

Amsterdam University of Applied Science

Digital Driven Business

June 2021

Master Project, Digital Driven Business

A description of text mining technique to extract Corporate Sustainability Reputation from online reviews

Author:

Arta Demaj, Student Nr. 500849744

Supervisors:

Frederik Situmeang,

Jesse Weltevreden

Table of Contents

1	Introduction	3
2	Literature review	6
3	Methodology	9
3.1	Text refining	10
3.1.1	Selection and grouping of indicators.....	10
3.1.2	Data collection	10
3.1.3	Approximate string matching.....	11
3.1.4	Data pre-processing	11
3.1.5	Sentiment analysis	12
3.1.6	Aggregation to the CSI.....	13
3.2	Knowledge distillation	14
3.3	Reliability and validity	14
3.3.1	Reliability.....	14
3.3.2	Validity	14
3.4	Ethics	14
4	Results	15
5	Discussions	17
6	Limitations and future research	18
7	References.....	19
8	Certification of authorship	25

List of tables, figures and equation

Table 1: Summary of available sustainability metrics	8
Table 2: Text pre-processing steps.....	11
Table 3: Environmental sustainability word list	12
Table 4: Social sustainability word list.....	12
Table 5: Economic sustainability word list	12
Table 6: Part of the results	15
Table 7: Correlation matrix	15
Table 8: Detailed statistics results.....	16
Figure 1: Three pillars of sustainability (Jiang et al., 2018)	6
Figure 2: Framework for extracting sustainability and sentiment analysis.....	9
Figure 3: Visual correlation matrix	15
Equation 1: Corporate sustainability index	13
Equation 2: Social index	13
Equation 3: Environmental index.....	13
Equation 4: Economic index	13

1 Introduction

Environmental and social concerns have gained interest and importance over the last 20 years. Stakeholders are increasingly demanding that companies take responsibility for the environmental, economic, and social impacts of their activities (Miralles-Quiros et al., 2017). This is reflected in the expansion of corporate sustainability (CS) initiatives. There are many definitions of CS, but the most widely used framework is based on Freeman's (1984) stakeholder theory. This theory states that companies have obligations to individuals and legal entities or groups both inside and outside the company. In addition, it is widely accepted that sustainable development includes environmental, economic, and social dimensions (Hahn et al., 2014; Meadows, 1972), which is also referred to as the 3Ps (profit, people, planet) approach to business or the "triple bottom line" (Antolín-López et al., 2016). Sustainability requirements and goals vary by stakeholder and include requiring companies to minimize pollution, use resources more efficiently, improve community relations, or ensure economic progress (Benn et al., 2014). This close linkage between stakeholder requirements presents a complex problem for CS initiatives for which there are no universal approaches (Searcy, 2012).

Nevertheless, corporate sustainability performance measurement (CSPM) and international rankings have emerged to quantify sustainable development (e.g., Dow Jones Sustainability Index [DJSI], Global 100, Newsweek Green Rankings) (Antolín-López et al., 2016). Despite their importance for companies, CSPM instruments have been little researched (Chelli & Gendron, 2013; Maas & Reniers, 2014). The literature on sustainability indicators, assessments, and reports is currently very extensive, however, systems for measuring sustainability performance are rarely mentioned (Searcy & Pryshlakivsky, 2017). Most sustainability studies focus on macroeconomic aspects such as industries (Labuschagne et al., 2005), cities (Shen et al., 2011), or regions (Wallis, 2006). Furthermore, the available studies on sustainability performance at the firm level mainly analyze a single dimension of CS – environmental sustainability (Delmas et al., 2013; Herva et al., 2011; Winn & Pogutz, 2013). Antolín-López et al. (2016) noted in their analysis that despite the wide acceptance of the three dimensions (environmental, economic, and social), both in the literature and among practitioners, there is still a great deal of variability in the way different CSPM tools operationalize the three dimensions, which leads to compartmentalization between the CS dimensions in measuring sustainable development. According to their analysis, one of the reasons for these differences is the way social and economic (socioeconomic) aspects are linked. Another aspect that they identified is the way companies contend with different stakeholder demands, ever-changing priorities, and a variety of alternatives to addressing their sustainability challenges. According to their study, this results in different implications for companies to consider with regard to both internal and external stakeholders. For example, managers weighing a sustainability strategy are confronted with sustainability measurements and strategies. The measures taken and sustainability targets set will influence investors who, especially in Europe, are increasingly making investment decisions based on companies' environmental and social concerns. Moreover, policy makers will consider sustainability challenges and set new sustainability benchmarks for companies.

The issue of sustainability also impacts stakeholders, such as customers or activists, who learn about or advocate for social and environmental concerns (Miralles-Quiros et al., 2017). For these reasons, it is important for companies to understand what CS means and how to measure and report their performance. The importance of addressing environmental and social concerns is also evident in the positive effect that sustainability investments have had on firm performance (Lai et al., 2010; Price & Sun, 2017) and equity (Heinberg et al., 2018; Wang & Sengupta, 2016).

Due to intensifying attention on CS and corporate social responsibility (CSR), CSPM and sustainability indices have emerged that have been linked to financial markets (López et al., 2007). According to Cowan and Guzman (2020), investments in CSR or CS not only affect market value but also affect consumers' perceptions of sustainability. There are also other studies that have found that both positive and negative sustainability-related information on social media significantly influences consumers' purchase intentions for sustainable products (Saeed et al., 2019). However, studies on consumer perceptions of sustainability are limited (D'Acunto et al., 2020). Since consumers' perceptions play a considerable role in CSR initiatives, this paper calculates a corporate sustainability index (CSI) based on consumer perceptions and analyzes whether the CSI is related to stock market performance.

Online reviews are important sources for obtaining consumer information. As several studies have shown, online reviews have become an important part of consumers' decision-making processes and have an economic impact (Chevalier & Mayzlin, 2006; Moe & Trusov, 2011; Schoenmueller et al., 2020). Online retailers have recognized the effectiveness of customer reviews and use them to convert online customer visits into online sales (Ludwig et al., 2013). Online reviews have been used by researchers, mainly through text analysis or numerical quality diagnostics (such as review volume or star ratings), to show their impact on sales or business reputation (Babic Rosario et al., 2016; Floyd et al., 2014; Hollenbeck, 2018; Ludwig et al., 2013; Zhang et al., 2013). However, there are few studies that have used online reviews to analyze consumer perceptions of sustainability (D'Acunto et al., 2020). Since the analysis of online reviews truthfully depicts customer perspectives, this work examines online reviews to calculate a sentiment-based CSI for companies. The importance of online reviews is reflected in the large number of review platforms where consumers provide other consumers with information about goods, services, brands, or companies. In marketing, online communication between consumers is referred to as "electronic word of mouth" or eWOM (Babic Rosario et al., 2016). One of the most popular review platforms is Trustpilot (www.trustpilot.com), which has more than 124 million published reviews as of 2021 and is among the top 1 % most-popular websites according to Alexa ranking (Trustpilot, 2021a). However, few studies have considered such online review platforms as sources for CSR research. Given the increasing consumer attention on CSR, the consideration of online reviews is relevant (D'Acunto et al., 2020). The aim of this project is to determine to what extent, if at all, online reviews are relevant measures of CSR.

