Predicting companies' ESG rating from their 10-K filings using a text mining approach

Auteur : Roufosse, Benjamin
Promoteur(s) : Ittoo, Ashwin
Faculté : HEC-Ecole de gestion de l'Université de Liège
Diplôme : Master en ingénieur de gestion, à finalité spécialisée en Supply Chain Management and Business Analytics
Année académique : 2023-2024
URI/URL : http://hdl.handle.net/2268.2/21642

PREDICTING COMPANIES' ESG RATING FROM THEIR 10-K FILINGS USING A TEXT MINING APPROACH

Jury :
Promoter :
Ashwin ITTOO
Reader :
CHUOR PORCHOURNG

Dissertation by
Benjamin ROUFOSSE
For a Master's degree in Business Engineering with a specialization in Supply Chain Management and Business Analytics
Academic year 2023/2024

Table of contents

List of tables

List of figures

Table of abbreviations

CDP: Carbon Disclosure Project

CSRD: Corporate Sustainability Reporting Directive

ESG: Environmental, Social and Governance

FN: False Negative

FP: False Positive

GRI: Global Reporting Initiative

LR: Logistic Regression

ML: Machine Learning

NB: Naïve-Bayes

NFRD: Non-Financial Reporting Directive

NLP: Natural Language Processing

RF: Random Forest

SASB: Sustainability Accounting Standards Board

SEC: Securities and Exchanges Commission

SMOTE: Synthetic Minority Over-sampling Technique

SVM: Support Vector Machine

TCFD: Task Force on Climate-related Financial Disclosures

TF-IDF: Term Frequency-Inversed Document Frequency

TN: True Negative

TP: True Positive

UNPRI: United Nations Principles for Responsible Investments

Acknowledgment

First and foremost, I would like to warmly thank my supervisor, Pr Ittoo, for accepting to oversee my master's thesis. His availability and advice have been of great value.

I would also like to express my gratitude to Mr Porchourng for the time he devoted to reading my master's thesis.

A special thank you goes to my parents, who have always supported me whenever I needed it. I could not have reached this point without their help.

I also have a special thought for my former classmates, my friends, and everyone who made these five years at HEC Liège a particularly enriching experience, where hard work and enjoyment went hand in hand.

Lastly I would like to thank HEC Liège and their professors, for giving me the opportunity to grow humanly and helping me finding my way into my professional life.

1. Introduction

1.1. Research Context

Since the term ESG first appeared in the 2004 'Who Cares Wins' report, interest in the topic has steadily increased. This report, published by the United Nations Global Compact, aimed to promote the disclosure of sustainability performance in corporate reports. Following the success of this report and other similar initiatives, the United Nations launched the Principles for Responsible Investments in 2006, with the goal of encouraging institutional investors to integrate ESG issues into investment decision-making and ownership practices. Signatories of these principles have grown steadily since then and represent $128.4 trillion in assets under management by 2024 (Principles for Responsible Investment, 2024), reflecting the widespread adoption of more responsible investing principles.

Other ESG-related trends are growing steadily and are expected to continue this trajectory in the coming years. Furthermore, various studies predict that ESG assets, which are way more restrictive than signatories of the PRI, will keep rising throughout this decade (PwC, 2022), potentially reaching $40 trillion by 2030 (Bloomberg, 2024).

These trends reflect a global increase in appetite among both private and institutional investors for ESG-related investments. A major consequence of this trend has been the growing demand for non-financial information from investors (Avetisyan et al., 2016), who now seek not only financial data but also ESG related data to assess the companies during their investment decision-making. In fact, recent years have shown that investors now not only require more detailed information, but that they have become very meticulous when treating this aspect of investments (Taylor et al., 2018).

It is clear that companies have responded to investors' calls for greater transparency and more comprehensive data on their ESG risks and opportunities. Over the past two decades, there has been a notable rise in ESG reporting, either through annual reports or dedicated sustainability reports. From a quantitative perspective, more than 10,000 publicly listed companies reported on their ESG performance in 2019, compared to fewer than 20 in the early 1990s (Grewal et al., 2019). This dramatic increase reflects not only a desire to integrate ESG issues into corporate strategy but also a deliberate effort by companies to position themselves favorably in the rapidly growing ESG rating market, which has been expanding for the past two decades (Escrig-Olmedo et al., 2019).

Accessing reliable information on companies' ESG performance can be a daunting task for investors. To facilitate this, so-called ESG ratings have emerged over the last two decades. These ratings aim to provide an objective measurement of a company’s performance against environmental, social, and governance criteria (Corporate Finance Institute, n.d.). Recognizing the need for this new type of product, many companies have begun developing their own methods to assess ESG performance. The ESG data industry, that started as an isolated niche market with a limited number of players, has evolved in recent years into a thriving industry (Escrig-Olmedo et al., 2019). In fact, it has undergone a so called “consolidation phase”, during which each major rating agency has either acquired or developed their own ESG data services (Escrig-Olmedo et al., 2019). The traditional rating agencies, which until then focused on

providing financial data only, have become social actors with an increased responsibility towards investors.

The consolidation phase in the ESG market, along with the increasing amount of ESG data disclosed by companies, has led to rapid evolution in ESG ratings and their methodologies. While this has benefited investors by providing more and increasingly accessible ESG data, some investors have expressed concerns about a potential decline in the quality of the data provided by rating agencies (Avetisyan et al., 2016).

Furthermore, despite their widespread use by investors seeking ESG information, the ESG ratings have been the source of various controversies, and their consistency has been questioned since their introduction. For instance, many researchers have explored whether the ESG scores provided by different agencies align or show any convergence. The divergence between the ratings provided by different ESG raters is not uncommon; it is often seen that a company rated highly by one agency is rated less favorably by another. These discrepancies arise because there is no standardized method for constructing an ESG rating—each ESG rating provider has its own framework and set of techniques to assess the companies' ESG performance.

Researchers have identified these roots of ESG rating divergence at different levels of the methodology. First, there is a lack of common theorization on what ESG should measure, meaning ESG rating providers do not agree on what should be assessed to determine a company's ESG performance (Chatterji et al., 2014). Second, they use various indicators and methodologies to measure the same ESG criteria (Berg et al., 2022). Furthermore, ESG rating agencies may introduce biases in their assessments, particularly regarding the size of the company (Drempetic et al., 2020). These divergences in the ESG ratings provided by various agencies can confuse investors and lead to a breakdown of trust between them and the rating agencies. Given the importance of ESG ratings today, addressing these issues is crucial. In fact, their widely usage by investors is likely to influence of the allocation of lots of investments (Chatterji et al., 2009).

We can also add a practical consideration when discussing ESG ratings: although ESG data is now available to anyone interested in assessing the ESG performance of companies, it is important to note that access comes at a significant cost. For instance, a subscription to a financial data platform from a major rating agency, such as Bloomberg or Refinitiv, ranges between $20,000 and $30,000 annually.

These various controversies surrounding ESG ratings have led researchers and investors to explore new methods for improving and standardizing the assessment of companies' ESG performance. Recent studies highlight the potential of machine learning algorithms in predicting ESG ratings of companies. The significant advancements in this field, coupled with the increasing availability of corporate data, allow models to provide accurate predictions of ESG ratings, potentially addressing some of the current shortcomings in the existing methodologies.

In this section, we have seen that ESG has evolved significantly over the past 20 years, with the rise of ESG ratings developing in parallel. However, we have also identified concerns from investors, particularly regarding the lack of transparency and standardization in these ratings, leaving considerable room for improvement. Given the vast amount of available corporate ESG

data and the advancements in technology, there is great potential to enhance both the accuracy and transparency of ESG ratings.

1.2. Research Objectives and Knowledge Contribution

Having recognized the various concerns surrounding ESG ratings and the methodologies behind them, this master's thesis is dedicated, at our humble scale, to developing a systematic method for accurately predicting the ESG ratings of companies. Specifically, we aim to train several supervised machine learning classifiers on ESG-related corporate data, and evaluate the extent to which these models can accurately classify the companies into their respective ESG rating category.

To achieve this, we will leverage the textual ESG-related information contained in the annual document of US companies, the 10-K filings. These documents, which are required to each US listed company, are publicly accessible via the Security and Exchange Commission's website. They are presented in a highly standardized format and their content can be easily extracted and analyzed using text mining techniques. Once their content is transformed in a structured format, we aim to train various supervised classification models on the ESG-related information contained in them. In these models, the ESG-related information will serve as the independent variables while the target variable will be the ESG ratings provided by Refinitiv, a well-known ESG rating agency. The machine learning models include Support Vetor Machine, Gaussian/Multinomial Naïve-Bayes, Logistic Regression and Random Forest. Each model will be trained on five distinct dataset, corresponding to five different 10-K filing years.

By developing this methodology we aim to answer the following research question:

"Can the ESG textual information disclosed by US listed companies in their 10-K filings be used to accurately predict their ESG rating?"

While various studies aimed to predict ESG ratings based on financial data (Garcia et al., 2020; Agosto et al., 2023; D'Amato et al., 2021), such as credit ratings for instance, less studies focused on determining if ESG corporate data could be used to predict them. Furthermore, to our knowledge, no other study has specifically used the ESG information contained in the 10-K filings of U.S. listed companies. We believe that the standardized structure, easy accessibility, and increasing quantity of ESG information in these filings make them a valuable data source for training machine learning classifiers to predict companies' ESG ratings. Furthermore, by training the various ML models on five separate dataset (corresponding to the 10-K filing years), we aim to evaluate whether the evolution of ESG-related information in these filings enhances the performance of the ML models.

1.3. Thesis Structure

The second chapter will be devoted to the literature review and will be divided into two parts. The first part aims to provide the reader with a comprehensive overview of the broad ESG topic. It will cover the history of the concept, offer insights into the methodologies behind ESG ratings, delve into the limitations and critiques of the subject, and present the current trends in the field. The second part will focus on text mining. It will provide general definitions and guide the reader

through the essential steps for successfully mining documents. Additionally, it will provide an overview of applications of text mining in the context of ESG analysis.

The third chapter will be devoted to the developments of this thesis and will consist of three parts. The first part will introduce the data used to train the ML models. This will involve a justification of the data sources selected, followed by a detailed explanation of the steps carried out to collect the data and how the various datasets were constituted. Descriptive statistics of the datasets used will be provided to give the reader a clear understanding of the data. The second part will detail the procedure that will be used to build ESG rating prediction models. It will begin with an explanation of the two approaches that will be followed to answer the research question. An important section will then describe the datasets that will be used, as well as the resampling strategies that will be applied. It will then describe the pre-processing steps that will be applied to transform the 10-K filings into usable data. This will be followed by an introduction to the different ML models on which the data will be trained with. Additionally, the methods used to evaluate the performance of these models will be outlined. Also, we will briefly discuss the procedure that will be put in place to ensure that the ML models are optimal.

The fourth chapter will be devoted to presenting the results of the prediction models implemented in accordance with the methodology. For each approach, we will first provide an overall assessment of the classifiers' performance. Following this broad overview, we will delve into more specific and detailed results, enabling a finer analysis. This chapter should allow us to answer the research question and identify which models perform best, as well as determine which datasets were most adapted for training.

