NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

DEVELOPING A DATA DRIVEN CSR TOOL FOR IMPACTFUL PERFORMANCE ANALYSIS AT A DUTCH BANK.

Leading towards the exclusion of investments in controversial activities of (potential) clients.

Marcella Marie Kneppers
Innovation Lab Intern at a Bank, the Netherlands

Internship report proposal presented as partial requirement for obtaining the Master's degree in Information Management.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

MGI

Title: DEVELOPING AN AUTOMATED CUSTOMER DUE DILIGENCE TOOL THAT LEADS TO IMPACTFUL CSR PERFORMANCE ANALYSIS.

Subtitle: Leading towards the exclusion of investments in controversial activities of (potential) clients.

Marcella Marie Kneppers
M20190529

2021

NOVA Information Management School

Instituto Superior em Gestão de Informação

Universidade Nova de Lisboa

DEVELOPING A DATA DRIVEN CSR TOOL FOR IMPACTFUL PERFORMANCE ANALYSIS AT A DUTCH BANK.

by

Marcella Marie Kneppers

Internship Report Draft presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence & Knowledge Management.

Advisor: professor Doutor Flávio Luís Portas Pinheiro

Nov 2021

ABSTRACT

The internship project is performed at a Dutch Bank in the role of Data Scientist Intern in the Innovation Lab located in The Hague, the Netherlands. I will support the creation of a Python based text analysis tool that gathers data on clients and monitors their performance in terms of Corporate Social Responsibility (CSR) related topics and activities. The tool will be used to support customer due diligence processes as well as review analysis purposes for clients to compare CSR scores with peers. This will help NIBC bank to obtain better insights into CSR performance and activities of their clients and potential clients. This results in the mitigation of risks associated with non-compliant CSR levels.

An important objective for the development of the CSR tool is upcoming (EU) regulation to increase transparency in third-party suppliers. This would increase the relevance of the CSR text analysis tool enormously. NIBC Bank wants to be the front-runner within sustainability in the financial sector. The internship will last for a period of six months (September 2020 till February 2021) and will take place in the Netherlands.

KEYWORDS

Corporate Social Responsibility, CSR, Environmental Social and Governance, ESG, Text Mining, Web Scraping.

INDEX

TABLE OF FIGURES

TABLE OF TABLES

ABBREVIATIONS

CSR	Corporate Social Responsibility
ESG	Environmental Social and Governance
IPCC	Intergovernmental Panel on Climate Change
CLO	Collateralized Loan Obligation
TF-IDF	Term Frequency-Inverse Document Frequency
SME	Small and Medium size Enterprises
FTE	Full Time Employee
PCAF	Platform Carbon Accounting Financials
GDPR	General Data Protection Regulation
EBIT	Earnings Before Interest and Taxes
EBITDA	Earnings Before Interest, Taxes, Depreciation and Amortization
LSA	Latent Semantic Analysis

1. INTRODUCTION

Corporate Social Responsibility (CSR), and social responsibility in general, are rapidly transforming our society and way of business. In particular the financial market, in which CSR has been a debated topic ever since the burst of the financial crisis in 2008. This crisis arose due to unethical and irresponsible behavior of mainly financial institutions. To avoid such a crisis in the future, the market must set a responsible, sustainable approach.

CSR addresses focus areas within the economical, legal, ethical, and social welfare domain. One of the most known and used definition of CSR comes from Carroll (2008, p. 33): "*The social responsibility of business encompasses the economic, legal, ethical and discretionary expectation that a society has of organizations at a given point in time*". The European Commission defines CSR more broadly and refers to CSR as companies taking responsibility for their impact on society.

CSR in businesses shows the possibilities of transparency and completeness in its services and towards its customers. The impact of implementing CSR in your business is a decrease in financial and governmental risks resulting in an increase in legal compliance. This is becoming an important aspect of businesses adopting CSR. The IPCC report of 2021 (IPCC, 2021), among others, makes businesses, governments, and the public in general aware of the need for a more responsible and sustainable way of life. Big nations such as the United States of America (USA), China and Europe feel the responsibility of acting now. In 2017 the European Union accepted a regulation for all European organizations (greater than 500 FTE and euro 40 million revenue) to publish a due diligence report on sustainability and transparency of business activities. Financial institutions in the EU are required to perform thorough due diligence (i.e. finance, governance, sustainability) on (new) clients and investors. NIBC Bank noticed that EU organizations fall short on this law due to a lack of in-depth due diligence as well as acting upon findings concerning sustainability. NIBC Bank expects forthcoming EU regulations regarding transparency and sustainability to be stronger due to the outcome of the IPCC report. A problem is identified here: how can NIBC Bank perform correct due diligence assessments on clients and investors when information regarding sustainability reports is falling short, non-transparent and CSR regulation is becoming stricter?

The mechanisms that are put into place today to perform due diligence on sustainability and transparency levels of EU businesses are survey-based. This entails that organizations themselves are required to complete the survey whereafter a sustainability report is published. Next to the time-consuming behavior of these reports, there is room for subjectivity and human errors. Hence, NIBC Bank is creating a data driven CSR Tool to assess activities of clients and third parties in a fast, safe,

and objective manner. The development of this data driven CSR Tool is based upon text analysis techniques and could become a business opportunity for NIBC Bank as a service to its clients.

In 2019, NIBC bank initiated a project to build a data driven Corporate Social Responsibility (CSR) tool. The tool should extract data from text documents and corporate websites whereafter it would transform the unstructured data into an in-depth CSR analysis of an organization. The driving force behind the start of this project was the forthcoming EU regulation regarding the reporting and transparency on CSR related topics. This, as well as the wish to be front-runner in sustainable banking and investments, is NIBC Banks main motivation for developing a data driven CSR Tool.

The CSR Tool will contribute to the closing of a gap between current and upcoming EU regulations regarding sustainability by its ability to detect and diminish non-CSR compliant organizations. The goal of this project is to develop a holistic CSR Tool that can detect CSR gaps and facilitate discussions with clients to co-influence more sustainable business activities that will result in awareness for sustainable reporting. Reporting on sustainability is necessary to start moving towards more transparency in business process activities.

1.1. NIBC BANK

NIBC Bank is a Dutch merchant bank that offers integrated solutions in the Benelux (i.e. Belgium, Netherlands, and Luxemburg) and Germany through a combination of advising, financing, and co-investing. Clients of the bank are mostly small- to medium-sized (SME) companies.

The Bank is headquartered in The Hague, the Netherlands and counts 660 employees. They serve more than 600 mid-market businesses together with 400,000 retail clients throughout Europe. The bank recognizes the impact of their actions on the world and tries to act in a responsible and sustainable manner. They focus on four key areas that are monitored and measured on an ongoing basis. These four areas are trust and integrity, people, environment, and society (NIBC Bank, 2021). By continuously improving these areas to reduce the company's carbon footprint, they increased awareness for CSR within the financial sector (and beyond) (NIBC Bank, 2021). Nowadays, NIBC bank's activities are carbon neutral, and all electric facilities are running on clean, renewable sources.

NIBC was the first bank to issue a fully ESG compliant Collateralized Loan Obligation (CLO) fund in 2019 (NIBC, 2019). This entails that the fund complies with ESG (Environmental, Social and Governance) investment criteria and excludes investments in controversial sectors (e.g. coal and mining, extreme fossil fuels, arms, tobacco, gambling). As a result, sustainability is fully embedded in the business

strategy and decision making of NIBC bank. The CSR Tool supports this sustainable strategy by its ability to co-influence sustainable developments.

The development of the CSR Tool is a popular case within the bank and stimulates interest in the topic. EU regulation drives most of their interest. Due to this, there are multiple stakeholders involved in the project. A distinction can be made between internal and external focused stakeholders. The stakeholders with an internal focus are interested in due diligence and audit processes. On the other hand, stakeholders with an external focus are noticing the growing importance of investor and corporate client due diligence in terms of sustainability. Providing this new tool to clients could attract investors with a CSR responsible mindset and prevent investments in non-CSR compliant clients.

2. THEORETICAL FRAMEWORK

This chapter aims to compare and critically evaluate the different theories and approaches available for this project. A literature review was conducted to define key concepts and draw relationships between these concepts. The theory of data mining, a process model methodology as well as web- and text mining techniques are discussed. Firstly, a theoretical introduction to CSR tooling is explained.

2.1. INTRODUCTION CSR TOOL

Over the last few years, CSR has evolved "*from philanthropy to a more complex concept*" (Diez-Cañamero et al., 2020, p. 6). The interconnection between social and environmental variables as well as the relationships among stakeholders is resulting in a need for complex, data-driven tools to measure CSR according to the article by Diez- Cañamero et al. (2020).

CSR tooling identifies social and environmental impacts of business operations of an organization or portfolio (e.g. collection of financial investments) and assesses its significance (Iatridis & Schroeder, 2016). Often, CSR assessments range from 1 to 100 based upon a company's scoring on different CSR criteria. The result of measuring CSR commitment leads to the identification of companies that consider and manage material factors of CSR such as Environmental, Social and Governance (ESG) factors. According to VigeoEiris (2020), organizations with high CSR scores indicate better relationship management with stakeholders.

Most organizations that perform sustainability scoring are based on frameworks holding specific CSR criteria. These inputs come from questionnaires the to-be-analyzed organization has to complete. Other types of input sources are web pages, documents (annual reports, factsheets, sustainability reports) and the consultation of scientific articles (Diez- Cañamero et al., 2020).

Organizations that perform CSR scoring with existing tooling are Sustainalytics, Bloomberg, MSCI, RepRisk, and Thomson Reuters. Specialist firms in CSR research are VigeoEiris, Oekom, Trucost and CDP. VigeoEiris created a visual flow of their CSR assessment process (Figure 1). This approach is used to generate an overall ESG score across different sectors and industries. In this CSR model, organizations are assessed based upon 38 distinct ESG criteria.

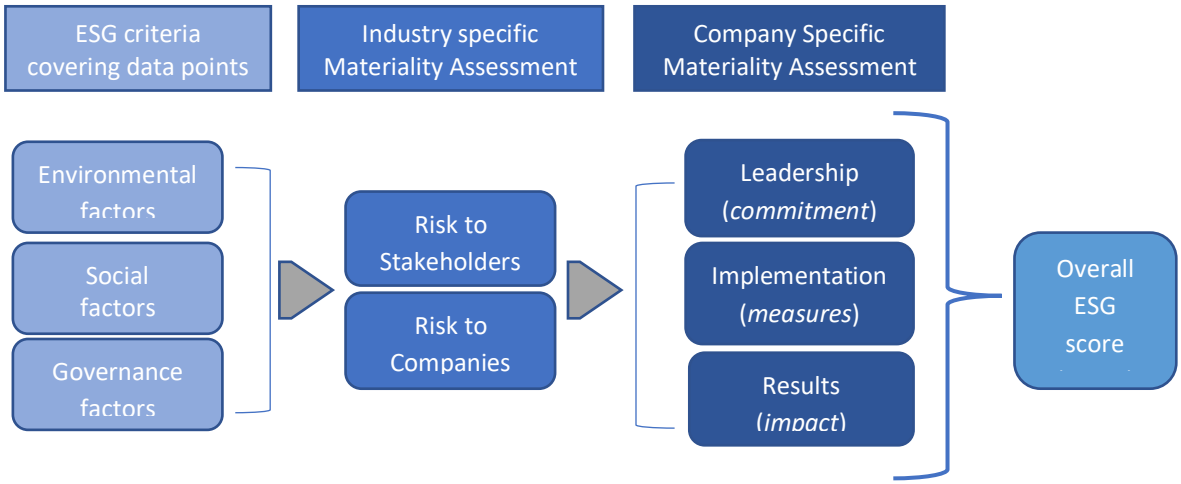

Figure 1. The ESG assessment approach by Vigeo Eiris showing three fases in order to obtain an overall ESG score when assessing text documents of organizations.

The outcome of a CSR assessment is used to identify areas of strong and weak management approaches which are connected to risks and opportunities for a company. Due to these insights, one can mitigate the impact of risks and exploit opportunities. The criteria that are used differ per assessment. However, they show overlap due to a few popular CSR frameworks. These frameworks are Global Reporting Initiative (GRI), OECD (Organization for Economic Co-operation and Development) Guidelines, ESG and CSR. Evaluating organizations in different sectors and industries becomes interesting due to this overlap because trends and patterns can be researched more easily (Vigeo Eiris, 2020).

The main difference between existing tooling and the CSR text analysis Tool developed by NIBC, is the data driven nature of the Tool. This entails that CSR commitment levels are determined based upon data analysis performed on text documentation originating from organizations. An organization or portfolio is assessed based on data it retrieves from text documents and corporate websites. The data driven character is due to a fundament of text mining and web mining techniques. These two techniques complement each other and are part of the overarching technique data mining.