Given that there is insufficient research in this area, theoretical and business insights into the influence of customer reviews on the CSI are provided in this work. In this context, online consumer reviews from Trustpilot were analyzed, a CSI was calculated, and the scores were compared with the stock prices of listed companies. A new method for calculating CS indices based on online ratings was used to investigate whether and to what extent online ratings influence CSR. This method will allow companies to analyze large amounts of unstructured data and quantify qualitative information. According to Miralles-Quiros et al. (2017), politics plays a relevant role in sustainability. Their study demonstrated that there are differences between national governments regarding sustainability standards and uniform public guidelines have not been developed. Due to this constraint, and to strengthen the comparison between companies, the tool in this work was constructed using online reviews, sustainability indices, and stock prices of publicly-traded companies based only in the United States. This tool is ideal for companies in English-speaking countries that want to gain information about consumers' sustainability perceptions from online customer reviews.

This work answers the following main question (MQ) and three additional sub-questions (SQs): (MQ) To what extent can online ratings and reviews of a U.S. company be used to measure its sustainability reputation, and are there relationships between the calculated CSI and market performance?

(SQ1) How can the different types of sustainability mentioned in online reviews be classified to measure a company's sustainability reputation?

(SQ2) How can online reviews be quantified to calculate the CSI for all three dimensions (environmental, social, and economic)?

(SQ3) What is the impact of the organization's sustainability index on the organization's market performance?

To answer these questions, this work is structured as follows. First, the background and theoretical underpinnings of the study are presented. Then, the methodology is discussed, including the process for obtaining the variables and the analytical framework. Finally, the results and implications are presented.

2 Literature review

Corporate sustainability reputation

Businesses play a major role in sustainable development because they claim the resources of society (Bansal, 2002). In order for resources to be used in a sustainable manner and for companies to be held accountable for the resources they use, the concept of CS has emerged (López et al., 2007). Sustainable development is a concept that defines desirable developments in society (Bansal, 2002; Dyllick & Hockerts, 2002; Gladwin et al., 1995; Jennings & Zandbergen, 1995; Shrivastava, 1995). CS requires companies to pursue not only economic development but also sustainable development (Hahn et al., 2014). This means that CS, in terms of the triple bottom line, requires companies to consider environmental, social, and economic goals simultaneously, which results in many different interdependent goals and measurements (Elkington, 1998; Hahn et al., 2014). Since all three dimensions must be included, decision makers are confronted with complex situations with numerous measurements, constraints, and challenges. Furthermore, each goal involves different problems and the solution to one problem can have an impact on the solution to another problem (Newton, 2002). Another challenge for corporate management is the simultaneous and often conflicting demands of multiple stakeholders (Hahn, 2012). Moreover, sustainability demands often come from "secondary" stakeholders such as social activists, nongovernmental organizations, and local communities (Clarkson, 1995; Fineman & Clarke, 1996), which creates conflicting pressures that cannot be resolved through conventional market transactions (Hall & Martin, 2005). In summary, CS refers to the interdependent environmental, social, and economic concerns at different levels of society that companies are expected to address simultaneously while managing tensions and contradictions between different aspects of sustainability (Hahn et al., 2014).

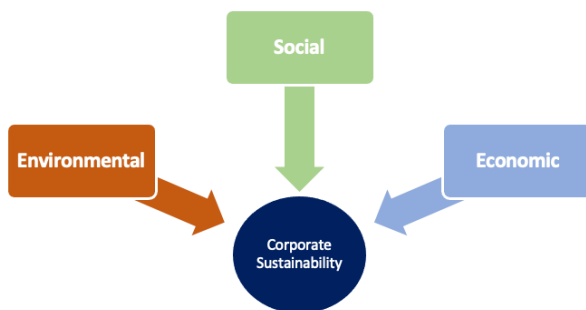


Figure 1: Three pillars of sustainability (Jiang et al., 2018)

Corporate sustainability assessment

Sustainability indicators are used to identify and monitor the three dimensions of CS (Azapagic & Perdan, 2000). Some authors see CSPM as the basis for the decision-making process for CS management strategies. Others see CS performance as a stand-alone component within the broader corporate sustainability assessment (CSA; (Pranugrahaning et al., 2021). The public can use CSI to understand the extent to which companies achieve economic profit, consider environmental and social responsibility, and whether stakeholders are considered in their business activities (Jiang et al., 2018). In addition, CSPM portrays sustainability issues in quantifiable terms and helps companies manage important sustainability concerns (Azapagic, 2004). Sustainability performance indicators are used to capture, measure, monitor, and disclose a company's CS performance.

It has become a trend for companies to voluntarily disclose not only their financial performance but also their environmental and social performance in corporate reporting (Searcy, 2012). The indicators quantifiably explain sustainability issues in terms of the three dimensions and help manage important short- and long-term sustainability concerns (Azapagic, 2004; Azapagic & Perdan, 2000; Searcy, 2012). Qualitative or quantitative data can be used to calculate indicators but quantitative indicators are needed for measurement because they are standardized and comparable (Schneider & Meins, 2012). As a result, many sustainability assessment frameworks have emerged that focus on business performance. According to Krajnc and Glavič (2005), only individual indices, such as social selling index and global salmon initiative, consider all three dimensions. It has been noted that many consider only one or two dimensions at a time. For this reason, Krajnc and Glavič (2005) have collected and developed a set of sustainability indicators that represent the important aspects of sustainable development. However, their framework is imperfect because different units of calculation are included, which makes comparison across companies and industries difficult. To combine many sustainability aspects and reduce the number of crucial factors, Krajnc and Glavič (2005) suggest the use of a composite sustainable development index. However, it is challenging to create an index that can be used both to efficiently assess CS and to benchmark companies within a given industry (Krajnc & Glavič, 2005).

