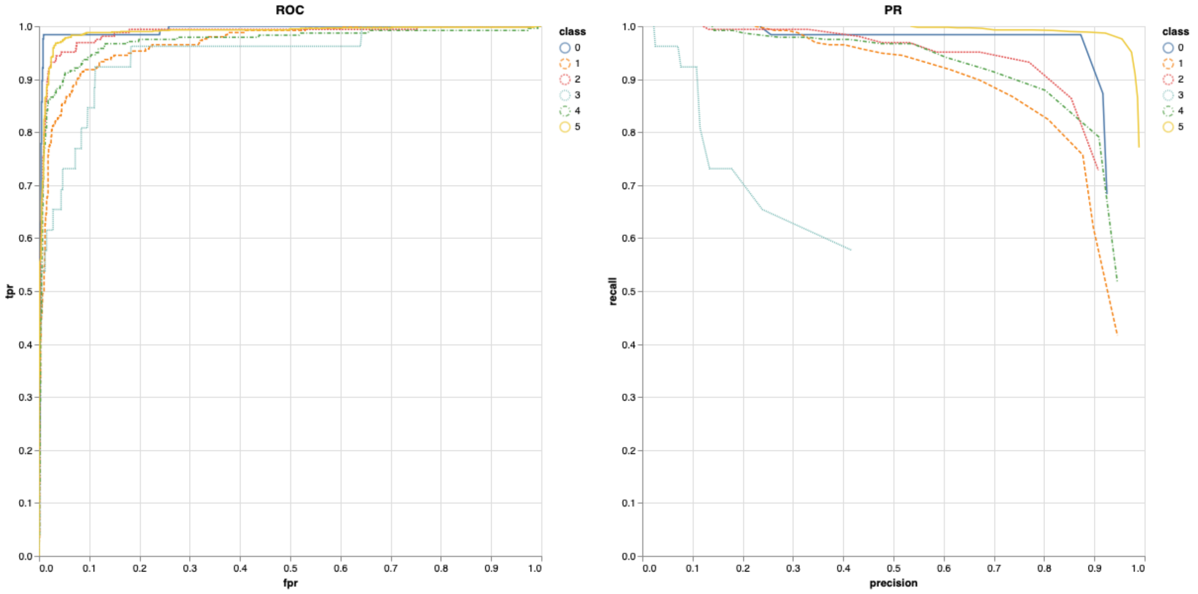

Figure 5.2: ROC Curve and Precision Recall Curve for ESG-issue Classifier.
Classes: 0 = Environment, 1 = Social Capital, 2 = Human Capital, 3 = Business Model & Innovation, 4 = Leadership & Governance, 5 = None

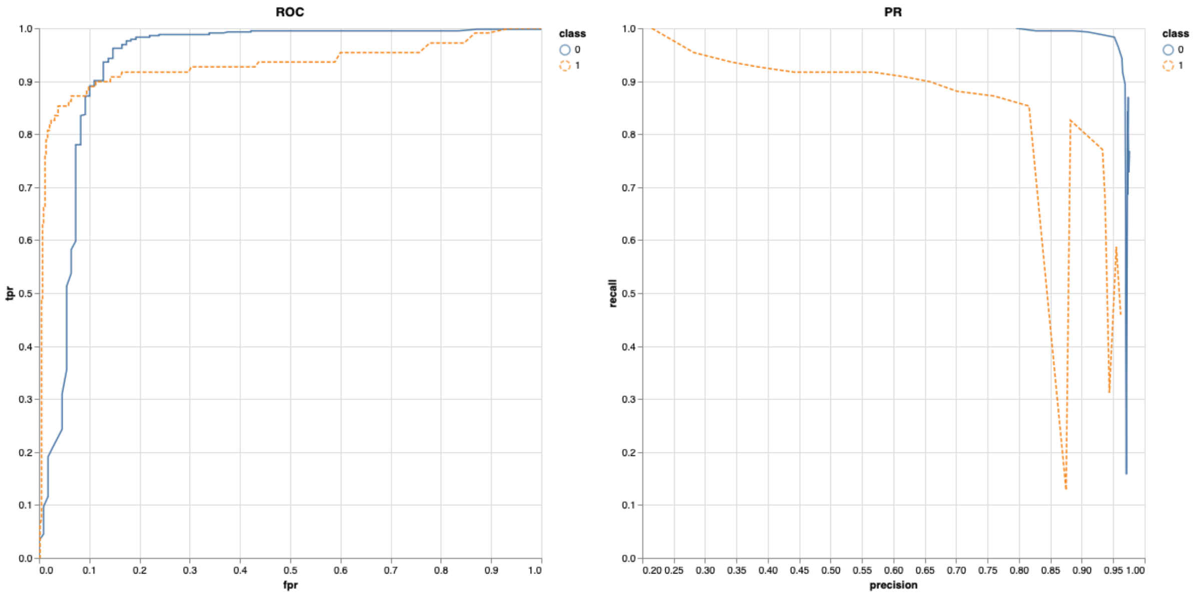

Figure 5.3: ROC Curve and Precision Recall Curve for Polarity Classifier.
Classes: 0 = Negative, 1 = Positive

A common rule of thumb for interpreting *ROC-AUC* states that scores above 0.9 are outstanding for classifiers [HLS13], which we achieve with both models across all classes. For reference, a classifier with a score of 0.5 is as good as a random decision, e.g. from a coin toss. As can be seen in Figures 5.2 and 5.3, the *ROC curves* for all classes lean very closely to the top-left corner, which represents the point where the model only predicts *true positives* and *true negatives*, i.e. makes no mistake. Furthermore, we also judge our models with respect to *Precision-Recall curves.* Tables 5.3 and 5.4 reinforce the great performance of both models with high values for the *PR-AUC.* As can be seen in Figure 5.2, except for the class *Business Model & Innovation*, where we lack training examples, we have a *precision* of over 80% at an 80% *recall.* To further ensure the generalizability of our model, we look for any warning signs of overfitting during training, e.g. a dramatically increasing evaluation error in combination with improving performance on the training data set. However, we do not find any indication of that.

These are very promising results which already showcase the potential for applying these models in production. However, we will discuss how to further improve their performance after we have evaluated the models in a case study.

5.2 Evaluation

In order to test the real-world applicability of the designed models, we classify tweets from 2018 with respect to their ESG-relevance and polarity. Here, we leverage the *Spritzer Twitter Stream Grab*, a 1%-sample of all tweets, as provided by *Archive.org Team.* We use the year 2018 for our analysis, since it provides the best-sustained archive in terms of longitudinal tweet data (still missing: *26.-30. April, 1. May, 9.-31. May, 1.-5. June, 4.-10. July, 9.-23. December*). Furthermore, we classify the last quarter of 2017 to enable a rolling window for our tweet distribution to judge *materiality* and *popularity.*
For the following case study, we proceeded as follows:

- For a case-study based model evaluation, we choose a sample of companies for which a yearly-updated score, the *Refinitiv* Controversies Score, is available. The Controversies Score is calculated in an objective and automated manner and reflects the events continuously as new information enters the media. As it resembles our approach of media monitoring to detect score changes, we use it as a benchmark for our case-study.
- We collect the official twitter handles of the companies of interest (namely *Adidas, Nike, Apple, Google, Facebook, Nestle, Starbucks, Exxon Mobil, ENI, Royal Dutch Shell, Johnson&Johnson, Boeing, H&M, Pepsi, Coca-Cola* and *PG&E*) and filter for English tweets, which feature a mention of at least one of the companies. Although this restricts our set of tweets by not including those in which the company is referenced by name or hashtag, it does ensure that the tweet is actually related to the company and not to a similar named object (e.g. apple).
- After filtering as described and cleaning for duplicates in the archive, we classify 104,731 tweets from 2018 and 28,532 from 2017 with respect to their polarity and their ESG dimension.
- We then aggregate the tweets per company, day and ESG dimension to detect events and derive a change in score with our proposed methodology.
- For the *material relevance detection*, we choose the frequency-based approach with a dynamic materiality threshold as a 2σ-event with a 3-month backward looking window. Furthermore, we decide to compute the *material intensity* through accounting for a *negativity bias* with a negativity shrinkage factor $\lambda = 0.2$ and a *popularity bias* given by the median number of documents generated per day in the last 3-months:

$$I_{f,t,k} = \frac{|\mathcal{M}^{+}_{f,t,k}| - |\mathcal{M}^{-}_{f,t,k}| * (1 - 0.2)}{median(\{|\mathcal{D}_{f,t-90}|, \dots, |\mathcal{D}_{f,t-1}|\})}$$

In this following section, we present the well-known firm Facebook as an example of the case studies and show its sustainabilty score changes based on stakeholder reports on Twitter throughout the year 2018. We present the score aggregated over all ESG categories (Figure 5.4) as well as a table with the respective material events which induced the score changes (Appendix D). We describe major events in further detail to highlight the advantages and flaws of our approach. It is important to note that the following analysis is subject to the 1% sample of Tweets provided by the *Spritzer Twitter Stream Grab*. This means, that intensities and events might not be a perfect image of the real stakeholder discourse and should be viewed as a proof-of-concept.

Looking at Facebook, an American online social media and social networking service, we can clearly see that there was a lot of ESG-related stakeholder discourse (Figure 5.4).

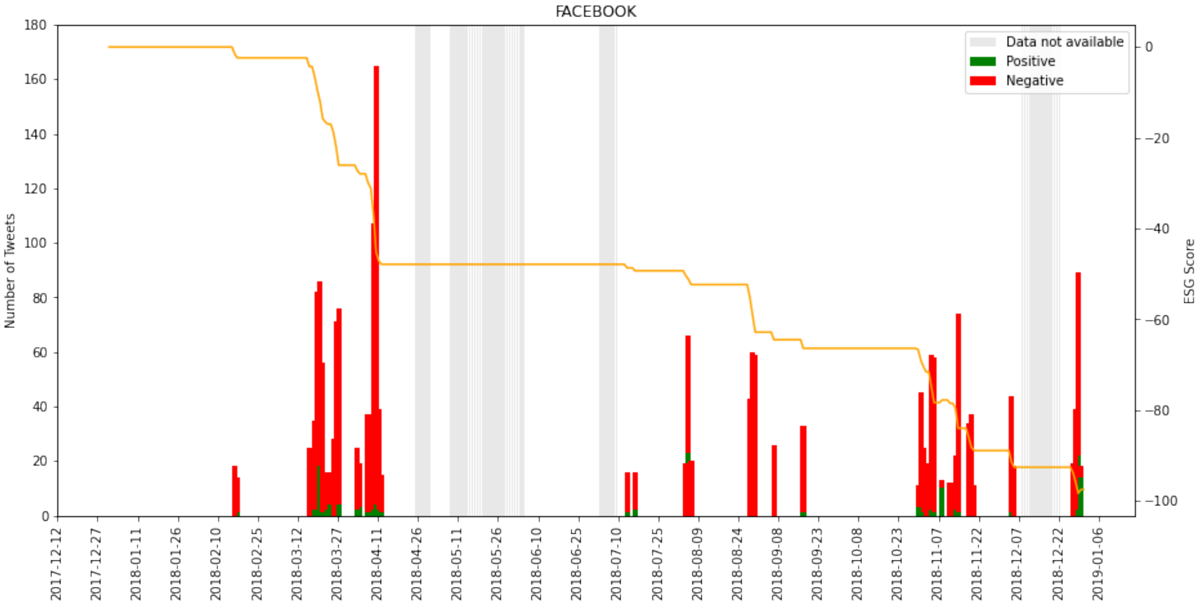

Figure 5.4: Tweet-based ESG Score

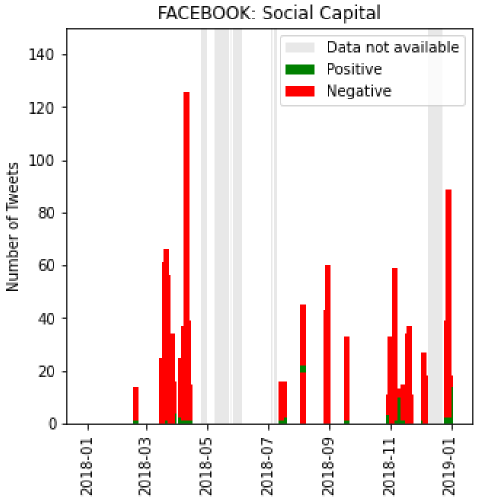

Figure 5.5: Social Capital Tweets

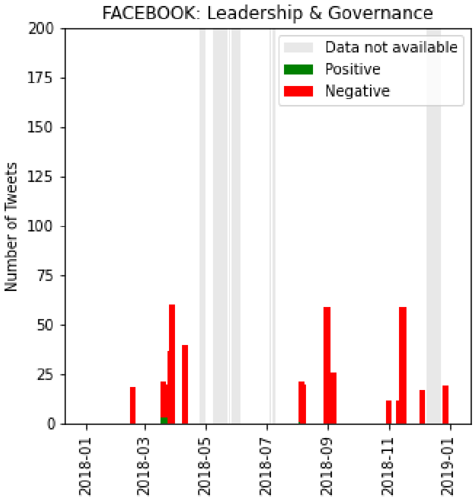

Figure 5.6: Leadership & Governance Tweets

The first event that we want to highlight is one of the biggest corporate scandals in 2018: Facebook's Cambridge Analytica affair in March (see Event #3 in Table D.1). Millions of Facebook users' personal data was harvested without the individuals' consent by Cambridge Analytica, a British data analytics firm, primarily to be used for political inference [The18b]. As a consequence, user growth slowed down leading to a loss in Facebook's market capitalization of more than $100 billion in days [The18a] and the negative public response eventually led to Facebook CEO Mark Zuckerberg agreeing to testify in front of the United States Congress [The18c]. For this event, our model detects Tweets talking about the data breach, demanding federal investigation and the connection to the U.S. election in 2016, correctly classifying it as a negative, material event in *Social Capital* since it is related to customer privacy and data security as well as socio-economic impacts. Here, the score-update happens in near real-time as the first day of the model-detected event is the same day this scandal was made public, namely the 17th of March 2018. For all that, we can also see a spike in the category *Leadership & Governance* (Figure 5.6) in the same time period. Here, however, the Tweets focus on the bigger picture of Facebook's responsibility towards the "free flow of information in our democracy" (see Event #5 in Table D.1), very well fitting towards the topic of ethical conduct of business and, hence, the ESG-dimension *Leadership & Governance*.

The second event we want to highlight (see Event #9 in Table D.2) again addresses the relationship between the business and a community in which they operate: conservatives. A big discussion was started, when the duo "Diamond and Silk", otherwise known as Lynette Hardaway and Rochelle Richardson, received a note from Facebook saying their "content and brand" were "unsafe to the community" [Tim18]. Our model detects Tweets voicing opinions about discrimination of conservative views and calls to boycott Facebook, correctly classifying it as a negative, material event in *Social Capital.* However, we can also see a spike in the category *Leadership & Governance* (see Event #10 in Table D.2), where the Tweets clearly focus on the topic of censorship and bias because, allegedly, followers were not notified of new content, limiting the spread of their posts [Was18]. Here, our model classifies Tweets, which are related to the same event, into different ESG-categories. This can be explained by the ambiguity of the topic and will be further discussed in the limitations section.