2.2. Data Mining

According to Hand & Adams (2015, p. 2) data mining is "*the discovery of structures and patterns in large and complex data sets*". This means that data mining is the practice of exploring data that is already collected through diverse methods to obtain new insights that can support in decision making. Through several types of algorithms and methodologies, unique findings are retrieved via data mining processes. In general, there are two aspects of data mining: model building and pattern detection (Hand & Adams, 2015). Model building is a secondary data analysis process in which a (statistical)

model is created out of (a) large data set(s). Pattern detection, on the other hand, lays emphasis on algorithms that seek local structures or anomalies in the data.

Within all data mining projects, the main activity constitutes of filtering and reducing (irrelevant) data to find "diamonds" (i.e. valuable insights organizations can act upon to optimize their business). Data and information are growing exponentially, due to Big Data, which brings possibilities but also increases the chances of data and thus useful knowledge getting lost. It is necessary to explore and analyze all this data and be aware of the possibilities for patterns, rules, anomalies, and relationships that might not appear at first but can be discovered after asking the correct questions, applying the right methodologies, and using domain knowledge. The data mining practice evolves every day with the rise of new knowledge and discovery of technologies, also by adopting techniques from other domains (Figure 2).

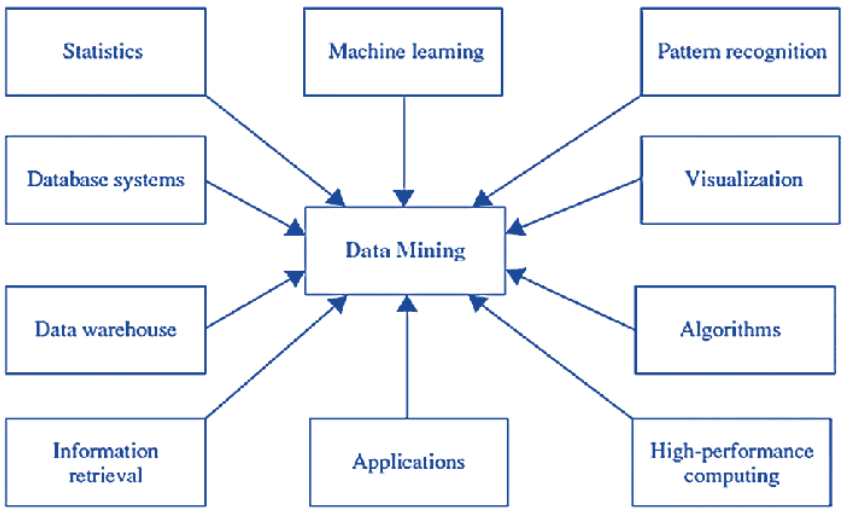

Figure 2. Graphical representation of adopted techniques from diverse domains into the data mining practice (Han, Kamber, & Pei., 2012).

2.2.1. Data Mining Process Model

Carrying out data mining projects can be different from one another due to the way processes are organized and activities performed. However, most data mining processes involve the following steps: understanding the environment, preparing the data, modeling, evaluating results and implementing the outcomes. A process model can help to understand and manage interactions within a complex project to ensure effective project management. One of the most popular data mining process models is CRISP-DM by Wirth & Hipp (2010).

In order to standardize data mining processes, Wirth & Hipp (2010) created the CRISP-DM (Cross Industry Standard Process for Data Mining) model. This model aims to make large data mining projects

more reliable, repeatable, manageable, and faster with less costs. The CRISP-DM model provides an overview of a data mining project by a life cycle holding six main steps. The steps are performed iteratively and require interactive feedback from (a) user(s). These main steps are: (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation, and (6) deployment. Figure 3 shows the CRISP-DM model.

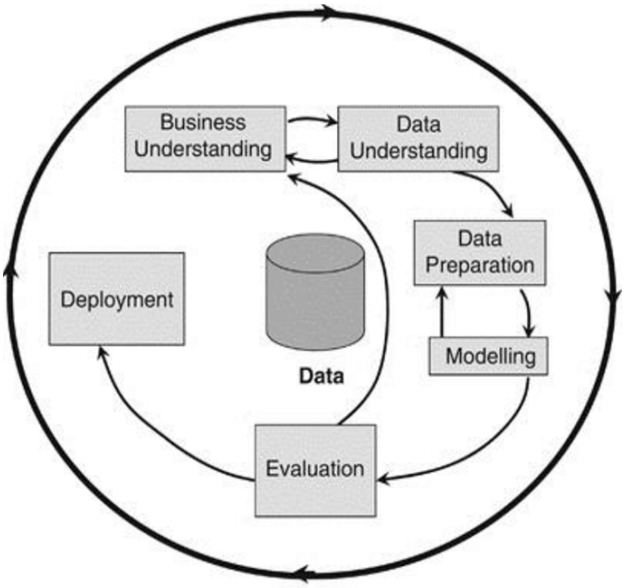

Figure 3. The CRISP-DM life cycle which can be used in data mining projects including arrows representing the direction of the dependencies between steps.

In the following, each step is briefly outlined:

- *Business Understanding*

  This first phase focuses on understanding the objectives and requirements of the project from a business perspective. Then, this knowledge is transformed into a data mining process definition accompanied by a draft version of the project plan to achieve the objectives.
- *Data Understanding*

  This phase starts with collecting data. Once the data is collected, the data is explored by activities such as getting familiar with the data, identifying data quality problems, and detecting first insights into the data to form hypotheses.
- *Data Preparation*

  Besides the challenge of understanding the overall task (the first two steps), one of the most time-consuming steps is data preparation. In this phase the data is preprocessed to fit the modeling tool(s). One could perform several tasks, multiple times if desired, and there is no

prescribed order for these tasks. Task could include data cleaning, transformation of data, and/or adding new attributes.

- *Modelling*

  This phase represents the various modeling techniques that have been selected and applied including parameters that optimize the outcomes. In general, there are multiple techniques used for the same data mining problem.

- *Evaluation*

  In this phase, the modelling phase is reviewed and evaluated to detect if all business objectives are properly met. At the end of this phase, a decision is made concerning the use of the results of the data mining model.

- *Deployment*

  At this phase, the outcome(s) of the model is presented is such a way that the customer can use it. In most cases, the user will carry out the deployment steps – and not the data analyst. It is important to set up a list of actions to make sure one will use the created model(s).

2.3. Web Mining

Web mining is the process of various data mining techniques to extract knowledge from web data (Ali, 2015). Data is moved from the World Wide Web towards a more comprehensible environment in which users can quickly and easily find information (Gupta, 2014). Web data can be categorized as web content, web structure and web usage (Figure 4). In short, web mining includes *'the discovery and analysis of data, document and multimedia from the World Wide Web'* according to Gupta (2014).

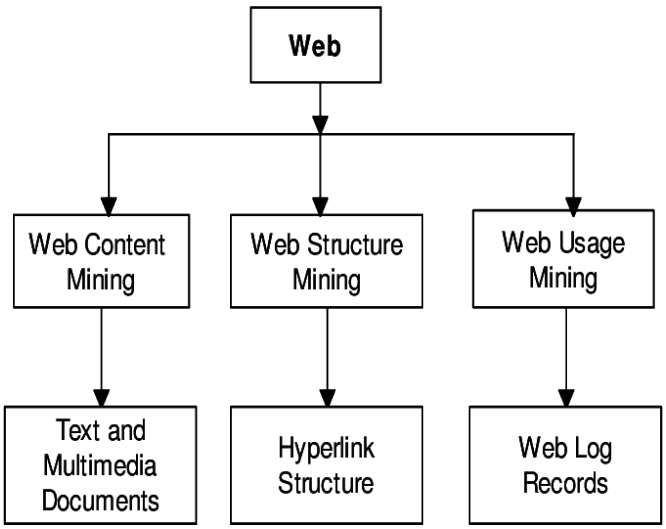

Figure 4. A graphical presentation of three sub-groups including further specification within the overarching domain of web mining.

In this project, web content mining is performed. Raw data from specified web pages is converted into useful information. In figure 4, Web Content Mining is further referred to in this project as web scraping. Text and multimedia documents are copied from the internet and therefore the term web scraping is continued.

According to the book of Mitchell (2018), web scraping is in theory "*the practice of gathering data through any means other than a program interacting with an API or a human using a web browser".* Therefore, a web scraper is an automated program that queries a web server. It requests data from the web server and then parses the data to extract relevant information (Mitchell, 2018). In other words: web scraping is a technique of programmatically extracting data from websites without the need for user interaction. A web scraper can interact with a website like a user would do.

2.3.1. Scraping and Crawling

There are differences between a web scraper and a web crawler – also called a spider – which the book of Jarmul & Lawson (2017) explains. Which type to choose depends on what one wants to extract from websites. The differences between a crawler and scraper are explained below.

A web scraper is typically developed to target (a) particular website(s) to extract specific information from the website(s). In the case a website changes, for example, the location of the desired information, the web scraper needs to be modified to still extract the same information (Jarmul & Lawson, 2017).

Then, in contrast to a web scraper, a web crawler is built to gather more specific information by picking up small bits of information from many different websites. Usually, a crawler is developed in a more generic way so that it can target websites from a series of top-level domains or for the entire web (Jarmul & Lawson, 2017). A web crawler can also crawl pages that are linked to other sites (via URL) or documents. Table 1 summarizes the differences between the two types.

Table 1. Explaining the differences between a web scraper and a web crawler.

Web Scraper	Web Crawler
The tool used is web scraper.	The tool used is web crawler/spider.
The purpose is downloading information.	The purpose is indexing web pages.
Does not visit every page of the website.	Visit every page of the website.
Activity is on a small and large scale.	Activity is (mostly) on a large scale.
Requires a crawl agent and parser for parsing a response.	Requires a crawl agent only.

To conclude, the main difference between a scraper and crawler is that a crawler is often used by search engines for indexing purposes as it easily browses through links on websites and a scraper extracts content from a website and stores it locally (Chapagain, 2019). BeautifulSoup and Scrapy are the most popular and commonly used libraries for web scraping (Mitchell, 2018). Selenium is a popular library when scraping JavaScript codes and has a different set-up than BeautifulSoup and Scrapy.

Scrapy is a Python framework that can be explained as an "*open source and collaborative framework for extracting the data you need from websites*" (Mitchell, 2018). The framework is used for large scale web scraping as it can extract data from websites in an efficient manner where after it processes and stores the data in the preferred format (Zaki Rizvi, 2017).

The Scrapy framework is also called web spider as it requires a root URL in order to start "crawling". One can specify constraints regarding the URLs, namely how many to crawl and fetch. Hence, a Scrapy web spider can crawl multiple URLs that are linked to the root URL. The spider automatically handles cookies which makes it easy to access other URLs without problems (Mitchell, 2018). The limitation of Scrapy is regarding JavaScript and dynamic content websites. To overcome this limitation, one could use the packages Splash or Selenium for Python.

The BeautifulSoup library fetches data from a given URL and allows one to parse specific parts of the website. BeautifulSoup is a popular library for obtaining data from JavaScript websites and/or dynamic contents (Mitchell, 2018). This library differs from Scrapy as BeautifulSoup is a parsing library in Python. BeautifulSoup can parse HTML and XML files.

A dynamic website does not necessarily mean that figures are moving or includes embedded media. The Dynamic HTML (DHTML) is HTML code that changes scripts on the client side so that when a user moves the cursor, it triggers a button to appear (Mitchell, 2018). Examples of websites with dynamic content are e-commerce websites, blogs or any websites that include information that needs to be updated regularly.

After fetching the contents of the given URL(s) main page, the crawler stops. If one wants the crawler to fetch more content, new URLs must be added. Furthermore, the handling of cookies, managing of proxy's and feed export options (e.g. CSV, JSON, XML etc.) are manually added to the code (Mitchell, 2018). This differs from the Scrapy framework as Scrapy is a more complete web crawler used to extract structured data on a large scale.

Fetching data from the web could include personal data. In Europe, the GDPR (General Data Protection Regulation) is a regulation withholding organizations from collecting and storing information of natural persons. Therefore, most web scrapers are equipped with a Python anonymization package: anonymization 0.1.2. Before loading the data into a local file, this package modifies the original names, e-mail addresses, and social security numbers into randomly selected alternate names or numbers. The tool still obtains personal data in this way, but only in such a way that it is not traceable to an individual. Thus, companies act in a compliant way with regards to scraping adhering to the General Data Protection Regulation (GDPR).

2.3.2. Web technologies

The most widely used web technologies are HTML (Hypertext Markup Language), CSS (Cascading Style Sheets), and JavaScript. All three are different and require diverse (web scraping) techniques to correctly extract content. This is due to the numerous ways websites can be built by developers. According to Chapagain (2019, p. 9), a web page is "*a document that contains blocks of HTML tags"*. These blocks can have sub-blocks linked as dependent or independent components from different interlinked technologies. These include JavaScript and CSS.