Research exists that focuses on the CSA performance of individual companies or industries. These indicators measure and monitor actual sustainability performance, which companies have typically identified in CS strategies (Arthur et al., 2017). However, this will not enable the calculation of the overall sustainability of industries. To make this possible, studies also exist that aim to measure and evaluate aggregate sustainability performance. To achieve this, several indicators are evaluated together to obtain an overall score (Jiang et al., 2018). The method presented in this work is more informative because this CSA provides for the measurement and comparison of consumer reviews of multiple U.S. companies on Trustpilot. Table 1 shows the existing sustainability metrics that are used to measure the three dimensions. In order to quantify the interrelationships of the environmental, social and economic performance of a company, various methods such as life cycle assessment, life cycle costing, social life cycle assessment or principal component analysis (PCA) have been used (Jiang et al., 2018). However, these integration methods are not practical due to the subjective weightings and the enormous time required to calculate them. The three-dimensional sustainability assessment model of Jiang et al. (2018) forms a useful framework because it refers to the coordinated development of benefits from different levels such as resources, economy, environment, and society, and enables the consideration of stakeholders. The PCA method has been integrated into their model, which supports objectivity and information integrity (Mainali & Silveira, 2015; Nardo et al., 2005). Moreover, they combine a PCA and a coordination degree into a mathematical model, which merges the individual indicator results into a single CSI. Through the proposed CSA, economic, environmental, and social dimensions are assessed and the relationship between these three dimensions is analyzed.

The CSA models described in this chapter form the basis for the creation of Table 1. The sustainability indicators used in this work were chosen based on this table. To reflect all three dimensions in the results, the sustainability indicators were differentiated as shown in Figure 1, above. Once the reviews were analyzed based on these indicators, the next step was to calculate the obtained datasets to construct the CSI. Thus, in this work, a three-dimensional sustainability assessment model was used for the first time in calculating a CSI based on online customer reviews for companies in the United States.

Metrics	Sustainability dimension	Used method	Data sources	References
Three-dimensional sustainability assessment methodology	Environmental, social, and economic	PCA	Literature review and industry survey	(Jiang et al., 2018)
Holistic sustainability index	Environmental, social, economic, and manufacturing	Analytic hierarchy process (AHP)	Literature review	(Harik et al., 2015)
Energy technology sustainability index	Environmental, social, economic, technical, and institutional	PCA	Literature research and expert forums	(Mainali & Silveira, 2015)
Sustainability indicators at EPA	Environmental, social, and economic	Searching, sorting and filtering the database	Literature review	(Fiksel et al., 2012)
Operational sustainability framework	Environmental, social, and economic	MCDA and monetary valuation	Literature review, CSR reports review from companies, analysis current frameworks & guidelines and industry survey	(Labuschagne et al., 2005)
Composite sustainable development index	Environmental, social, and economic	AHP	CSR reports review from two companies and analysis current frameworks & guidelines	(Krajnc & Glavič, 2005)

Table 1: Summary of available sustainability metrics

3 Methodology

The number of available online reviews is too large to process manually. Therefore, text mining, which attempts to uncover patterns and non-trivial information and knowledge from unstructured text, was applied in this project. Data mining, machine learning, natural language processing, information retrieval, and knowledge management techniques are used to regularize information overload (Feldman & Sanger, 2007). Text mining consists of two main elements: text refining and knowledge distillation (Kantardzic, 2019). In text refining, the text to be analyzed is processed and prepared for knowledge distillation. The original and unstructured text is transformed into a data form suitable for analysis. In the second element, knowledge distillation, text patterns are derived and information is identified. This step can uncover new knowledge from multiple customer reviews. The goal of this project was to determine whether or not a CSI could be calculated by considering the frequency of sustainability-related terms and customer perceptions found in online customer reviews. The higher the frequency of a particular term, the higher the CSI will be and the better the sustainability reputation of the rated companies will be in the eyes of consumers. Text mining and term frequency analysis can also be used to identify in which of the three dimensions a company has a particularly good or bad reputation. The basic hierarchy for the calculation of the CSI from online reviews is shown in Figure 2. The procedure for calculating the CSI from online reviews is divided into several parts: selection and grouping of sustainability words, data collection, approximate string matching, data pre-processing, determining occurrence of indicators according to sustainable word lists, calculation of sentimentality and aggregation to the CSI. These procedural judgments are presented below under the classification of text refinement and knowledge distillation.