The fifth chapter will aim to discuss the results presented in the fourth chapter.

The sixth chapter will conclude this master thesis. It will allow conclusions to be drawn through the discussion in the fifth chapter. It will also provide suggestions for improvement and explore potential directions for future work on the subject.Literature Review

2. Literature Review

2.1. Review of ESG Topic

The term ESG first appeared in 2004 in a report published by the United Nations Global Compact. This initiative, backed by 23 financial institutions, aimed "to develop guidelines and recommendations on how to better integrate environmental, social, and corporate governance issues in asset management, securities brokerage services, and associated research functions" (United Nations Global Compact, 2004). Before delving deeper into the challenges associated with ESG criteria, it is important to understand what each pillar encompasses.

Each pillar focuses on a specific set of issues that companies should address:

The environmental pillar focuses on the assessment of a company’s behavior towards the natural environment (Senadheera et al., 2022). It includes key issues such as climate change and carbon emissions, air and water pollution, biodiversity, deforestation, energy efficiency, waste management and water scarcity (CFA Institute, n. d.).

The social pillar emphasizes the relationships a company maintains with people and institutions in the communities where it operates (Henisz et al., 2019). This includes aspects such as gender policies, protection of human rights, labor standards, workplace and product safety, public health, and income distribution (Billio et al., 2020).

At last, the governance pillar focuses on how a company deals with its external stakeholders (Deutsche Wealth, n. d.) This pillar is essential for ensuring the effective implementation of both the environmental and social pillars (World Economic Forum, 2022). It emphasizes issues such as the independence of the board of directors, the protection of shareholders' rights, the executive compensation practices, the internal control procedures, and the prevention of anti-competitive practices. Additionally, it covers issues related to bribery and corruption, as well as the company's adherence to legal and regulatory standards (Billio et al., 2020; CFA Institute, n.d.)

2.1.1. ESG Performance

The ESG performance of a company refers to the effectiveness of its management in integrating ESG criteria into its business operations, by using a set of ESG metrics for instance. Just as financial metrics help traditional investors assess a company's financial performance, ESG metrics can be used to evaluate a company's exposure to risks and opportunities related to ESG issues (Kay, 2020).

2.1.2. From SRI to ESG

As previously stated, the term ESG appeared for the first time in 2004. However, considerations of non-financial information in investment decision processes have existed since the 19th century (Eccles et al., 2019). For instance, early faith-based organizations such as the Quaker Friends Fiduciary Corporation already restricted their investment decisions by avoiding the ownership of so called "sin stocks", which consisted in excluding weapons, alcohol and tobacco from their portfolio (Roselle, 2016). In the late 1920s investment funds were created in the USA to meet the needs of religious groups by avoiding investments that did not align with their values (Eccles et al., 2019). In the 1970s, investors seeking to align their portfolio with their values widely adopted this investment approach, which eventually became known as Socially Responsible Investment. The excitement around SRI peaked in the 1980s, driven by initiatives such as the Vietnam anti-war movement and anti-apartheid campaign (Krantz, 2024). Nonetheless, the focus on ESG within this investment approach only became mainstream in the 1990s (Krantz, 2024) as an increasing number of investors became aware of climate change (Camilleri, 2020).

In 1995, the first comprehensive measurement of sustainable investments by the US Sustainable Investment Forum showed that investors were gradually shifting toward more sustainable choices (Krantz, 2024). Over the following years, investors recognized the potential for improving financial performance by addressing ESG issues, leading them to use metrics to measure companies' sustainability performance. Additionally, several initiatives, such as the GRI and the CDP, were created to help companies address and report on these issues (Krantz, 2024). This growing trend culminated in the release of the "Who Cares Wins" report in 2004, where the term ESG was officially introduced. Following this publication, another initiative, the UN Principles for Responsible Investments, was created in 2006 to link ESG factors to investment

performance (Caplan et al., 2013). This initiative provides guidelines that investors can voluntarily follow to integrate ESG factors into their decisions (Principles for Responsible Investments, 2024). The increased emphasis in ESG-related issues for two decades is also evident from a quantitative point of view, as mentioned in the introduction, since there are now more than 5000 institutional signatories of the PRI, accounting for more than 128$ trillion in assets under management.

2.1.3. What is Driving ESG?

Given the growing interest in these investment practices, it is worth asking why investors are turning to them. This section aims to identify the various drivers of this upward trend.

The first driver of ESG practices is linked to the growing social awareness of investors (Reneboog et al., 2008; Deloitte, 2024), showing their willingness to address climate change (Micilotta et al., 2018). Furthermore, it seems that the emergence of younger generations of investors tends to amplify this driver. In fact, GenZ and millennials particularly are more likely to look at the social impact of a company before investing in it (Miler, 2023) and more likely to support environmental and social issues (Haber et al., 2022). One should also note that several studies found that investors are also "ready to earn less in order to do good". In fact, various studies outline the fact that some investors are willing to lose some of their financial returns to be in line with their values (Kräussel et al., 2024; Giglio et al., 2023; Riedl et al., 2017).

The second driver is linked to the investors' belief in the pertinence of ESG investments as credible financial opportunities (Micilotta et al., 2018). Those investors believe that companies with strong ESG practices are better positioned to manage risks and achieve long-term value as these companies are seen as more resilient to various challenges (OECD, 2020), including regulatory changes, environmental risks, and social pressures (Bell, 2021). Also, Giglio et al. (2023) suggest that investors believing in the outperformance of ESG investments tend to allocate a greater part of their assets in ESG related assets. Lastly it is important to bring some nuance to this driver as studies focusing on the relationships between ESG and financial performance do not deliver consensual findings. In fact, since the 1970s, researchers have studied the link between ESG performance and financial performance, with mixed results (Friede et al., 2015). Early studies (Vance, 1975; Boyle et al., 1997) found a negative relationship, suggesting investors did not see ESG efforts as beneficial. Later research showed neutral or insignificant impacts (McWilliams et al., 2000; Reneboog et al., 2008). More recent studies, however, suggest a positive relationship between ESG and financial performance (Dimson et al., 2020). Meta-analyses by Friede et al. (2015) and Whelan et al. (2020) also support this positive connection, though more data is required for confirmation.

The third driver is linked to the regulatory requirements that companies face (Bondar et al.). This driver is experiencing a quick evolution as many regulations have been launched lately or will be launched in the coming years. However, those regulations are still heterogeneous and strongly depend on the country a company operates. For instance, the European Union already has an ambitious plan for regulating ESG investments while the US still lacks a clear agenda Matos et al.). A subsequent chapter will delve deeper into the regulations imposed by the United States and the European Union.

We can therefore see that the increase in these investment practices is primarily due to investors' desire to be in tune with their values. This is followed by the belief that these investments can bring an interesting financial return, and finally the desire to comply with increasing regulations.

2.1.4. ESG Reporting

Investors can rely on a set of tools to asses a company's ESG performance, among those are CSR and sustainability reports disclosed by companies (Antolín-Lopez et al., 2016). In fact, the last two decades have seen an increase in this practice along with the growing trend of more responsible investment practice (Vartiak, 2016). In fact, companies have recognized that investors require more non-financial information (Bose, 2020), and data shows that they have responded accordingly to these signals. In fact, ESG reporting has become a common practice across the world's largest corporations, with 96% of the 250 largest companies reporting on sustainability or ESG matters according to a KPMG (2022) study. In the United States, Bloomberg estimated in 2022 that 96% of S&P companies reported on their ESG activities, reflecting a voluntary effort by these companies to disclose such information. This is especially important considering that, at the time, there were no mandatory reporting requirements on ESG issues for American companies. This indicates that they were driven to report on these activities for reasons beyond regulatory compliance. However, discussing those reasons falls outside the scope of this literature review.

This increasing demand for more reporting from the companies on their ESG performance has led to the development of various standard-setters and sustainability accounting frameworks (Bose., 2020), which aim the companies at reporting in a standardized way. In fact, companies can rely on a large number of frameworks and standards to report on their ESG performance. While frameworks focus more about principles, standards provide detailed requirements, such as specific metrics to report about the different ESG matters (Byrne, n.d.). The most widely adopted framework is the GRI (Bose, 2020). However, various non-governmental organizations have also designed their own reporting framework. Among those, the ISO 26000 designed by the International Organization for Standardization or the B-Corp certification for sustainable companies (Antolín-Lopez, 2016).Other utilized reporting frameworks include frameworks whose target audience are the investors rather than the stakeholders in general, for instance the International Integrated Reporting Council the Sustainability Accounting Standards Board, or the EFFAS (Bose et al., 2020, La Torre et al., 2018). Lastly, there are climate change-related frameworks that focus more on the environmental dimension of ESG, such as the Climate Disclosure Standard Board, the Carbon Disclosure Protocol and the Task Force on Climate-related Financial Disclosures (Bose et al., 2020).

Lastly, we must talk about the regional differences regarding the ESG reporting requirement regulations. Particularly, the situation is very different in the European Union and in the United States.

The European Union is the most strict economic region regarding ESG disclosure. In fact, it possesses an advanced legislation on the matter and it concerns many companies. The EU started its sustainability disclosure in 2014 by putting in law the Non-Financial Reporting Directive (European Parliament, & Council of the EU, 2014). This directive applied to large companies of more than 500 employees and required the disclosure of certain information

related to environmental, social, employee, human rights, corruption and bribery matters (Hummel et al., 2024). It came into law in 2017. However, this directive was the source of various critiques, notably for its shortcomings in ensuring the comparability, consistency and reliability of the required information (Hummel et al., 2024). Also, the scope of applications was limited on too few companies. As a result, the EU reviewed the CSRD, which gave birth to the Corporate Sustainability Reporting Directive in 2022 (European Parliament and Council of the EU, 2022). The key features of this new directive are the larger companies that are in scope, an extension of the required information as well as a strengthening of the double materiality concept (Hummel et al., 2024). An important step towards the standardization of sustainability reporting standards was also achieves by the EU by adopting in 2023 the European Sustainability Reporting Standards, which will require every company in scope of the CSRD to report about its ESG issues according to these standards (Hummel et al., 2024).

The situation in the U.S. is quite different. In fact, there is still no mandatory requirement for companies to disclose information about their ESG performance. However, as mentioned earlier, most large companies report on their ESG issues, primarily following the SASB, TCFD, and GRI disclosure frameworks. U.S. companies are required to comply with SEC regulations, which traditionally mandate the disclosure of information material to investors—typically financial information. In recent years, however, there has been an increase in the disclosure of non-financial information in SEC-required documents, such as the 10-K filings (Gez et al., 2022). Notably, these disclosures have largely been made on a voluntary basis. More recently, the SEC decided to adopt its Climate Disclosure Rule, which would require companies within its scope to disclose their climate-related risks (SEC, 2024). However, this new rule has been paused due to legal challenges following its release (ESG Today, 2024).