An application protocol widely used on websites is Hyper Text Transfer Protocol (HTTP). HTTP transfers resources such as HTML documents between a web browser and a web server (Chapagain, 2019). HTTP is a stateless, non-interactive protocol. Web browsers and servers exchange information using HTTP requests and responses. Figure 5 visualizes the information exchange with HTTP.

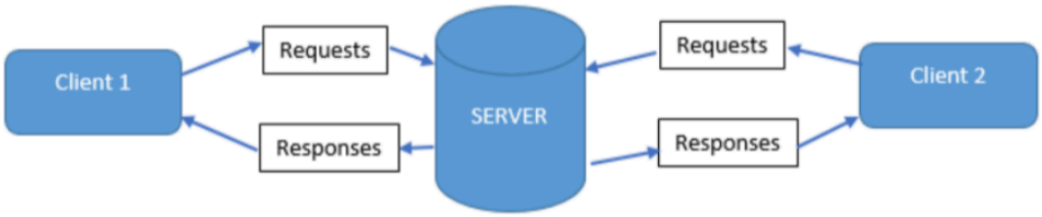

Figure 5. Visualization of the information exchange between web browsers and servers using HTTP requests and responses (Chapagain, 2019).

The first step in any web scraping process is to send an HTTP request to the website's server to obtain the data on the target web page. A principal element of a website set up is the markup language such as HTML and Extensible Hypertext Markup Language (XHTML). HTML is the primary markup language for web pages, whereas XHTML is an advanced and extended version of HTML. These markup languages include text, figures, style sheets and scripts. The markup language determines and contains the contents of a web page. Web scrapers can easily access information and other data sources inside HTML pages due to a predefined instruction set called 'tags' (Chapagain, 2019).

Another commonly used technology is CCS (Cascading Style Sheets) and describes the *"display properties of HTML elements and the appearance of web pages"* (Chapagain, 2019, p. 21). Web-based technologies like JavaScript and HTML are focused on content (i.e. content binding, content development and content processing) whereas CSS provides the styling (Chapagain, 2019). It presents the appearance of HTML elements which are read by web scrapers. CSS can be controlled by web developers as well as designers.

A third, and widely used, web technology is JavaScript. JavaScript is a programming language that is mostly used to add dynamic features to a web page. An example of this dynamic nature of JavaScript is the user-based interaction which allows log-in functionality (Chapagain, 2019). JavaScript contradicts HTML as the latter is framed as a static interface. However, JavaScript can be incorporated into HTML script when adopting the programming logic of JavaScript (Mitchell, 2018).

Most modern web pages are not solely HTTP based as elements of JavaScript are incorporated to dynamically populate pages. This creates a challenge for web scrapers when built with low-technical complexity. The dynamic nature of JavaScript disables such scrapers from accessing and downloading data on JavaScript based webpages (Mitchell, 2018). Additionally, scraping JavaScript is a slow process due to the complexity of web pages which results in a decrease in speed of the crawler. Hence, one

requires a more technical scraper to access JavaScript pages which results in the use of the Python library Selenium.

Selenium works by automating browsers to read and execute JavaScript code before making it available for a scraper to parse the data. Selenium can take screenshots to assert that certain actions have taken place on the website. The library requires an integration with third-party browsers in order to run (Mitchell, 2018). Examples of browser options are Firefox, Safari and Microsoft Edge. Depending on the scale and time restrictions of a web scraping project, one can decide to develop a Selenium scraper. Alternatives are performing HTTP requests or developing a more sophisticated approach to scrape websites with JavaScript in a fast and comprehensible way (Chapagain, 2019).

2.4. TEXT MINING

Text mining is the process of detecting interesting, non-trivial patterns from large text-based data sources (both structured as unstructured sources). Unstructured data sources contain data that cannot be readily indexed or mapped into a standard database field. By contrast, structured data sources can. Text mining adopts techniques from information retrieval, information extraction as well as natural language processing (NLP) to connect them with algorithms and methods of data mining, statistics, and machine learning (Hotho & Paaß, 2005). In contrast to data mining, text mining takes a linguistic approach when evaluating data from sources (Bengford, Bilbro & Ojeda, 2018).

Considering language as ambiguous and arbitrary is required to leverage data encoded in language (Deitel & Deitel, 2019). Two elements of text are *tokens* and *words*. "*A token is a string of encoded bytes that represent text*" according to Deitel and Deitel (2019, p. 12) whereas words represent meaning and textual construct. Words are considered as symbols when analyzing text.

Text mining gained popularity as over 80% of the world's information is stored as text (Gupta & Lehal, 2009). Synonyms for text mining are text analysis or text data mining.

2.4.1. Text preprocessing

A text document is split into a string of words and to obtain all words that are used in a given text, tokenization is required. This process removes all punctuation marks and replaces non-text characters (i.e. tabs) by single white spaces. After tokenization, the text's representation can be used for further processing. The following preprocessing step, after tokenization, is reducing the size of the dictionary. A dictionary is defined by Hotho and Paaß (2005, p. 6) as "*the set of different words obtained by merging all text documents of a collection*". Filtering, lemmatization and stemming methods can be applied to reduce size. See table 2 for techniques on reducing size.

Table 2. Text preprocessing techniques explained that are used in text preprocessing.

Preprocessing technique	Explanation
Convert to lowercase	The conversion of text to all lower-case letters creates a state in which text becomes case-insensitive for algorithms and thus easier to group and sort (Deitel & Deitel, 2019).
Filter stop words	Stop words bear little or no content information and appear very frequently in text. Examples of stop words are the, and, it, is, on. These words have little statistical relevance due to the high frequency of occurrences in text and therefore no distinguishment between documents is possible (Hotho & Paaß, 2005). The filtering of stop words can improve analysis significantly as there is less noise in the data.
Stemming and Lemmatization	In most text documents there is the possibility that the root of a word is mentioned multiple times in several ways throughout the document (e.g. happy, happier, happiest). Such words can be considered as duplicates. Lemmatization is a technique in text mining that links words with the same root to minimize redundancy but saving the original words' context (Deitel & Deitel, 2019). For example, the lemmatized form of "*happier*" is "happy". Stemming is a technique that reduces words to their stems. A stem is a group of words with equal (or very similar) meaning. It strips the plural 's' from nouns, the 'ing' from verbs, or other affixes (Hotho & Paaß, 2005; Deitel & Deitel, 2019).

2.4.2. Relationship and pattern identification

Several text analyses techniques can be used to discover relationships and patterns from unstructured text documents. The first technique to discuss is frequency analysis. This analysis can analyze a corpus in detail, and more specifically, into four levels: (1) frequency of word classes, (2) frequency of words, (3) combinations of word classes, and (4) combinations of words (Johansson & Hofland, 1989; Bengford et al., 2018). These analyses are often used to identify patterns in text (frequent co-occurrences) and the relationship words and word classes have in a corpus.

Frequency of word classes (1) presents insight on the frequency of nouns, verbs, and prepositions in the text body as well as within a topic category (e.g. food, news). A list (alphabetically) presents the frequency of words (2) in text documents. The frequency of word class combinations (3) provides information on word classes that immediately precede each other and/or follow a specific word class. Such information helps to gather context of words and text categories. The last frequency analysis type, (4), is often referred to as n-gram analysis and shows the combinations of individual words that are often mentioned together (Johansson & Hofland, 1989).

Deitel & Deitel (2019, p. 479) defines n-gram analyses as "*the creation of consecutive words in a corpus for use in identifying words frequently appear adjacent to one another*". The name n-gram is deduced from the fact that an argument of n-words is tolerated. As a result, an analysis of two consecutive words is called a bi-gram and for three consecutive words it is called a tri-gram analysis. The argument n can produce n-grams of any desired length.

Figure 6. The n-gram pipeline when analyzing a text body using TF-IDF technique resulting in an output which is text classification or clustering of words.

Figure 6 represents the pipeline of an n-gram analysis including output: the classification or clustering of words. This will be discussed in the next paragraph, 2.4.3. In the middle, TF-IDF (Term Frequency – Inverse Document Frequency) normalizes the frequency of words in a document. By doing this, one accentuates words that are truly relevant to a specific instance or paragraph. As a result, this leads to better classification and/or clustering of words.

TF-IDF is an approach that weighs each word in a text document according to its uniqueness (Al-Talib & Hassan, 2013). The method is based upon a statistical measure that multiplies the frequency of a word in one document with the inverse article frequency of the word across the set of documents. TF-IDF measures the relevancy of words resulting in good differentiation ability of this technique. The equation (1) shows the formula of TF-IDF.

$$TD - IDF(ti, dj) = TF(ti, dj)\, LOG\, \frac{N}{Ni} \quad (1)$$

The part tf(ti,dj) represents the term frequency of term i in document j, N represents the total number of documents in the dataset, and n represents the number of documents where the term i appears (Zhang, Yoshida & Tang, 2008). Hence, the more times a word appears in a document, the more value it will (proportionally) gain.

However, the methodology is also criticized on two main points. The first is that TF-IDF is considered as 'ad-hoc' and not derived from a mathematical model for term distribution. Secondly, the dimensionality of the corpus affects the size of the vocabulary used across the entire document set (Al-Talib & Hassan, 2013). This results in a high computation effort for weighing the terms occurring in the document set.

2.4.3. Topic classification and modelling

Topic classification and topic modelling are both methods to classify text documents with (a) label(s). The main difference between the methods is "nature" of learning: supervised or unsupervised. Topic classification is a supervised machine learning technique which entails that the method requires training. On the other hand, topic modelling is an unsupervised machine learning technique that abstracts topics from a set of documents (Bengford et al., 2018).

2.4.3.1. Topic classification

The aim of topic classification is to assign pre-defined topics to text documents (Hotho & Paaß, 2005). A classification task starts with a training set D = (d1, …, dn) of text documents that have been labelled with a topic L ∈ L (food, news, etc.). Hereafter, a classification model is determined that can assign the correct topic to a (new) text document d:

$$f : D \rightarrow L \quad f(d) = L \tag{2}$$

The performance of the classification model is measured by a sub-set of the documents that have been labelled by topic and were not used for training the model. A test set is required when working with supervised machine learning models in order to keep performance measurement standards accurate (Hotho & Paaß, 2005). The test set is classified by the trained model whereafter one can determine the accuracy of the model by comparing the estimated topics with the true topics.

The accuracy of the performance is measured by *precision* and *recall*. The precision of the topic classification quantifies the number of documents labelled correctly, i.e. belonging to the target class. The recall of the performance quantifies the relevant documents retrieved. The overall performance of classifiers is measured by the F-score. This withholds biases in case only documents with a high degree of the target class are assigned to the set resulting in a high precision, or when the search for documents was extensive and the recall increased, but precision decreases.

$$precision = \frac{\#\{relevant \cap retrieved\}}{\#retrieved} \tag{3}$$

$$recall = \frac{\#\{relevant \cap retrieved\}}{\#relevant} \quad (4)$$

$$F = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \quad (5)$$

2.4.3.2. Topic modeling

Topic modeling is used to identify and label the topics of a set of documents by detecting patterns and recurring words (Deitel & Deitel, 2019). The identification of words and expressions in text leads to the grouping of similar words from diverse documents together. The grouping of similar word patterns creates the possibility to infer topics from the unstructured data.

Topic modelling is often referred to as clustering modeling. Clustering, however, aims to build groups of documents within a text body, whilst topic modeling seeks to abstract core topics from a set of expressions. As it is an unsupervised machine learning technique, topic modelling does not require training.

One of the most frequent topic modeling methods is Latent Semantic Analysis (LSA) (Pascual, 2019). LSA creates vector-based representations of text that captures their semantic content (Wiemer-Hastings, 2004). The semantics of words are linked to the context the words appear in. Thus, the semantics of two words will be similar if they have a high probability of occurring in similar contexts although being in different documents.

LSA treats each document as a bag of words that can be computed to detect how frequent words occur in documents. The method assumes that similar documents tend to contain the same distribution of word frequencies for certain words. The standard method within LSA to compute word frequencies is TF-IDF (described in the previous paragraph 2.4.2).

The vector-based approach is extended by LSA using Singular Valeu Decomposition (SVD) to reconfigure text documents. A set of underlying, independent latent variables compass the meanings that would otherwise be expressed in a language. The SVD technique (re-)orients and ranks dimensions in a vector space. Dimensions in SVD are ordered by significance (i.e. from most to least significant). LSA assumes that the top 300 dimensions are useful for capturing the meaning of texts. Due to this method of reducing dimensions, words that frequently occur in similar contexts have similar vectors and will obtain high similarity ratings in the LSA methodology.

2.4.4. Sentiment analysis

According to Deitel & Deitel (2019), sentiment analysis is one of the most common and valuable Natural Language Processing (NLP) tasks. Data originating from social media sources (e.g. Twitter, Facebook) or text including an opinion (e.g. blogs) are specifically used for sentiment analysis projects. These types of sources hold subjective data and can be connected to a statement or sentence holding a sentiment classification.