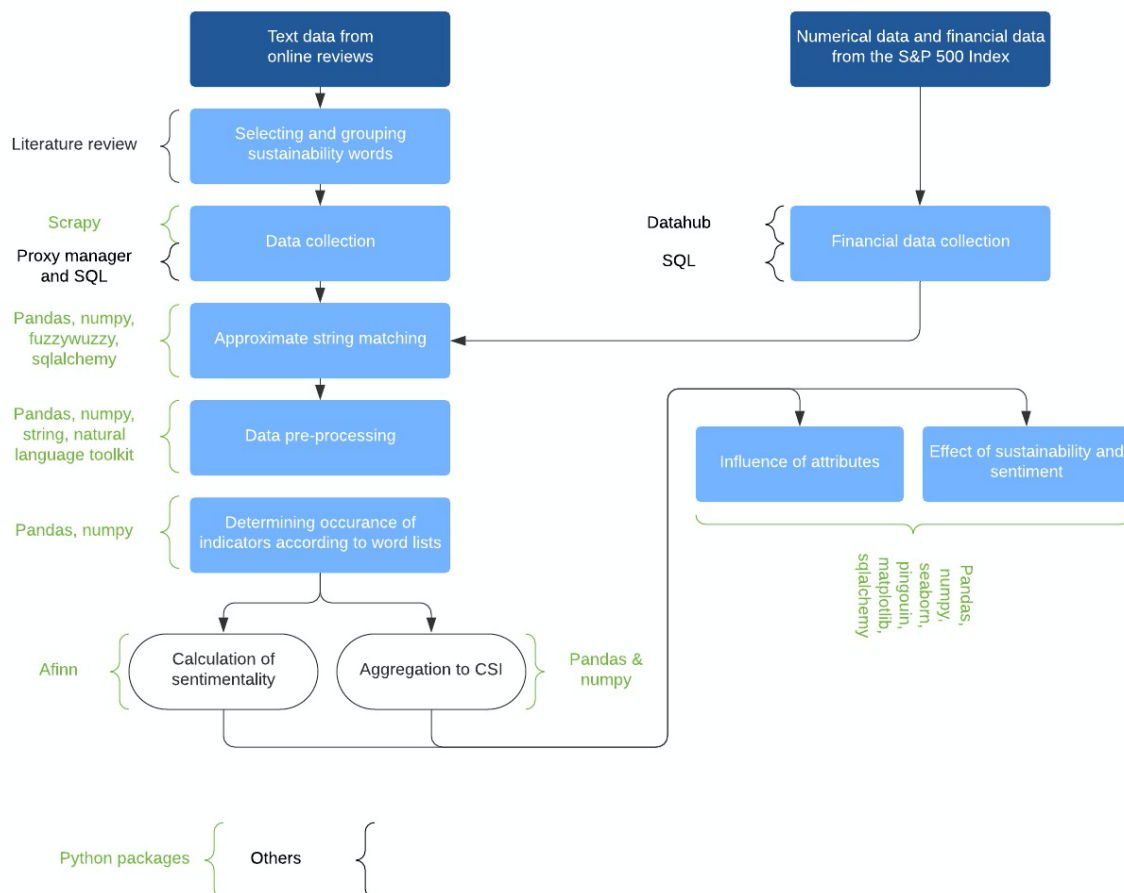


Figure 2: Framework for extracting sustainability and sentiment analysis

3.1 Text refining

3.1.1 Selection and grouping of indicators

First, indicators were selected and grouped into the three dimensions of sustainability. For this purpose, a literature review of papers focusing on the calculation of the CSI in relation to all three dimensions was undertaken. To guarantee reliable scientific literature, only articles published in Q1 level scientific journals were used for the determination and grouping of the indicators. The ranking of journals was checked on the Scimago Journal Ranking website (*Scimago Journal & Country Rank*, 2021).

3.1.2 Data collection

This work targeted U.S. companies listed in the S&P 500 stock market index. The primary data for this work came from the customer-review website, Trustpilot. To successfully collect all online reviews from Trustpilot, the Python package Scrapy and the Proxy Manager from Luminati Networks were used. The use of a proxy and therefore Proxy Manager was necessary to make Trustpilot more reliable to crawl. With a proxy, it was possible to make more requests to the target website without it being blocked or blocked because the proxy bypasses blanket IP blocks that some websites impose. Thus, a proxy helped maximize the accessibility and accuracy of information by storing the same data that a private user would see (*Luminati Networks*, n.d.). After successfully connecting the scraper to the proxy, a scraper was developed that automatically submit the data from online reviews on Trustpilot to a CSV file. A CSV file is a plain text file that stores tables and table information. CSV was chosen because this files can be easily imported and exported with Python or any other program language (Carnes, 2019). The files were used for text analysis and to facilitate text processing. Due to the high number of online reviews and limited storage capacity, the scraper was created and launched separately per category or industry, resulting in separate CSV output files, which were then cached in a database via SQL. SQL stands for Structured Query Language and is a language used by databases. With this language it is possible to manage information about tables and process data. With SQL, data or data objects can be added, deleted or updated faster than in CSV files (Larson, 2020). Only relevant information such as company name, category or industry, total reviews, and individual review text was collected. The dataset from Trustpilot reviews consisted of 22 CSV files in which a total of 33'889 companies and 30'545'722 million reviews were collected across the United States. Primary data from Datahub's website was also necessary for this project (*DataHub*, 2021). Datahub gives people access to high-quality data and tools. It is a website where high-quality datasets can be discovered and shared. In addition to providing data, it also develops tools and applications for data (Kariv & Pollock, 2018). The primary data involves directly downloading a CSV file containing financial information about the companies in the S&P 500 stock market index to analyze whether or not online reviews regarding sustainability can have an impact on company's market performance. To ensure that the reviews from subsidiaries of the listed companies were also included, the CSV file with financial information was manually expanded with a subsidiary list. The subsidiary information was retrieved from the U.S. Securities and Exchange Commission website (*U.S. Securities and Exchange Commission*, 2021). After that, the new CSV file was also cached in a database using SQL.

3.1.3 Approximate string matching

The aim of this research was to determine if and to what extent a sentiment-based CSI impacts the stock market. Python was used to combine the collected data and perform further steps as shown in Figure 2. Python was chosen because of the numerous and diverse approaches to data analysis and the availability of the latest Python tools, techniques, and open-source libraries (Taieb, 2018).

First, the Python package sqlalchemy was used to transfer the created SQL tables with online reviews and financial information into Python. Second, the dataset with online ratings from Trustpilot (T_df) and the dataset with financial data from companies listed in the S&P 500 stock market index (S&P_df) must be combined. By combining the two CSV datasets, a third data frame would be created where all reviews of parent companies and subsidiaries listed in the S&P 500 stock market index were recorded with their corresponding stock value. However, the company name column in T_df was sometimes different from the company name column in S&P_df and could not be merged. For example, reviews of McDonald's in T_df are listed under "McDonald's" and McDonald's financials in S&P_df are listed under "McDonald's Corp." Furthermore, Trustpilot sometimes only lists the names of the subsidiaries. For example, reviews of KFC can be found in T_df under "KFC" but the financial data of KFC in S&P_df is listed under "Yum! Brands, Inc" because KFC is a subsidiary of Yum! Brands, Inc. The python package fuzzywuzzy (Fuzzy Matching) was used to compile the data. With this package, the SQLite full-text search function is used, which allows two different data sets to be matched. After defining the correct columns for the necessary matching of the two data frames (i.e., company names), Fuzzy Matching determines the best match for each combination and stores the results in a new column showing the quality of the linkage. The linkage of the company names is displayed only if a quality rating of at least a 90% match is shown. All match qualities below 90% were excluded to prevent ambiguous results. To ensure the linkage of the companies that have longer company names in one data frame than in the other data frame, Fuzzy Matching was run twice: once using the first 10 letters and once using the last 20 letters. For the latter, 20 letters were used because the legal form is also noted in the company name in S&P_df and if fewer letters were used, in some cases only the legal form name would be compared, which would lead to ambiguous results.

3.1.4 Data pre-processing

For text processing, the open-source program pandas and numpy was used. The steps involved are listed in Table 2. After performing these steps, a new filtered column containing the cleaned review text was added to the dataset, which was crucial for the sentiment analysis.

Table 2: Text pre-processing steps.

Step 1: Classify datatypes	The scraped CSV files were saved in a SQL database and then connected to Python. The datatypes from each column were set correctly.
Step 2: Transform cases	The review text has been converted to lowercase to ensure that the same strings are equivalent in different cases.
Step 3: Tokenize	The text is broken down into smaller parts, so-called tokens, for processing. There are different ways to split the text. In this work, the text is split at non-letters, which means that a new token is generated at each punctuation mark or space. For example, the sentence "I ordered" can be decomposed into two-word tokens "I" and "ordered".
Step 4: Remove English stop words	English stop words are filtered from the text by using RapidMiner's built-in list of stop words. This reduces the number of tokens, since English function words make up a large portion of the texts. Examples of stop words are 'about', 'and', 'above', 'across', etc.

3.1.5 Sentiment analysis

Sentiment analysis was used in this report to find the emotional inventory behind the words used, with the goal of understanding consumer attitudes, opinions, and emotions (Thelwall & Buckley, 2013). In this work, the python package *afinn* was used to analyze the cleaned review dataset. The package is a word list-based approach for sentiment analysis. The *afinn* method estimates the strength of positive and negative sentiment in texts using a predefined sentiment word list (Miazga & Hachaj, 2019). The *afinn* lexicon contains scored words that are rated using a scale from -5 (meaning that the sentiment of the sentence is strongly negative) to +5 (which corresponds to a positive sentiment).