We have seen in this section that companies have numerous frameworks at their disposal to evaluate the sustainability performance of companies. However, this multiplicity of ways to report about sustainability can lead to confusion among the investors (Bose, 2020, Brown et al., 2009). Furthermore, assessing the ESG performance of accompany by analyzing would be a repetitive and time-consuming task, particularly for retail investors. That is why there exist other tools that investors can use to assess a company's ESG performance, such as the ESG ratings.

2.1.5. ESG Ratings

Not surprisingly, the demand for ESG ratings by investors has risen over the last decade, such that a 2020 study shows that it is now the preferred source of ESG information among investors (SustainAbility, 2020).

This section aims to give the reader an extensive understanding of the ESG ratings landscape by explaining the different rating providers and the methods they use to create these ratings. It will also cover why these ratings have been criticized in recent years and highlight the latest trends in the field.

2.1.5.1. Definition

An ESG rating or score can be defined as "an opinion regarding an entity, issuer, or debt security's impact on or exposure to ESG factors, alignment with international climate agreements or sustainability characteristics, issued using a defined ranking system of rating

categories" (Mazzacurati, 2021). The purpose of an ESG rating is to provide a scoring framework that investors can rely on during their investment decision-making process, allowing them to compare companies against one another (Pagano et al., 2018).

2.1.5.2. ESG Rating Providers

As we mentioned in the introduction, the rise of SRI investments in the past two decades has given birth to companies specialized in the assessment of other company's ESG performance (Berg et al., 2022). These companies, the ESG rating agencies, offer services similar to those provided by credit rating agencies, but with a focus on ESG criteria (Avetysian et al., 2017; Scalet et al., 2009; Berg et al., 2022). By providing an accurate reflection of a company's management of its ESG issues, they ensure that the investors use consistent data that can be used during investment decision. ESG rating agencies offer most of the time quantitative information over the companies, but some agencies also provide narrative discussion (Scalet et al., 2010). It's important to note that the data provided by ESG rating agencies primarily focuses on the assessment of companies, but some agencies also offer data on institutions or countries. For example, MSCI provides ESG government ratings.

The expertise of the ESG rating agencies has been recognized not only by companies, but also by finance professionals and academics (Escrig-Olmedo et al., 2019). As a result the size of its market has been rising since the appearance of the first ESG rating agency (EIRIS) in France in 1983 (Berg et al., 2022). At the time, it was a niche market that very specific clients, wishing to align their values with their investment decisions. Since 2005 particularly, the ESG rating industry has undergone an important phase of consolidation with many mergers and acquisitions that were driven by financial motives mainly (Avetysian et al., 2017). However, we can note that it has brought some negative impacts, such as a decrease in the ESG data quality.

Even if this period of consolidation has slow down but we can still except the industry to evolve (Escrig-Olmedo et al., 2019). In 2021 we could identify among the various ESG rating agencies different classes of providers (Bouyé et al., 2021):

- First, we have the major players in the financial industry, such as S&P, Moody's, Bloomberg, FTSE Russell, MSCI, Thomson Reuters, and Morningstar. These "big players" have recognized the growing demand for ESG data from investors and have either acquired or established their own ESG data services.
- Next, there are specialized firms whose core business focuses on providing ESG research. Examples include RepRisk, Arabesque, Covalence, CSRHub, Ethos, Inrate, RobecoSAM, Oekom Research, Vigeo Eiris, and Sustainalytics.
- Finally, there are consultancy companies that offer ESG research services with a focus on only one of the three ESG pillars. For instance, the Carbon Disclosure Project focuses on climate change and water, Trucost specializes in environmental risks, and ISS is known for its governance research.

Even though each ESG rating agency has its own methodology for providing investors with ESG ratings, the major ESG rating agencies follow similar processes. The aim of the next section is to provide the reader with an overview of these processes.

2.1.5.3. ESG Ratings Methodologies

ESG ratings are developed by ESG rating agencies through a comprehensive process that involves several steps. The process typically begins with data collection, which is then analyzed and used to calculate a set of metrics. These metrics are subsequently aggregated according to a defined weighting system. Finally, the different pillars are combined to form the overall ESG rating.

Data

The ESG data providers typically begin their analysis by collecting ESG data. They rely on various sources to build a robust database. These sources can be public, quasi-public or private (Larcker et al., 2022). Public sources include any information accessible without restriction, such as company filings with the SEC, company-produced sustainability reports, press releases, newswires, and media reports. Quasi-public sources include data obtained from government, regulatory bodies, or NGO datasets. Private sources involve information directly collected from companies, often through questionnaires or similar methods (Larcker et al., 2022). When necessary, ESG ratings providers may also rely on third-party data providers to further enhance their database.

However, collecting data does not come without issues according to Larcker et al. (2022), who identifies three challenges when collecting data:

- ESG rating agencies often encounter challenges with missing data. To address this, agencies might omit the missing data, make assumptions based on industry averages (as MSCI does), or assume the worst-case scenario (as FTSE does) to encourage transparency.
- Companies often report ESG data using different scales. For instance, the performance of a certain issue might be reported using raw numbers, rates or percentages, which makes direct comparison challenging. As a result, the agencies must ensure that the data they use is standardized to calculate consistent overall ESG scores across companies.
- To maintain consistency in their models, ESG ratings providers may retroactively adjust historical data as better information becomes available. While this practice can improve the accuracy of the models, it can also make past data appear more predictive than it was at the time.

Framework

In order to provide a rating, the major agencies divide their overall rating into several pillars. A common approach consists in assessing each ESG pillar individually. ESG rating agencies then typically select a set of theme for each ESG pillar:

- Environmental: Factors such as carbon emissions, energy efficiency, waste management, and sustainability initiatives.
- Social: Issues including workforce diversity, human rights, community engagement, and data privacy.
- Governance: Aspects like board diversity, executive compensation, shareholder rights, and corporate ethics.

For each of these themes, various metrics are computed and aggregated to obtain a score. The resulting scores are then combined to produce a pillar score using a specific weighting system, which varies depending on the ESG rating agency. This crucial step involves determining which factors are most relevant and important to each company. Since this step is highly subjective and depends on the methodology of each data provider, it becomes a significant source of variation in ratings (Boffo et al., 2020). For example, Sustainalytics relies on GRI metrics to define its materiality metrics, while MSCI uses a set of industry-specific metrics from the SASB.

This process allow ESG rating agencies to provide traditional ESG ratings that therefore reflect the ESG performance. In recent years, some agencies have also set up tools to incorporate controversies which might hurt the company's reputation into their ESG scoring model. Those ratings help investors spot companies whose involvement in controversies in controversies could negatively impact them.

In a previous section, we discussed that the ESG rating industry comprises many different ESG rating agencies. This has a significant consequence; the methodologies used to build ESG ratings vary among these agencies. There is an absence of a common framework meaning that each ESG rating provider has its own structure, hierarchy, weighting system, and set of metrics (Köbel et al., 2022). This has led researchers to interest themselves in the topic.

2.1.5.4. Limitations and Critiques

A very common critique addressed to ESG rating lies in the lack of standardization across the various agencies. In fact, we have seen in the previous section that each agency had its own framework to build their ESG ratings. This has led researchers to interest themselves in the possible consequences of this constatation. Several studies have examined whether the ratings provided by different agencies converge. Chatterji (2014) found that the ratings from various large ESG rating agencies were not correlated, revealing a surprising lack of agreement among them. The study also emphasized that this disagreement was not simply due to the use of different frameworks, but rather stemmed from deeper issues in how ESG was measured. These findings were confirmed by another study, which identified substantial divergence in ESG ratings across different agencies (Berg et al., 2022). This research further noted that the primary source of divergence lies in the use of different metrics to measure the same issues. Furthermore, the disclosure of more ESG data does not reduce the divergence, it makes it more important, notably for the environmental and social pillars (Christensen et al., 2021). A last study found the same results, highlighting minimal correlation between the different ESG rating agencies (Dimson et al., 2020). This last paper argues that the divergence might be rooted, for some part, in the different weighting schemes. Kotsantonis and Serafeim (2019) also discussed the reasons behind the divergence in a paper. They identify four reasons: data discrepancies, benchmarking differences, data imputation methods and information overload. The fact that ratings of various agencies are not correlated has significant implications for both investors and companies. On the side of the it increases the risk of misallocated capital due to inconsistent company assessments, which is concerning given the important amount of investments ESG can impact (Chatterji, 2014). On the side of the companies, those are not getting real incentives to improve their ESG performance (Berg, 2022).

Another common critique of ESG rating agencies is the lack of transparency in their methodologies. This problem does not allow the investors to understand what lie behind the

ratings. The fact that the agencies are reluctant to fully disclosing their methodology (Scalet et al., 2009) lie in the commercial character of the ratings agencies (Windolph, 2011). In fact, this is an important barrier to uniformity (Stubbs, 2013). Among the things for which the lack of transparency is problematic, there is often little transparency about the peer group with which a company is compared during its assessment (Kotsantonis et al., 2019), which has an important impact on its assessment. This lack of transparency leads to frustration among the investors, who clearly prefer agencies with more transparent practices (SustainAbility, 2021). For improving their transparency, Windolph (2011) suggest disclosing their methods as well as the content of their surveys. However, the slowly changing regulation over ESG disclosure, in EU notably, will oblige the raters to adapt their level of transparency (The SustainAbility Institute by ERM, 2023). Greater transparency on the ESG ratings agencies methods would benefit investors and other stakeholders as they could compare agencies ratings more easily (Berg et al., 2022).

A last subject of controversy that ESG ratings have suffered from is the existence of potential bias in ESG ratings. In fact, the literature has identified, among others a firm size bias, meaning that the larger a company, the better is the rating it gets. The reasons that drive this bias are to lie in the fact that larger companies have more resources they can spend in managing their ESG issues. Or it could also be due to the greater amount of ESG information they disclose (Larcker et al., 2022). This hypothesis was tested by another study, that found a positive relationships between the quantity of ESG data a company disclose in the public domain and the rating it gets (Chen et al., 2021; Hughey et al., 2012), outlining the fact that investors should thus look further than the ratings to judge if a company is a good pick or not. It can also be noted that the higher concentration of English news sources in the ESG controversies data of various agencies results in a selection bias. For example, if an ESG rating agency mainly relies on media sources from English-speaking countries, the resulting data may disproportionately represent companies from those regions while underreporting issues in non-English-speaking regions (Barkemeyer et al., 2023). A last study also identifies an industry as well as a geographical bias (Bruder et al., 2019).

2.1.5.5. Recent trends

Recent trends in the broader ESG landscape include new EU regulations on ESG ratings, which aim to address trust issues that have emerged around them. Specifically, these regulations seek to improve the transparency of the methodologies used to construct ESG ratings, thereby making the ratings more reliable and comparable (Council of the European Union, 2024). For the ESG rating agencies, this will bring some changes as they will be required an authorization of the ESMA to sell its data. It marks an important step into more standardized ESG ratings

2.2. Text Mining: Definition and Applications in ESG

Since the first part of the literature review highlighted the increasing quantity of ESG information available over the past two decades, it is essential to have techniques to analyze this data efficiently. Analyzing different textual sources individually can be repetitive and time-consuming. This is where text mining comes into play, allowing for the automatic processing of textual information. This chapter is organized as follows: first, we will provide an overview of text mining, including the various steps it involves and its common applications. Next, we will

explore how text mining has been applied specifically to ESG data by examining relevant studies and use cases.