Sentiment analysis is a classification method that derives the opinion from a text document, formulates a sentiment and based on this, performs sentiment classification (Gupta et al., 2017). A classifier is described as a machine learning model that is used to differentiate objects on certain features. According to Gupta et al. (2017, p. 30) a feature is "*a piece of information that can be used as a characteristic which can assist in solving a problem*". Features' quality and quantity are important as they are influential for the results generated by the model. A two-class sentiment classification holds a positive and negative sentiment, whereas a three-class sentiment classification shows positive, negative, and neutral dimensions. The sentiment analysis can apply to a process, product, person, or company (Educba, 2019).

Next to the sentiment, the level of subjectivity can also be measured. The sentiment (or polarity) is a float and lies in the range of [-1, 1] where 1 is a positive statement and -1 is a negative statement. A sentiment of 0 refers to a neutral statement when having a three-class sentiment classification. The subjectivity of a document or article refers to a personal opinion, judgement, or emotion. Subjectivity is a float and lies in the range of [0,1] where 0 means objective and 1 is of subjective nature (Deitel & Deitel, 2019).

A sentiment analysis can be based on roughly two approaches: lexicon based (i.e. the vocabulary of a language), and machine learning based (Gupta et al., 2017). A lexicon based approach performs the sentiment analysis by using lexicons (or word forms) and a scoring method to evaluate opinions. On the other hand, the machine learning approach uses feature extraction methods and trains the model with a set of features and a dataset.

Prior to performing sentiment analysis, basic steps must be taken to prepare the data for analysis. Important steps are data collection and data pre-processing whereafter feature extraction and sentiment detection can be performed. The methodology of sentiment analysis is described in figure 7.

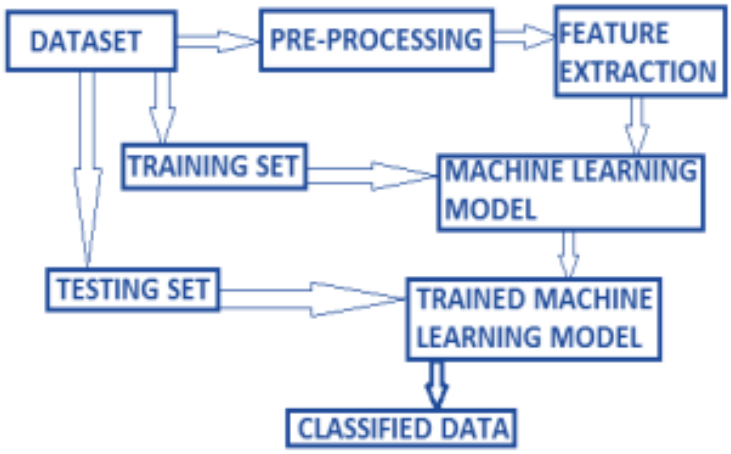

Figure 7. The general methodology for sentiment analysis visualized in steps in order to prepare data for analysis by preprocessing and splitting the data in a training- and test set.

As mentioned above, sentiment analysis is a classification method. Therefore, there are multiple sentiment classification approaches. These classifiers are discussed below.

The Bayesian Logistic Regression selects features and optimizes text categorization. It is a classification approach based upon logistic regression, but with the addition of prior knowledge (or beliefs) about the parameters. Prior knowledge (or belief) is based upon real-life domain knowledge and common sense and prevents overfitting (Gupta et al., 2017).

$$P(c|f) = \frac{1}{z(f)} \text{EXP}((\sum_i \lambda i, cFi, c(f,c))) \quad (6)$$

The Bayesian Logistic Regression approach follows the above formula where Z(f) is the normalization function, the λ is a vector of weight parameters for a feature set and Fi,c represents a binary function with input of a feature and class label. The formula is triggered when a feature exists, and the sentiment is similar to the hypothesized one.

Naïve Bayes classifier is based upon the Bayes theorem and is a machine learning model of probabilistic nature. The theory finds the probability of A happening, given that B became a fact. B is the evidence and A is the hypothesis which assumes that the features (i.e. predictors) are independent. The classifier is called naïve since the presence of one feature does not affect the other (Gupta et al., 2017.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (7)$$

The Naïve Bayes classifier is a simple classifier to understand and apply, although the results are not as satisfying as other classifiers. The Support Vector Machine (SVM) classifier, on the other hand, is a classifier most chosen due to significant accuracy levels and fast processing time.

The SVM algorithm has the objective to find a hyperplane in an n-dimensional space that precisely classifies the data points. In this, n is the number of features. Hence, the algorithm uses decision boundaries (i.e. hyperplanes) to classify data points (Gupta et al., 2017). In case data points are falling on either side of the hyperplane, the points are clustered to different classes.

$$g(X) = w^T \emptyset(X) + b \tag{8}$$

The formula supporting the SVM classifier shows X as feature vector, w as weights of the vector, and b as bias vector. The Ø is the input space from non-linear mapping to high dimensional feature space.

A final classifier to describe is the Artificial Neural Network (ANN). ANN is based on the neural structure of the human brain. This entails those records are processed one by one and compares the classification of the record with the actual classification label (Deitel & Deitel, 2019). This way ANN can learn from errors as initial classifications are fed back into the network and used to improve the networks algorithm for future iterations (Gupta et al., 2017). An ANN model used for supervised learning is Multi-Layer Perceptron (MLP). As shown in figure 8, training data comes in via the input layer, is processed in the hidden layer in the middle, and becomes output in the last layer (Gupta et al., 2017).

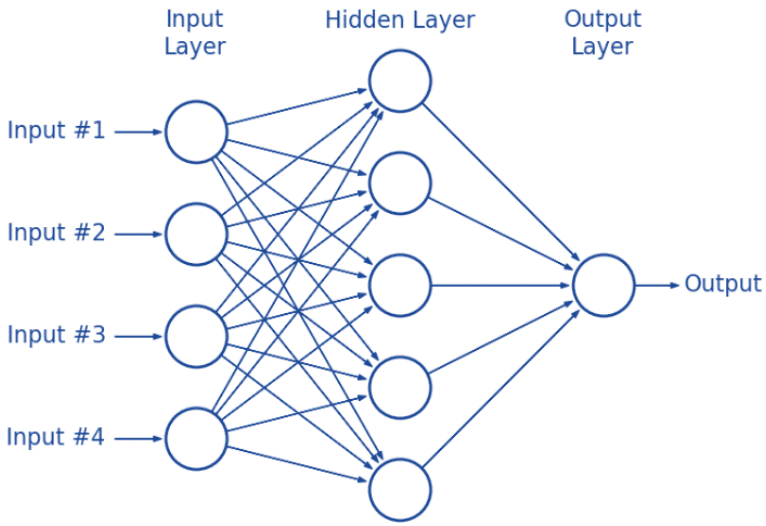

Figure 8. The Multi-Layer Perceptron (MLP) model graphically explained in which the input layer represents the training data, is processed in the hidden layers whereafter the output is generated in the output layer.

2.5. DASHBOARD FRAMEWORK

Within this project, the creation of models and dashboards was done in close collaboration with stakeholders. The User Requirement Model of Lavalle et al. (2019) is the framework behind the dashboards created for the CSR project. This model is chosen due to its emphasis on the end-user: to provide them with the right information to co-influence organizations towards higher commitment on CSR and detect red flags in an early stage.

The User Requirements Model is a goal-based, iterative framework that focuses on the user to assist the user in defining and achieving their goals (Lavalle et al., 2019). The framework can determine which elements of the data sources the user wishes to represent in the visualizations. By defining a series of guidelines, users can present the User Requirements Model. After this, the specification of Key Performance Indicators (KPIs) is done to measure the degree of achievement towards the goals. The above steps also support in determining which elements of the data sources the users wish to represent in the dashboard's visualizations.

2.5.1. Measuring performance

The output of the analyses are consolidated into one overview dashboard, named the CSR Scorecard. The performance of the Scorecard is measured in terms of reliability and validity of the outcomes. Joppe (2000, p. 1) defines reliability as "*the extent to which results are consistent over time and an accurate representation of the total population under study is referred to as reliability*". Hence, reliability concerns whether the result of the Scorecard is replicable. Validity on the other hand, is defined as "*whether the research truly measures that which it is intended to measure or how truthful the research results are*" (Joppe, 2000 p. 1). In other words, the validity of the CSR output is the degree to what one intended to measure and succeeds.

3. CSR PROJECT

Compliant investments and clients are an essential part of being a responsible bank nowadays. Therefore, this project is of high importance for NIBC Bank. The project process was designed for long-term value and using sustainable techniques that would stay relevant and available.

The project follows the framework of the CRISP-DM model. The main steps of the model are (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation, and (6) deployment and will form the basis of this chapter. This framework is applied on a portfolio NIBC holds to illustrate the way of researching CSR levels. The portfolio exists out of 15 organizations in different industries, countries, and sizes which is an all-round example for testing the CSR tool. Both the name of the portfolio and the organizations within the portfolio are confidential information.

The bank has been working on the CSR project for one year now. The project consists of two parts: a web scraping part and a text mining part. Both parts are required to serve an integrated and comprehensive approach towards CSR commitment. The output of the project is called CSR tool; however, the actual output is a dashboard presenting a CSR scorecard and several supporting dashboards with an in-depth focus.

3.1. BUSINESS UNDERSTANDING

This first step of the CRISP-DM focuses on understanding the objectives and requirements of the stakeholders. After that, a data mining process definition accompanied by a draft version of the project plan is created to achieve the objectives set.

The CSR project has two main stakeholders: internal stakeholders with internal interests and internal stakeholders with external interests. Stakeholders with external interests want to share data outputs with external parties whereas the internal interest stakeholders do not. For simplicity reasons, the first group of stakeholders is considered ‘internal’ and the second group of stakeholders is ‘external’. This split in stakeholder groups arose due to different requirements in data output (figure 9). Internal stakeholders make use of the output in the form of dashboards and research reports together with reports for due diligence and monitoring usage. External stakeholders, to the contrary, make use of the output in the form of a CSR scorecard that is shared with (potential) clients and makes use of a dashboard for a thorough company analysis.

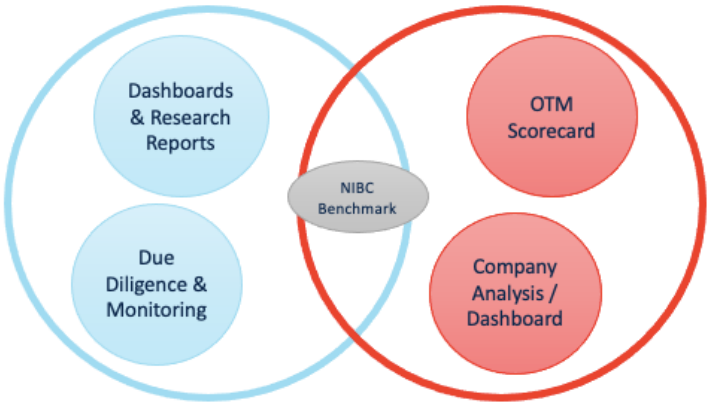

Figure 9. The data outputs of the CSR Tool are customized to the wishes of internal (blue) and external (red) stakeholders within NIBC Bank.

The stakeholders created objectives and requirements for the CSR project. The objective of the project is to obtain reliable insights into best-scoring CSR criteria and worst-scoring CSR criteria from a portfolio (or a single organization) that could become a risk in terms of CSR commitment to base investment decisions on.

The User Requirement Model framework was applied to the project to obtain requirements and preferences of stakeholders. The main requirement all stakeholders agreed upon was the inclusion of company profile information, quick insights into the overall ESG commitment (micro- and macro-level) and incorporating a CSR commitment timeline to see if organizations are improving their CSR levels over time.

These requirements lead to KPIs (Key Performance Indicators) that measure the degree of achievement towards the requirements and objectives of the stakeholders. The KPIs were defined using the User Requirement Model framework and described below:

1) Show 90% compliance with the needs of governmental regulations.
   This is an important KPI for Corporate teams within NIBC bank as they expect that upcoming (EU) regulations will be pushing towards more transparency regarding CSR topics.
2) Meeting the pillars in sustainability by showing 50% commitment towards:
   a) Environmental impact metrics. These metrics are incorporated in the 300 sustainability criteria in the dictionary of the text mining Tool. The output file provides a numeric indication of commitment towards these metrics. Amongst others, these metrics are the CO2 emission and carbon footprint of organizations.
   b) Social impact metrics. Includes the organization's commitment towards labor best practices inside as well as outside their organization (i.e. third-party supplier).

c) Governance impact metrics. Governmental impacts include many aspects that are hard to quantify but need to be mentioned in reports to acknowledge their impact. These aspects are incorporated in the dictionary of 300 sustainability criteria.

3) An industry-level improvement towards CSR levels each year.
   Incorporating a timeline of CSR commitments by industry shows which industries are front-runners and which ones are behind. This helps could end-users to determine which industries to avoid or prefer.