A third event detected by our model happened on the 29th of August 2018, when Facebook removed a post by the Anne Frank Center calling for the need for Holocaust Education [Bus18] (see Event #16 in Table D.4). This event mainly consists of Retweets of the official reaction tweet by the Anne Frank Center, raising the question of hypocrisy, since, allegedly, Holocaust Denial pages still exist on Facebook. As this event highlights a feeling of mistreatment of the Jewish community, we consider the classification as a negative, material event in *Social Capital* as accurate.

Overall, the events detected in the ESG-dimension *Social Capital* mostly cover the topic of communities feeling misrepresented or mistreated by Facebook's actions (e.g. racist and violent content (see Event #27 in Table D.5, D.4), women (see Event #19 in Table D.4), pro-life (see Event #21 in Table D.4)) and privacy topics (see Event #29 in Table D.5). Yet, events detected in the ESG-dimension *Leadership & Governance* tend to also cover the topic of communities feeling misrepresented or mistreated by Facebook's actions. However, they predominantly discuss a bias towards democratic views (see Events #15,17,20,23,28,30 in Table D.2,D.4,D.5).

5.3 Results

Our goal in this chapter was to introduce a new ESG scoring methodology, which yields a score fulfilling the following criteria:

- dynamic, automatically updated in near real-time
- objective, independent, and consistent assessment of firms
- efficient in its assessment process

In terms of a dynamic evaluation, we compare the 2018 scores from our case study to the Controversies Score from *Refinitiv* (see Facebook as an example in Figure 5.7). It becomes clear that the Controversies Score does not change throughout the year 2018, whereas our score has successfully detected material events in the 1% Twitter sample and has adjusted accordingly. Particularly the Cambridge Analytica Affair in March, which culminated in a congressional hearing for the CEO Mark Zuckerberg, should have been reflected in the Controversies Score. In contrast to our score behaviour, it only updates at the end of the year, also penalizing them with respect to ESG-related controversies - as our score does, but without indication of what happened throughout the year.

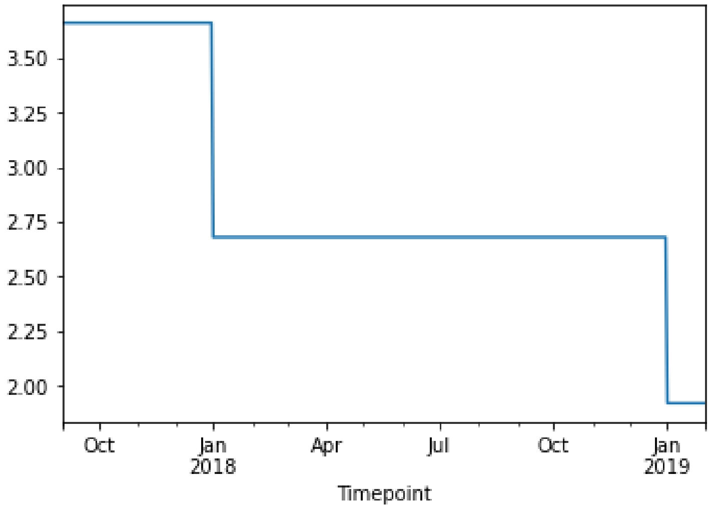

Figure 5.7: *Refinitiv*'s Controversies Score - Facebook

In terms of an objective, independent assessment of firms, we find that our NLP-powered score methodology successfully disengages the scoring body from the evaluated firm by solely relying on stakeholder observations and opinions from social media. Furthermore, information is no longer handled manually, which largely increases the degree of objective and consistent processing as the same input Tweets yield identical output score change. However, we acknowledge that full independence is not yet reached as the dependence on information from a party other than the scoring body still exists and has now moved from the evaluated company itself towards its stakeholders. Furthermore, the handling of information still depends on the manual labelling process before training the NLP models. The vulnerabilities of our approach will be discussed in detail in the next section. In terms of efficiency, we find our approach to drastically streamline the scoring process when it comes to information retrieval and evaluation. In a productive use of such a methodology, the Twitter data would be accessed through an API and would therefore minimize the manual effort to mere setup of the models in a productive environment and regular updates of the fine-tuning data body to cover new language and large societal changes (e.g. Covid-19 pandemic).

5.4 Limitations

Having brought up some of the limitations in previous sections, we now give a structured overview of their three main dimensions and discuss them in more detail:

- **Data**
  Looking back at section 5.1.5, we can see very promising results when it comes to the performance of the underlying NLP-classifiers. However, in order to successfully use such a model structure in production, the performance could be even further improved. We are confident, that by obtaining a bigger corpus of labelled data (especially for the ESG dimension *Business Model & Innovation*), the model performance could be increased even further [HNP09]. The phenomenon of Tweets from different, but similar events and Tweets from the same, ambiguous event being classified in different ESG dimensions (especially *Social Capital* and *Leadership & Governance*) could also be mitigated by more training data giving the model a clearer understanding of the subtle differences. This leads us directly to the next point - data labelling. The models are highly susceptible to the labelling process, meaning that a clearly-defined framework needs to be in place and objectivity must be ensured. Moving to a further part of the data pipeline, the documents used as input for the ESG scoring methodology in this paper are only a limited representation of the true stakeholder discourse taking place in the year 2018 as they only contain a 1% Twitter sample. Again, to be able to have a production-ready solution, the document body should be increased.

- **Score Design**
  Our score methodology clusters events as all documents related to the same ESG dimension on a given day. It follows that there is a chance that a set of documents will be falsely flagged as materially relevant because it consists of more than one event (i.e. different contexts in the same ESG-dimension on the same day, see events #4,5 in Table D.1), and the events, treated separately, would not exceed the material relevance threshold. Furthermore, the score is highly dependent on the choice of n, x (section 5.1.2) and λ (section 5.1.3). On the one hand, these parameters grant the user extensive flexibility to tailor the score to individual requirements. On the other hand, this makes it subjective with no single, optimal choice of parameters obtainable.

- **Choice of NLP models**
  The polarity classifier we use in this paper is not able to detect the polarity of the writer towards the company in the document, but rather the document's polarity in

general as expressed by the author. This leads to falsely classified Tweets, because the polarity of the document does not translate to the polarity towards the company of interest. One way of overcoming this challenge could be to leverage targeted aspect-based sentiment analysis (TABSA), which aims to identify a document's polarity towards a specific aspect (i.e. ESG dimension) associated with a given target (i.e. company) [SHQ19].

As a last point, we want to acknowledge the vulnerability towards fake news and manipulation through systematic document creation. Our model and methodology in its current form cannot account for such documents, which spread false information, and campaigns with the goal of generating documents (e.g. with bots) to skew a company's score. One could, for example, try to overcome this problem by regulating the input data (e.g. only audited news outlets) or accounting for it in the score design with a checking mechanism.

Chapter 6

Conclusion

This thesis investigates the relationship between *Refinitiv's* ESG Scores and financial performance and introduces a new approach for updating ESG Scores more frequently using stakeholder data and leveraging state-of-the-art natural language processing models.

First, we compare the explanatory power of ESG scores based on company reports and global media, respectively, with respect to the valuation multiple *Price-to-Book*. Here, we find that, while accounting for firm charcteristics, public ESG sentiment picked up by the media (as reported in the Controversies Score) has a considerable positive effect on how a company is valued in the market. On the contrary, ESG ratings based on company reportings (as reported by the ESG Score) seem to have a negative effect.

Second, we explore the connection between ESG Scores and financial perfomance through a portfolio-based approach, i.e. we build portfolios dependent on differently focused sustainability scores and construct a factor to examine their performance through return metrics and a linear regression on well known market anomalies. In these analyses, we find that the market allocates a higher risk-premium to companies with low sustainability activities, possibly due to hazards these companies face and cause. Furthermore, in a more granular analysis, we discover that this phenomenon is concentrated on exactly those firms that also have a rather positive sentiment portrayed in society by global media. Finally, by constructing an ESG Factor and studying its performance and relationship with the Fama French 5 Factors, we see a significant outperformance when creating a negative spread on sustainability activities as reported by the company and a positive spread on sustainability activities as reported by the media and conclude that the market overvalues bad sustainability activities in the presence of positive sentiment.

Building on these results and leveraging the ever growing body of stakeholder data, we propose an ESG scoring methodology based on Twitter discourse and powered by a pipeline of state-of-the-art NLP models, which dynamically and automatically represents ESG-relevant events in near real-time. In detail, we build two classification models to evaluate daily Tweets and introduce methods to detect and compute the *material relevance* and *material intensity* based on the conducted classification. We test the real-world applicability of the designed models and scoring methodology by classifying Tweets from 2018 with respect to their ESG-relevance and polarity and by performing a case study. Here, we find that our framework has successfully detected material events in a 1% Twitter sample and adjusts the score accordingly, providing an efficient, objective and independent assessment of firms.

To conclude this thesis, we want to outline some areas for further research. In order to overcome the constraint of only using Tweets with mentions of the official company Twitter-handle, a model using *named entity recognition* could be introduced to have a preceding classification for relevant company Tweets. Furthermore, *question answering* could be leveraged to more accurately read the relationship between company and ESG issue in a Tweet (e.g. "Is the author talking about ESG issue X regarding company Y?"). Finally, *targeted aspect-based sentiment analysis* [SHQ19] could be a highly promising approach to also tackle the limitation of not being able to detect the polarity of the writer towards the company-ESG issue pair in the Tweet.

Appendix A

Variables

Table A.1: List of Variables

Name	TR Eikon Code	Details
ESG Combined Score	TRESGCS	Refinitiv's ESG Combined Score is an overall company score based on the reported information in the environmental, social and corporate governance pillars (ESG Score) with an ESG Controversies overlay.
ESG Controversies Score	TRESGCCS	ESG controversies category score measures a company's exposure to environmental, social and governance controversies and Negative events reflected in global media.
ESG Score	TRESGS	Refinitiv's ESG Score is an overall company score based on the self-reported information in the environmental, social and corporate governance pillars.
E Score	ENSCORE	Refinitiv's Environment Pillar Score is the weighted average relative rating of a company based on the reported environmental information and the resulting three environmental category scores.
S Score	SOSCORE	Refinitiv's Social Pillar Score is the weighted average relative rating of a company based on the reported social information and the resulting four social category scores.
G Score	CGSCORE	Refinitiv's Governance Pillar Score is the weighted average relative rating of a company based on the reported governance information and the resulting three governance category scores.

Price-to-Book	PTBV	This is the share price divided by the book value per share.
Return on Equity (ROE)	WC08301	All Industries: (Net Income – Bottom Line - Preferred Dividend Requirement) / Average of Last Year's and Current Year's Common Equity * 100 For Insurance companies, Policyholders' Surplus is substituted where Net Income – Bottom Line is not available and Policyholders' Equity where Common Equity is not available.
Net Sales Growth	WC08631	All Industries: Annual Time Series: (Current Year's Net Sales or Revenues / Last Year's Total Net Sales or Revenues - 1) * 100 Interim Time Series: (Current Year's Trailing 12 Months Net Sales or Revenues / Last Year's Trailing 12 Months Total Net Sales or Revenues - 1) * 100 This calculation uses restated data for last year's values where available
Dividend Yield	DY	The dividend yield expresses the dividend per share as a percentage of the share price. The underlying dividend is calculated according to the same principles as datatype DPSC (Dividend per share, current rate) in that it is based on an anticipated annual dividend and excludes special or once-off dividends. Dividend yield is calculated on gross dividends (including tax credits) where available. Note that dividend yield for UK, Irish and French stocks is calculated on gross dividends (including tax credits), although dividends per share for these countries are displayed net.
Leverage (Total Debt % Common Equity)	WC08231	All Industries: (Long Term Debt + Short Term Debt & Current Portion of Long Term Debt) / Common Equity * 100 Insurance Companies: If Common Equity is not available, Policyholders Equity is substituted.

Capital Expenditures (CapEx)	DWCX	Capital Expenditures represent the funds used to acquire fixed assets other than those associated with acquisitions.
Price-Earnings Ratio (PE)	PE	This is the price divided by the earnings rate per share at the required date.
Market Value	MV	Market value on Datastream is the share price multiplied by the number of ordinary shares in issue. The amount in issue is updated whenever new tranches of stock are issued or after a capital change. § For companies with more than one class of equity capital, the market value is expressed according to the individual issue. § Market value is displayed in millions of units of local currency.

Appendix B

Regression Results - Market Valuation

B.1 Data Preparation

Originally, our data set consists of 10100 observations, i.e. annual data of 505 companies for the past 20 years gathered from *Thomson Reuters Datastream*. In detail, we look at the following variables:

- ESG Score
- Controversies Score
- Size (= ln(Market Cap)) (+,*)
- Return-on-Equity (*)
- Net Sales Change (in %)(*)
- Annual Return (discrete) (*)
- Leverage (*)
- Capital Expenditures (+,*)
- Price-Earnings Ratio (*)
- Dividend Yield (*)

since they are commonly used in valuation analyses in research ([Ser20], [TVY18], [HTW16]). In this analysis, we aim to explain the companies' Price-to-Book ratio as an indicator for market valuation.

As a first step, we scale the absolute variables (+) by market capitalization to make them more comparable. Then, we normalize all market variables (or rather their scaled equivalent) (*) by computing the respective z-score (for each year as to not introduce a forward-looking bias)

$$z(X)|_{year=t} = \frac{X - \mu(X)|_{year=t}}{\sigma(X)|_{year=t}} \tag{B.1}$$

in order to detect outliers/extreme values. Basically, the z-score is a statistical measure that indicates how far an observation is from the rest of the sample, by measuring the distance from the mean in standard deviations. In this master thesis, we choose 3.5 standard deviations to denote the arbitrary threshold for extreme values [CC10]. In detail, this means that we exclude all observations with a z-score of less than −3.5 or larger than 3.5. Figures B.1 - B.8 show the distribution of the observed values for our variables of interest before cleaning for outliers (left) and after (right).