2.2.1. Text Mining

2.2.1.1. Definition

Text mining, also known as text analytics, is a process that aims to discover new and valuable information by automatically extracting and analyzing data available under textual format (Segall et al., 2007). This process typically begins with gathering data in an unstructured (and thus initially unusable) format, which is then transformed into a structured format. Once structured, algorithms and methods from the field of machine learning can be applied to the data to identify useful patterns (Hotho et al., 2005).

As it has just been mentioned, most textual data comes in a format that computers cannot directly understand, making it difficult to process and analyze. In fact, textual data is typically either unstructured or semi-structured. Structured data is organized in a fixed schema and typically resides in relational databases or spreadsheets (Vijayarani et al., 2015). Unstructured data lacks a predefined structure and is therefore more challenging to analyze; it typically includes text documents, images, or emails. Semi-structured data falls somewhere between raw and structured data (Abiteboul, 1997). Well-known semi-structured data formats include HTML and JSON. Semi-structured data combines aspects of both structured and unstructured formats by using tags or markers to provide some organization without enforcing a rigid structure (Abiteboul, 1997). Consequently, raw data, whether unstructured or semi-structured, must first undergo preprocessing steps to be usable (Hotho et al., 2005). The next section aims to provide the reader with an overview of the most common methods used to preprocess textual data.

2.2.1.2. Text Preprocessing Steps

Mining textual data typically starts with some preprocessing steps (Vijayarani et al., 2015). The goal of this process is to transform each textual document contained in a collection into a structured representation. In text mining, we can use different approaches to represent documents in a structured format.

The simplest approach to represent a text document is the bag-of-words (BoW) model. This approach involves representing each document as a set of words contained in the document, ignoring the syntactic structure and semantics of the text (Hotho et al., 2005).

A more commonly adopted approach in text mining is the vector space model (VSM) (Salton, 1975). In this model, each document is described by a vector, where each cell of this vector represents a word from the entire document collection (Hotho et al., 2005). There are different methods to encode the value of the vector elements. One approach is binary: a value of one is assigned if the word is present in the document and zero if it is not. Another approach is to compute the frequency of each word in the document, or to use a technique called term-frequency inverse document frequency, which will be discussed later in this thesis.

Although this model particularly suits the purposes of this thesis, it has certain limitations. The VSM assumes that the position of words is independent, meaning it does not account for semantic relationships (Manning et al., 2008). For instance, a word like “car” is semantically

similar to “automobile,” but the VSM treats them as different words. Additionally, the VSM does not handle polysemy, meaning it does not take context into account, which can lead to misclassification. For example, the word “bank” can refer to a financial institution or the side of a river, depending on the context. Finally, the term independence assumption prevents the VSM from capturing concepts formed by combinations of terms. For example, “climate change” is treated as “climate” and “change,” missing the fact that they represent a single concept when combined. Overcoming these problems can be achieved by using embeddings (e.g., Word2Vec) or conceptual embeddings (e.g., BERT models or ELMO). In conclusion, even though the vector space model has its shortcomings, it is still commonly used for text classification (Miner et al., 2012).

As mentioned at the beginning of this section, documents need to be preprocessed to be represented either as a BoW or as a vector in the VSM. The next section provides an overview of the commonly applied steps to preprocess documents.

This section will provide an overview of the steps that need to be done in order to transform a text document into a bag-of-word. This process is essential in order transform the raw, unstructured or semi-structured format into a structured format. Hence, it will be used to reduce the size by removing words that carry little to no explanatory power such as articles, pronouns or punctuation (Hotho et al., 2005).

Preprocessing a text document starts with the tokenization of the text, i.e. the splitting of the text into discrete items called tokens. For the English language, it can be achieved by considering white space and punctuation as delimiters (Miner et al., 2012)

After this first step, filtering methods can be applied to the documents in order to further decrease the size of the dictionary (Hotho et al., 2005). Common filtering methods include the removal of stop words like articles conjunctions and prepositions (Hotho et al., 2005). In fact, those words usually have little explanatory power and can thus be removed.

The next preprocessing step consists in applying a stemming algorithm to the tokens, i.e. transforming related tokens into a single form. The aim of this algorithm is to remove prefixes, suffixes and pluralizations (Miner et al., 2012), such that after the algorithm has been applied each token is represented by its stem (i.e. its simplest form). For instance words like “walk”, “walks”, “walking”, “walked” or “walker” will be transformed into “walk”. This preprocessing step will thus further decrease the dimensionality of the vector space model (Hotho et al., 2005). This concept is similar to another more advanced concept called lemmatization. It involves reducing words to their root form, which is called the lemma. Lemming algorithms consider the context of each word as well grammatical information (e.g. part of speech) in order to reduce each word to its lemma. For instance, “meeting” can either be a noun or a verb depending on the context it is being used. The lemming algorithm will reduce “meeting” to “meet” in case it serves as a verb but will keep “meeting” it if it serves as a noun (Miner et al., 2012).

The last preprocessing steps usually involve removing the punctuation and case normalization, i.e. either lowercasing or uppercasing all words (Miner et al., 2012).

Once all steps have been applied to the text document, we can “vectorize” the tokens. There exist two main forms of vectorizations. The first one consists in assigning a value to each token

present in the preprocessed text document and this value will be equal to the frequency of the token in the document. The second technique, known as TF-IDF, adjust the term frequency by how commonly the term appears across all documents. In other words (Miner et al., 2012), it penalizes terms that are very common across all documents. For instance, the word “the” is likely to appear very often in the English language. Thus, it is very likely that it will appear many times in a single document, but it will also appear very often in all documents which will lower the weight of the term (Miner et al., 2012).

2.2.1.3. Text Mining Techniques

After a text document has been preprocessed and presented in a structured format, such as a vector in a VSM, various data mining techniques can be applied to accomplish different tasks. These techniques include text classification, information extraction, information retrieval, clustering, text summarization, natural language processing, and more (Allahyari et al., 2017; Talib et al., 2016; Jusoh et al., 2012).

In this thesis, we will focus on one specific technique: text classification. Given that the empirical part of this thesis involves accurately classifying company documents into their respective ESG categories, it is essential to review this technique in detail.

Text classification, also called text categorization, is a supervised machine learning task which aims to assign pre-defined classes to text documents (Hotho et al., 2005). Typical text classification tasks include email routing, spam filtering and fraud detection (Miner et al., 2012). Classification is a well-known application in data mining and its process can be replicated in text mining. It involves choosing a classifier, i.e. a classification model that will be used to train the data. Commonly used classifiers are Naïve-Bayes, Support Vector Machines, Logistic Regression, Random Forests , k-Nearest Neighbors and Gradient Boosting Machines. Among those, some seem to suit better than others to text classification tasks. In fact, NB models perform well both with small and large vocabulary size (McCallum et al., 1998). Furthermore, SVM is also considered as a robust method that deals particularly well with high dimensional feature spaces (Joachims, 1998), i.e. that suits well documents with large dictionaries. Also, LR models seem to be at least as good as SVM for classifying texts (Genkin et al., 2007). A more recent study carried out in 2020 by Shah et al. compared the performance of three models (LR, k-NN and SVM) when applied on BBC new data. The researchers found out that the LR model was the most accurate of the three models. Even if it the literature seems to prefer the mentioned models, other models can also be considered if they improve the performance of the classification task.

2.2.2. Text Mining applied to ESG

Text mining techniques have been applied in various fields to achieve ESG-related tasks, such as the automatic identification of ESG disclosure, the monitoring of regulatory compliance, sentiment analysis applied to various sources and the automatic sustainability reporting among others. In this section, we aim to briefly mentioned studies that prove the utility of text mining.

A comparison between content analysis and text mining for analyzing CSR disclosures has demonstrated the usefulness of both techniques, emphasizing that text mining is particularly valuable when dealing with large volumes of CSR disclosures (Aureli, 2017). However, it also highlights that text mining may not be the best approach if the goal is to obtain nuanced insights.

A study carried out on Korean ESG management reports using various text mining analyses techniques also found text mining to be effective for this purpose (Yoon et al., 2023). Another study also found that text mining techniques could be used to automatically classify the content of ESG reports (Castellanos et al., 2015).

Text mining, coupled with named entity recognition allowed for the analysis of the level of compliance of Spanish companies with the TCFD (Moreno et al., 2020), highlighting a lack of standardization in the climate-related disclosure of those companies.

Sentiment analysis has also been used extensively on ESG disclosures, for example in order to study the effect of ESG information on stock prices (Schmidt, 2019) or to gain insights into the opinion of ESG analysts (Mandas et al., 2023).

Last we can mention studies that aimed to construct ESG ratings, using text mining on a dataset comprising both corporate and ESG reports (Caudron et al., 2022), or using a random forest algorithm on financial indicators of various companies (D'amato et al., 2022). Other studies leveraged more advanced AI tools to build ESG rating predictors (Krappel et al., 2021).

3. Developments

3.1. Research question

The aim of this master thesis is to develop a predictive tool which could be used for accurately predicting the ESG rating of U.S. listed companies. The literature review has highlighted that ESG ratings provided by traditional ESG data providers suffered from several shortcomings, including a significant divergence in the ratings provided as well as a lack of standardization in the methodologies used to develop them. Additionally, there has been an increasing demand from investors for credible and consistent ESG data. These various findings justify the objective of this master's thesis, i.e. developing a systematic approach to accurately predict ESG ratings. To build this predictive tool, we will leverage the ESG information that those companies disclose in their 10-K filings in order to train various machine learning classification model. More formally, this thesis aims to answer the following research question:

"Can the ESG textual information disclosed by US listed companies in their 10-K filings be used to accurately predict their ESG rating?"

In summary, the question is whether the ESG data that companies disclose in their 10-K filings is sufficient and consistent enough for classification models to learn patterns and accurately predict their ESG ratings.

In order to answer this question, we will assess and compare the predictive performance of various supervised machine learning models trained on the 10-K filings of a set of selected companies over a defined research horizon. This analysis will help us understand if any of those classifiers can learn patterns from the ESG information disclosed by the companies in their 10-K reports and thus see if some of the classifiers can serve as proxies of the ESG performance of the companies.

In the literature review, we noted that the amount of ESG information disclosed by U.S companies in their SEC filings had increased in recent years (see section 2.1.4). Consequently, this thesis also aims to evaluate whether the performance of the various classifiers remains stable or evolve when they are trained on datasets from different years within the research horizon. We believe it is important to investigate whether an increase in ESG disclosure leads to better prediction accuracy. However, it will be necessary to first verify the assumption that more ESG information is indeed being disclosed.

3.2. Data

This section will present an overview of the data used in the various experiments. It will describe the different datasets and outline the methods employed for their collection. Additionally, It will show how we selected the companies constituting the different dataset. Lastly, it will allow to explain which specific data points will be treated as independent variables and which will be treated as dependent variables.