The bank developed a web scraper due to its ease of use and completeness of information gathering. The web scraper is used to systematically retrieve data (raw text) from (potential) clients and suppliers' company websites, process it and then store the information in a local file. The use of an API would have been easier and less time consuming, however, there is no API that can gather small sets of data across many websites. Therefore, NIBC decided to build a web scraper themselves.

The web scraping scripts were written in Python using Scrapy. The Scrapy framework was preferred over BeautifulSoup due to the fit towards the type of usage of the web scraper. Scrapy does an excellent job with static HTML websites and allows manual input URLs. The input is homepage URLs of corporate websites that are analyzed into detail concerning CSR and sustainability commitment. The URLs can be collected as random series of organizations or per portfolio.

In order to compare organizations in terms of CSR performance, the collected data should include basic company information as well as press and media information. Basic company information entails the location of the organization, sector, industry, financial ratings, number of FTEs, total revenues, total balance sheet and emission data. By combining the balance sheet and revenues one can estimate emissions using PCAF methodology. PCAF stands for Platform Carbon Accounting Financials and is an initiative of Dutch financial institutions to calculate their investments' carbon footprints (Warmerdam et al., 2019).

Press and media information are required to detect CSR opinions related to the portfolio. It proactively monitors the portfolio for mentions in press, media, NGO reports or other public data in terms of controversies. Figure 10 presents the process flow of the CSR project: data collection via diverse data sources, data processing via text mining techniques and application of the sustainability dictionary, data analyses and modelling resulting in data outputs. The data outputs are split to meet the requirements of both groups of stakeholders.

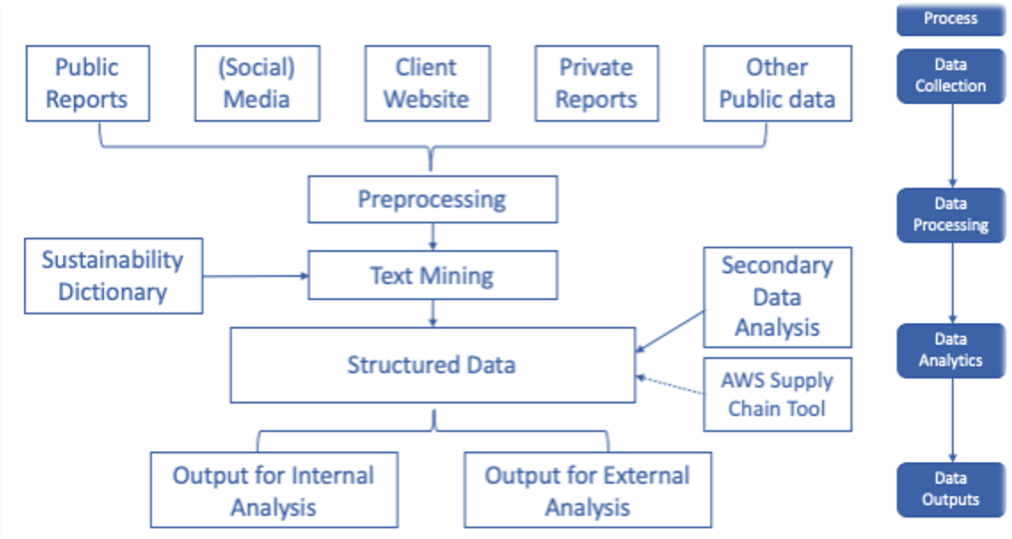

Figure 10. Process flow outline of the CSR project with five main processes which are data collection, data processing, data analysis, and data outputs.

3.2. DATA UNDERSTANDING

This step of the CRISP-DM model starts with the collection data. Once the data is collected, the data is explored to get familiar with the data and identify data quality problems. Lastly, the first insights can be obtained from the data by running basic analyses.

The collection of data for the CSR project is done via diverse data sources. These data sources have been selected due to the diversity of content (formal, informal), source location (web, media, report), and the type of end-user. The motivation for such a diverse selection of data sources is to obtain a wide spectrum of information related to an organization to be able to filter CSR related content only. This way, NIBC can grasp CSR content from several types of stakeholder's organizations deal with. The data sources are described in table 3 including the selection objective per data source.

Table 3. The data sources founding the CSR too described per data type and selection objective.

Data Source	Data type	Selection objective
Public Reports (listed organizations)	Annual reports, code of conduct, sustainability reports	Level of CSR commitment through the eye of the organization.
(Social) Media	Twitter, Google News	Detecting red flags and allegations a company does not (want to) report.
Client Website	Home page URL	CSR commitment on public website.
Private Reports (non-listed organizations)	Annual reports, code of conduct, sustainability reports	Level of CSR commitment through the eye of the organization.
Other Public Data	Yahoo Finance	Company-profile data to establish fair comparison requirements.

The data collection of client website information is performed via a web scraper. The scraper requires a list of URL's and company names in order to perform a web crawl. These URL's need to be saved in a separate document called 'seeds.txt'. The scraper crawls through each of these websites and checks each web page for keywords related to sustainability. These keywords are defined in a whitelist called 'web.py' within the folder called 'spiders'. The web scraper performs a level two scrape which means that it can navigate to other pages that are linked to the input URL and can open PDF documents from the website to extract all text.

Public and private reports are analyzed using text mining techniques. NIBC's existing text mining scripts are developed in Python and require three inputs to run: a set of to-be analyzed .txt files, a concept vector list as .txt file, and a CSR dictionary as .csv file.

- The folder 'sets' includes a series of .txt files from public or private documents from one organization or a portfolio (i.e. set of organizations). Public documents such as annual reports and code of conduct reports are gathered here, as well as the output of the web scraper. Every .txt file will form a column in the output of the Text Mining Tool. This way it becomes clear where organizations place statements concerning CSR.

- A CSR dictionary holding over 300 sustainability criteria based on the frameworks of ESG and SASB. Each criterion in the sustainability dictionary is linked to a unique code. This unique code is the identifier of the sustainability criterion and will be used as primary ID. Per criteria, a timestamp and company name are fixed columns in the output file. The timestamp functions as evidence trail to track CSR performance over years.

- The concept vector is a document holding the regular expressions (RegEx) of the 300 sustainability criteria. The expressions make it possible for the Tool to detect these sustainability criteria in the unstructured text documents from the folder 'sets'.

The output of the text mining Tool is an .xlsx file. The output file contains over 300 sustainability criteria as rows and holds a column for every document in the folder 'sets'. All input files hold the same structure after the text mining tool. This facilitates an easy integration into one text corpus for analyses on a portfolio instead of approaching the files as single organizations.

3.2.1. Data collection

Data sources have been added to develop a more holistic data collection foundation. At first, only public documents and website contents were active data collection mechanisms. By introducing (social) media sources (i.e. Google News and Twitter) and other public documents like Yahoo Finance data, the data sources are extended in a comprehensive manner. Depending on the type of analysis one wants to conduct with the CSR Tool, big batches of data can be extracted and collected via these sources.

First, the data sources collected by the web scraper are discussed, whereafter the other data sources are explained. Public and private reports as well as client websites are data sources accessed via the internet (web). Home page URLs are collected and used as input before running the scraper.

- Public and private report
- Client website

Other public data sources are collected in a different manner. The (social) media sources collect their data via Application Programming Interfaces (API) in which the API functions as an open door to media data (e.g. tweets, posts). An API functions as a servant between the media source and your local environment.

- Twitter

  Twitter requires registration when one requests access to their API (Deitel & Deitel, 2019). The

API request is based on a specific hashtag, username, or topic but could also be a combination of the mentioned forms. The objective for extracting Twitter data focuses on the detection of sentiment and trending topics. These support the banks' research into clients and suppliers in order to detect (potential) allegations.

The available Twitter applications in the development area explain the many ways the APIs can extract information. The way of extracting tweets via the API is defined in the specification settings. The first specification was set to not include retweets. Retweets (RT) are original tweets that are reposted by other users. By only including original tweets in a search string, the output quality is controlled.

Secondly, the language of the tweets can be set. English is the default setting and furthermore the language chosen for the CSR project. In case of researching a non-English tweeting organization, the Python library of Google Translate *'googletrans'* can directly translate the tweets into English if desired.

Lastly, the maximum number of tweets to load into the data frame can be set. The API extracts both real-time and historic tweets, which could result in long processing times when no maximum is chosen. To ensure fast processing times, the maximum number of tweets to be extracted was set to 500.

One of the most popular Python libraries that interacts with the Twitter APIs is Tweepy. Tweets are extracted based upon a search query on username, hashtag, or search term. The search queries were as follows: @OrganizationName sustainability; @OrganizationName Corporate Social Responsibility #CSR; @OrganizationName UN Guidelines; @OrganizationName accused.

In the output data frame, an additional column is added containing a timestamp of the current date. This timestamp will contribute to the evidence trail of analysis performed by the bank regarding the CSR Scorecard.

- Google News

  Google News possesses a Python library that can extract Google news in a fast and easy to manage manner. The Python libraries necessary to run Google News are *newspaper* and the *tldextract* module. The data that can be collected via Google News is enormous, meaning that

a search query is optimized when it is extremely specific. This way, less noise data will enter the data frame, hence data quality will be higher.

The library of Google News includes Google Search and Google Trends. The library is user-friendly in such a way that once you run it, it will ask for a search query. This can be a company name combined with a (sustainability) topic. The search queries given in this project were: OrganizationName sustainability; OrganizationName Corporate Social Responsibility; OrganizationName UN Guidelines; OrganizationName accused. The columns exported to Excel are publication, title, article text, publication year, publication month, publication day, URL and timestamp.

The objective for adding Google News as a data source was due to a lack of news-oriented documents in the current data collection sources. The news documents will be used for sentiment analyses purposes as well as topic classification when handling NGO blogs, newspaper articles and commercial publications.

- Yahoo Finance

  Similar to Google News, Yahoo Finance is a Python library called *yfinance* which can extract desired information by performing lookups based on tickers. Tickers are abbreviations (or nicknames) of the organization's name that is trading stocks. For example, the ticker for Apple Inc. is AAPL in Yahoo Finance.

  Distinguishing one organization from another requires differentiating variables. These variables could be size (revenue, FTEs) as well as industry, location (country) and an organization's main activity (i.e. NACE code). This type of information is defined as profile information in Yahoo Finance (table 4). Next to profile information, *yfinance* extracts financial statements such as income statement, cash flow, and balance sheet. These statements were extracted in the same way as profile information.

Table 4. Profile information of tickers requested in yfinance including the attribute, ticker and recent (i.e. output per attribute).

Attribute	Ticker	Recent
Sector	NIBC.AS	Financial
fullTimeEmployees	NIBC.AS	460
City	NIBC.AS	The Hague
Country	NIBC.AS	Netherlands
longBusinessSummary	NIBC.AS	NIBC Bank is a merchant bank situated..
companyOffices	NIBC.AS	5

Only specific fields from the financial statements are extracted to minimize noise in the data. These fields in the income statement are revenue, profit, tax, EBIT, and EBITDA. The balance sheet includes total assets, total capitalization, and total debt. Lastly, the cash flow position is a required field from the cash flow statement. Concerning basic profile information, the bank is interested in size (i.e. FTEs, total capitalization), location (i.e. address, country), sector and industry. All information from *yfinance* is triggered using tickers.

One can tweak the Python module in such a way that only the desired data comes out. This minimizes unnecessary data outputs and increases data quality. NIBC determined which information was desired for the CSR project to extract from Yahoo Finance. For example, financial statements were gathered to determine emission numbers and revenue numbers. Emission numbers are helpful indicators for the Corporate Sustainability teams.

A disadvantage of using yfinance as source for the CSR project is that yfinance can only find profile information and financial statements of listed organizations. Thus, when a (potential) client of the bank is not listed, one can (most probably) not find the organization and will get no output. Most of NIBC's clients are small to medium sized enterprises. This means that if Yahoo Finance does not provide information on an organization, the missing data needs to be filled in manually.

The portfolio holds 15 organizations with the following data observations collected:

- Public reports from the last 3 to 5 years for all 15 organizations. All organizations in the portfolio are listed and thus hold an annual report, code of conduct and sustainability report. The reporting of non-financial information has been obligated for listed organizations since 2017 by EU regulation. Organizations are obligated to act upon this directive when they have more than 500 employees and a net turnover of 40 million euro.
- Media data holds Twitter and Google News resulting in the collection of 200 tweets and 200 news articles per organization in the portfolio.
- Each organization in the portfolio possesses its own website resulting in 15 home page URLs as input for the web scraper. The web scraper performs a level two scrape which entails the scraper can navigate to other web pages that are linked to the input URL including opening PDF documents.
- Private reports have not been collected as all organizations in the portfolio are listed.
- Yahoo Finance collects per organization in the portfolio the profile information, balance sheet and income statement. Profile information includes country, offices, FTEs, and NACE code (business activity).