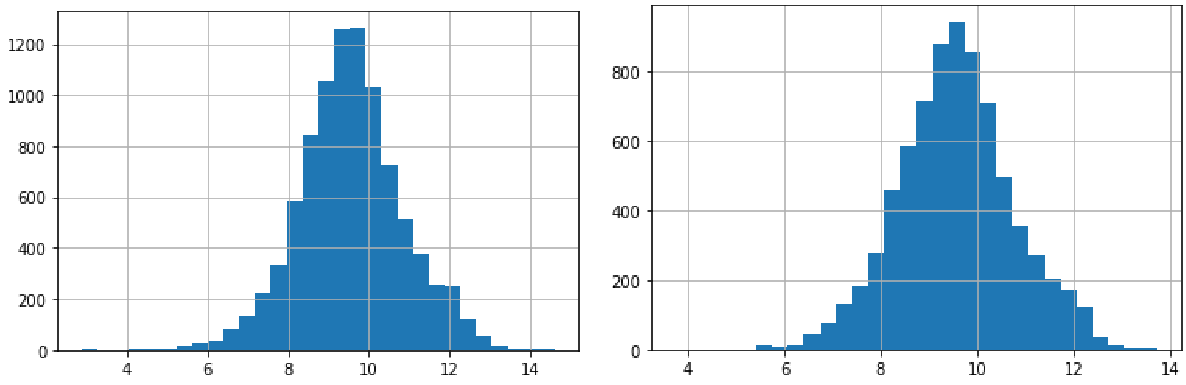

Figure B.1: Histogram of Size. Left = uncleaned, Right = cleaned.

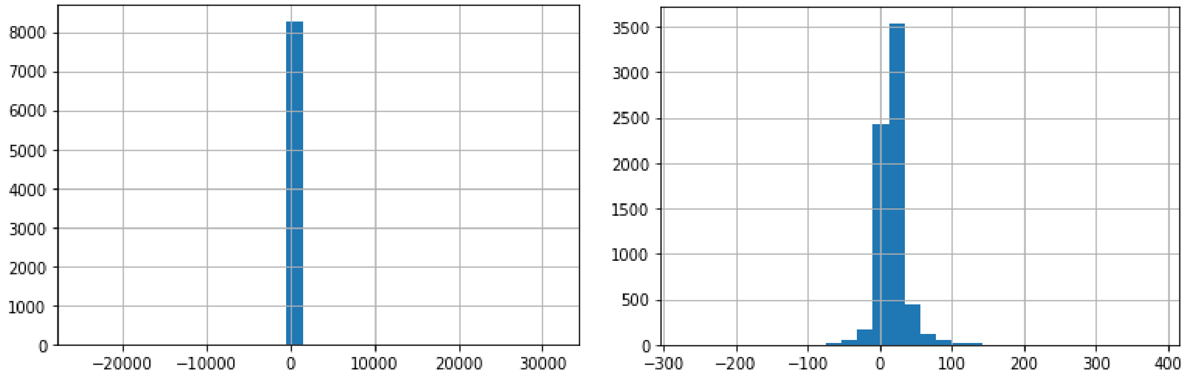

Figure B.2: Histogram of Return-on-Equity. Left = uncleaned, Right = cleaned.

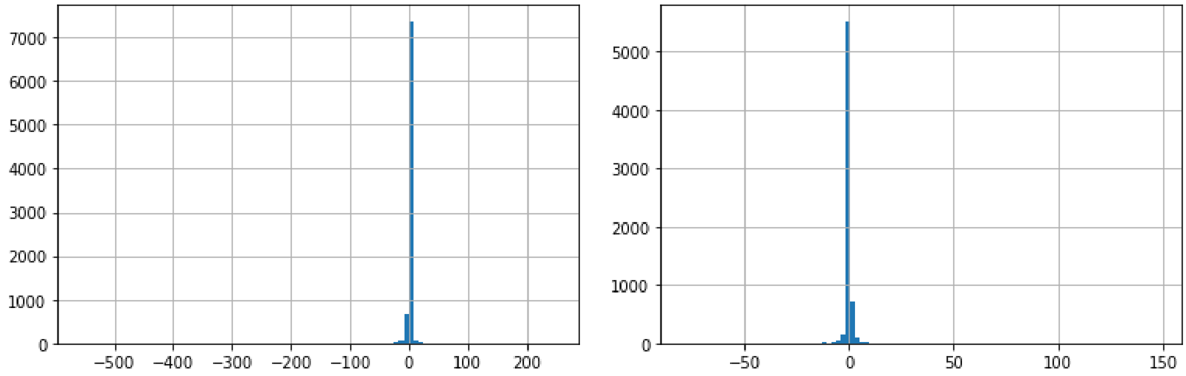

Figure B.3: Histogram of Net Sales Change. Left = uncleaned, Right = cleaned.

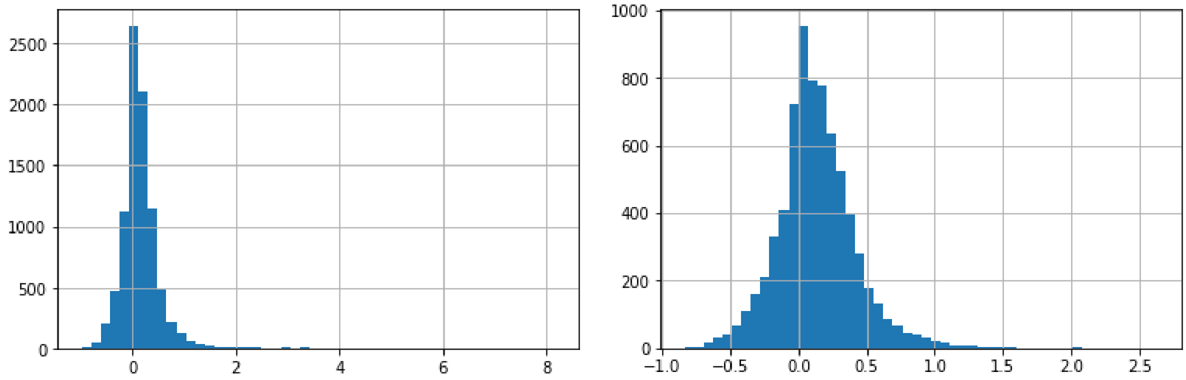

Figure B.4: Histogram of Annual Return. Left = uncleaned, Right = cleaned.

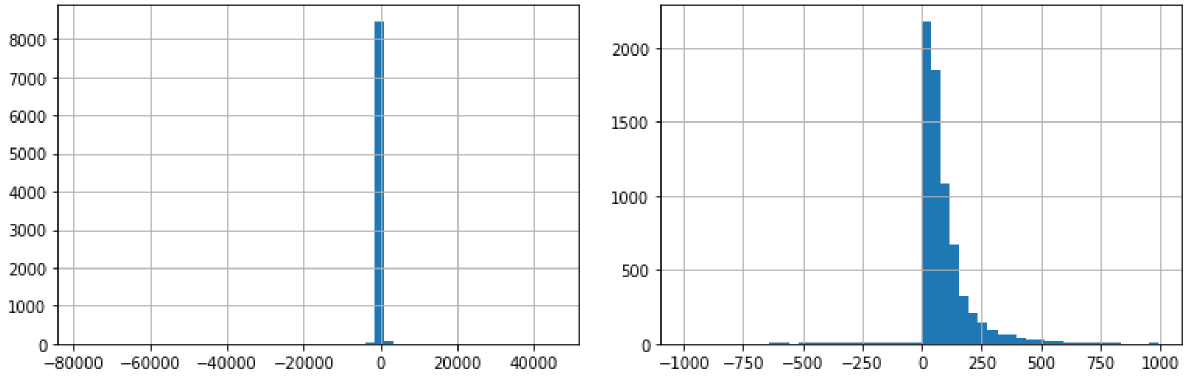

Figure B.5: Histogram of Leverage. Left = uncleaned, Right = cleaned.

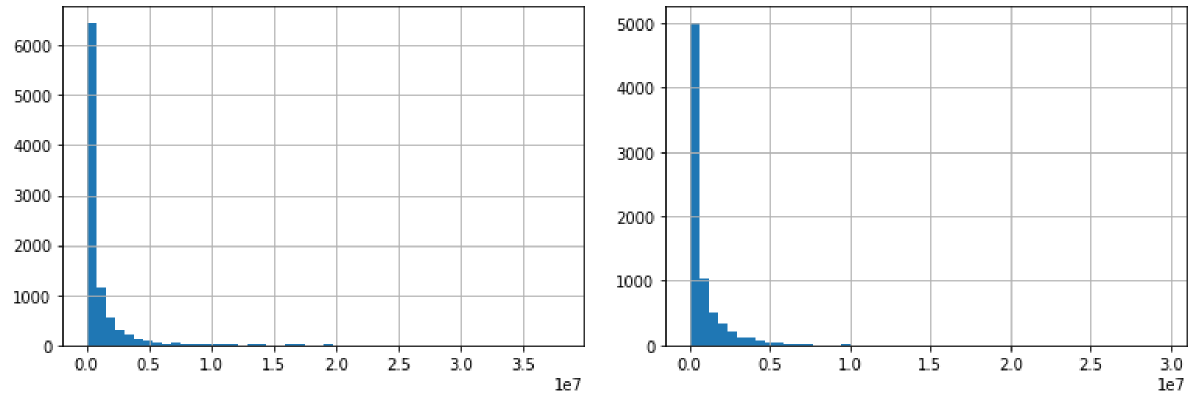

Figure B.6: Histogram of Capital Expenditures. Left = uncleaned, Right = cleaned.

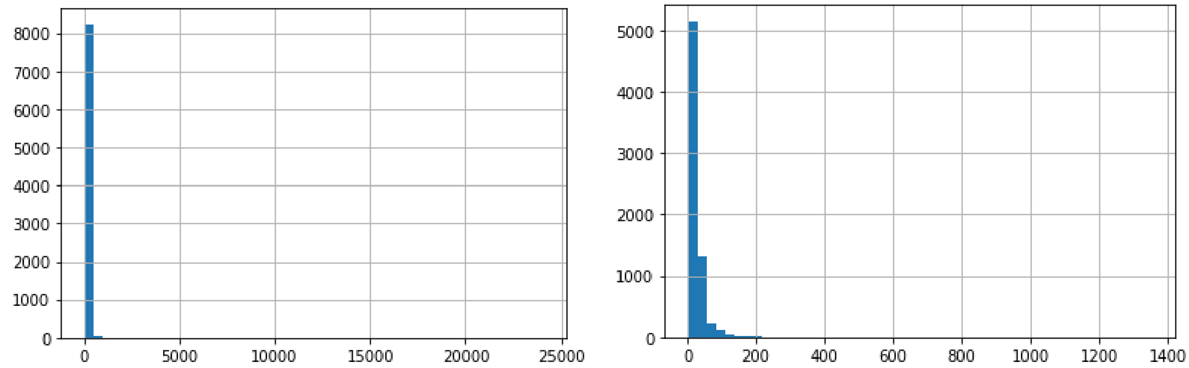

Figure B.7: Histogram of Price-Earnings Ratio. Left = uncleaned, Right = cleaned.

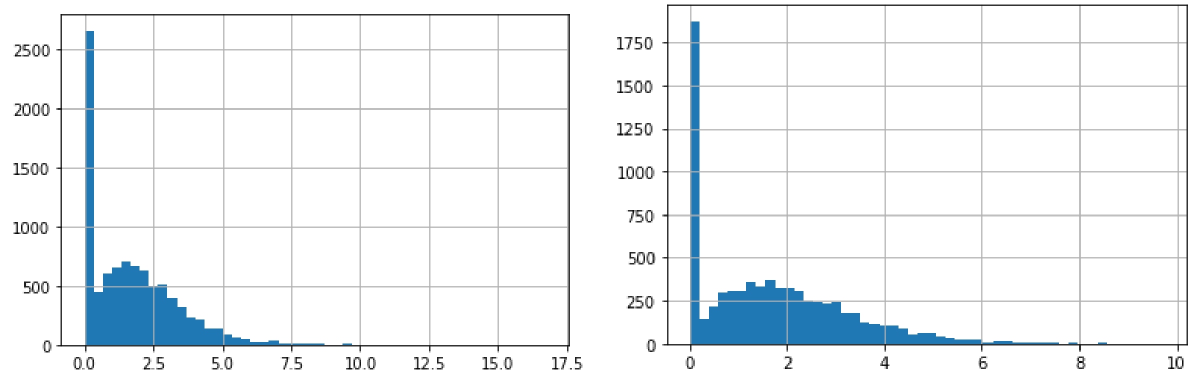

Figure B.8: Histogram of Dividend Yield. Left = uncleaned, Right = cleaned.

Now, our sample consists of 8339 observations. Furthermore, we need to drop all observations with missing values for one of the variables of interest. This is mostly the case for the ESG Score and Controversies Score in earlier years, leaving us with a final sample size of 5006 complete observations.

B.2 Exploratory Data Analysis

In order to investigate the relationship between the independent variables and the dependent variable, we look at scatter plots, correlations and the Variance Inflation Factor next.

As a first step, we explore the Variance Inflation Factor, i.e. whether any of the explanatory variables can themselves be explained by any set of the remaining variables. An introduction into the method was given in Definition 3.1.10. Table B.1 below shows the Variance Inflation Factor for all explanatory variables. The top row computes the VIF without having sector and yearly fixed effects, the bottom row includes the fixed effects in its computation of the VIF. We can clearly see that there is multicollinearity for the variable *Size*. In order to reduce this, we compute its yearly z-score and use *Z-norm: Size* as our new explanatory variable.

	ESG Score	**Controversies Score**	**Size**	**ROE**	**Net Sales Change Pct**	**Annual Return**	**Leverage**	**CapEx**	**Price-Earnings Ratio**	**Dividend Yield**
VIF - no fixed effects	10.86	8.29	24.21	2.22	1.01	1.25	1.83	1.62	1.59454	2.8313
VIF - fixed effects	15.01	10.30	65.08	2.52	1.03	1.91	2.03	2.51	1.70	4.71

Table B.1: Variance Inflation Factors

Table B.2 below shows the Variance Inflation Factor for all explanatory variables, after having normalized *Size* with a z-score transformation. Again, the top row computes the VIF without having sector and yearly fixed effects, the bottom row includes the fixed effects in its computation of the VIF. We can see, that the ESG Score and Controversies Score still have a very high VIF when including the fixed effects in our model. Hence, we split our analysis into two models, each including only one score, to interpret their influence on the market valuation multiple separately without having to deal with them influencing each other and possibly giving misleading results in our regression analysis. This reduced their respective Variance Inflation Factors to around 10.

	ESG Score	**Controversies Score**	**Z-norm: Size**	**ROE**	**Net Sales Change Pct**	**Annual Return**	**Leverage**	**CapEx**	**Price-Earnings Ratio**	**Dividend Yield**
VIF - no fixed effects	6.35	5.44	1.62	2.18	1.01	1.21	1.88	1.70	1.52	2.82
VIF - fixed effects	14.33	10.41	2.23	2.51	1.03	1.85	2.03	2.69	1.68	4.68

Table B.2: Variance Inflation Factors

Next, Figure B.9 shows the scatter plots of all variables, where the last row depicts the connection between the explanatory variables and the variable of interest, i.e. the Price-to-Book ratio.