3.2.1. ESG Ratings – Dependent Variables

We begin by talking about the data points that will serve as the dependent variables of our classification models: the ESG ratings. Those are the variables we aim to predict in this thesis. We chose to rely on the ESG ratings provided by Refinitiv, the ESG data provider of the London Stock Exchange Group. This specific ESG data provider was selected primarily because of the free access to its database granted by HEC Liège.

To collect data from Refinitiv, we used an Excel add-in available through Eikon, the software platform developed by Refinitiv. Eikon is designed to provide easy access to data resources of Refinitiv. The Eikon add-in integrates directly with Excel, enabling us to quickly retrieve the essential ESG data for our research from Refinitiv's comprehensive database.

Refinitiv's database is extensive and it offers, for each company, several metrics about their ESG performance. It provides, for instance, a separate rating for each pillar E, S and G, as well as an ESG controversy rating. However, for the purposes of this thesis, we will only collect and use the combined ESG rating which aims to replicate the overall ESG performance of a company.

Refinitiv's ESG ratings are represented under three distinct formats, corresponding to three levels of detail. It offers notably a continuous rating, a grade rating and a category rating, the continuous format being the most precise. **Table 1** offers a summary of the three ESG rating formats. Note that the ESG ratings provided by Refinitiv are updated on a yearly basis.

In this thesis, we collected the ESG ratings of companies under both the grade and the category format. Explanations of this choice will be covered in section 3.3.2.

Name	Rating Range
Continuous	[0: 100]
Grade	{D-, D, D+, C-, C, C+, B-, B, B+, A-, A, A+}
Category	{D, C, B, A}

Table 1: Refinitiv ESG Ratings Format

3.2.2. 10-K Filings – Independent Variables

ESG ratings aim to quantify the ESG performance of companies. Therefore, the classification models we will use need to be trained on data that approximates well the ESG performance of the companies. Furthermore, we must either choose to rely on data disclosed by the companies themselves, such as CSR reports or annual reports, or external information such as news or social media. In this research, we will build upon the work of Caudron (2022) and rely on company disclosed information to approximate the ESG performance of companies.

As already mentioned, the type of information disclosed by companies we will be using in this research are 10-K filings, also known as 10-K forms. These documents are comprehensive and highly standardized reports that contain detailed information about a company's business and financial condition. They provide investors with an extensive overview of a company's activities, highlight the various risks it faces, and present its operating and financial results in detail. The SEC mandates that all publicly traded companies listed on American stock exchanges submit a

10-K filing annually. This requirement ensures transparency and allows investors to make informed decisions based on reliable and up-to-date information. Each 10-K filing is composed of four parts, with each part containing several items, as depicted in **Table 2**: 10-K Filing Structure.

A primary objective of 10-K filings is to provide investors with all material information necessary for making informed investment decisions. In other words, each 10-K filing should disclose all information that an investor might consider important when making an investment decision. Material information usually encompasses mainly financial information. However, the concept of materiality has evolved over time, and an increasing number of investors now regard ESG information as material (Amel-Zadeh, 2018). Companies have recognized the evolution of how materiality was perceived by the investors. Also, companies are willing to anticipate regulations changes in the future, as many of them believe the SEC will require to disclose more ESG information in the future. Consequently, an increasing number of companies have started to voluntarily disclose ESG information in their 10-K filings in recent years (Gez et al., 2022).

The increasing amount of ESG information disclosed in 10-K filings has led many researchers to focus on this area. For instance, Baier et al. (2020) created a dictionary of ESG terms based on the most frequent ESG words appearing in 10-K filings. Ignatov (2023) studied the relationship between ESG disclosures in 10-K filings and stock returns. Rouen et al. (2024) compared ESG disclosures in ESG reports and 10-K filings, finding that while ESG reports offer more financially relevant ESG content, the volume and importance of ESG disclosures in 10-K filings have increased over time. The authors of this study also suggest that ESG information is very likely to be found in specific items of the 10-K filings. In fact, they conclude that ESG information can be included in the Item 1 (Business Description), Item 1A (Risk Factors) and Item 7 (Management's Discussion and Analysis of Financial Condition and Results of Operations). Consequently, we will focus on the ESG information contained in those sections in this research.

As we have just seen, various studies demonstrate that 10-K filings contain substantial ESG information. Therefore, we can make the assumption that a company's ESG performance can be approximated by the ESG information disclosed in its 10-K filing. Particularly, we make the assumption that this information is disclosed in Item 1, Item 1A and Item 7 (Rouen et al., 2024). Following a specific process that we explain later, we will extract a fixed set of ESG-related features from each 10-K. These features will then serve as the independent variables to train the classification models.

The 10-K filings of any company can be found on the EDGAR platform of the SEC. The access to this database is free and it can be accessed by anyone. To collect data from this platform, we use an informatic tool to gather information from webpages called "web crawler". A web crawler is a program that aims to index and collect information from websites. For the purpose of this thesis, using a web crawler is essential as it automates the downloading of the 10-K filings, eliminating the need to manually download each document individually and thereby saving a significant amount of time.

We initially considered implementing a web crawler ourselves. However, our limited programming skills made this option too challenging. Instead, we decided to clone an existing

crawler, named edgar-crawler (details about the crawler can be found in the related Github[1] repository). This crawler allowed us to easily download the necessary 10-K filings and efficiently extract relevant items. We crawled the 10-K filings of the selected companies and the five selected years of analysis. For each 10K-filing, the output of the crawling process is a JSON file whose structure is presented below (**Table 3**: Structure of The Extracted 10-K Filings).

The other items that are usually found in a 10-K filing were not extracted as we decided to solely focus on 3 items to collect ESG data about each company. The complete crawling process took around 24 hours.

Item	Heading
Part 1	
Item 1	Business
Item 1A	Risk factors
Item 1B	Unresolved staff comments
Item 1C	Cybersecurity
Item 2	Properties
Item 3	Legal proceedings
Item 4	Mine safety disclosures
Part 2	
Item 5	Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities
Item 6	Selected financial data
Item 7	Management's Discussion and Analysis of Financial Condition and Results of Operations
Item 7A	Quantitative and Qualitative Disclosures about Market Risk
Item 8	Financial Statements and Supplementary Data
Item 9	Changes in and Disagreements with Accountants on Accounting and Financial Disclosure
Item 9A	Controls and Procedures
Item 9B	Other Information
Item 9C	Disclosure Regarding Foreign Jurisdictions that Prevent Inspections
Part 3	
Item 10	Directors, Executive Officers and Corporate Governance
Item 11	Executive Compensation

[1]https://github.com/nlpaueb/edgar-crawler

Item 12	Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters
Item 13	Certain Relationships and Related Transactions, and Director Independence
Item 14	Principal Accountant Fees and Services
Part 4	
Item 15	Exhibit and Financial Statements Schedules
Item 16	Form 10-K Summary

Table 2: 10-K Filing Structure

Marker	Description
cik	Central Index Key. Unique key to identify each company in the EDGAR database
company	Company name
filing_type	Type of report that has been extracted. In this case: 10-K
filing_date	The date on which the 10-K form was filed with the SEC
filing_period	Period covered by the extracted report
sic	Standard Industrial Classification. It is used to categorize companies into industry sectors
state_of_inc	State of Incorporation. The state where the company is legally incorporated
state_location	The primary location of the company's operations
fiscal_year_end	Date on which the fiscal year ends
filing_html_index	Link to the HTML index page of the filing on the SEC's EDGAR database
htm_filing_link	Link to the HTML version of the complete 10-K filing
complete_text_filing_link	Link to the plain text version of the complete 10-K filing
filename	Local filename of the document
item_1	Business
item_1A	Risk factors
item_7	Management's Discussion and Analysis of Financial Condition and Results of Operations

Table 3: Structure of The Extracted 10-K Filings

3.3. Methodology

3.3.1. Methodology overview

This section aims to briefly describe the methodology that will be carried out to answer the research question.

As stated in section the introduction of this chapter (see section 3.1), we want to leverage the ESG information contained in 10-K filings in order to predict the companies ESG rating. Before processing the documents, we first construct the different datasets that will be used in this experiment. Then we undergo a resampling procedure of the datasets, as some datasets had important shortcomings. Then for each dataset, we preprocess each 10-K filing in order to

represent each of them in a structured format. Once structured, we will extract the ESG-related features from each document. In this way, the ESG performance of each company will be summarized by a vector. Each dataset, constituted of the vectors from all companies previously selected, will be divided in a training and testing set. Each training set will be used to train various classifiers. Finally, each testing dataset will be tested with each optimal classifier, i.e. each classifier with its optimal parameters.

3.3.2. Approaches

In order to assess the capability of our classification models to identify patterns within 10-K filings, we propose to evaluate ESG ratings at two distinct levels. By conducting this experiment at both a macro and a micro, we aim to gain a better understand of how well the models can capture broader versus more nuanced patterns in the ESG data.

3.3.2.1. ESG Category Classification – Macro Approach

This first approach can be referred to as the "macro" approach. The goal here is to classify each company into one of the broader ESG rating categories. Specifically, the classification models will be tasked with predicting whether a company falls into Category A, B, C, or D.

3.3.2.2. ESG Grade Classification – Micro Approach

The second approach is more granular and can be described as the "micro" approach. Here, the focus shifts to classifying each company into its precise ESG rating grade, which represents a more challenging task due to the finer distinctions between grades. The classification models will need to predict one of the twelve possible grades, ranging from D- to A+.

We will follow nearly the same procedure to predict each type of ESG rating. The only difference is that there will be no resampling of the datasets for the micro approach.

3.3.3. 10-K Classification Implementation

3.3.3.1. Dataset Selection

In this section, we aim to show how we selected the companies that constitute our different datasets. At the end of the chapter, we also provide an overview of the distribution of the ESG ratings (category and grade) of each dataset.

For this thesis, we aimed to base our research on the largest and most diverse data samples possible, allowing the classification models to benefit from a more extensive dataset and learn patterns more effectively.

However, our goal of having the largest possible dataset (within the limits of our computer's capabilities, of course) was constrained by two key requirements when selecting the companies for our dataset. First, we needed companies that are required to submit a 10-K filing to the SEC, which limited our selection to U.S. publicly traded companies. Although there is no precise data on the exact number of companies required to file a 10-K, there were approximately 4,000 U.S. publicly traded companies in 2020 (Gupta et al., 2021). Second, we considered the availability

of ESG ratings from Refinitiv, which covers nearly 16,000 companies globally, including around 3,300 companies in the U.S. Therefore, the best option would be to rely on the set of companies with both an ESG rating offered by Refinitiv and required to submit a 10-K filing.

Selecting this subset appeared to be challenging since there is no access to the full list of U.S. companies for which Refinitiv offers an ESG rating. Therefore, we decided to refer to the constituents of a large American index: the Russel 3000 index. The choice of this index is pertinent since it is a measure of the performance of the largest 3,000 US traded companies (representing approximately 96% of the US equity universe) (Hayes, 2024). However, another issue arose: we were unable to retrieve the constituents of this index from the Eikon platform due to our Eikon HEC student account lacking permission to access this specific data. As a result, we had to download a free list of companies included in the iShares Russell 3000 Exchange Traded Fund (ETF), which replicates the performance of the Russell 3000 Index.