All data observations are saved in a local database and form the foundation of the NIBC Universe benchmark. Both national and international organizations are collected for the benchmark as the bank actively operates in several European countries. Two indexes have been used as benchmark in this project, namely the Dutch Small Cap (AScX) and the German Small Cap index (SDAX). The SDAX holds 70 German SME enterprises whereas the AScX holds 25 SME enterprises. An organization is considered SME when they represent a stock market value between €300 million to €2 billion (Semmie, 2021)

3.2.2. Data exploration

Data exploration is performed to identify data quality problems. The identification of quality problems in textual data begins with detecting miss-spelled words and handling random white spaces. If this is not handled correctly, the analyses built upon the data may also include faults. Data is collected from multiple sources, is unstructured and includes many punctuations, stop words and numbers. In order to correctly handle these issues, preprocessing is a key step in building reliable, structured data. Paragraph 3.3. explains the preprocessing steps taken in this project.

A word cloud provides a first visual insight into the data by showing the frequently used words and topics from the text corpus. Words that appear frequently gain a bigger font size in the word cloud. Thus: the bigger the font size, the more often the word is used. Figure 11 shows a word cloud of the

total portfolio and was based on public reports and website content. In this example, the words 'impact', 'environmental', 'https' and 'human rights' are most frequently used among all 15 organizations in the portfolio.

Figure 11. Word cloud of the portfolio's public data sources and website content showing the most used words in the text documents of the portfolio.

3.3. DATA PREPARATION

Once all the input data is collected, the pre-processing part starts. This step is the most time-consuming within the CSR project' process. The tasks performed within data preprocessing include data cleaning and data transformation. By doing so, the data is being preprocessed to fit the modelling tools in the next step (paragraph 4.4.).

The goal of these pre-processing steps is to reduce the vocabulary size of the corpus without removing any important content (Deitel & Deitel, 2019). A small vocabulary leads to a lower memory complexity and as a result the estimated parameters become more robust.

3.3.1. Data cleaning

The data sources (social) media and other public data are preprocessed by cleaning and filtering raw text. After cleaning and filtering, the text becomes valuable for secondary text analysis outcomes as noise is erased from the corpus and it becomes more structured. The NLTK library in Python includes diverse methods to clean and filter text to prepare a corpus for analysis. The filtering and cleaning of text includes the removal of stop words, punctuations and numbers as well as splitting the words and stemming.

- One case letter

  The TextBlob module within the NLTK library converted text into all lowercase letters. This conversion is also referred to as the normalization of text. This cleaning step was chosen to

withhold algorithms from treating words differently due to capitalized letters. The data collected will hold large numbers of company names which means that these will be harder to distinguish. The function *lowerstrip* in the textblob.utils module performed the conversion. For example, '*Apple'* versus '*apple'* after the *lowerstrip* function.

- Punctuations

  The TextBlob module also removed punctuations from the corpus. The list of punctuations removed are: !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~ (,. The TextBlob function *strip_punc* removes the list of punctuations in one step. Contractions of words are eliminated tue to punctuation removal. The contraction '*What's'* has now become '*Whats'*. The meaning of words got lost due to this and is a minor disadvantage for CSR text analysis.

- Stop words

  The NLTK library can download stop words and specify the language of the stop words. Most of the collected data is in English, hence English is chosen as language for stop word removal. Ater the removal, the corpus decreased in size. This is a good sign as now only words with meaning stay.

- Split words

  The NLTK library was used to perform *word_tokenize* in order to splits words based on white spaces and punctuations. After this, the corpus was no longer one long piece of text, but exists out of single words. This is a step that is preferred for many modelling tasks performed in the secondary analysis, like word frequency analysis and n-gram analysis.

- Dates and numbers

  Regular Expression is used to detect and delete numbers in the corpus. Numbers (including dates) do not have meaning for text analysis purposes and are therefore eliminated. The Regular Expression used to remove numbers was re.sub(r'\d+', '', text) meaning the range of numbers [0-9] is substituted by '' (meaning nothing) in text.

- Stemming

  The Porter stemming method was chosen to stem words due to the ease of application. The NLTK library includes a stemming algorithm based on Porter's stemming method. Prefixes and

suffixes are removed by the algorithm resulting in a more focused vocabulary and sentiment of the text corpus.

The mentioned data cleaning steps are applicable to text documents collected via the web scraping tool or private sources (i.e. client websites and private documents). Data originating from a Google News source followed the same cleaning and filtering options as web- and private sources. The media source Twitter is cleaned with the same steps, but due to the collection of this data via an API, the cleaning can be performed in one task.

Basic tweet cleaning can be performed via the library *tweet-preprocessor* where one can call the 'set_options' function for specifying which cleaning tasks to be executed. To achieve similar normalization of text, the same steps as above are applied in the 'set_options' function with an addition of removing RT (retweets), FAV (favorites), duplicates and URLs.

Google News provides a great quantity of data, however only specific fields are interesting to obtain. These fields are determined by their contribution to detecting CSR compliant behavior. The following fields are filtered from Google News and presented as columns in a data frame:

- Publisher (i.e. name of website, newspaper, blog etc.);
- Title of the publication;
- Article content (i.e. the full article's text);
- Polarity [-1 to 1] and
- Subjectivity [0 to 1].

3.3.2. Data transformation

No substantial data transformation has been made as the data needs to be as authentic and transparent as possible. A sole transformation was made to the file extensions of public reports and website sources. Those file extensions needed to be converted to .txt to align with the CSR dictionary extension in order to run. The CSR dictionary holds 300 sustainability criteria matched with the content of the .txt files. The output of running the text mining tool is structured data.

3.4. MODELLING

In the previous step, data preparation, the data has been modified to fit various modelling and analyses techniques. The modelling of data creates visual outputs and is the phase in which it becomes interesting for stakeholders to view results of the CSR project. Secondary analyses and models are

performed to explore the preprocessed data and gain insights. The modelling is performed in Microsoft Power BI dashboards.

3.4.1. Secondary data analyses

Discovering the core CSR focus points within the portfolio is done by an N-gram analysis with bi- and tri-grams (i.e. two and three-word combinations). Table 5 shows the N-gram analysis. The most frequently mentioned word combinations are at the top. A timestamp is added for evidence reasons as well as the ability to place CSR insights on a timeline and map the portfolio's commitment over years.

Table 5. The N-gram analysis output of a client's portfolio showing the frequency of the word combination, the word combination (bi- or trigram) and a timestamp of the analysis.

Frequency	Bigram/Trigram	Timestamp
1380	climate change	2021-01-13
810	net impact	2021-01-13
765	greenhouse gas	2021-01-13
646	supply chain	2021-01-13
480	sustainable aviation	2021-01-13
468	oil gas	2021-01-13
405	see human rights	2021-01-13

The same analysis is done on Twitter data collected on the organizations within the portfolio. To maintain speed in the process of analyses, the maximum number of articles per company was set at 200. In total, this means that 3000* text documents have been extracted and analyzed in this N-gram analysis. The 3000 text documents are a sum of the 200 text documents per analysis times 15 organizations in the portfolio.

A sentiment analysis can provide context on the opinion(s) the organization has on specific CSR topics. A sentiment analysis obtains the average sentiment of the full portfolio as well as per individual organization within the portfolio. Text documents collected via social media sources are prone to

subjectivity and thus interesting for sentiment analyses. Input data from Google News and Twitter has been used for the analyses.

Figure 12 presents the sentiment analysis on a single organization within the portfolio. The x-axis represents the sentiment (negative, neutral, or positive) and the y-axis shows the total number of tweets. As described above, the maximum was set to 200 tweets that have been extracted per organization in the portfolio. The outputs of the N-gram analysis and sentiment analysis are saved into a Python data frame to export for further usage and deployment.

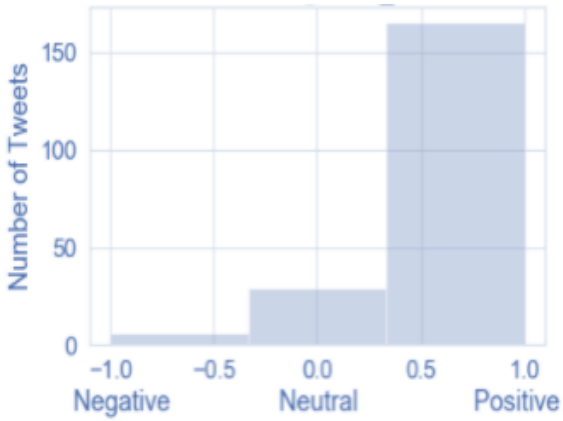

Figure 12. Sentiment analysis output based on tweets from one of the organizations in the portfolio showing the number of negative, neutral and positive tweets.

Topic modeling is the next analysis performed. Documents are grouped together into one text corpus ensuring the modelling algorithm can handle the text as one and group parts of text based on similarity. The text corpus is extensive with ESG being limited to 3 topics: environment, social and governance. To stay focused and not obtain a too broad spectrum of topics, a maximum of nine groups were chosen. These groups do not have a topic name yet. Hence, the words within each group will form an overall theme.

The output of the topic modeling for news articles found on Google News searches. Initially, the column which contains the 'topic' was empty. The topics selected for the nine buckets were chosen manually and presented in table 6.

Table 6. Topic modeling output showing similar word buckets (keywords) based on a topic.

Topic	Keyword
Financial context	Financial, cash, assets, total, eur, liabilities, income, year, net, equity
Risk impact	Risk, management, economic, growth, impact, financing, rate
Governance	Report, board, compliance, corporate, annual, company, governance, audit
Organizational	Co, gmbh, ltd, se, uk, inc, kg
Country	Islands, republic, saint, guinea, samoa, korea, st
Sustainability	Energy, sustainability, climate, emissions, strategy, environment, quality
Supervisory	Mail, phone, estate, report, supervisory, board
Data policy	Data, constant, search, reduced, immaterial, privacy, protection, terms
Int. Standards	ISO, environmental, recycling, plant, zinc, aluminum, policy, water

After topic modelling, topic classification can calculate the probability of text belonging to a specific topic based upon the similarity of words in the text. A low probability means that the text does not fit the topic in an adequate way. A high probability indicates a match between the topic and the content of the text. If the text column is empty, this indicates that the company does not have data on this specific sustainability topic.

Figure 13 shows a remarkably high population of text documents with low probability belonging to a topic. This can be related to the empty columns for specific sustainability topics searched upon when collecting the data. On the other hand, there are 180 text documents with a hundred percent probability belonging to one of the nine topics.

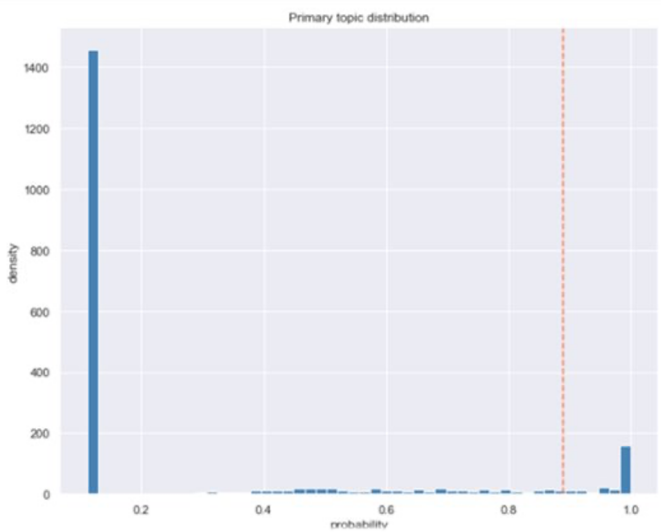

Figure 13. Topic classification model on Google News data from the portfolio showing the probability a piece of text belonging to a specific topic.

An additional step in topic classification is the correlation between one of the nine topics and an organization within the portfolio. The color of the boxes is important when reading figure 14 as the darker the box, the higher the correlation between topic and organization. Such correlation matrix shows ESG strengths and pitfalls in an organized, structured way.

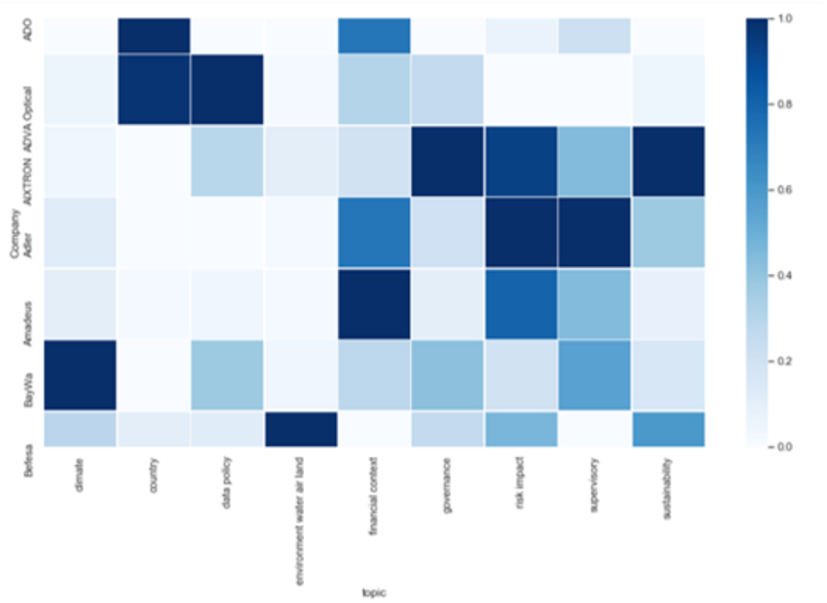

Figure 14. A topic correlation matrix from the full text corpus showing the correlations among text documents and the topics extracted in topic modelling analysis.