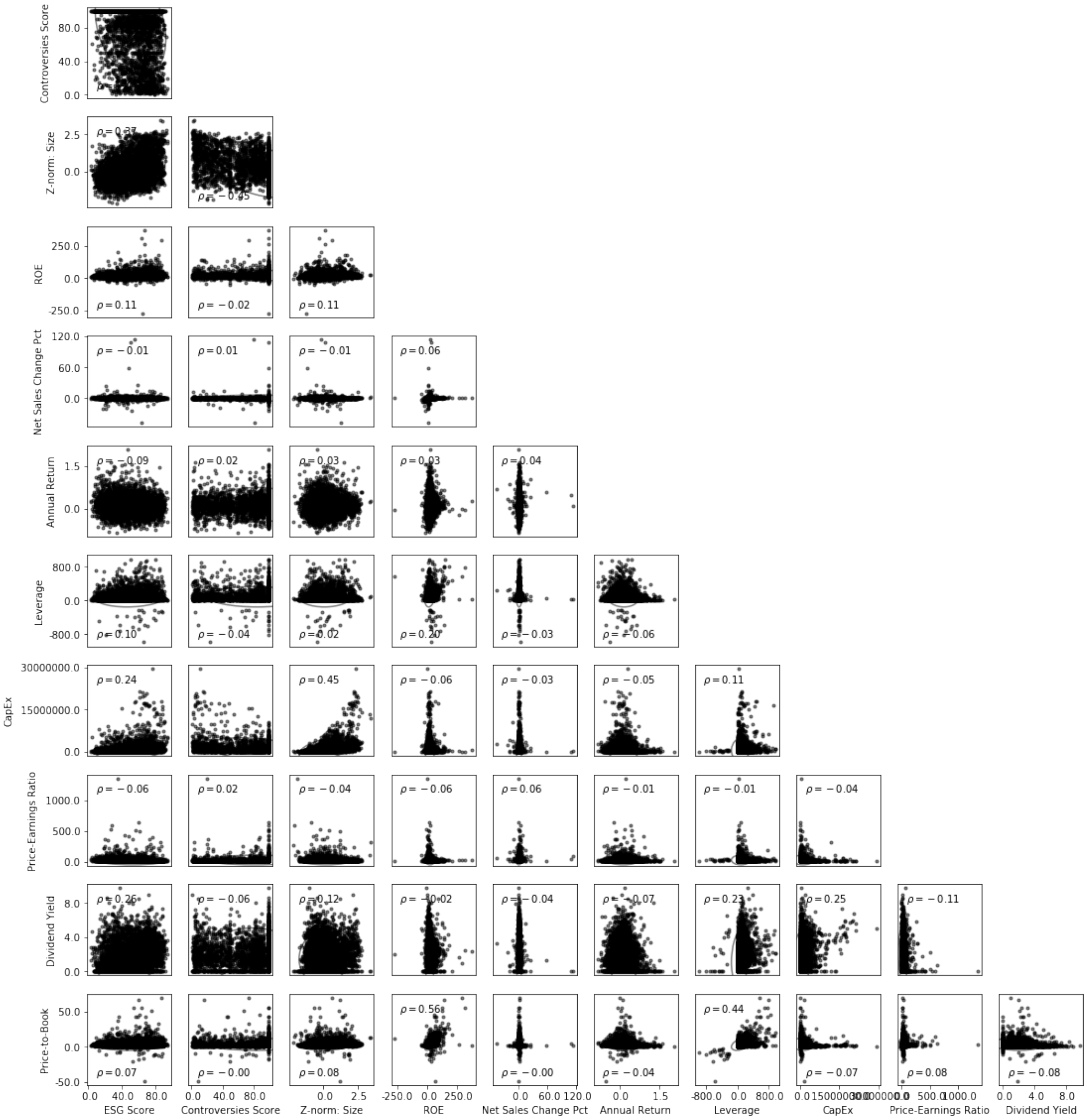

Figure B.9: Scatter Plots - All Variables

We can see quite clearly, that there is a logarithmic connection between the Price-to-Book ratio and the Annual Return, Capital Expenditures, PE ratio and Dividend Yield. Furthermore, the other variables' observations also seem to be concentrated around 0. Hence, we log-transform our dependent variable for the following analysis in order to improve linearity. The resulting scatter plots are shown in Figure B.10.

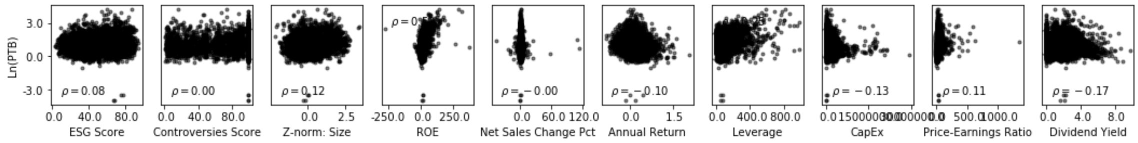

Figure B.10: Scatter Plots - All Variables

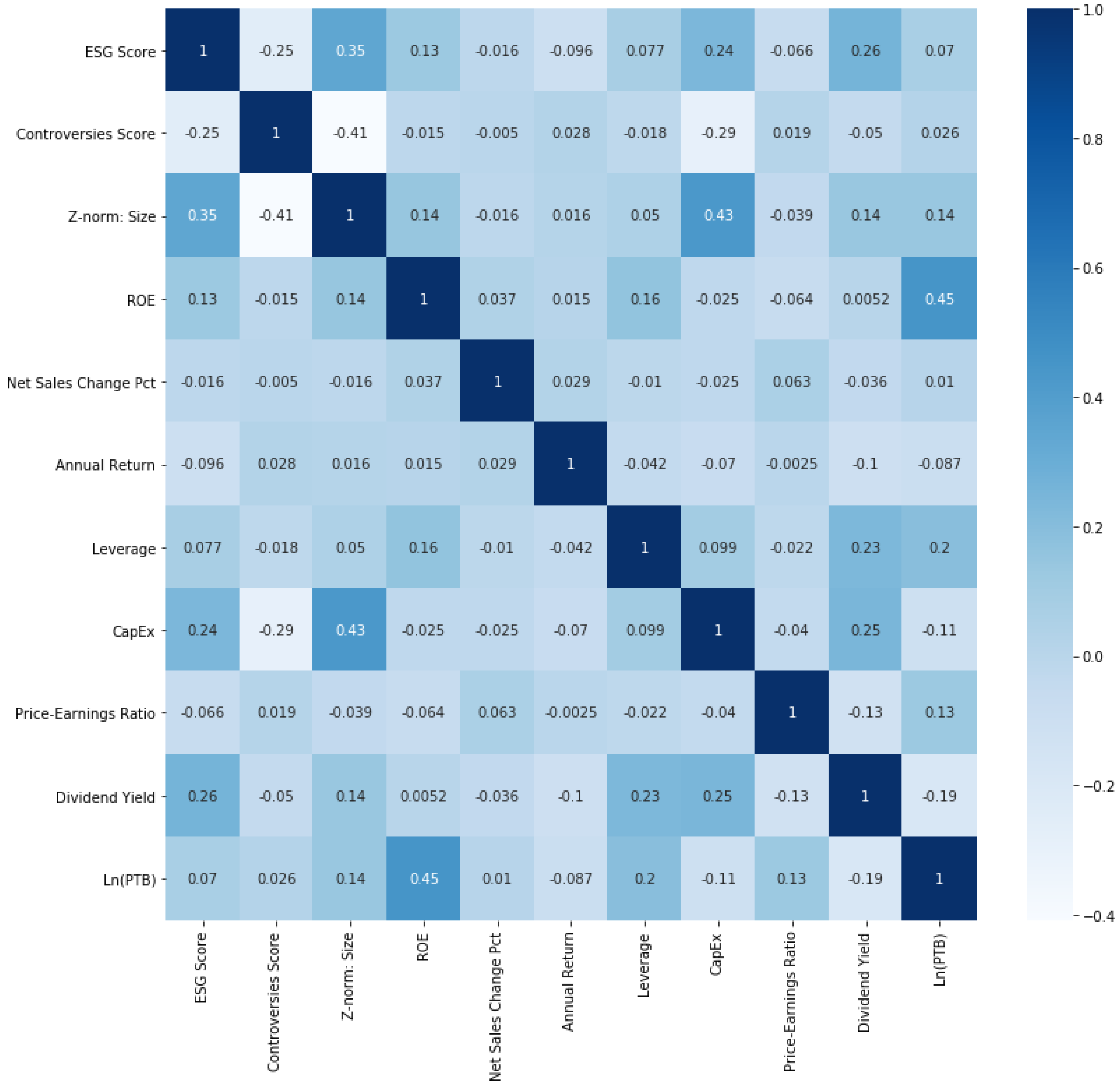

Figure B.11: Correlations - All Variables - Transformed

Figure B.11 shows a heatmap of the correlations between each variable. We observe a rather high negative correlation of −0.41 between *Z-norm: Size* and the Controversies Score, which indicates that larger companies tend to have a lower Controversies Score. In contrast, the correlation between the ESG Score and *Z-norm: Size* is positive with a value of 0.35, suggesting the opposite for this score. Furthermore, there is a negative correlation of -0.25 between the ESG Score and the Controversies Score.

In addition, there is a rather high positive correlation between the Return-on-Equity and our dependent variable $Ln(PTB)$ and between capital expenditures and size. All in all, the correlations are relatively low, further ensuring we have no multicollinearity.

B.3 Linear Regression Assumptions

As introduced in Section 3.1 Assumption A, the regressand is assumed to be a linear function of the regressors specified in the model, i.e. the specification must be linear in its parameters.

In order to detect nonlinearity we investigate two plots: If our model fulfills the linearity assumption, the points should be symmetrically distributed around (1) a diagonal line when plotting the observed vs. predicted values, (2) a horizontal line when plotting residual vs. predicted values [Gre00]. In detail, if we observe a ‘bowed’ pattern in any of the formerly mentioned plots, this indicates that the model makes systematic errors whenever it is making exceptionally large or small predictions and would hint at the functional form of our model not being correctly specified.

Figure B.12 shows the resulting plots for Model 4.2, Figure B.13 for Model 4.3 and Figure B.14 for Model 4.4, where the ”Full Model” corresponds to the model with added explanatory variables Capital Expenditures, Price-Earnings Ratio and Dividend Yield.

We can see that the plots for the small models do not exhibit a pattern that significantly deviates from a diagonal (horizontal) line, i.e. we can conclude that the linearity assumption is satisfied. For the full models there seems to be a slight trend in the residuals, indicating that large values tend to be predicted too small and small values tend to be predicted too large. However, this observed pattern is minimal and should not interfere with the assumption.

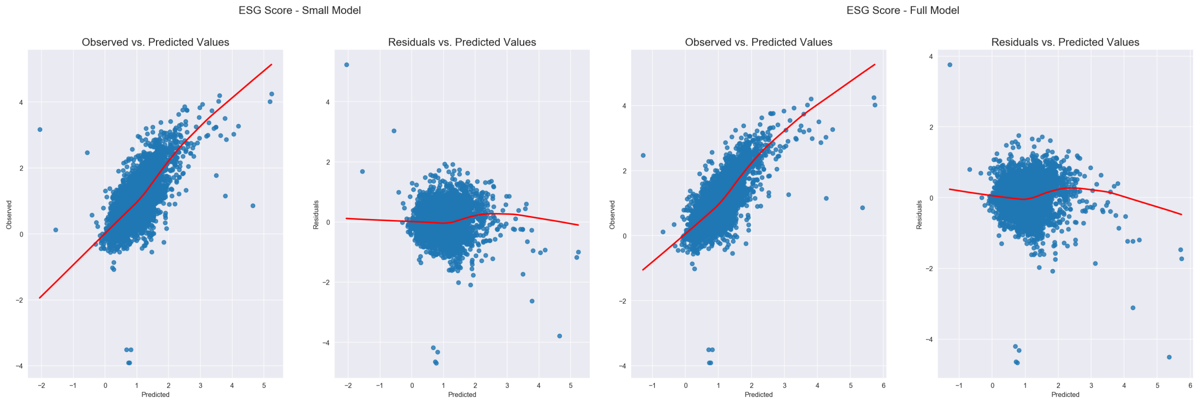

Figure B.12: Residual Plots - ESG Score Models

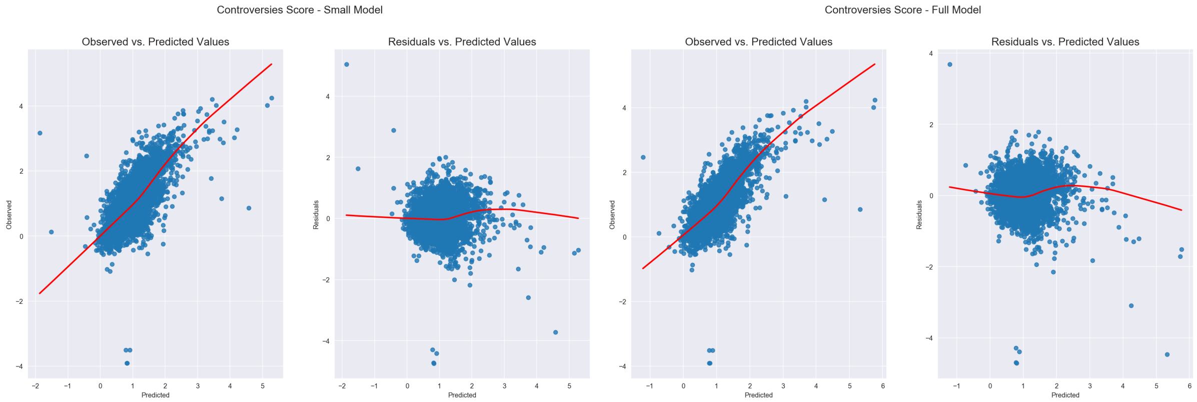

Figure B.13: Residual Plots - Controversies Score Models

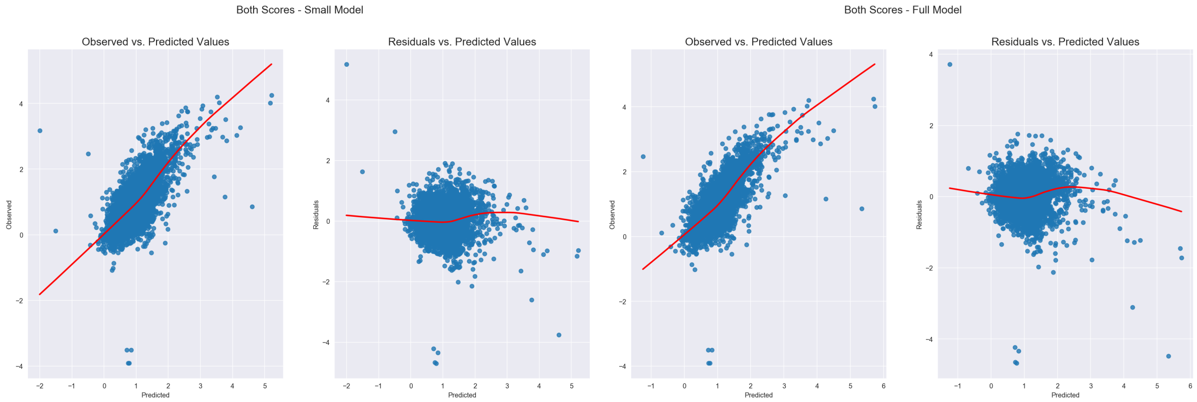

Figure B.14: Residual Plots - Complete Models

Another assumption that can be checked in these plots is the assumption of homoscedasticity (Assumption D). In detail, when residuals exhibit heteroscedasticity, they have non-constant variance. Looking at the residuals vs. predicted values plots in Figures B.12 - B.14, we cannot see a trend in variance, but we have some outliers that skew the plot. We further investigate this assumption by performing a Breusch-Pagan-Test. Here we see, that the assumption of homoscedasticity is rejected with a p-value smaller than 0.01 for all models. This could stem from the few extreme residuals, indicate ommitted variable bias or most likely emerge since our model is based on panel data. Even though we include sector and yearly fixed effects to tackle this problem, it does not seem to vanish.