This option was ultimately chosen to constitute the set of companies included in our dataset. While this approach is not entirely optimal—since the companies in the ETF may not exactly match those in the original index—it provided a practical solution under the circumstances. Note that the constituents are those present in the index in 2024 since it was impossible to retrieve earlier constituents. A list of those companies can be found in the appendices (Appendix 1)

We decided to collect ESG ratings of five years: 2018, 2019, 2020, 2021 and 2022. We did so given that they were the only accessible years of data with our student account.

Lastly, for each year of the surveyed time horizon, we created a dataset consisting of companies that were both constituents of the iShares Russell Index ETF and had a 10-K filing that could be crawled and extracted. By following this methodology, we were left with a dataset of 2045 companies representing 11 industries for the year 2022. Note that the size of the dataset might slightly vary from year to year. In fact, the constituents of the Rusell index are those of 2024. Also there has been a very little cases where it was impossible to effectively crawl the 10-K filing of some companies. **Table 4** provides an overview of the number of companies belonging to each ESG rating category for each year of the research horizon. **Table 5** does the same for the ESG rating grade. Additional data such as an overview of the industries of the selected companies (Appendix XX) can be found in the appendices.

ESG Rating Category	2018	2019	2020	2021	2022
A	114	130	167	184	191
B	450	556	639	746	831
C	956	980	906	881	831
D	487	376	274	225	191
Support	2007	2042	1986	2036	2045

Table 4: Number of Companies per ESG Rating Category and Year

ESG Rating Grade	2018	2019	2020	2021	2022
A+	2	2	3	3	2
A	36	38	40	31	32
A-	76	90	124	150	161
B+	120	160	186	223	250
B	158	169	213	255	293
B-	172	227	240	268	288
C+	231	279	279	284	289
C	331	337	315	336	300
C-	394	364	312	261	239
D+	318	274	206	167	146
D	152	93	65	57	42
D-	17	9	3	1	3
Support	2007	2042	1986	2036	2045

Table 5: Number of Companies per ESG Rating Grade and Year

As mentioned in the previous paragraph, we can see that the size of the dataset slightly vary due to the explained reason.

We can also see in **Table 4** that approximatively 80% of the companies fall into the categories B and C, which suggests that most of the American companies get a low to very low ESG rating. However, it is not that surprising since we saw in section 2.1.5.4 that a geographical bias might exist, resulting in lower grades in the U.S., notably. As a result, we can see the remaining categories A and B are significantly underrepresented.

This disproportionate distribution results in what is referred to in data mining as a class imbalance problem (Longadge et al., 2013). The next section will be devoted to describing which strategies can be put in place to tackle this issue.

The situation for the ESG rating grades is even more important. In fact, we can see on **Figure 2**: Barchart of the Distribution of ESG Rating Grades Across Years that there is an imbalance between the ESG rating grades. While grades such as C- and C- have significant higher counts, others like A+, A, and D- have very low counts. In general, we can say that the distribution of the ESG rating grades is skewed towards the middle ESG rating grades. We can also say that the lower frequencies are located in the extremes.

What is also interesting to remark is the upward trend over the years. In fact, we see a noticeable increase in the ESG rating grades over the years, especially in the A- to C+ range. On the other hand, grades like D+, C- and D see a decrease in frequency over time. This suggests that companies have improved how they integrate ESG into their businesses, or at the very least, that Refinitiv has evaluated it as such.

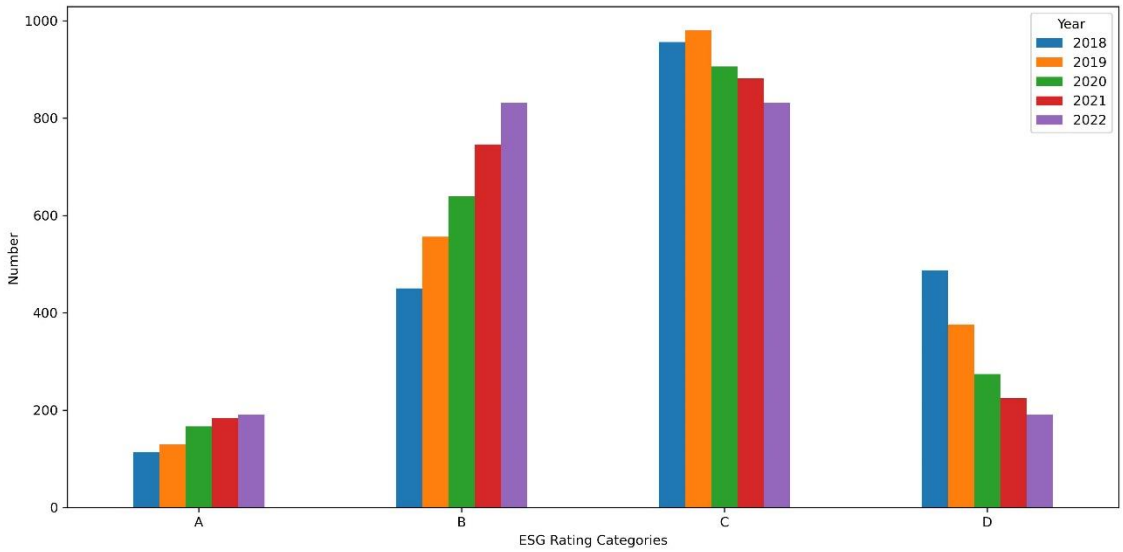

Figure 1: Barchart of the Distribution of ESG Rating Categories Across Years

The situation for the ESG rating grades is even more important. In fact, we can see on **Figure 2**: Barchart of the Distribution of ESG Rating Grades Across Years that there is an imbalance between the ESG rating grades. While grades such as C- and C- have significant higher counts, others like A+, A, and D- have very low counts. In general, we can say that the distribution of the ESG rating grades is skewed towards the middle ESG rating grades. We can also say that the lower frequencies are located in the extremes.

What is also interesting to remark is the upward trend over the years. In fact, we see a noticeable increase in the ESG rating grades over the years, especially in the A- to C+ range. On the other hand, grades like D+, C- and D see a decrease in frequency over time. This suggests that companies have improved how they integrate ESG into their businesses, or at the very least, that Refinitiv has evaluated it as such.

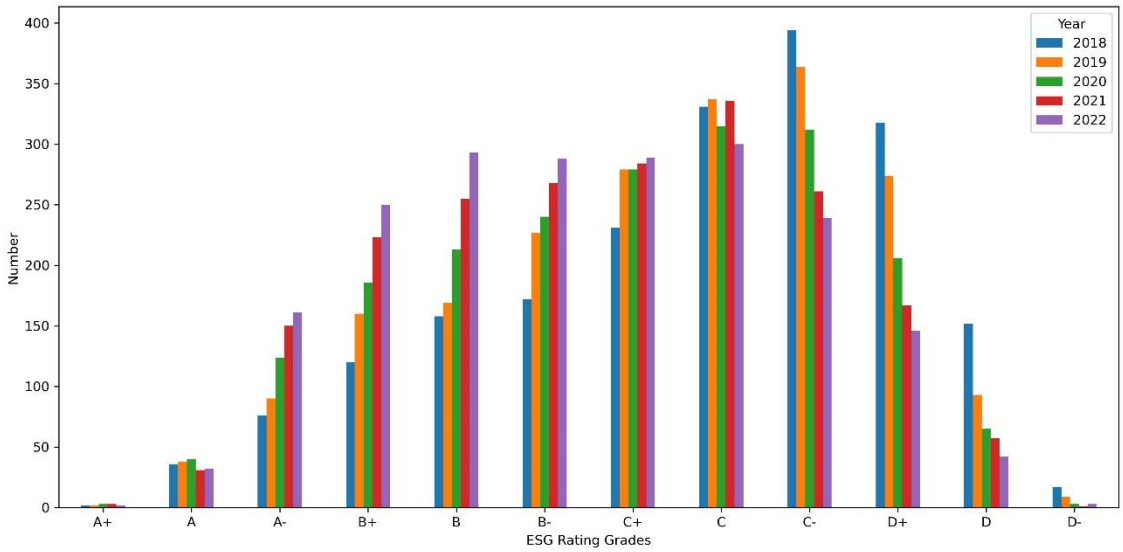

Figure 2: Barchart of the Distribution of ESG Rating Grades Across Years

3.3.3.2. Dataset Resampling

Before delving into the description of the preprocessing steps, we have to go through a resampling procedure of our datasets. This decision has been taken because we have significant imbalance in our datasets.

In fact, looking back to the distribution of the ESG rating grades and categories over the years (**Figure 1** and **Figure 2**), we see a clear skewed distribution towards the extremes.

Table 6 shows particularly well the categories that are underrepresented for each year. Categories A and B belong each year to the minorities. We can also see that their part decrease over the years due to large number of companies who have seen their category upgraded.

ESG Rating Category	2018	2019	2020	2021	2022
A	5.68%	6.37%	8.41%	9.04%	9.34%
B	22.42%	27.23%	32.18%	36.64%	40.64%
C	47.63%	47.99%	45.62%	43.27%	40.64%
D	24.27%	18.41%	13.80%	11.05%	9.34%

Table 6: Percentage of ESG Ratings Categories Across Years

If we take at the values of the most extreme ESG rating grades in **Table 5**, we see that they are nearly no company with those ESG rating grades.

Training classification algorithms with imbalanced datasets presents significant challenges. On one hand, the algorithms may favor the majority classes, performing well with those but failing to accurately capture the characteristics of the minority classes. As a result, while a classifier might appear to perform well, it may still be ineffective at recognizing the minority class. For example, consider a classifier trained to detect fraudulent transactions in a dataset where 98% of the transactions are legitimate and only 2% are fraudulent. If the classifier predicts all

transactions as legitimate, it would achieve high accuracy, but completely fail to detect the fraudulent ones. This highlights the importance of being cautious when evaluating classifier performance, especially in the context of imbalanced data.

This situation makes that we need to consider resampling our datasets in order to avoid the aforementioned problems linked to imbalanced datasets. We consider, for that, two options: oversampling and downsampling the datasets. While the first option consider enriching the minority classes with additional samples, the latter considers reducing the majority classes (see **Figure 3**).

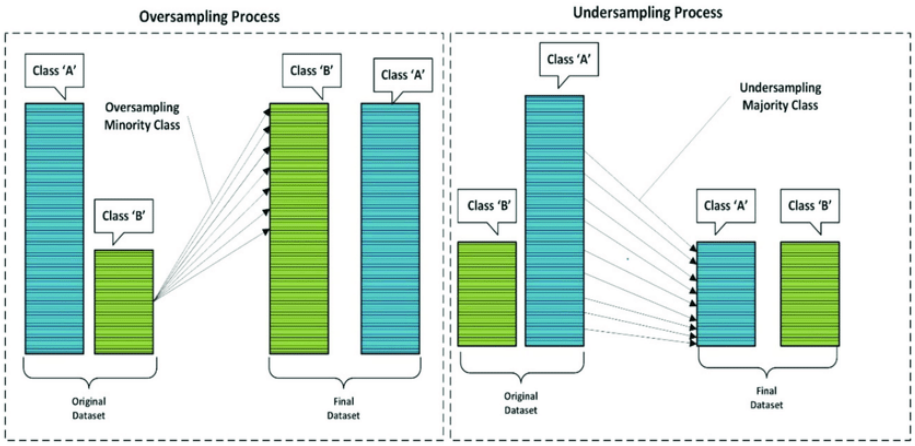

Figure 3: Comparison Oversampling vs. Downsampling

Source: Kumar, Vinod. (2022). Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques.