At last, the data that was extracted from Yahoo Finance was not analyzed or processed but is directly incorporated in dashboard models (see the next paragraph 3.4.2.). This is due to the filtering options the *yfinance* library provides to extract only desired profile information and financial statements. Therefore, the output files do not contain redundant data which requires no preprocessing tasks.

3.4.2. Dashboard modelling

Dashboard modelling is applied to rate CSR commitment levels on the portfolio. The output of the secondary data analyses was focused on detecting red flags and addressing CSR topics within a portfolio showing high or low commitment levels towards specific topics. Dashboard models show the overview of CSR due diligence and are used for monitoring purposes. The main overview presenting insights into the portfolio's commitment levels is the CSR Scorecard (see figure 15).

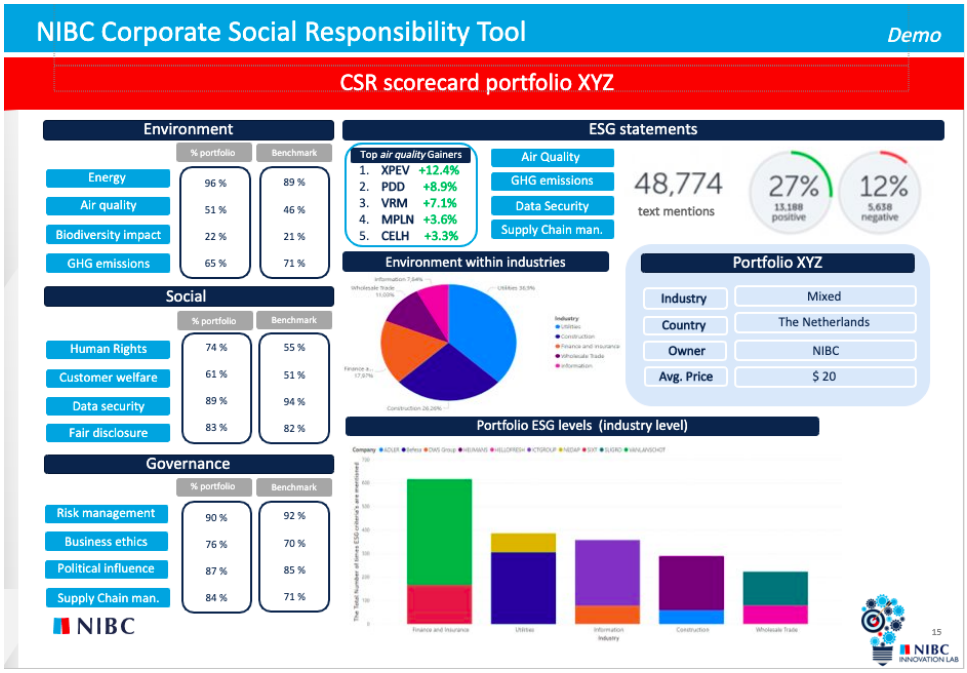

Figure 15. The CSR scorecard presenting the CSR commitment levels of a client's portfolio based on text analysis highlighting the most interesting findings.

The CSR scorecard and its accompanied dashboards were created in Microsoft Power BI, filled with data from the Python data frames originating from secondary data analyses. Due to the broad collection of data, interactive models with various drilldown possibilities were created. Each visual within the CSR scorecard is explained below, from left to right.

- On the left side of the dashboard (figure 15), from top to bottom, the fields E, S and G are mentioned. The first box holds information on E (environment) commitments with the blue boxes holding four words (or word combinations) most frequently used in the portfolio and

their commitment levels. These levels are determined by a threshold of 50% (based on KPI number 2) and connected with the level of commitment from a benchmark. The benchmark is a set of similar organizations combined as a portfolio which the system selects from a data lake. This data lake is filled by the web scraper. In theory this means that if a commitment level of the portfolio is higher than the benchmark's commitment level, the portfolio outperforms the benchmark. This entails that the portfolio is more committed to this topic than their peers.

By clicking on one of the E, S or G field topics, one is redirected to a new board (figure 16) in which ESG levels are modelled in a graph per industry. This graph also highlights two ESG topics based on importance to the portfolio. The importance is measured in the N-gram analyses by high-frequency and in the model's threshold level compared to the benchmark.

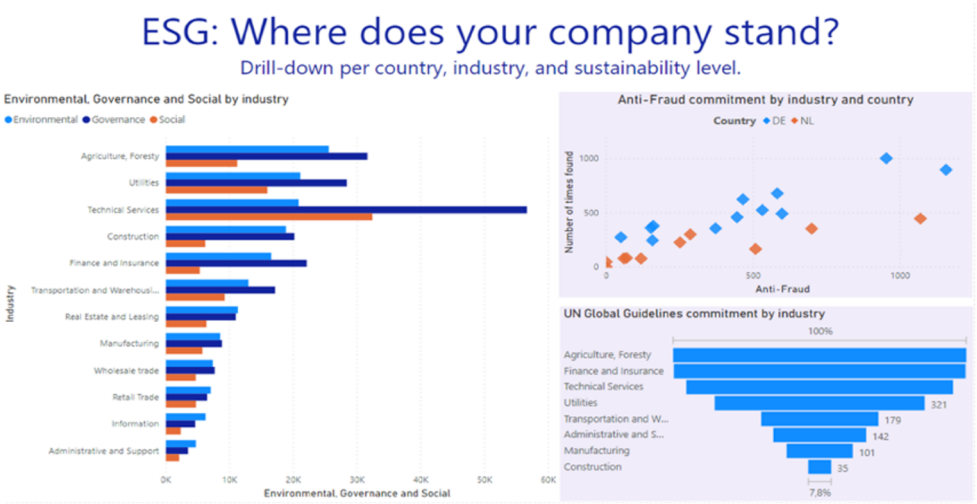

Figure 16. Dashboard presenting ESG classification of a specific organization in the portfolio in order to understand and zoom into a certain CSR specification.

- The upper, middle box states the four most important ESG topics of the portfolio together with the total number of text classifiers analyzed in the sentiment analysis. The ESG topics present which are the portfolio's most valued topics by using the topic classification analysis outcome. The sentiment of the text is gained from the social media sources collected on the portfolio. A total of 3000 articles shows the percentage of positive and negative commitment towards the selected CSR topics.

Figure 17 presents the drill-down dashboard when clicking on the middle box. This dashboard presents findings on ESG topics and their sentiment in more detail. The timestamp is added to ensure data quality and reliability by using up-to-date information.

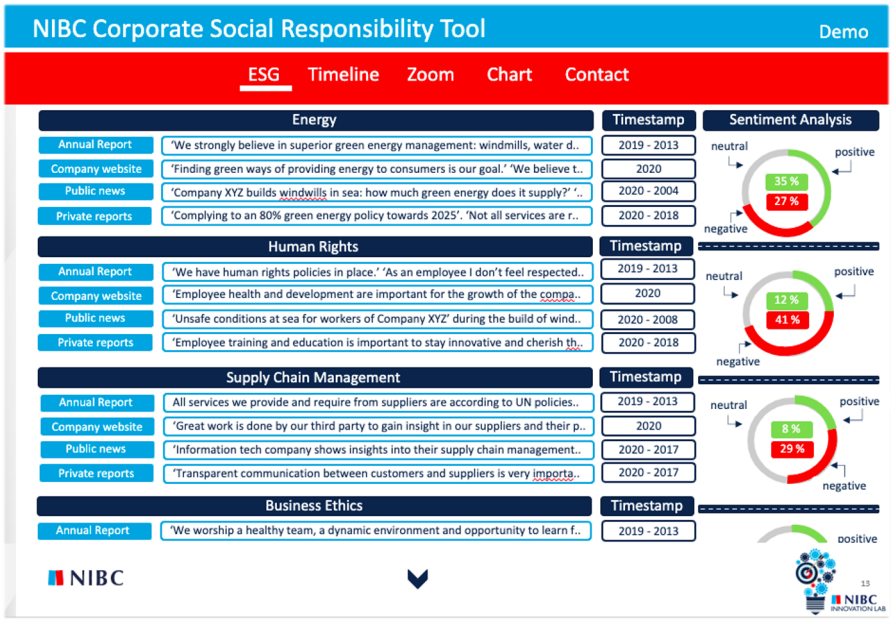

Figure 17. Dashboard presenting the sentiment analysis per ESG topic for the portfolio based on sentiment analyses performed and expressed in visuals and text.

- The pie chart in the middle is a graphical presentation of one of the popular ESG topics from the box above. It presents the industries within the portfolio that are most committed to this specific topic. By clicking on the graph, a new dashboard opens named 'ESG criteria per level' (see figure 18). This dashboard presents a drilldown into organizational ESG commitment over time. It can highlight the country and size (in revenue) of organizations to make a fair and justified conclusion on commitment levels. By clicking on the bar chart, this information becomes visible.

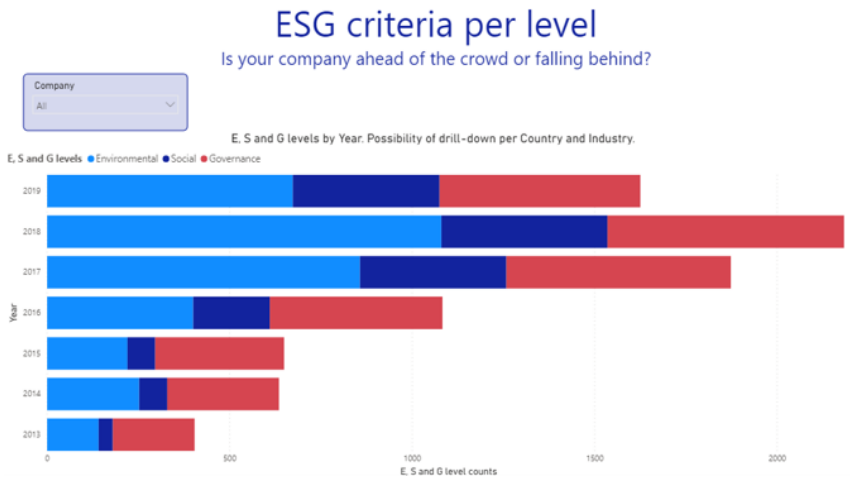

Figure 18. Dashboard presenting the E, S, and G criteria over time of the portfolio.

- The light blue box, on the middle right side, contains the portfolio profile information. This box discloses financial- and background information on the organizations within the portfolio. This data is pulled from Yahoo Finance. Upon clicking the box, a new dashboard will open with more details on the portfolio (figure 19).

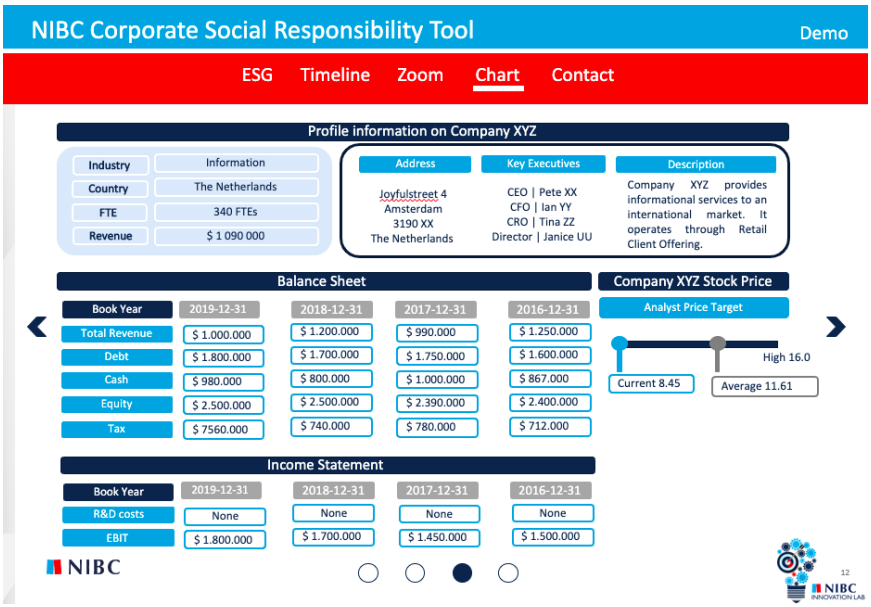

Figure 19. Profile information dashboard showing generic information such as industry, FTE and financial statements on an organization within the portfolio.