In total, there are four steps involved in producing a sentiment-based CSI. In the first step (1), the *afinn* sentiment score was created for all online reviews in the dataset. One of the disadvantages of *afinn* scores is that longer texts can yield higher scores because they contain more words. To compensate for this, the scoring value was divided by the number of words in the text and considered for further analysis. In the next step (2), word lists (Table 3 to 5) were developed, based on previous research (according to Table 1), to provide for distinctions between the three sustainability dimensions. In a further step (3), the sustainability of the companies was quantified using the relative occurrence of sustainability words. More details about this step can be found in the next chapter, 3.1.6. Finally (4), companies and industries were compared with each other.

Table 3: Environmental sustainability word list

energy efficiency, material efficiency, water consumption, resources, pollution, tree waste, emissions, air emissions, water runoff, solid waste, environmental protection, environmental investment, environmental expenditures, material use, energy use, environment prevention, air pollution, water pollution, solid waste, land use, gas emission, non-renewable material used, renewable material used, fuel consumption, water recycled, water reused, emission, biodiversity, atmosphere, land, oceans, seas, coast, fresh water, spills, hazardous waste, global warming, nuisance, product toxicity

Table 4: Social sustainability word list

interaction workers, interaction customers, interaction community, social impact, product liability, worker well-being, worker status, employment stability, human rights, health, safety, job satisfaction, education, training, customer satisfaction, local community, profit, taxes, justice, diversity, labor, social inclusion, indigenous people, employee turnover, strikes, lockouts, injury, disease, absenteeism, fatalities, discrimination, corruption, non-monetary sanctions, Product Stewardship Programme, equity, housing, security, population, job opportunity, employment compensation, career development, mobility infrastructure, social pathologies, social cohesion, fatalities, local communities, bribery, business ethics

Table 5: Economic sustainability word list

company size, growth capacity, assets, production, sales, profit, asset turnover, profitability, costs, equity return, raise cash, net working capital, debt repayment, debt-to-assets, technical innovation capability, research, development, market control, fiscal stewardship, internal controls, investment, independent directors, female directors, competences, stakeholder representation, consumption, financial benefits, exploration cost, safety fines, wealth creation, distribution wealth, pollution prevention

3.1.6 Aggregation to the CSI

To enable aggregation to the CSI, several variables had to be recognized and calculated. The first required variable, "Total number of reviews" ($Total_R$), was already captured by the data collection in the data frame. To enable the calculation of the three sustainability dimensions, the occurrence of each indicator (according to Tables 3 to 5) in the review text was counted and stored. This was done using a count vectorizer from the Python package Scikit-learn. A count vectorizer converts the words used into numbers. For each word used in a review, the count vectorizer creates a new column and this column is numbered with the number of words that match the sustainability words. If there is no match between the verification of the review and sustainability words, the column is numbered with a 0. Thus, the sustainability words used could be counted per review. Only words from the sustainability word list (Tables 3 to 5) were filtered out and stored with the calculated sentiment of each review (positive or negative).

By identifying the sustainability terms in the review text, further calculations could be made. First, for each company, the number of reviews containing at least one sustainability indicator was summed and stored as "Total of Sustainable Reviews" ($Total_S$). With these new columns and the following equation, the CSI of each company was calculated:

$$Corporate\ Sustainability\ Index = \frac{Total_S}{Total_R} \times 100$$

Equation 1: Corporate sustainability index

Then, for each dimension, the number of reviews containing at least one sustainability indicator from the respective dimension was also totaled and stored as "total social reviews" ($Total_{S_s}$), "total environmental reviews" ($Total_{S_{env}}$), or "total economic reviews" ($Total_{S_{eco}}$). With these new columns, subsequent values were calculated:

$$Social\ Index = \frac{Total_{S_s}}{Total_S} \times 100$$

Equation 2: Social index

$$Environmental\ Index = \frac{Total_{S_{env}}}{Total_S} \times 100$$

Equation 3: Environmental index

$$Economic\ Index = \frac{Total_{S_{eco}}}{Total_S} \times 100$$

Equation 4: Economic index

3.2 Knowledge distillation

The above steps for text refinement resulted in 22 final datasets that allowed for various interpretations and further calculations. The indices (Equations 1 to 4) form the basis for the analysis of the sustainability reputation of the companies that are included in the S&P 500 stock market index. To get an overview of these companies, the 22 final datasets were appended with each other. The sustainability indices were compared on a company level as well as on an industry level. Furthermore, it is possible to filter out the differences of the sustainability dimensions and examine them within an industry. Lastly, the relationship between the CSI and stock value was also analyzed with a correlation matrix to see if the variables had an impact on the stock value.

3.3 Reliability and validity

In addition to checking the quality of the selected literature, other measures were used to ensure the reliability and validity of the results of this study.

3.3.1 Reliability

According to Saunders et al. (2009), reliability is achieved when data collection techniques and analysis produce consistent results. Since all online English reviews of U.S. companies were collected on Trustpilot by the scraper, there are no uncertainties regarding data collection, and random or systematic measurement errors are not expected (Cheliotis et al., 2015). Another aspect to support reliability is the calculation of a subjectivity score with a pattern package when performing a sentiment analysis. Furthermore, the Twitter-based sentiment lexicon *afinn* is a good method for determining sentiment level from social media posts (Ribeiro et al., 2016).

3.3.2 Validity

In this work, consumers of different organizations were included for the calculation of the CSI. To obtain meaningful values from the calculated sustainability indices, organizations were compared within the same stock market index. However, due to the short reporting period, the sustainability indicators were only compared with the average stock value once. To obtain explicit validity, consumer reviews would have to be compared with the stock value applicable on the reporting date.

3.4 Ethics

Several aspects must be considered in order for research to be ethical, such as objectivity, respect, or non-maleficence (Saunders et al., 2009). To ensure ethical research, the rights, dignity, welfare, and safety of all participants in this work were protected by ensuring the personal integrity of the collector, maintaining confidentiality, and avoiding invasion of privacy. Since no personalized data was collected, the privacy of all consumers who placed a review on Trustpilot is guaranteed.