Data Resampling Strategy 1: Oversampling

In order to oversample a dataset, we can either use a technique called random oversampling or a technique called SMOTE (Synthetic Minority Oversampling Technique).

The first technique simply involves duplicating existing samples from the minority classes within each dataset in order to balance each dataset. This technique is likely to lead to overfitting as the algorithm will see the same examples multiple times (Kotsiantis, 2006). Overfitting is a phenomenon where a ML model becomes too closely fitted to the training data. As a result, it struggles making accurate prediction on unseen samples.

The other technique, SMOTE, does not simply duplicates existing samples; instead it creates synthetic samples of the minority class by "operating in feature space rather than data space" (Chawla et al., 2002). Practically, SMOTE first selects a set of instances from the minority class. For each selected instance, the algorithm identifies a specified number of nearest neighbors within the minority class. Finally, the new synthetic samples are generated by interpolating between the selected instance and its neighbors (Chawla et al., 2002). **Figure 4** illustrates this process.

The values represented in **Table 4** suggest that using smote to resample these dataset might potentially improve the performance of the classification models, as the imbalance is not too severe. However, the class distributions shown in **Table 5**, particularly the extreme grades of A+

and D-, raise concerns about the effectiveness of oversampling. In fact, such a significant imbalance could very likely result in overfitting (Chawla et al., 2002).

Consequently, we decided not to apply SMOTE in the micro approach since it would have led to the mentioned issues.

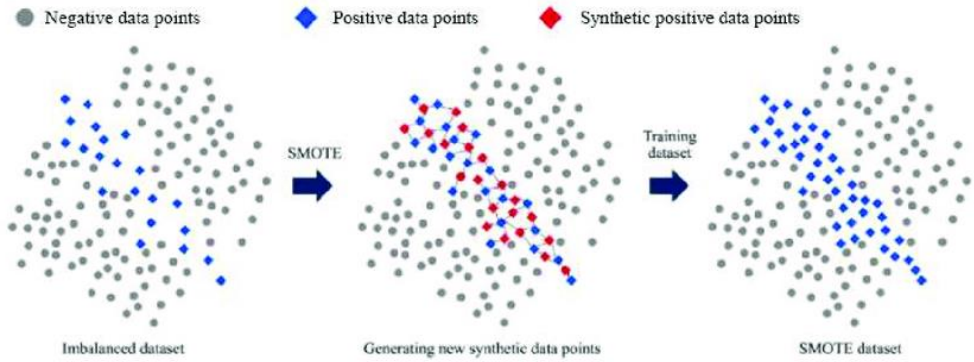

Figure 4: SMOTE Process

Source: Chen et al. (2022). Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning

Data Resampling Strategy 2: Downsampling

The second resampling strategy carried out is a simple random downsampling within each dataset. It means that each class is reduced to the size of the smallest minority class by randomly removing samples in each class.

However, the extreme imbalance once again makes resampling impractical for the micro approach. In the best-case scenario, downsampling the datasets shown in **Table 5** would result in a new dataset containing only three samples per ESG rating grade, which would make it impossible for the machine learning models to effectively learn from such limited data.

Consequently, we do not resample any of the datasets for the micro approach, which means that the models will only be trained on the imbalanced datasets.

For the macro approach, applying this resampling procedure leads to the creation of two new datasets per year. **Table 7** provides a summary of the size of the new datasets. It also means that, the classification models will be trained on 15 different datasets in the macro approach, but on only five datasets for the micro approach.

Sampling Strategy	2018	2019	2020	2021	2022
Imbalanced	2007	2042	1986	2036	2045
Oversampling	3824	3920	3624	3524	3324
Downsampling	456	520	668	736	764

Table 7: Size of the Dataset for Each Sampling Strategy and Year

3.3.3.3. Preprocessing steps

As discussed in the literature review, any data text mining process involves some text preprocessing steps. This section describes how each 10-K filing is transformed into a bag-of-words. Note that we used various python packages to perform the preprocessing steps.

For this text preprocessing procedure, we follow the common steps identified in the literature review that have proven to be effective. The preprocessing of the text starts with the removing all non-ASCII characters. For instance, characters such as "™" or "©". The process continues with the tokenization, i.e. the breaking of the text into individual words, also called tokens. The tokenization of each 10-K is carried using the nltk.word_tokenize function. Right after we bring all tokens to their lowercase and remove both non-alphabetic characters and the punctuation. We further reduce the dimensionality of the documents by removing all stop words using the python NLTK library of stop words. Finally we apply a lemmatizer on the remaining tokens using the python function WordNetLemmatizer from NLTK.

Recall that this procedure is applied for each 10-K filing only on the 10-K filings items likely to contain ESG data: the items 1, 1A and 7.

At the end of this procedure, we obtain, for each 10-K filing in each dataset, a simple bag-of-words model.

3.3.3.4. Features Extraction

However, at this stage, each BoW still contains many words that have little ESG explanatory power. That is why we apply a last filter on each BoW, keeping only the words that are related to ESG. We chose to keep the words contained in the ESG glossary developed by Baier et al. (2020). They created an initial list of 482 ESG terms, which was updated later into a list of 492 ESG words. Using this glossary of ESG words seems particularly pertinent for the purposes of this thesis since it was built by analyzing the ESG content of several 10-K filings. The full list of ESG words can be consulted in the appendix.

At the end of the filtering procedure, each Bag of Words is reduced to a set of features where each unique word is represented as a feature, and each feature is constrained to the ESG glossary as defined by Baier. To enhance the effectiveness of the model, we then compute the Term Frequency-Inverse Document Frequency (TF-IDF) for each word in this filtered set. This process transforms the BoW into a vector representation for each 10-K filing, with each cell representing the TF-IDF value of an ESG-related word. By doing so, we ensure that the features used to train the classification models are both relevant and meaningful.

For clarification purpose, we recall the definition of term inversed-inversed document frequency.

$$tf - idf_{t,d} = tf_{t,d} * idf_t$$

Where:

- $tf_{t,d}$ is the term frequency of the term *t* in document *d*.

$$tf_{t,d} = \frac{n_{t,d}}{\sum_{k \in d} n_{k,d}}$$

With $n_{t,d}$ representing the number of times term t appears in document d and $\sum_{k \in d} n_{k,d}$ representing the total number of terms in document d.

- idf_t is the inverse document frequency of the term t.

$$idf_t = log \frac{N}{df_t}$$

With N denoting the total number of documents in a collection and df_t the number of documents where the term t appears.

In practice, we use the vectorizer from the Python sklearn package to transform each Bag of Words into a vector. In this vector, each element represents the TF-IDF score of an ESG-related word. The list provided by Baier et al. (2022) contains 491 ESG words, meaning that the ESG information within each 10-K filing is now represented by a vector of size (1 x 491). This transformation allows us to capture the relevant ESG content within the filings in a structured and consistent way.

3.3.3.5. Classification Models

Now that each 10-K filing is represented in a structured format, we will describe the classification models that will be used to predict both ESG categories and grades. The selection of models was guided by the literature review conducted on the subject. For each model, we will provide a definition, highlighting its strengths and weaknesses. Finally, we will briefly discuss the practical implementation of each model.

Naïve Bayes Classifier

Naive Bayes classifiers are a family of probabilistic models that are based on the Bayes' theorem. This theorem assumes that the features are conditionally independent given the class label (McCallum, 1998). Even if this strong assumption appears to be often false in the real world, Naive Bayes classifiers have been widely adopted due to their simplicity and proven performance. In practice, Naive Bayes classifiers return the class that maximizes the posterior probability:

$$\hat{c} = \arg\max_{c \in C} P(c|X)$$

More formally, Bayes' theorem is expressed as follows:

$$P(c|X) = \frac{P(X|c)\, P(c)}{P(X)}$$

In this equation:

- $P(c|X)$ is the probability of class c given the features X, called the posterior probability.
- $P(X|c)$ is the probability of the features X given class c, called the conditional probability.
- $P(c)$ is the prior probability of c before observing the evidence.
- $P(X)$ is the probability of observing the features X, called the evidence.

To achieve the classification of a new document, the classifier will thus need to compute these three probabilities. However, Naive Bayes benefits from simplifying assumptions that make calculating the posterior probability more manageable. In fact, since we consider that the features *X* are independent of each other given a class *c* and because $P(X)$ is independent of class *c*, the posterior probability can be rewritten as follows:

$$P(c|X) \propto P(c) \prod_i P(x_i|c)$$

There are three well-known Naive Bayes classifiers commonly used for classification tasks: Bernoulli, Gaussian, and Multinomial. While the Bernoulli classifier was not used in this research due to its limitation of only accepting binary features as input, both the Multinomial and Gaussian Naive Bayes classifiers were selected.

In practical, we used two models of the python package sklearn: MultinomialNB and GaussianNB.

Support Vector Machine (SVM)

Support Vector Machines (SVM) are supervised learning models that can be both used for classification and regression tasks (Cortes et al., 1995). The basic SVM is designed to distinguish between two classes by finding a hyperplane that best separates the data. This hyperplane aims to maximize the margin between the closest points of two , which are called the support vectors (see **Figure 6**) When new data points, represented in a m-dimensional space for example, are presented to the classifier, it will classify them by comparing their position in the same space with the position of the hyperplane.

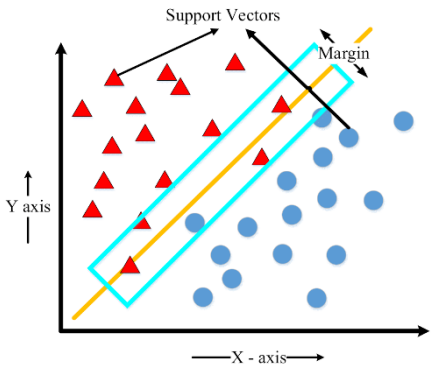

Figure 6: Support Vector Machine

Source: Muzzammel, R., & Raza, A. (2020). A support vector machine learning-based protection technique for MT-HVDC systems

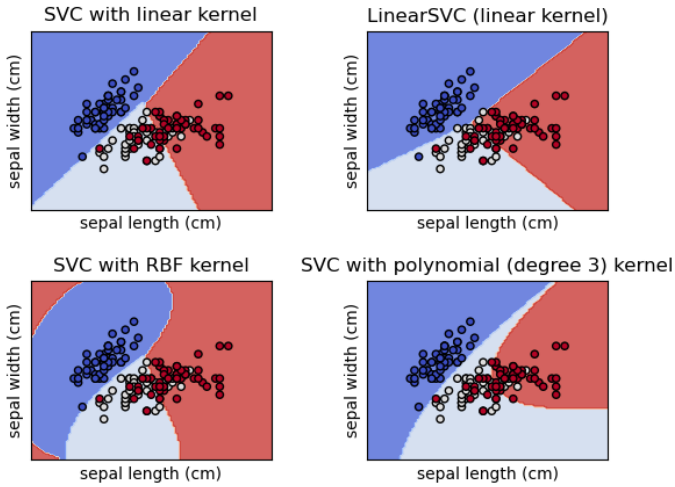

Figure 5: sklearn SVM with Different Kernel Functions

Source: https://scikit-learn.org/stable/modules/svm.html

One of the major advantages of the SVM is that they are “universal learners” (Joachins, 1998), i.e. they can use different kernel functions. This allow us to handle cases where data cannot be linearly separated, as it is depicted in **Figure 6**.