We expand our analysis to a generalized regression model, where Assumption D is dropped and, instead, a robust estimator for the covariances of the residuals is used (i.e. $Var[\epsilon] = \sigma^2\Omega$) [Gre00]. The resulting parameter estimates and significance levels in the generalized regression model are very similar and do not change our interpretation.

Assumption B is fulfilled, since we have more observations than explanatory variables and we have checked the Variance Inflation Factor and correlations in the previous section.

Furthermore, Assumption C can easily be verified by computing the mean of the models' residuals: here, we get values smaller than 10^{-8}.

Table B.3: OLS Regression Results - Market Valuation Inference ESG Score

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.7428	0.052	14.181	0.000	0.640	0.845
ESG Score	-0.0040	0.000	-8.873	0.000	-0.005	-0.003
Z-norm: Size	0.0798	0.010	8.338	0.000	0.061	0.099
ROE	0.0099	0.000	28.000	0.000	0.009	0.011
Net Sales Change Pct	-0.0018	0.002	-1.081	0.280	-0.005	0.001
Annual Return	-0.2574	0.028	-9.286	0.000	-0.312	-0.203
Leverage	0.0014	6.91e-05	20.332	0.000	0.001	0.002
Automobiles and Parts	-0.0188	0.072	-0.261	0.794	-0.160	0.123
Banks	-0.6271	0.052	-12.066	0.000	-0.729	-0.525
Beverages	0.5436	0.065	8.422	0.000	0.417	0.670
Chemicals	0.3075	0.055	5.617	0.000	0.200	0.415
Construction and Materials	-0.1992	0.062	-3.214	0.001	-0.321	-0.078
Consumer Services	0.4700	0.101	4.650	0.000	0.272	0.668
Electricity	-0.4211	0.046	-9.219	0.000	-0.511	-0.332
Electronic and Electrical Equipment	0.4408	0.053	8.366	0.000	0.337	0.544
Finance and Credit Services	0.7818	0.157	4.971	0.000	0.473	1.090
Food Producers	0.3014	0.052	5.779	0.000	0.199	0.404
Gas, Water and Multi-utilities	-0.4741	0.058	-8.172	0.000	-0.588	-0.360
General Industrials	0.2095	0.050	4.181	0.000	0.111	0.308
Health Care Providers	-0.0987	0.058	-1.704	0.089	-0.212	0.015
Household Goods and Home Construction	-0.1960	0.059	-3.305	0.001	-0.312	-0.080
Industrial Engineering	-0.0500	0.068	-0.735	0.462	-0.183	0.083
Industrial Materials	0.2168	0.093	2.322	0.020	0.034	0.400
Industrial Metals and Mining	0.2515	0.088	2.871	0.004	0.080	0.423
Industrial Support Services	0.3766	0.048	7.870	0.000	0.283	0.470
Industrial Transportation	0.2236	0.052	4.339	0.000	0.123	0.325
Investment Banking and Brokerage Services	-0.2212	0.049	-4.533	0.000	-0.317	-0.126
Leisure Goods	0.2991	0.076	3.918	0.000	0.149	0.449
Life Insurance	-0.3925	0.071	-5.491	0.000	-0.533	-0.252
Media	0.0164	0.061	0.270	0.787	-0.103	0.136
Medical Equipment and Services	0.2944	0.044	6.658	0.000	0.208	0.381
Non-life Insurance	-0.3379	0.049	-6.845	0.000	-0.435	-0.241
Oil, Gas and Coal	-0.1842	0.049	-3.796	0.000	-0.279	-0.089
Personal Care, Drug and Grocery Stores	0.3373	0.053	6.340	0.000	0.233	0.442
Personal Goods	0.6110	0.060	10.249	0.000	0.494	0.728
Pharmaceuticals and Biotechnology	0.4248	0.051	8.361	0.000	0.325	0.524
Precious Metals and Mining	0.0101	0.138	0.073	0.942	-0.261	0.282
Real Estate Investment Trusts	-0.0048	0.046	-0.106	0.916	-0.094	0.085
Real Estate Investment and Services	0.3770	0.139	2.721	0.007	0.105	0.649
Retailers	0.4993	0.048	10.495	0.000	0.406	0.593
Software and Computer Services	0.5670	0.046	12.336	0.000	0.477	0.657
Technology Hardware and Equipment	0.2768	0.045	6.200	0.000	0.189	0.364
Telecommunications Equipment	0.1772	0.072	2.475	0.013	0.037	0.318
Telecommunications Service Providers	-0.3201	0.074	-4.337	0.000	-0.465	-0.175
Tobacco	0.6680	0.131	5.103	0.000	0.411	0.925
Travel and Leisure	0.2362	0.052	4.585	0.000	0.135	0.337
Waste and Disposal Services	-0.0672	0.096	-0.701	0.483	-0.255	0.121
2004	0.0488	0.047	1.039	0.299	-0.043	0.141
2005	0.0628	0.046	1.361	0.174	-0.028	0.153
2006	0.1139	0.046	2.486	0.013	0.024	0.204
2007	0.0970	0.046	2.109	0.035	0.007	0.187
2008	0.0496	0.049	1.022	0.307	-0.046	0.145
2009	-0.1690	0.045	-3.774	0.000	-0.257	-0.081
2010	-0.0744	0.045	-1.642	0.101	-0.163	0.014
2011	-0.0318	0.046	-0.695	0.487	-0.121	0.058
2012	-0.0394	0.045	-0.870	0.384	-0.128	0.049
2013	0.0746	0.045	1.648	0.100	-0.014	0.163
2014	0.2165	0.046	4.750	0.000	0.127	0.306
2015	0.2370	0.046	5.164	0.000	0.147	0.327
2016	0.1841	0.046	4.031	0.000	0.095	0.274
2017	0.2899	0.046	6.337	0.000	0.200	0.380
2018	0.3114	0.047	6.650	0.000	0.220	0.403
2019	0.2421	0.048	5.073	0.000	0.149	0.336
2020	0.6553	0.123	5.321	0.000	0.414	0.897
R-squared	0.506					
Adj. R-squared	0.500					
No. Observations	5435					

Table B.4: OLS Regression Results - Market Valuation Inference Controversies Score

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.4257	0.056	7.604	0.000	0.316	0.535
Controversies Score	0.0022	0.000	8.299	0.000	0.002	0.003
Z-norm: Size	0.0689	0.009	7.537	0.000	0.051	0.087
ROE	0.0096	0.000	27.008	0.000	0.009	0.010
Net Sales Change Pct	-0.0015	0.002	-0.883	0.377	-0.005	0.002
Annual Return	-0.2387	0.028	-8.658	0.000	-0.293	-0.185
Leverage	0.0014	6.93e-05	20.526	0.000	0.001	0.002
Automobiles and Parts	0.0434	0.072	0.603	0.546	-0.098	0.184
Banks	-0.6276	0.052	-12.058	0.000	-0.730	-0.526
Beverages	0.5643	0.065	8.729	0.000	0.438	0.691
Chemicals	0.2651	0.055	4.848	0.000	0.158	0.372
Construction and Materials	-0.1987	0.062	-3.203	0.001	-0.320	-0.077
Consumer Services	0.5505	0.101	5.451	0.000	0.353	0.748
Electricity	-0.4540	0.046	-9.928	0.000	-0.544	-0.364
Electronic and Electrical Equipment	0.4206	0.053	7.972	0.000	0.317	0.524
Finance and Credit Services	0.8118	0.157	5.156	0.000	0.503	1.121
Food Producers	0.2593	0.052	4.990	0.000	0.157	0.361
Gas, Water and Multi-utilities	-0.4805	0.058	-8.274	0.000	-0.594	-0.367
General Industrials	0.1996	0.050	3.980	0.000	0.101	0.298
Health Care Providers	-0.0799	0.058	-1.378	0.168	-0.194	0.034
Household Goods and Home Construction	-0.1687	0.059	-2.842	0.004	-0.285	-0.052
Industrial Engineering	-0.0793	0.068	-1.165	0.244	-0.213	0.054
Industrial Materials	0.1637	0.093	1.753	0.080	-0.019	0.347
Industrial Metals and Mining	0.2542	0.088	2.898	0.004	0.082	0.426
Industrial Support Services	0.4174	0.048	8.767	0.000	0.324	0.511
Industrial Transportation	0.2493	0.052	4.836	0.000	0.148	0.350
Investment Banking and Brokerage Services	-0.2250	0.049	-4.604	0.000	-0.321	-0.129
Leisure Goods	0.2997	0.076	3.921	0.000	0.150	0.450
Life Insurance	-0.4005	0.072	-5.593	0.000	-0.541	-0.260
Media	0.0558	0.061	0.917	0.359	-0.063	0.175
Medical Equipment and Services	0.2968	0.044	6.705	0.000	0.210	0.384
Non-life Insurance	-0.3241	0.049	-6.563	0.000	-0.421	-0.227
Oil, Gas and Coal	-0.1857	0.049	-3.821	0.000	-0.281	-0.090
Personal Care, Drug and Grocery Stores	0.3314	0.053	6.222	0.000	0.227	0.436
Personal Goods	0.5998	0.060	10.052	0.000	0.483	0.717
Pharmaceuticals and Biotechnology	0.4410	0.051	8.671	0.000	0.341	0.541
Precious Metals and Mining	-0.1184	0.138	-0.858	0.391	-0.389	0.152
Real Estate Investment Trusts	-0.0221	0.046	-0.482	0.630	-0.112	0.068
Real Estate Investment and Services	0.2500	0.138	1.810	0.070	-0.021	0.521
Retailers	0.4960	0.048	10.415	0.000	0.403	0.589
Software and Computer Services	0.5682	0.046	12.346	0.000	0.478	0.658
Technology Hardware and Equipment	0.2560	0.045	5.711	0.000	0.168	0.344
Telecommunications Equipment	0.1637	0.072	2.286	0.022	0.023	0.304
Telecommunications Service Providers	-0.1887	0.074	-2.565	0.010	-0.333	-0.044
Tobacco	0.6818	0.131	5.201	0.000	0.425	0.939
Travel and Leisure	0.2596	0.052	5.024	0.000	0.158	0.361
Waste and Disposal Services	-0.1031	0.096	-1.073	0.283	-0.291	0.085
2004	0.0494	0.047	1.050	0.294	-0.043	0.142
2005	0.0544	0.046	1.175	0.240	-0.036	0.145
2006	0.0998	0.046	2.176	0.030	0.010	0.190
2007	0.0701	0.046	1.527	0.127	-0.020	0.160
2008	0.0150	0.048	0.309	0.757	-0.080	0.110
2009	-0.2189	0.044	-4.959	0.000	-0.305	-0.132
2010	-0.1152	0.045	-2.576	0.010	-0.203	-0.028
2011	-0.0902	0.045	-2.004	0.045	-0.178	-0.002
2012	-0.1028	0.044	-2.314	0.021	-0.190	-0.016
2013	0.0083	0.044	0.188	0.851	-0.078	0.095
2014	0.1454	0.045	3.267	0.001	0.058	0.233
2015	0.1394	0.045	3.130	0.002	0.052	0.227
2016	0.0856	0.044	1.948	0.051	-0.001	0.172
2017	0.1766	0.044	4.034	0.000	0.091	0.262
2018	0.2123	0.045	4.716	0.000	0.124	0.300
2019	0.1396	0.046	3.066	0.002	0.050	0.229
2020	0.5765	0.123	4.695	0.000	0.336	0.817
R-squared	0.505					
Adj. R-squared	0.499					
No. Observations	5430					

Table B.5: OLS Regression Results - Market Valuation Inference ESG Score - Added Control Variables