4 Results

After looking only at online ratings of companies in the S&P 500 stock market index, in this work a total of 337 out of 500 companies were compared with one another using a newly calculated CSI that recognized predetermined sustainability words and sentiment in consumer reviews from Trustpilot to create a relative weighting. While the CSI represents the overall sustainability reputation of each company, the social, environmental, and economic indices show in which sustainability dimension the company has a particularly good (if the CSI is positive) or bad (if the CSI is negative) reputation. Table 6 shows a part of the results. Previous researchers have illustrated that it is very important to consider all three dimensions of sustainability when calculating a CSI, and this work builds on these dimensions and classifies a total of 197 words into these dimensions (see Table 3 to 5).

	company_market	category	Sustainability Index	Social Index	Environmental Index	Avg stock price	price	Sentiment
0	Cisco Systems	animals_pets	1.39	100.00	0.00	0.00	38.77	-130.12
1	Wynn Resorts Ltd	animals_pets	0.00	NaN	NaN	NaN	169.28	172.91
0	Advanced Micro Devices Inc	beauty_wellbeing	0.00	29.03	0.00	70.97	11.22	26142.46
1	BB&T Corporation	beauty_wellbeing	0.25	75.00	18.75	6.25	51.07	4194.96
2	Zions Bancorp	beauty_wellbeing	25.00	100.00	0.00	0.00	50.71	83.19
...
25	Affiliated Managers Group Inc	vehicles_transportation	0.00	NaN	NaN	NaN	179.11	116.42
26	Progressive Corp.	vehicles_transportation	0.00	NaN	NaN	NaN	51.07	58.69
27	Royal Caribbean Cruises Ltd	vehicles_transportation	0.16	25.45	69.09	5.45	122.45	204.23
28	Southern Co.	vehicles_transportation	0.04	3.45	0.00	96.55	43.40	6407.28
29	Viacom Inc.	vehicles_transportation	1.02	0.00	0.00	100.00	32.71	342.74

Table 6: Part of the results

Figure 4 and Table 7 show the correlation matrix (visual and with numbers) between the calculated indices, stock value and sentiment score. The average CSI score was 1.67 out of 100 and the average stock price score was 110.2. As Table 8 on the next page shows, there is no strong correlation between the CSI and stock prices ($r = -.06$, $p < .33$). Moreover, all three dimensions show a neutral correlation with the calculated CSI (social index: $r = -.01$, $p < .89$; environmental index: $r = -.08$, $p < .25$; economic index: $r = .06$, $p < .36$). Overall, the correlation between the three dimensions exceeds the level of $r = .50$, confirming the interdependence of the constructs. Furthermore, it can be seen in Figure 4 and Table 7 that the social and environmental dimensions have a negative correlation with the economic dimension ($r = -.75$ and $r = -.46$). Moreover, the values of the three dimensions were summed separately, and the value for the economic dimension was the highest. This leads to the conclusion that economic affairs are most prevalent in consumer ratings of companies listed in the S&P 500 stock market index.

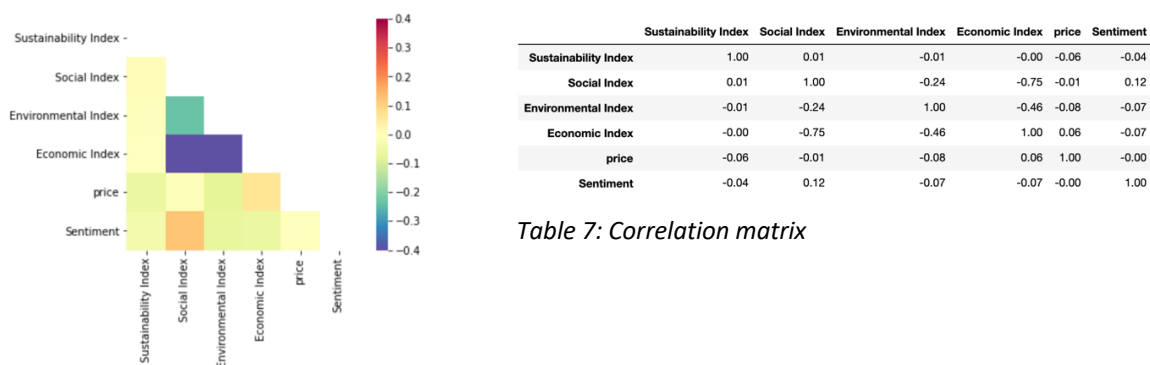


Table 7: Correlation matrix

Figure 3: Visual correlation matrix

Variables	Outcomes					
	n	r	CI95%	p-val	BF10	power
CSI and price	pearson 267	-0.059754	[-0.18, 0.06]	0.330717	0.123	0.163587
	n	r	CI95%	p-val	BF10	power
Social index and price	pearson 212	-0.009651	[-0.14, 0.13]	0.888899	0.087	0.052166
	n	r	CI95%	p-val	BF10	power
Environmental index and price	pearson 212	-0.079193	[-0.21, 0.06]	0.250947	0.165	0.209727
	n	r	CI95%	p-val	BF10	power
Economic index and price	pearson 212	0.062618	[-0.07, 0.2]	0.364287	0.129	0.148451

Table 8: Detailed statistics results

Finally, to answer the main research question (MQ), the sub-research questions are addressed below.

(SQ1) How can the different types of sustainability mentioned in online reviews be classified to measure a company's sustainability reputation?

To classify the sustainability of an online review, a sustainability lexicon must first be identified, as described above in Section 3.1.1. Furthermore, the data (review texts) must be preprocessed, as described above in section 3.1.4. If one or more sustainability indicators according to the sustainability word lists are present in the online reviews, they can be classified as sustainable.

(SQ2) How can online reviews be quantified to calculate the CSI for all three dimensions (environmental, social, and economic)?

Once online reviews are classified, the CSI can be calculated. This is done by first calculating the total number of sustainable indicators per review per organization, which results in the total number of sustainable reviews per organization. The CSI is calculated as the percentage of the total number of sustainable ratings per company out of the total number of online ratings per company (see Equation 1). This value can be compared with other companies.

(SQ3) What is the impact of the organization's sustainability index on the organization's market performance?

After defining the CSI per organization, the correlation of the CSI and the average stock price of the organization was analyzed to see if customer perception regarding sustainability has an impact on the market performance of the companies.

5 Discussions

Sustainability is becoming increasingly important when considering corporate reputation (Miralles-Quiros et al., 2017). The literature review discussed that there are many studies that examine corporate reports for sustainability indices. However, no study has examined consumers' actual opinions regarding CS reputations. This work enriches the literature by showing that consumer reviews that can be used to calculating CS reputation already exist. Other researchers can use the methodology and equations of this work to calculate a sentiment-based CSI and indices for each of the three dimensions of CS for other companies and in different markets. After considering only online reviews of companies from the S&P 500 stock market index, the analysis of 303,491 reviews shows that researchers need to develop a sustainability vocabulary to evaluate companies, and that consumers are most concerned with the economic dimensions of sustainability.