- The graph at the bottom of the board presents the overall ESG commitment per country in the portfolio. The dark blue levels are Dutch organizations whereas the light blue ones visualize the German organizations in the portfolio. Such visualizations create a good overview of legislation challenges and ESG gaps within countries. By clicking on a specific industry, it becomes possible to drilldown until company-name level within the main CSR scorecard board.

3.5. EVALUATION

In this step of the CRISP-DM model, stakeholders review and evaluate the outputs from the modelling phase and determine if objectives and requirements are properly met. The evaluation of data models and output is a key step in the project as stakeholders express their trust in the outputs. Based upon this, a decision is made to use the results.

The output types are driven by the objective of this project and the expected upcoming (EU) regulations to increase reporting on sustainability. The main objective of this project is fulfilled when the CSR scorecard and accompanied dashboards present reliable insights into CSR commitment levels

and identify (potential) red flags. The quality of the output has been measured by the validity and reliability of the insights. Reliability is about the consistency of measures used in the dashboards, and validity is about the accuracy of measures used. These quality measurements were held against reports presented by other CSR rating reports by independent organizations. An example of a CSR reporting organization is VigeoEiris.

Due to the incorporation of a diverse selection of data sources, the reliability and validity of the CSR scorecard has increased. This is caused by the addition of media sources which highlight different angles of an organization as well as a representable benchmark to compare results to. The inclusion of these sources of data proved to be beneficial when detecting allegations and helpful to gain insights into the company's accountability in supply- and process chains.

The set of requirements and KPIs are assessed in this step. The first KPI was 90% compliance with the needs of government regulations. This KPI is met in this project. All organizations within the portfolio had mentions of governmental regulations which indicated that they are aware of these regulations and most probably adhere to them. The second KPI was a 50% commitment level towards the E, S, G pillars. This KPI was not fully met as the portfolio held organizations from different industries and countries resulting in uneven distributions of commitment levels in pillars E and G. Not meeting this KPI is not alarming, but rather interesting to identify the cause of these differences in commitment levels. This led to the insight of low standardization levels within Dutch company reporting on the E pillar whereas German companies have frameworks that require them to report on Environmental topics. Low standardization levels cause differences in the number of mentions even though organizations might be active in such topics.

The last KPI, industry-level improvement towards CSR commitment each year, was met by all organizations in the portfolio. Regulations drive most of these growing commitments, however, stakeholders and consumers value CSR responsible organizations more than less CSR responsible organizations. This entails organizations improving brand image when associated with high CSR commitment in business activities. The identification of industry front-runners in terms of CSR is explored this way.

The stakeholders concluded that the portfolio performed better than the benchmark in terms of CSR commitment. The drilldown into the sentiment of media news did not result in the detection of red flags or other concerns that could result in a risk for the bank. The sentiment of the portfolio is interesting for the stakeholders to base conclusions on. Negative sentiments can be identified quickly to explore the nature and severity of the sentiment.

3.6. DEPLOYMENT

The last step in the CRISP-DM methodology is deploying the tool for customers to use. Deployment is the action of bringing resources into effective action which entails bringing the CSR scorecard into use. This entails that the dashboard will be used by the internal teams which will present the findings to clients of NIBC Bank. NIBC Bank would like to be known as a 'green-bank' so investors can trust investments in non-controversial and CSR compliant portfolios.

Internal users could consult this tool when researching an organization or supplier. As the dashboard shows the comparison between the organization itself and a suitable benchmark, it becomes clear whether commitment levels are sufficient for the bank. Discussions with the client could be initiated based upon the output of the CSR scorecard which could lead to more understanding and disclosure of CSR related topics. This is an important outcome for NIBC Bank as co-influencing (potential) clients towards a more non-financial reporting style creates opportunities to increase the transparency in CSR related topics. More transparency in CSR topics could mitigate risks involved in the supply chain and start open conversations to disclose these risks.

In 2020, more than a thousand companies have been mined by the CSR tool of NIBC Bank. This number will only increase as the bank wants to create a solid benchmark to compare organizations' sustainability levels. In order to provide the dashboard with data, the web scraper needs to run on input URLs whereafter the documents need to be preprocessed. These tasks need to be performed by the Innovation Lab who is the owner of the CSR tool. These steps could become a bottleneck in the deployment process. The CSR tool runs on local computers resulting in overwhelmed computers due to large numbers of text documents and processing power required.

Therefore, the need for a shared database is increasing especially when deployment for customers is the next step. Setting up a central data environment where the CSR tool can run and store its documents and dashboards would be a solution. NIBC bank makes use of Azure Cloud which could function as a database for the CSR tool.

The presence of a Cloud database will increase flexibility with regards to automatic dashboarding as well. Ideally, NIBC Bank would like to deploy the CSR scorecard with an input function for a customer (e.g. to-be analyzed portfolio) but running independently (i.e. no manual tasks) and delivering the dashboard a few minutes later. To create a tool like this, all processes must be connected and be triggered by one another.

Another condition for creating a connected CSR scorecard is the type of web scraper. The current framework (Scrapy) scrapes static HTML websites. However, in the future the bank would like to scrape all kinds of websites including JavaScript and dynamic contents. This challenge could be overcome by the development of a second web scraper for input URLs that are JavaScript based. Such a web scraper could be built upon the framework Selenium. Selenium is a slower crawler than Scrapy as it mimics a user's interface on a website.

4. CONCLUSION

The internship was a positive experience and essential for the beginning of my career in the area of Data Science. It has formed the fundament of my professional growth and taught me how and when to apply theories obtained at the University. Next to that, I discovered how to run and complete a project that involved several areas of the data field (from back-end to front-end). Throughout the project, many new ideas, insights and business opportunities were gained and undertaken to advance the project and my practical knowledge as a data scientist. The internship was a kick-start of my career in data, and I am thankful for the opportunities and responsibilities NIBC Bank gave to me.

The most challenging part of the project was consolidating all the input data and adjusting it to the correct ESG framework. In order to judge an organization on its ESG levels, one must ensure the right and relevant data is available. Correct URL's and financial reports as input are required, as well as relevant keywords to search news and (social) media websites for relevant information.

The biggest success during the project was the alignment and connection between the gathering of the data (via web scraping tools and APIs) and the dashboard. The data mining classes at the University were essential for this project. Performing complex analysis and being challenged to develop your skills in a practical way (not only by theory), helped me to develop a holistic view of Data Science. Detecting a problem or challenge, which possibilities there are, and techniques to apply. I also learned that simplicity is important in data analytical projects as clients as well as stakeholders are more engaged then. Most companies are not yet aware nor prepared for detailed data analytical applications which means that even small adjustments can make a significant impact on organizations.

4.1. LESSONS LEARNED

The main lessons I learned during the project at NIBC Bank are the technical challenges I ran into when developing the CSR tool. These challenges are the understanding of the existing web scraping code, the interpretability of the ESG frameworks, and lastly, the relationships between the data sources and variables.

During the project, less time was spent on familiarizing yourself with the existing code and understanding the objective and future-planning of the project within the existing IT infrastructure of the bank. Next to that, it took some time to collect and align the stakeholder requirements. The requirements were formulated in a non-technical way thus required a transformation towards a data driven strategy. Due to my minor experience in data science, this was quite challenging. I have experienced that learning by doing is a way towards professional development.

Undoubtedly the biggest eye-opener during this project, was the lack of understanding from IT-side towards business, and business towards IT. For me, working with data and information infrastructures is amazing, because it combines technical knowledge (mathematics, statistics, computing) with human knowledge (psychology, organizational structures). The combination makes it interesting, dynamic and different in every situation. I am glad and thankful for joining this master at NOVA IMS and the internship at NIBC Bank.

REFERENCES

Ali, D. (2015). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, Github, and More, by Matthey A. Russell: (2013). *Sebastopol, CA: O'Reilly Media, Inc*. 448 pp.

Al-Talib, G. A., & Hassan, H. S. (2013). A study on analysis of SMS classification using TF-IDF Weighting. *International Journal of Computer Networks and Communications Security,* 1(5), 189-194.

Bengford, B., Bilbro, R. & Ojeda, T. (2018). Applied Text Analysis with Python. *O'Reilly*.

Carroll, A.B. (2008); A history of Corporate Social Responsibility: concepts and practices; In A. Crane, A. McWilliams, D. Matten, J. Moon, D.S. Siegel; The Oxford Handbook of Corporate Social Responsibility; p. 19- 46, *Oxford University Press, United States*

Chapagain, A. (2019). *Hands-On Web Scraping with Python: Perform advanced scraping operations using various Python libraries and tools such as Selenium, Regex, and others*. Packt.

Deitel, P. J., & Deitel, H. (2019). *Intro to Python for Computer Science and Data Science*. PEARSON EDUCATION.

Diez-Cañamero, B., Bishara, T., Otegi-Olaso, J. R., Minguez, R., & Fernández, J. M. (2020). Measurement of Corporate Social Responsibility: A review of Corporate Sustainability Indexes, Rankings and Ratings. *Sustainability*, *12*(5), 2153.

Educba (2019). *Text Mining vs. Text Analytics*. Retrieved January 3, 2021 from https://www.educba.com/text-mining-vs-text-analytics/

Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P., & Tech, B. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. I*nternational Journal of Computer Applications, 165*(9), 29-34.

Gupta, R. (2014). Journey from Data Mining to Web Mining to Big Data. arXiv preprint *arXiv:1401.4140*

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, *1*(1), 60-76.

Hand, D. J., & Adams, N. M. (2015). Data Mining. *Wiley StatsRef: Statistics Reference Online,* 1-7.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques*. Waltham, MA: Morgan Kaufmann.

Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A brief survey of text mining. In *Ldv Forum* (Vol. 20, No. 1, pp. 19-62).

Iatridis, K., & Schroeder, D. (2016). Responsible research and innovation in industry. *Doi*, 10, 978-3.

IPCC (2021). Climate Change: The Physical Science Basis. *IPCC*, AR6 WGI.

Jarmul, K., & Lawson, R. (2017). *Python Web Scraping.* Packt Publishing Ltd.

Johansson, S., & Hofland, K. (1989). *Frequency analysis of English vocabulary and grammar*.

Joppe, M. (2000). *The Research Process*. Retrieved January 10, 2021, from http://www.ryerson.ca/~mjoppe/rp.htm

Lavalle, A., Maté, A., Trujillo, J., & Rizzi, S. (2019, September). Visualization Requirements for Business Intelligence Analytics: A Goal-Based, Iterative Framework. In *2019 IEEE 27th International Requirements Engineering Conference (RE)* (pp. 109-119). IEEE.

Li, X., Zaiane, O. R., & Li, Z. (2006). Advanced data mining and applications. In *Proceedings of Second International Conference, ADMA* (pp. 14-16).

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Mitchell, R. (2018). *Web Scraping with Python: Collecting more data from the modern web*. O'Reilly Media, Inc.

NIBC brings the first CLO fully compliant with ESG best practices (2019). Retrieved November 19, 2020, from https://www.financialinvestigator.nl/nl/nieuws-detailpagina/2019/11/19/NIBC-First-CLO-fully-compliant-with-ESG-best-practices

NIBC Leading the Way in ESG. (2019). Retrieved November 22, 2020, from https://twentyfouram.com/eur/2019/11/22/nibc-leading-the-way-in-esg/

Pascual, F. (2019). *Introduction to topic modeling.* Retrieved on March 9, 2021, from https://monkeylearn.com/blog/introduction-to-topic-modeling/

Scrapy (2020). *Scrapy*. Retrieved December 7, 2020, from https://scrapy.org/

Semmie (2021). *Wat is AScX?* Retrieved March 10, 2021, from https://semmie.nl/wiki/AScX.html

Sustainabile Insights Capital Markets (2016). *Sustainable Perspective for the Mainstream Investor.* Retrieved March 7, 2021, from https://www.sicm.com/docs/who-rates.pdf

UNEP, F. (2011). UNEP FI guide to banking & sustainability. *Geneva: United Nations Environment Programme*.

Vigeo Eiris (2020). *ESG Assessment Methodology – Executive Summary*.

Warmerdam, W., de Wilde, J., van Gelder, J. W., & Christopoulou, A. (2019). *ABP's carbon footprint: Trend analysis per asset class and sector.*

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (2004, November). Latent semantic analysis. In *Proceedings of the 16th international joint conference on Artificial intelligence* (pp. 1-14).

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International conference on the practical applications of knowledge discovery and data mining* (Vol. 1). London, UK: Springer-Verlag.

Zaki Rizvi, M. S. (2017). *Web Scraping in Python using Scrapy (with multiple examples).* Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/

Zhang, W., Yoshida, T., & Tang, X. (2008, October). TFIDF, LSI and multi-word in information retrieval and text categorization. In *2008 IEEE International Conference on Systems, Man and Cybernetics* (pp. 108-113). IEEE.