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.7137	0.053	13.443	0.000	0.610	0.818
ESG Score	-0.0028	0.000	-6.136	0.000	-0.004	-0.002
Z-norm: Size	0.1190	0.010	11.366	0.000	0.098	0.140
ROE	0.0119	0.000	31.550	0.000	0.011	0.013
Net Sales Change Pct	-0.0076	0.002	-3.263	0.001	-0.012	-0.003
Annual Return	-0.2932	0.028	-10.302	0.000	-0.349	-0.237
Leverage	0.0014	7.05e-05	19.181	0.000	0.001	0.001
CapEx	-4.628e-08	5.01e-09	-9.243	0.000	-5.61e-08	-3.65e-08
Price-Earnings Ratio	0.0015	0.000	8.025	0.000	0.001	0.002
Dividend Yield	-0.0218	0.006	-3.656	0.000	-0.034	-0.010
Automobiles and Parts	0.0495	0.071	0.699	0.484	-0.089	0.188
Banks	-0.6174	0.057	-10.902	0.000	-0.728	-0.506
Beverages	0.5552	0.064	8.633	0.000	0.429	0.681
Chemicals	0.2782	0.054	5.121	0.000	0.172	0.385
Construction and Materials	-0.2578	0.062	-4.172	0.000	-0.379	-0.137
Consumer Services	0.4140	0.099	4.189	0.000	0.220	0.608
Electricity	-0.2822	0.047	-5.948	0.000	-0.375	-0.189
Electronic and Electrical Equipment	0.4316	0.052	8.297	0.000	0.330	0.534
Finance and Credit Services	0.6405	0.150	4.259	0.000	0.346	0.935
Food Producers	0.2765	0.052	5.360	0.000	0.175	0.378
Gas, Water and Multi-utilities	-0.3408	0.060	-5.665	0.000	-0.459	-0.223
General Industrials	0.2045	0.050	4.123	0.000	0.107	0.302
Health Care Providers	-0.1547	0.057	-2.713	0.007	-0.267	-0.043
Household Goods and Home Construction	-0.1607	0.060	-2.677	0.007	-0.278	-0.043
Industrial Engineering	-0.0281	0.066	-0.424	0.671	-0.158	0.102
Industrial Materials	0.2884	0.094	3.069	0.002	0.104	0.473
Industrial Metals and Mining	0.2536	0.086	2.939	0.003	0.084	0.423
Industrial Support Services	0.3597	0.048	7.569	0.000	0.267	0.453
Industrial Transportation	0.2243	0.051	4.425	0.000	0.125	0.324
Investment Banking and Brokerage Services	-0.2260	0.049	-4.642	0.000	-0.321	-0.131
Leisure Goods	0.2577	0.077	3.356	0.001	0.107	0.408
Life Insurance	-0.2466	0.087	-2.823	0.005	-0.418	-0.075
Media	0.0528	0.062	0.853	0.394	-0.069	0.174
Medical Equipment and Services	0.2570	0.044	5.787	0.000	0.170	0.344
Non-life Insurance	-0.2923	0.050	-5.865	0.000	-0.390	-0.195
Oil, Gas and Coal	-0.1159	0.051	-2.294	0.022	-0.215	-0.017
Personal Care, Drug and Grocery Stores	0.2955	0.052	5.666	0.000	0.193	0.398
Personal Goods	0.4979	0.061	8.153	0.000	0.378	0.618
Pharmaceuticals and Biotechnology	0.3487	0.051	6.884	0.000	0.249	0.448
Precious Metals and Mining	0.0041	0.150	0.027	0.978	-0.290	0.298
Real Estate Investment Trusts	0.0097	0.048	0.202	0.840	-0.084	0.104
Real Estate Investment and Services	0.2671	0.144	1.861	0.063	-0.014	0.548
Retailers	0.4304	0.047	9.121	0.000	0.338	0.523
Software and Computer Services	0.4276	0.047	9.145	0.000	0.336	0.519
Technology Hardware and Equipment	0.2506	0.045	5.590	0.000	0.163	0.338
Telecommunications Equipment	0.1087	0.070	1.549	0.121	-0.029	0.246
Telecommunications Service Providers	0.0227	0.080	0.284	0.776	-0.134	0.179
Tobacco	0.5709	0.127	4.504	0.000	0.322	0.819
Travel and Leisure	0.2449	0.051	4.782	0.000	0.145	0.345
Waste and Disposal Services	-0.0828	0.092	-0.897	0.370	-0.264	0.098
2004	0.0205	0.047	0.435	0.663	-0.072	0.113
2005	0.0228	0.046	0.497	0.620	-0.067	0.113
2006	0.0931	0.046	2.045	0.041	0.004	0.182
2007	0.0762	0.046	1.669	0.095	-0.013	0.166
2008	0.0252	0.048	0.524	0.600	-0.069	0.120
2009	-0.1329	0.045	-2.950	0.003	-0.221	-0.045
2010	-0.0834	0.046	-1.804	0.071	-0.174	0.007
2011	-0.0570	0.046	-1.251	0.211	-0.146	0.032
2012	-0.0394	0.045	-0.878	0.380	-0.127	0.049
2013	0.0625	0.045	1.383	0.167	-0.026	0.151
2014	0.1976	0.045	4.348	0.000	0.108	0.287
2015	0.2136	0.046	4.673	0.000	0.124	0.303
2016	0.1816	0.046	3.973	0.000	0.092	0.271
2017	0.2658	0.046	5.786	0.000	0.176	0.356
2018	0.2700	0.047	5.769	0.000	0.178	0.362
2019	0.2442	0.048	5.138	0.000	0.151	0.337
2020	0.6293	0.121	5.199	0.000	0.392	0.867
R-squared	0.544					
Adj. R-squared	0.538					
No. Observations	5006					

Table B.6: OLS Regression Results - Market Valuation Inference Controversies Score - Added Control Variables

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.4932	0.057	8.585	0.000	0.381	0.606
Controversies Score	0.0016	0.000	5.985	0.000	0.001	0.002
Z-norm: Size	0.1129	0.010	11.131	0.000	0.093	0.133
ROE	0.0117	0.000	31.030	0.000	0.011	0.012
Net Sales Change Pct	-0.0076	0.002	-3.275	0.001	-0.012	-0.003
Annual Return	-0.2823	0.028	-9.965	0.000	-0.338	-0.227
Leverage	0.0014	7.06e-05	19.351	0.000	0.001	0.002
CapEx	-4.457e-08	5.03e-09	-8.857	0.000	-5.44e-08	-3.47e-08
Price-Earnings Ratio	0.0016	0.000	8.544	0.000	0.001	0.002
Dividend Yield	-0.0288	0.006	-4.929	0.000	-0.040	-0.017
Automobiles and Parts	0.0925	0.071	1.312	0.190	-0.046	0.231
Banks	-0.6074	0.057	-10.729	0.000	-0.718	-0.496
Beverages	0.5707	0.064	8.867	0.000	0.445	0.697
Chemicals	0.2483	0.054	4.584	0.000	0.142	0.355
Construction and Materials	-0.2550	0.062	-4.127	0.000	-0.376	-0.134
Consumer Services	0.4565	0.099	4.623	0.000	0.263	0.650
Electricity	-0.2950	0.048	-6.208	0.000	-0.388	-0.202
Electronic and Electrical Equipment	0.4174	0.052	8.024	0.000	0.315	0.519
Finance and Credit Services	0.6623	0.150	4.404	0.000	0.367	0.957
Food Producers	0.2522	0.051	4.905	0.000	0.151	0.353
Gas, Water and Multi-utilities	-0.3298	0.060	-5.483	0.000	-0.448	-0.212
General Industrials	0.1999	0.050	4.031	0.000	0.103	0.297
Health Care Providers	-0.1495	0.057	-2.619	0.009	-0.261	-0.038
Household Goods and Home Construction	-0.1402	0.060	-2.338	0.019	-0.258	-0.023
Industrial Engineering	-0.0460	0.066	-0.694	0.488	-0.176	0.084
Industrial Materials	0.2639	0.094	2.808	0.005	0.080	0.448
Industrial Metals and Mining	0.2630	0.086	3.049	0.002	0.094	0.432
Industrial Support Services	0.3856	0.047	8.147	0.000	0.293	0.478
Industrial Transportation	0.2391	0.051	4.717	0.000	0.140	0.338
Investment Banking and Brokerage Services	-0.2308	0.049	-4.741	0.000	-0.326	-0.135
Leisure Goods	0.2471	0.077	3.218	0.001	0.097	0.398
Life Insurance	-0.2520	0.087	-2.884	0.004	-0.423	-0.081
Media	0.0761	0.062	1.228	0.220	-0.045	0.198
Medical Equipment and Services	0.2508	0.044	5.651	0.000	0.164	0.338
Non-life Insurance	-0.2798	0.050	-5.619	0.000	-0.377	-0.182
Oil, Gas and Coal	-0.1237	0.051	-2.445	0.015	-0.223	-0.025
Personal Care, Drug and Grocery Stores	0.2902	0.052	5.566	0.000	0.188	0.392
Personal Goods	0.4808	0.061	7.886	0.000	0.361	0.600
Pharmaceuticals and Biotechnology	0.3594	0.051	7.094	0.000	0.260	0.459
Precious Metals and Mining	-0.1029	0.149	-0.690	0.490	-0.395	0.190
Real Estate Investment Trusts	0.0073	0.048	0.152	0.879	-0.087	0.101
Real Estate Investment and Services	0.1587	0.143	1.112	0.266	-0.121	0.438
Retailers	0.4229	0.047	8.970	0.000	0.330	0.515
Software and Computer Services	0.4166	0.047	8.916	0.000	0.325	0.508
Technology Hardware and Equipment	0.2338	0.045	5.216	0.000	0.146	0.322
Telecommunications Equipment	0.0961	0.070	1.371	0.170	-0.041	0.234
Telecommunications Service Providers	0.1053	0.079	1.331	0.183	-0.050	0.260
Tobacco	0.6024	0.127	4.750	0.000	0.354	0.851
Travel and Leisure	0.2577	0.051	5.020	0.000	0.157	0.358
Waste and Disposal Services	-0.1007	0.092	-1.092	0.275	-0.282	0.080
2004	0.0207	0.047	0.438	0.661	-0.072	0.113
2005	0.0141	0.046	0.307	0.759	-0.076	0.104
2006	0.0798	0.045	1.757	0.079	-0.009	0.169
2007	0.0553	0.045	1.219	0.223	-0.034	0.144
2008	-0.0010	0.048	-0.020	0.984	-0.095	0.093
2009	-0.1626	0.044	-3.660	0.000	-0.250	-0.076
2010	-0.1097	0.046	-2.403	0.016	-0.199	-0.020
2011	-0.0985	0.045	-2.205	0.027	-0.186	-0.011
2012	-0.0820	0.044	-1.868	0.062	-0.168	0.004
2013	0.0172	0.044	0.391	0.696	-0.069	0.103
2014	0.1464	0.044	3.314	0.001	0.060	0.233
2015	0.1427	0.044	3.233	0.001	0.056	0.229
2016	0.1132	0.044	2.577	0.010	0.027	0.199
2017	0.1861	0.044	4.245	0.000	0.100	0.272
2018	0.1976	0.045	4.417	0.000	0.110	0.285
2019	0.1741	0.045	3.837	0.000	0.085	0.263
2020	0.5812	0.121	4.818	0.000	0.345	0.818
R-squared	0.544					
Adj. R-squared	0.538					
No. Observations	5006					

Appendix C

Keywords

- **GHG Emissions**
  GHG Emissions, greenhouse gas, emissions, climate change, CO2, Paris Agreement
- **Air Quality**
  air pollution, heavy metals, air quality
- **Energy Management**
  energy management, energy efficiency, energy mix
- **Water & Wastewater Management**
  water use, water consumption, wastewater generation, groundwater pollution
- **Waste & Hazardous Materials Management**
  hazardous waste, non-hazardous waste, waste treatment, waste storage, waste disposal
- **Ecological Impacts**
  ecosystem, biodiversity, land exploration, natural resource extraction, deforestation, habitat destruction,
- **Human Rights & Community Relations**
  human rights, treatment indigenous, community impacts, community engagement, environmental justice, local workforce, local businesses
- **Customer Privacy**
  customer privacy, personally identifiable information, customer data, user data
- **Data Security**
  data security, data breaches, data collection, IT infrastructure, data policies, security training

- **Access & Affordability**
  universal need, accessability, affordability, accesable, affordable

- **Product Quality & Safety**
  health risk, safety risk, product quality, product safety, recalls, market withdrawal

- **Customer Welfare**
  customer welfare, antibiotic animal production, nutrition

- **Selling Practices & Product Labeling**
  transparency, marketing statements, product label, product labeling, selling practices, incentive structures

- **Labor Practices**
  Labor practices, labor standards, labor law, child labor, forced labor, bonded labor, exploitation,fair wages, organized labor, freedom of association, union, overtime pay, minimum wage, benefits

- **Employee Health & Safety**
  workplace injuries, fatalities, illness, safety plans, safety audit, physical health, mental health, protective equipment

- **Employee Engagement, Diversity & Inclusion**
  culture, hiring practice, hiring, promotion, talent, discrimination, discriminatory

- **Product Design & Lifecycle Management**
  environmental friendly resources, environmental friendly product, environmental friendly packaging, environmental friendly shipping, sustainable product, recycling

- **Business Model Resilience**
  risk, opportunity, social transitions, environmental transitions, political transitions, low-carbon economy, climate-constrained economy, new market, evolving market, underserved market

- **Supply Chain Management**
  Sustainable supply chain, suppliers, operational activities, environmental impact, social impact

- **Materials Sourcing & Efficiency**
  resources, sourcing, resource availability, recycled material, reused material, dematerialization, resource efficiency, substitution

- **Physical Impacts of Climate Change**
  climate change, climate risk, climate exposure, climate change impact, incorporate climate change, extreme weather, adapt to climate change, shifting climate, sea level

- **Business Ethics**
  Business ethics, ethical conduct, fraud, corruption, bribery, facilitation payments, fiduciary, business norms, ethical standards, professional standards, conflict of interest, bias, negligence

- **Competitive Behavior**
  monopoly, excessive price, poor quality, anti-competitive practice, collusion, bargaining power, price fixing, manipulation, patent, intellectual property

- **Management of the Legal & Regulatory Environment**
  Regulator, regulations, conflicting interest, regulatory policy, subsidies, taxes, lobbying, compliance, alignment

- **Critical Incident Risk Management**
  High-impact accident, accident, incident, critical, emergency, safety, controls, internal control

- **Systemic Risk Management**
  systemic risk, system collapse, system weakening, financial system, natural resource system, technological system, safeguards, shocks, financial stress, economic stress, complexity, interconnectedness

Appendix D

Case Study - Facebook

#	date	event description	#tweets	ESG dimension	example tweets	polarity
1	17.02.18	Russian manipulation of election	18	Leadership & Governance	This Facebook executive argues that Moscow's primary goal was to divide Americans, not help Trump. But dividing populations is an essential part of the Kremlin's playbook for installing friendly, corrupt leaders in other countries. It's critical for @Facebook to understand this. https://t.co/HMMrjKbds6	Negative
2	18.02.18	False statements by Facebook executive	14	Social Capital	RT @wikileaks: Thread from the Vice President of @Facebook ads on U.S. election contradicts media narrative: https://t.co/NLTKPVmnhf	Negative
3	17.03.18 19.-24.03.18 26.-28.03.18	Cambridge Analytica	318	Social Capital	Members of @CamAnalytica exfiltrated data from @Facebook to create the ”Alamo dataset”, which was used to microtarget and manipulate Americans during the 2016 presidential election. Today, Zuckerberg banned @CamAnalytica and @TheSCLGroup from his platform for abuse. Video below. https://t.co/K1DnANvNXs	Negative
					Want to know why @Facebook suddenly banned a Trump-linked data firm last night? They were trying to preempt this from @nytimes: Trump Consultants Exploited the Facebook Data of Millions https://t.co/naUY2jsAmO	Negative
					Here it is: one of the largest, most consequential data breaches in US history occurred at the hands of @Facebook, @CamAnalytica, and @TheSCLGroup. 50 million people is nearly 25% of the US electorate. Demand a federal investigation. #Resistance #Resist https://t.co/zmttGPn0Uc https://t.co/p1k3jDemhX	Negative
					FACEBOOK @Facebook signals that Mark Zuckerberg *will* testify before Congress Three committees invited Zuckerberg to testify, including @ChuckGrassley & @SenFeinstein's Senate Judiciary Cmte at an *April 10* hearing on data privacy. https://t.co/T3puBHJxKe	Positive