These results also show that a CSI based on consumer reviews does not appear to affect the S&P 500 stock values. In this paper, we proposed a method for extracting consumer reviews of S&P 500 stock market companies and their subsidiaries to see if and how the relationship between the CSI and average stock value is affected. Although the results do not show any relationships between the two variables, other researchers can apply this framework to compare the CSI with other financial dimensions such as daily stock prices. This work proposed three lists of words that customers have used to express their perceptions of sustainability. Other researchers can apply these word lists and assign weights to them based on their degree of sustainability.

The results of this work have important implications for managers. Consumer reviews are an important source of how customers view corporate reputation (Schoenmueller et al., 2020). In addition, sustainability is playing an increasingly larger role in consumer choices (Miralles-Quiros et al., 2017). Therefore, it is necessary for managers to understand how their products and services are perceived by consumers in terms of sustainability. Managers can therefore use this CSI to better understand their own sustainability reputation from a consumer perspective and make any necessary changes.

6 Limitations and future research

This work inevitably has limitations. First, it examined a specific stock market index for comparison. Replication of the work in other stock markets, including in other countries, is essential for future research.

Second, the average stock price of companies was used for comparison in this work. Although it is not certain whether there is a correlation between stock market and CSI, further research would be worthwhile as investments are increasingly based on the sustainability reputation of companies (Miralles-Quiros et al., 2017). Tracking stock prices and sentiment indices over time could show a correlation, especially if the date of the rating coincides with the date of the stock price. Dividing online ratings and stock price ratings into time periods could also help discover a correlation between CSI and stock price. The division into time periods could be interesting because stock prices sometimes react with a lag (Shutes et al., 2016). For example, if a set of ratings with a high environmental index was published, it could take a few weeks for the stock value to rise due to investors or increased sales.

Third, in this paper, the lists of sustainability words were developed based on literature. Accordingly, only the three lists of words (Table 3 to 5) have been considered, and other important sustainability words for specific industries have not been included. This introduces bias because sustainability reputation is an overall perception, and sentiment and sustainability words related to other attributes (e.g., industry, price) also influence this perception. Further research may identify additional attributes. In addition, when determining the occurrence of indicators using the sustainability word lists, misspellings and colloquial alternatives to each word were not considered. Adding more colloquial words and phrases to the created sustainability word lists (Table 3 to 5) could provide more results, as online reviews are often misspelled and colloquial.

Finally, this research did not consider fake ratings. Although Trustpilot has an algorithm to filter out fake reviews (Trustpilot, 2021b), future research could examine whether reviews on Trustpilot are fake and, if so, investigate their effects on sentiment.

7 References

- Antolín-López, R., Delgado-Ceballos, J., & Montiel, I. (2016). Deconstructing corporate sustainability: A comparison of different stakeholder metrics. *Journal of Cleaner Production*.
- Arthur, C. L., Wu, J., Yago, M., & Zhang, J. (2017). Investigating performance indicators disclosure in sustainability reports of large mining companies in Ghana. *Corporate Governance (Bradford)*.
- Azapagic, A. (2004). Developing a framework for sustainable development indicators for the mining and minerals industry. *Journal of Cleaner Production*.
- Azapagic, A., & Perdan, S. (2000). Indicators of sustainable development for industry: A general framework. *Process Safety and Environmental Protection Part B: Transactions of Institution of Chemical Engineers*.
- Babic Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. A. (2016). The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors. *Journal of Marketing Research*.
- Bansal, P. (2002). The Corporate Challenges of Sustainable Development. *The Academy of Management Executive*.
- Benn, S., Dunphy, D., & Griffiths, A. (2014). *Organizational Change for Corporate Sustainability*. Routledge.
- Carnes, B. (2019, November 1). *FreeCodeCamp*. What Is a CSV File and How to Open the CSV File Format. <https://www.freecodecamp.org/news/what-is-a-csv-file-and-how-to-open-the-csv-file-format/>
- Cheliotis, G., Lu, X., & Song, Y. (2015). Reliability of Data Collection Methods in Social Media Research. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*.
- Chelli, M., & Gendron, Y. (2013). Sustainability Ratings and the Disciplinary Power of the Ideology of Numbers. *Journal of Business Ethics*.
- Chevalier, J. A., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*.

- Clarkson, M. B. E. (1995). A stakeholder framework for analyzing and evaluating corporate social performance. *The Academy of Management Review*.
- D'Acunto, D., Tuan, A., Dalli, D., Viglia, G., & Okumus, F. (2020). Do consumers care about CSR in their online reviews? An empirical analysis. *International Journal of Hospitality Management*.
- DataHub. (2021, April 1). Find Data, S&P 500 Companies with Financial Information.
<https://datahub.io/core/s-and-p-500-companies>
- Delmas, M. A., Dror, E., & Nairn-Birch, N. (2013). Triangulating environmental performance: What do corporate social responsibility ratings really capture? *Academy of Management Perspectives*.
- Dyllick, T., & Hockerts, K. (2002). Beyond the business case for corporate sustainability. *Business Strategy and the Environment*.
- Elkington, J. (1998). Partnerships from cannibals with forks: The triple bottom line of 21st-century business. *Environmental Quality Management*.
- Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook*. Cambridge University Press.
- Fiksel, J., Eason, T., & Frederickson, H. (2012). *A Framework for Sustainability Indicators at EPA*. United States Environmental Protection Agency.
- Fineman, S., & Clarke, K. (1996). Green stakeholders: Industry interpretations and response. *Journal of Management Studies*.
- Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How Online Product Reviews Affect Retail Sales: A Meta-analysis. *Journal of Retailing*.
- Gladwin, T. N., Kennelly, J. J., & Krause, T.-S. (1995). Shifting Paradigms for Sustainable Development: Implications for Management Theory and Research. *The Academy of Management Review*.
- Hahn, T. (2012). Reciprocal Stakeholder Behavior: A Motive-Based Approach to the Implementation of Normative Stakeholder Demands. *Business & Society*.
- Hahn, T., Pinkse, J., Preuss, L., & Figge, F. (2014). Tensions in Corporate Sustainability: Towards an Integrative Framework. *Journal of Business Ethics*.