Table D.1: Facebook Events

#	date	event description	#tweets	ESG dimension	example tweets	polarity
4	20.-21.03.18	Meeting to combat online harassment and promote Internet safety	35	Social Capital	RT @FLOTUS: It was a very productive meeting on cyber safety & how to teach our children to be responsible digital citizens. Thank you @amazon @Facebook @FOSI @Google @m_Beckerman @Microsoft @Twitter & @snap for coming to the @WhiteHouse today & sharing your valuable insight and expertise. https://t.co/YY2srrbYhP	Positive
5	20.-21.03.18	Cambridge Analytica aftermath	34	Leadership & Governance	Facebook shapes the conversations we have every day. @Facebook has created a system that allows for people to corrupt the free flow of information in our democracy for power & money. They owe us an explanation & they must fix it. With great power comes great responsibility.	Negative
			7		As expected @Facebook recently joined @Google, @comcast, @Verizon and @ATT to contribute more than $1 million to a PAC set up to oppose a California Ballot Initiative that would allow consumers to opt-out of data sharing with firms like #CambridgeAnalytica https://t.co/DtxQmvhbCG	Positive
6	27.-28.03.18	Unethical data collection	97	Leadership & Governance	RT @Snowden: Thread: an exquisite breakdown using real-life examples of how @Facebook and @Google exploited your trust to quietly create a decade-long dossier of your most private activities. With a bonus: how to download a copy of your own. https://t.co/6AjcMfetMD	Negative
7	04.04.18	Deleted profile of shooter at YouTube HQ	25	Social Capital	Extremely annoying that @YouTube, @Facebook and @Instagram deleted the shooter's profiles. That stuff is important for journalists, historians, researchers, etc. Another invitation for big tech regulation... https://t.co/wGzW4XfgpF	Negative
8	05.04.18	Indian users affected by Cambridge Analytica	19	Social Capital	RT @BDUTT: Breaking : @Facebook submits official response on #CambridgeAnalytica to the of India. An FB spokesperson confirmed to me that 335 Indians downloaded the Kogan APP - 'This Is Your Digital Life'- which potentially enabled CA to access data of 562,455 Indians .	Negative

Table D.2: Facebook Events

#	date	event description	#tweets	ESG dimension	example tweets	polarity
9	08.- 13.04.18	Censorship of black Women-Duo supporting Trump	308	Social Capital	I’m sitting out @Facebook this week & I will not be posting in protest of their continued censorship. Yep, it’s a small protest, & I don’t expect my actions to move the needle one bit against the Facebook totalitarians, but my small one-week boycott is my way of fighting back. https://t.co/M7rtoGbhuF	Negative
					WATCH: TRUMP SUPPORTERS ”Diamond and Silk” Take On @Facebook After They’re Told They Are ”Unsafe For The Community @DiamondandSilk #Facebook #Zuckerberg #Censorship #MondayMotivaton https://t.co/uHkr02NFzN via @100PercFEDUP	Negative
					America: If you believe anything that comes out of the mouth of Facebook’s Mark Zuckerberg during his congressional hearings: I HAVE A PRIVACY POLICY TO SELL YOU. @Facebook: Liars, Cheats, and Creeps. #DeleteFacebook	Negative
					Leaked documents from @Facebook show that Facebook was training content moderators to protect white men but not Black children. How can we know that Facebook fixed this problem if they keep their training documents secret? #AuditFacebook https://t.co/RwRxwWiyDr	Negative
					Earlier this afternoon in HELP Committee I discussed @Facebook censoring conservative views. Russians trying to influence our elections? Kick them off. People trying to incite violence and hate? Goodbye. But @DiamondAndSilk They are hardly ”unsafe to the community”. https://t.co/ySfeqiJ0RO	Negative
					Facebook openly stands for discrimination, hate speech & censorship towards all Conservatives, especially black women. @SenatorTimScott during HELP Committee discussed @Facebook ”censoring” @DiamondAndSilk They are hardly ”unsafe to the community” https://t.co/GkGEI7xnez	Negative
					@DiamondandSilk on @Facebook censorship: ”They reached out via Twitter. We saw it this morning when we woke up...We’ve been in interviews all day but we will be reaching back out to them.” @TeamCavuto https://t.co/IrkjkM40dw	Negative
10	10.-11.04.18	Censorship of black Women-Duo supporting Trump	79	Leadership & Governance	Why is @Facebook stopping some people that have liked and followed our @DiamondandSilk page from seeing our content first even though they are following FB instructions? This is what we call undercover bias tactics. ”CENSORSHIP” Press play, here’s the proof. https://t.co/14WjcEe9g3	Negative
11	14.07.18	Pages with conspiracy theories about Sandy Hook Elementary School Shooting (2012) were not taken down	16	Social Capital	Not only did @Facebook allow for the Sandy Hook Hoax page to stay up- they suggested it as ”a page you may be interested in” based on geographic algorithms. To us. Victims, first responders, survivors and their families who still lived in Newtown. It’s just not good. https://t.co/35FXYMHiMw	Negative
12	17.07.18	Restricted engagement for conservative sites	17	Social Capital	SHOCK STUDY: @Facebook Has Eliminated 93% of Traffic to Top Conservative Websites Since 2016 Election https://t.co/TQH9XVLLmY via @gatewaypundit	Negative
13	05.-06.08.18	Unethical donation to politician	50	Social Capital	Facebook is trying to buy Devin Nunes’ vote on regulating @Facebook and other social media platforms. Thousands of dollars donated to Nunes already this cycle. Very disturbing. #SaveDemocracy https://t.co/oChIu4BCQo	Negative

Table D.3: Facebook Events

#	date	event description	#tweets	ESG dimension	example tweets	polarity
14	06.08.18	Removed InfoWars pages	24	Social Capital	YouTube just terminated Alex Jones' channel which had 2.4 million subscribers & Facebook has removed four of his pages. Spotify finally completely kicked him to the curb. Never doubt that your voice makes difference. Bravo @Facebook, @YouTube, @Spotify. https://t.co/TcY6jN7vXu	Positive
15	06.-07.08.18	Blocked House GOP candidate's campaign ad	46	Leadership & Governance	This is unacceptable. @ElizabethHeng's story should not be censored. @Facebook, how come you cant figure this out? And you wonder why conservatives think you are biased. Do better. https://t.co/1XjP9VAa6F	Negative
16	29.-30.08.18	Removed post about Holocaust Education	108	Social Capital	RT @AnneFrankCenter: Hi @Facebook, you removed our post promoting the need for Holocaust Education for apparently violating community standards. You haven't given us a reason, yet allow Holocaust Denial pages to still exist. Seems a little hypocritical?(the post was the exact same as the tweet below) https://t.co/H4bYTdEQp3	Negative
17	31.08.18	Pro-Trump meme creator gets shut down	61	Leadership & Governance	RT @Carpedonktum: Today @Facebook shut down my Facebook page without warning. All I do is create funny Trump videos. My videos are not explicit, they do not contain violence or sexuality, and they are not hateful. No specific explanation was given. #StopTheBias Example: https://t.co/WNleDdKx1e	Negative
18	07.09.18	Censorship of black Women-Duo supporting Trump	28	Leadership & Governance	RT @DiamondandSilk: If @Facebook didn't know who we were, why did they verify our page with a blue check mark in 2015? We talked about "TRUMP" and promoted several posts with the name TRUMP included. Now FB has a case of amnesia like they don't know who we are. Undercover Censorship at its finest.	Negative
19	18.09.18	Lawsuit filed by the American Civil Liberties Union about practices of gender discrimination on ads	32	Social Capital	RT @ACLU: BREAKING: We've filed charges against @Facebook and 10 employers for using the platform to target their job ads for positions in male-dominated fields only to younger men. Facebook is violating federal civil rights law. Period.	Negative
20	01.11.18	Removed political pages and accounts	13	Leadership & Governance	RT @BrianKolfage: My family lives the sacrifice daily after I lost 3 limbs in combat. @Facebook has taken away my ability to support my family by deleting my businesses SIGN THIS PETITION https://t.co/jrjefNxkdF @RyanAFournier @TIMENOUT @WeSupport45 @DeepStateExpose @andersonDrLJA @kwilli1046	Negative
21	31.10.-03.11.18	Removed pro-life ads	88	Social Capital	BREAKING: @Facebook just banned ANOTHER one of our ads - this time in Montana. This is getting ridiculous, @SherylSandberg. What's the deal? Is #ProLife speech welcome on your platform or not? #MTSen @MattForMontana #IVoteProLife Watch the ad Facebook banned: https://t.co/skSGlrglIG	Negative
					If you want to know why @Facebook and @Google are blatantly censoring ads promoting Marsha Blackburn and opposing Phil Bredesen, you don't have to look very hard. https://t.co/8x9RVXoR2J	Negative
					Why is Facebook so intent on disenfranchising Charlotte & Micah? They deserve to be allowed to tell their stories. But @Facebook is shutting down our ads that share how they survived premature births. Charlotte ad - banned for being "too graphic" - decide for yourself: https://t.co/DzRPl69phq	Negative
22	05.-06.11.18	Removed Donald Trump Campaign Ad	117	Social Capital	So, to summarize today's ad action: @CNN, @NBC, @FoxNews and @Facebook have all rejected an ad by the *President of the United States* because they consider it racist. Amazing.	Negative

Table D.4: Facebook Events

#	date	event description	#tweets	ESG dimension	example tweets	polarity
23	09.11.18	Information on Camp Fire	13	Social Capital	I don’t have a @Facebook account, but here is a good follower who has set up a page for info on #CampFire related events and people. #CampFireJamesWoods https://t.co/vATEtVGFOH	Positive
24	12.11.18	Fired top executive because of donation to anti-Hillary Clinton group	12	Leadership & Governance	Why did @Facebook fire a top executive? Because he was a @RealDonaldTrump supporter https://t.co/MVfIJs1A1Y via @WSJ	Negative
25	14.-15.11.2018	Unethical lobbying in aftermath of Cambridge Analytica	69	Leadership & Governance	Here’s the whole incredible story. Kudos to @nytimes. Without it, we’d never know @Facebook execs castigated employees for investigating Russian interference. Then smeared its critics with anti-semitic tropes that it lifted straight from the Kremlin https://t.co/mOXJERzFq9	Negative
26	14. -15.11.2018	Facebook groups for white supremacists and far-right	27	Social Capital	Thread. Two weeks ago, media began reporting that @Facebook removed Pages and Groups belonging to extremist ”Proud Boys”. How’s that going? Let’s look at the data! Over the past 18 months, I’ve tracked 156 Proud Boys (PB) and Alt-Knights (AK) groups recruiting on Facebook.	Negative
					White supremacists and other hate groups are using platforms like @Facebook, @Twitter, and @Google to organize, fund, and recruit online. Take action and tell tech companies to #ChangeTheTerms by adopting policies to stop dangerous hate before it spreads. https://t.co/3ndDlQWROc https://t.co/LEhcr1vBzv	Negative
					From demanding protection for Black leaders doxxed by white supremacists in a FB group, to ensuring transparent & just @Facebook moderation policies, to demanding an end to racially targeted digital voter suppression, we’ve sat across the negotiation table w/ @Facebook for years	Negative
27	19.-20.11.18	Allowed post of violent and racist content	71	Social Capital	That the right is mobilizing fake quotes, conspiracies + outright hatred, particularly on Congresswomen of color, shows how vapid they are on actual issues. The racism allowed towards Ilhan + others is completely unacceptable. @Facebook has clearly lost control of their product. https://t.co/vqRaM3iAjX	Negative
28	05.12.18	Bias towards democrats	17	Leadership & Governance	It’s time to #StopTheBias. RETWEET if you’re sick of conservative voices being silenced on @Twitter, @Facebook, and @Google. https://t.co/tzGh5YBtNZ	Negative
29	05.-06.12.18	Six4Three: internal Facebook documents about unethical data policies	45	Social Capital	2) Facebook engineered ways to access user’s call history w/o alerting users: Team considered access to call history considered ’high PR risk’ but ’growth team will charge ahead’. @Facebook created upgrade path to access data w/o subjecting users to Android permissions dialogue. https://t.co/Oth6WF2oVa	Negative
30	28.12.18	Bot campaign against Republican candidate	19	Leadership & Governance	BOOM! Cybersecurity Expert: It Looks Like @Twitter and @Facebook Were Complicit in Bot Campaign Against Roy Moore (Video) https://t.co/bQ9D0OKQFK via @gatewaypundit	Negative
31	29.-31.12.18	Banned evangelist Franklin Graham	108	Social Capital	RT @Franklin_Graham: Last week I was banned from posting on @Facebook for 24 hrs because of a 2016 post about NC’s House Bill 2 (bathroom bill). They said the post went against their ”community standards on hate speech. Facebook is making & changing the rules. 1/2 https://t.co/HYIgErnp3J	Negative
			38		RT @Franklin_Graham: I thank @Facebook for their apology and I accept it. All truth is in the Lord Jesus Christ, who is ”the Way, the Truth, and the Life.” I would encourage all Christians as well as Facebook to stand on God’s Word and His truth. https://t.co/rgEsOZna3C	Positive

Table D.5: Facebook Events

List of Figures

List of Tables

Bibliography

[20] *Report on US Sustainable and Impact Investing Trends*. Source. Nov. 2020.

[Alb13] Elisabeth Albertini. "Does environmental management improve financial performance? A meta-analytical review". In: *Organization & Environment* 26.4 (2013), pp. 431–457.

[Ant20] Madelyn Antoncic. "Uncovering hidden signals for sustainable investing using Big Data: Artificial intelligence, machine learning and natural language processing". In: *Journal of Risk Management in Financial Institutions* 13.2 (2020), pp. 106–113.

[AS18] Amir Amel-Zadeh and George Serafeim. "Why and how investors use ESG information: Evidence from a global survey". In: *Financial Analysts Journal* 74.3 (2018), pp. 87–103.

[BBC10] BBC. *US markets plummet amid eurozone debt crisis fears*. Source. 2010.

[BBC12] BBC. *Timeline: The unfolding eurozone crisis*. Source. 2012.

[BBC20] BBC. *A quick guide to the US-China trade war*. Source. 2020.

[BC05] Gregory W Brown and Michael T Cliff. "Investor sentiment and asset valuation". In: *The Journal of Business* 78.2 (2005), pp. 405–440.

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[Ber+11] James Bergstra et al. "Algorithms for hyper-parameter optimization". In: *25th annual conference on neural information processing systems (NIPS 2011)*. Vol. 24. Neural Information Processing Systems Foundation. 2011.

[BKR19] Florian Berg, Julian F Koelbel, and Roberto Rigobon. *Aggregate confusion: The divergence of ESG ratings*. MIT Sloan School of Management, 2019.

[Bol+96] Michael D Boldin et al. "A check on the robustness of Hamilton's Markov switching model approach to the economic analysis of the business cycle". In: *Studies in Nonlinear Dynamics and Econometrics* 1.1 (1996), pp. 35–46.

[Bra97] Andrew P Bradley. "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[Bus18] Business Insider. *Facebook apologises to the Anne Frank Center for removing image of naked child holocaust victims.* Source. 2018.

[Bus20] Business Roundtable. *Statement on the Purpose of a Corporation.* Source. 2020.

[BW06] Malcolm Baker and Jeffrey Wurgler. "Investor sentiment and the cross-section of stock returns". In: *The journal of Finance* 61.4 (2006), pp. 1645–1680.

[BWY12] Malcolm Baker, Jeffrey Wurgler, and Yu Yuan. "Global, local, and contagious investor sentiment". In: *Journal of financial economics* 104.2 (2012), pp. 272–287.

[CC10] Denis Cousineau and Sylvain Chartier. "Outliers detection and treatment: a review." In: *International Journal of Psychological Research* 3.1 (2010), pp. 58–67.

[CEG20] Costanza Consolandi, Robert G. Eccles, and Giampaolo Gabbi. "Better Fewer, But Better: Stock Returns and the Financial Relevance and Financial Intensity of Materiality". In: *"Available at SSRN"* (2020). DOI: `http://dx.doi.org/10.2139/ssrn.3574547`.

[CFN95] Mark A Cohen, Scott Fenn, and Jonathan S Naimon. *Environmental and financial performance: are they related?* Citeseer, 1995.

[Cha+16] Aaron K Chatterji et al. "Do ratings of firms converge? Implications for managers, investors and strategy researchers". In: *Strategic Management Journal* 37.8 (2016), pp. 1597–1614.

[Cho03] Gobinda G Chowdhury. "Natural language processing". In: *Annual review of information science and technology* 37.1 (2003), pp. 51–89.

[Cho20] Noam Chomsky. *Syntactic structures.* De Gruyter Mouton, 2020.

[CNB18] CNBC. *US stocks post worst year in a decade as the S&P 500 falls more than 6% in 2018.* Source. 2018.

[Con17] Cone Communications. *CSR Study.* Source. 2017.

[De 89] Ferdinand De Saussure. *Cours de linguistique générale*. Vol. 1. Otto Harrassowitz Verlag, 1989.

[Del20] Deloitte. *Millennials Survey*. Source. 2020.

[Dev+19] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: `1810.04805 [cs.CL]`.

[DHN15] Gregor Dorfleitner, Gerhard Halbritter, and Mai Nguyen. "Measuring the level and risk of corporate responsibility – An empirical comparison of different ESG rating approaches". In: *Journal of Asset Management* 16.7 (2015), pp. 450–466. DOI: `https://doi.org/10.1057/jam.2015.31`.

[DKL13] Xin Deng, Jun-koo Kang, and Buen Sin Low. "Corporate social responsibility and stakeholder value maximization: Evidence from mergers". In: *Journal of financial Economics* 110.1 (2013), pp. 87–109.

[DMC16] Ian Dewancker, Michael McCourt, and Scott Clark. "Bayesian optimization for machine learning: A practical guidebook". In: *arXiv preprint arXiv:1612.04858* (2016).

[Eis13] Jacob Eisenstein. "What to do about bad language on the internet". In: *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*. 2013, pp. 359–369.

[EK19] Robert G Eccles and Svetlana Klimenko. "The investor revolution". In: *Harvard Business Review* 97.3 (2019), pp. 106–116.

[FBB15] Gunnar Friede, Timo Busch, and Alexander Bassen. "ESG and financial performance: aggregated evidence from more than 2000 empirical studies". In: *Journal of Sustainable Finance & Investment* 5.4 (2015), pp. 210–233. DOI: `10.1080/20430795.2015.1118917`. eprint: `https://doi.org/10.1080/20430795.2015.1118917`. URL: `https://doi.org/10.1080/20430795.2015.1118917`.

[Fed09] Federal Reserve Bank of Atlanta. *Federal Reserve Bank of Atlanta - Stock Prices in the Financial Crisis*. Source. 2009.

[FF15] Eugene F Fama and Kenneth R French. "A five-factor asset pricing model". In: *Journal of financial economics* 116.1 (2015), pp. 1–22.

[FG67] Donald E Farrar and Robert R Glauber. "Multicollinearity in regression analysis: the problem revisited". In: *The Review of Economic and Statistics* (1967), pp. 92–107.

[FHT+01] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[Fin18] Financial Times. *Adidas vows to use only recycled plastics by 2024*. Source. 2018.

[For16] Forbes. *Everything You Need To Know About The 2016 Stock Market Selloff*. Source. 2016.

[Fre] Kenneth R French. *Website - Data Library*. Source.

[Ges21] Flora Geske. "Social-Media based, NLP-powered ESG Score Methodology". In: (2021).

[Goo+16] Ian Goodfellow et al. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.

[Gra13] Alex Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013).

[Gre00] William H Greene. "Econometric analysis 4th edition". In: *International edition, New Jersey: Prentice Hall* (2000), pp. 201–215.

[Ham10] James D Hamilton. "Regime switching models". In: *Macroeconometrics and time series analysis*. Springer, 2010, pp. 202–209.

[Ham89] James D Hamilton. "A new approach to the economic analysis of nonstationary time series and the business cycle". In: *Econometrica: Journal of the econometric society* (1989), pp. 357–384.

[HCB13] Bo Han, Paul Cook, and Timothy Baldwin. "Lexical normalization for social media text". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 4.1 (2013), pp. 1–27.

[Hen+19] Roy Henriksson et al. "Integrating ESG in portfolio construction". In: *The Journal of Portfolio Management* 45.4 (2019), pp. 67–81.

[HKN19] Witold Henisz, Tim Koller, and Robin Nuttall. "Five ways that ESG creates value". In: (2019).

[HKR15] Wolfgang Karl Härdle, Sigbert Klinke, and Bernd Rönz. *Introduction to statistics: using interactive MM* Stat elements*. Springer, 2015.

[HLS13] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[HM13] Haibo He and Yunqian Ma. "Imbalanced learning: foundations, algorithms, and applications". In: (2013).

[HNP09] Alon Halevy, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data". In: *IEEE Intelligent Systems* 24.2 (2009), pp. 8–12.

[HTW16] Yu He, Qingliang Tang, and Kaitian Wang. "Carbon performance versus financial performance". In: *China Journal of Accounting Studies* 4.4 (2016), pp. 357–378. DOI: 10.1080/21697213.2016.1251768. eprint: https://doi.org/10.1080/21697213.2016.1251768. URL: https://doi.org/10.1080/21697213.2016.1251768.

[Jeb19] Ruth Jebe. "The Convergence of Financial and ESG Materiality: Taking Sustainability Mainstream". In: *American Business Law Journal* 56.3 (2019), pp. 645–702.

[Jor07] Philippe Jorion. *Value at risk: the new benchmark for managing financial risk*. The McGraw-Hill Companies, Inc., 2007.

[KB14] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KÇ18] Emrah Keleş and Ayten Çetin. "Corporate Social Responsibility, Investor Sentiment, and Stock Returns". In: *Sustainability and Social Responsibility: Regulation and Reporting*. Springer, 2018, pp. 443–462.

[Kim94] Chang-Jin Kim. "Dynamic linear models with Markov-switching". In: *Journal of Econometrics* 60.1-2 (1994), pp. 1–22.

[KSY16] Mozaffar Khan, George Serafeim, and Aaron Yoon. "Corporate Sustainability: First Evidence on Materiality". In: *The Accounting Review* 91.6 (2016), pp. 1697–1724.

[Lid01] Elizabeth D Liddy. "Natural language processing". In: (2001).

[LPM15] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015).

[LW18] Tao Liu and Wing Thye Woo. "Understanding the US-China trade war". In: *China Economic Journal* 11.3 (2018), pp. 319–340.

[MA13] André Varella Mollick and Tibebe Abebe Assefa. "US stock returns and oil prices: The tale from daily data and the 2008–2009 financial crisis". In: *Energy Economics* 36 (2013), pp. 1–18.

[Mar20] Markets and Markets. *Plant-Based Meat Market Report*. Source. 2020.

[Mik+13a] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *arXiv preprint arXiv:1310.4546* (2013).

[Mik+13b] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[Mor19] Morgan Stanley Institute for Sustainable Investing. *Sustainable Signals - The Individual Investor Perspective*. Source. 2019.

[MSC] MSCI. *ESG Rating Methodology*. Source.

[Nem+19] Azadeh Nematzadeh et al. “Empirical study on detecting controversy in social media”. In: *arXiv preprint arXiv:1909.01093* (2019).

[NVN20] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. “BERTweet: A pre-trained language model for English Tweets”. In: *arXiv preprint arXiv:2005.10200* (2020).

[Oli03] David J Olive. *Linear regression analysis*. 2003.

[Pen20] Pensions&Investments. *Global ESG-data driven assets hit $40.5 trillion*. Source. 2020.

[Pri] Principles of Responsible Investing. *About the PRI*. Source.

[Pri13] Nicolas Privault. “Understanding Markov Chains”. In: *Examples and Applications, Publisher Springer-Verlag Singapore* (2013).

[PSK19] M Porter, George Serafeim, and Mark Kramer. “Where ESG fails”. In: *Institutional Investor* (2019), pp. 1–50.

[RB10] Jason M Ribando and George Bonne. “A new quality factor: Finding alpha with ASSET4 ESG data”. In: *Starmine Research Note, Thomson Reuters* 31 (2010).

[RBN20] Natraj Raman, Grace Bang, and Armineh Nourbakhsh. “Mapping ESG Trends by Distant Supervision of Neural Language Models”. In: *Machine Learning and Knowledge Extraction* 2.4 (2020), pp. 453–468.

[RE1] RE100. *Members*. Source.

[Ref] Refinitiv. *Next-level NLP and potential ESG controversies*. Source.

[Ref21] Refinitiv. *Refinitiv/Thomson Reuters Sustainability Scores*. Source. 2021.

[Reu16] Reuters. *What’s behind the global stock market selloff?* Source. 2016.

[RR01] Paul Rozin and Edward B Royzman. “Negativity bias, negativity dominance, and contagion”. In: *Personality and social psychology review* 5.4 (2001), pp. 296–320.

[RU+00] R Tyrrell Rockafellar, Stanislav Uryasev, et al. “Optimization of conditional value-at-risk”. In: *Journal of risk* 2 (2000), pp. 21–42.

[Rud17] Sebastian Ruder. *An overview of gradient descent optimization algorithms*. 2017. arXiv: `1609.04747 [cs.LG]`.

[Ser20] George Serafeim. “Public Sentiment and the Price of Corporate Sustainability”. In: *Financial Analysts Journal* 76.2 (2020), pp. 26–46. DOI: `10.1080/0015198X.2020.1723390`. eprint: `https://doi.org/10.1080/0015198X.2020.1723390`. URL: `https://doi.org/10.1080/0015198X.2020.1723390`.

[Sha94] William F Sharpe. “The sharpe ratio”. In: *Journal of portfolio management* 21.1 (1994), pp. 49–58.

[SHQ19] Chi Sun, Luyao Huang, and Xipeng Qiu. “Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence”. In: *”Available at arXiv”* (2019).

[SL09] Marina Sokolova and Guy Lapalme. “A systematic analysis of performance measures for classification tasks”. In: *Information processing & management* 45.4 (2009), pp. 427–437.

[SLL98] Mike EC P So, Kin Lam, and Wai Keung Li. “A stochastic volatility model with Markov switching”. In: *Journal of Business & Economic Statistics* 16.2 (1998), pp. 244–253.

[Sok+20] Alik Sokolov et al. “Building Machine Learning Systems to Automate ESG Index Construction”. In: (2020).

[SP a] S&P Global. *ESG Rating Methodology*. Source.

[SP b] S&P Global. *How can AI help ESG Investing?* Source.

[Spe19] Spencer Stuart. *US Board Index*. Source. 2019.

[Susa] Sustainability Accounting Standards Board. *Mission of the Sustainability Accounting Standards Board (SASB) Foundation*. Source.

[Susb] Sustainalytics. *ESG Rating Methodology*. Source.

[Sus20] SustainAbility. *Rate the Raters*. Source. 2020.

[Sus21] Sustainability Accounting Standards Board. *SASB Materiality Map®*. Source. 2021.

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *arXiv preprint arXiv:1409.3215* (2014).

[SYY12] Robert F Stambaugh, Jianfeng Yu, and Yu Yuan. “The short of it: Investor sentiment and anomalies”. In: *Journal of Financial Economics* 104.2 (2012), pp. 288–302.

[The11] The Guardian. *Debt crisis sends financial markets into turmoil*. Source. 2011.

[The18a] The Guardian. *"Over \$119bn wiped off Facebook's market cap after growth shock"*. Source. 2018.

[The18b] The Guardian. *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*. Source. 2018.

[The18c] The Guardian. *The key moments from Mark Zuckerberg's testimony to Congress*. Source. 2018.

[The20] The Business Research Company. *Ethical Fashion Market Report*. Source. 2020.

[Tho+20] Ph.D. Thomas Kuh et al. *Dynamic Materiality^{TM}: Measuring what Matters*. Source. Jan. 2020.

[Tim18] Time. *Mark Zuckerberg Was Just Asked About Diamond And Silk. Here's What to Know About the Trump-Supporting Sisters*. Source. 2018.

[TVY18] Joseph Taylor, Joseph Vithayathil, and Dobin Yim. "Are corporate social responsibility (CSR) initiatives such as sustainable development and environmental policies value enhancing or window dressing?" In: *Corporate social responsibility and environmental management* 25.5 (2018), pp. 971–980.

[Uni12] United Nations. *United Nations Sustainable Development Goals*. Source. 2012.

[US 99] U.S. Securities and Exchange Commission. *SEC Staff Accounting Bulletin No. 99, Materiality*. Source. 1999.

[Vas+17] Ashish Vaswani et al. "Attention is all you need". In: *arXiv preprint arXiv:1706.03762* (2017).

[Vel17] Patrick Velte. "Does ESG performance have an impact on financial performance? Evidence from Germany". In: *Journal of Global Responsibility* (2017).

[Was18] Washington Post. *Facebook told two women their pro-Trump videos were 'unsafe'*. Source. 2018.

[Wol+20] Thomas Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. 2020. arXiv: `1910.03771 [cs.CL]`.

[Yil20] Hakan Yilmazkuday. "COVID-19 Effects on the S&P 500 index". In: *Available at SSRN 3555433* (2020).

[You+18] Tom Young et al. "Recent trends in deep learning based natural language processing". In: *ieee Computational intelligenCe magazine* 13.3 (2018), pp. 55–75.

[Zag+10] Rudi Zagst et al. “Responsible Investing”. In: (2010).

[ZB17] Martina Zimek and Rupert Baumgartner. “Corporate sustainability activities and sustainability performance of first and second order”. In: *18th European Roundtable on Sustainable Consumption and Production Conference (ERSCP 2017), Skiathos Island, Greece*. 2017, pp. 1–5.