- Hall, J. K., & Martin, M. J. C. (2005). Disruptive technologies, stakeholders and the innovation value-added chain: A framework for evaluating radical technology development. *R & D Management*.
- Harik, R., EL Hachem, W., Medini, K., & Bernard, A. (2015). Towards a holistic sustainability index for measuring sustainability of manufacturing companies. *International Journal of Production Research*.
- Heinberg, M., Ozkaya, H. E., & Taube, M. (2018). Do corporate image and reputation drive brand equity in India and China? - Similarities and differences. *Journal of Business Research*.
- Herva, M., Franco, A., Carrasco, E. F., & Roca, E. (2011). Review of corporate environmental indicators. *Journal of Cleaner Production*.
- Hollenbeck, B. (2018). Online Reputation Mechanisms and the Decreasing Value of Chain Affiliation. *Journal of Marketing Research*.
- Jennings, P. D., & Zandbergen, P. A. (1995). Ecologically Sustainable Organizations: An Institutional Approach. *The Academy of Management Review*.
- Jiang, Q., Liu, Z., Liu, W., Li, T., Cong, W., Zhang, H., & Shi, J. (2018). A principal component analysis based three-dimensional sustainability assessment model to evaluate corporate sustainable performance. *Journal of Cleaner Production*.
- Kantardzic, M. (2019). *Data Mining: Concepts, Models, Methods, and Algorithms, 3rd Edition*. Wiley-IEEE Press.
- Kariv, A., & Pollock, R. (2018). *Datahub*. About DataHub. <https://datahub.io/docs/about>
- Krajnc, D., & Glavič, P. (2005). How to compare companies on relevant dimensions of sustainability. *Ecological Economics*.
- Labuschagne, C., Brent, A. C., & van Erck, R. P. G. (2005). Assessing the sustainability performances of industries. *Journal of Cleaner Production*.
- Lai, C.-S., Chiu, C.-J., Yang, C.-F., & Pai, D.-C. (2010). The Effects of Corporate Social Responsibility on Brand Performance: The Mediating Effect of Industrial Brand Equity and Corporate Reputation. *Journal of Business Ethics*.

- Larson, Q. (2020, February 1). *FreeCodeCamp*. SQL and Databases - A Full Course for Beginners.
<https://www.freecodecamp.org/news/sql-and-databases-full-course/>
- López, M. V., Garcia, A., & Rodriguez, L. (2007). Sustainable Development and Corporate Performance: A Study Based on the Dow Jones Sustainability Index. *Journal of Business Ethics*.
- Ludwig, S., de Ruyter, K., Friedman, M., Brüggem, E. C., Wetzels, M., & Pfann, G. (2013). More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates. *Journal of Marketing*.
- Luminati Networks. (n.d.). Retrieved June 20, 2021, from <https://brightdata.com/faq>
- Maas, S., & Reniers, G. (2014). Development of a CSR model for practice: Connecting five inherent areas of sustainable business. *Journal of Cleaner Production*.
- Mainali, B., & Silveira, S. (2015). Using a sustainability index to assess energy technologies for rural electrification. *Renewable & Sustainable Energy Reviews*.
- Meadows, D. H. (1972). *The limits to growth: A report for the club of Rome's project on the predicament of mankind*. Earth Island.
- Miazga, J., & Hachaj, T. (2019). Evaluation of Most Popular Sentiment Lexicons Coverage on Various Datasets. *Proceedings of the 2019 2nd International Conference on Sensors, Signal and Image Processing*.
- Miralles-Quiros, M. del M., Miralles-Quiros, J. L., & Arraiano, I. G. (2017). Are Firms that Contribute to Sustainable Development Valued by Investors? *Corporate Social Responsibility and Environmental Management*.
- Moe, W. W., & Trusov, M. (2011). The Value of Social Dynamics in Online Product Ratings Forums. *Journal of Marketing Research*.
- Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2005). *Handbook on Constructing Composite Indicators*.
- Newton, T. J. (2002). Creating the new ecological order? Elias and actor–network theory. *The Academy of Management Review*.

- Pranugrahaning, A., Donovan, J. D., Topple, C., & Masli, E. K. (2021). Corporate sustainability assessments: A systematic literature review and conceptual framework. *Journal of Cleaner Production*.
- Price, J. M., & Sun, W. (2017). Doing good and doing bad: The impact of corporate social responsibility and irresponsibility on firm performance. *Journal of Business Research*.
- Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). SentiBench—A benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*.
- Saeed, M. A., Farooq, A., Kersten, W., & Ben Abdelaziz, S. I. (2019). Sustainable product purchase: Does information about product sustainability on social media affect purchase behavior? *Asian Journal of Sustainability and Social Responsibility*.
- Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students*. Pearson education.
- Schneider, A., & Meins, E. (2012). Two dimensions of corporate sustainability assessment: Towards a comprehensive framework. *Business Strategy and the Environment*.
- Schoenmueller, V., Netzer, O., & Stahl, F. (2020). The Polarity of Online Reviews: Prevalence, Drivers and Implications. *Journal of Marketing Research*.
- Scimago Journal & Country Rank. (2021). Scimago Journal & Country Rank.
<https://www.scimagojr.com>
- Searcy, C. (2012). Corporate Sustainability Performance Measurement Systems: A Review and Research Agenda. *Journal of Business Ethics*.
- Searcy, C., & Pryshlakivsky, J. (2017). A Heuristic Model for Establishing Trade-Offs in Corporate Sustainability Performance Measurement Systems. *Journal of Business Ethics*.
- Shen, L.-Y., Jorge Ochoa, J., Shah, M. N., & Zhang, X. (2011). The application of urban sustainability indicators – A comparison between various practices. *Habitat International*.
- Shrivastava, P. (1995). The Role of Corporations in Achieving Ecological Sustainability. *The Academy of Management Review*.

- Shutes, K., McGrath, K., Lis, P., & Riegler, R. (2016). Twitter and the US stock market: The influence of micro-bloggers on share prices. *Economics and Business Review*.
- Taieb, D. (2018). *Data Analysis with Python: A Modern Approach*. Packt Publishing, Limited.
- Thelwall, M., & Buckley, K. (2013). Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*.
- Trustpilot. (2021a). <https://de.trustpilot.com/about>
- Trustpilot. (2021b). Trust Is the Core of Everything We Do.
https://images.ctfassets.net/b7g9mrbfayuu/6PKWxt87MoRfKz6ViG90Hi/5eee217deb16146d82e7ce1863b6687d/trust_at_the_core_ENG.png
- U.S. Securities and Exchange Commission. (2021). <https://www.sec.gov>
- Wallis, A. M. (2006). Sustainability indicators: Is there consensus among stakeholders? *International Journal of Environment and Sustainable Development*.
- Wang, H.-M. D., & Sengupta, S. (2016). Stakeholder relationships, brand equity, firm performance: A resource-based perspective. *Journal of Business Research*.
- Winn, M. I., & Pogutz, S. (2013). Business, Ecosystems, and Biodiversity: New Horizons for Management Research. *Organization and Environment*.
- Zhang, L., Ma, B., Cartwright, D. K., & Lingreen, A. (2013). The impact of online user reviews on cameras sales. *European Journal of Marketing*.

8 Certification of authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text. I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student:

Arta Demaj

Date:

Sunday, June 20, 2021