Recall that Support Vector Machines (SVMs) were initially designed for binary classification. In this thesis, we use the SVC class from the scikit-learn Python package, which is capable of handling multi-class classification. This class supports several kernels, including linear, polynomial, Gaussian (also known as RBF), and sigmoid. For this research, we exclusively used the RBF kernel, as it effectively handles nonlinear data and is the most widely used kernel function (Cervantes et al., 2020).

Random Forest

The random forest algorithm (Breiman, 2001) is a supervised learning procedure that follows the 'divide and conquer' principle (Biau et al., 2016) and for both classification and regression tasks. s an ensemble learning method, it constructs multiple decision trees during the training phase. For classification tasks, it predicts the class by aggregating the majority vote from all the decision trees, while for regression tasks, it predicts the output by averaging the results from all the trees.

In practice, the Random Forest algorithm begins by creating random subsets of data using bootstrap sampling (random sampling with replacement). Each subset is then used to train an individual decision tree. For each tree, a random subset of features is selected, a process known as feature bagging. The tree splits the nodes to increase the homogeneity of the data, but only considers the randomly selected features. Finally, for classification tasks, the algorithm aggregates the predictions from all the trees, with the class receiving the majority of votes being returned.

Random Forest is known for its high accuracy and robustness to noise and overfitting (Fawagreh et al., 2014). However, it is also considered a "black box" model, meaning its results are difficult to interpret.

In this thesis, we implemented the Random Forest algorithm using the RandomForestClassifier class from the scikit-learn Python package.

Logistic Regression

Logistic regression is a statistical model used for classification tasks. It models the probability that a given data point belongs to a specific class. Unlike linear regression, logistic regression fits an S-shaped logistic function, with the output being a probability between 0 and 1, as shown in **Figure 7**. This probability can be interpreted as the likelihood that the data point belongs to a particular class, and a classification decision is made by applying a threshold, typically 0.5.

In practice, logistic regression works by first finding a linear combination of the input features.

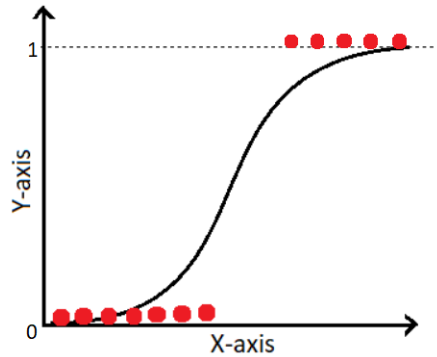

Figure 7: Logistic Function Curve

Source: https://medium.com/analytics-vidhya/the-math-behind-logistic-regression-c2f04ca27bca

This linear combination is then transformed using the logistic (sigmoid) function to produce a probability score between 0 and 1, representing the likelihood that the instance belongs to the positive class. The model is trained by adjusting the feature weights to minimize the error between the predicted probabilities and the actual class labels, typically using maximum likelihood estimation. Once trained, the model makes predictions by computing the weighted sum of the features and applying the logistic function. The classification decision is made based on a predefined threshold, usually 0.5, to determine the final class label.

In this thesis, we used the LogisticRegression class from the Python package scikit-learn, which supports multinomial logistic regression for cases where there are more than two possible outcomes.

3.3.3.6. Classification Models Evaluation

This section provides the reader with an overview of the procedure that has been put in place to evaluate the performance of the classification models used. For this purpose, several performance metrics have been used and are described below.

Note that before evaluating the classifications models, we decided to split each dataset in a training (80% of all samples) and testing (20% of all samples) set in order to ensure that the classifiers will be tested on unseen data samples.

For both prediction approaches, we then evaluate each classification model using a 10-fold cross validation strategy. It consists in dividing each training dataset in ten subsets (the folds), where one fold is kept as a test set, while the other nine subsets are used to train the model. This process is repeated ten times, each time using a different test set and the remaining nine folds as the training set. We do this process with all possible combinations of hyperparameter using GridSearch and compute, for each fold, the accuracy of the model. After the 10 folds, the average accuracy is computed. Once we have tested all combinations of parameters, we compare the average accuracies. The selected parameters are those whose accuracy is the highest. We can then evaluate the real performance of the model by testing it with the best parameters on the testing dataset.

Confusion matrix

A confusion matrix describes the performance of a classification algorithm by comparing the predicted classifications with the actual classifications. It allows to understand where the model performs well et where it performs poorly. In the case of a binary classification, it is composed of four components:

- True Positives (TP): The number of correct predictions that the instance is positive.
- True Negatives (TN): The number of correct predictions that the instance is negative.
- False Positives (FP): The number of incorrect predictions that the instance is positive.
- False Negatives (FN): The number of incorrect predictions that the instance is negative.

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

The confusion matrix can be extended to more classification classes.

Accuracy

The accuracy can be derived from the components of the confusion matrix as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

This measure aims to evaluate the global correctness of a classifier (Hossin et al., 2015). Accuracy is widely used due to its easy calculation and its ability to summarize well the performance of a model. However, it has limitations when evaluating imbalanced datasets, as the majority class can achieve high accuracy without identifying any instances of the minority class (Brownlee, 2021).

Precision

The precision metric indicates how many of the predicted positives instances are actually correct.

$$Precision = \frac{TP}{TP + FP}$$

Recall

The recall metric, also known as sensitivity, measures the proportion of actual positive instances that are correctly identified by the model. It indicates how well the model can capture all positive instances.

$$Recall = \frac{TP}{TP + FN}$$

Macro F1

The Macro F-1 score aims to provide a balanced evaluation by evaluating precision and recall in a single measure.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

4. Results

In this chapter, we present the results we obtained by applying the methodology we outlined in section 3.3. Through the analysis of these results, we aim to address the research question posed earlier in the thesis.

We present the results in two parts, corresponding to the two approaches: macro and micro.

For each approach, we will begin by providing a high-level overview of the results by combining the results from all datasets and classifiers. In this way, we will offer a general sense of the models' effectiveness without diving into details. Afterward, we will further break down the results by showing the performance for each year. Additionally, we will present the overall performance based on each dataset resampling strategy. Finally, we will also show the overall performance obtained for each classification model.

We will then delve into more detailed results by presenting the results for each resampling strategy—imbalanced, downsampling, and oversampling. For each strategy, we will compare the performance of the different classification models. Additionally, we will conduct the same comparison across the different years.

4.1. Macro Approach – ESG Rating Categories Prediction

We begin this section by examining the overall performance of the classification models. This is done by aggregating the results from all datasets and classifiers, which are presented in two boxplots in **Figure 8**.

From this figure, it is evident that the classification models performed poorly overall, with an average accuracy of 49.15% (see **Table 8**), meaning they were able to accurately predict the ESG category for only one out of every two companies.

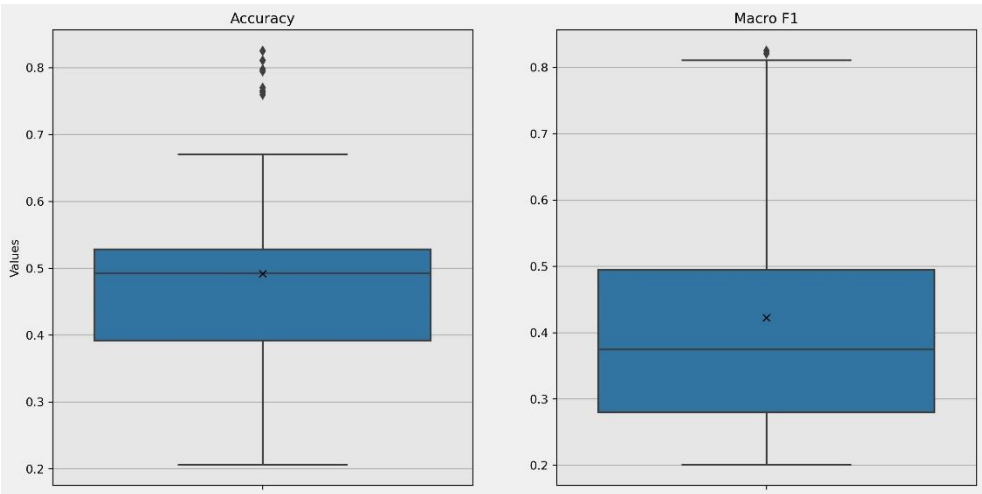

Figure 8: Boxplots of Overall Accuracy and Macro F1-score

Moreover, the left boxplot in **Table 8** reveals a wide dispersion in classifier performance, suggesting that the accuracies of the different models varied significantly. This variability raises concerns about the stability of the predictions and indicates that the models may not be consistently reliable. Further investigation is required to understand the factors contributing to this high variability in accuracy.

We can already say at the bgining of tis section that, based on the overall performance, the classification models we implemented clearly struggle to accurately predict ESG rating classes, as their average performance is poor, and their variability is considerable.

	2018	2019	2020	2021	2022	Average
MultinomialNB	0.4859	0.4520	0.4522	0.4630	0.4672	0.4641
Gaussian NB	0.3631	0.3341	0.3164	0.3324	0.3750	0.3442
Logistic Regression	0.5389	0.4958	0.5078	0.5282	0.5120	0.5165
Random Forest	0.5764	0.5825	0.5708	0.5631	0.5692	0.5724
SVM	0.5760	0.5499	0.5563	0.5737	0.5450	0.5602
Average	0.5081	0.4829	0.4807	0.4921	0.4937	0.4915

Table 8: Average Accuracy for Each Classifier by Year

This finding becomes even more evident when we consider the overall Macro F1 performance obtained (see right boxplot in **Table 8**). In fact, we see that we obtain an average Macro F1-score of 0.4222 (see **Table 9**) and that the F1-scores are more dispersed than the accuracies (we obtain an average variance of 0.249 for the accuracy vs. 0.0303 for the Macro F1-score). The fact that the average F1-score is lower than the average accuracy suggests that the classifiers may have struggled with the minority classes in the imbalanced datasets. This issue will be explored in more detail in the following sections.

	Precision		Recall		F1-score	
	Average	Variance	Average	Variance	Average	Variance
MultinomialNB	0.4395	0.0051	0.464	0.004	0.3807	0.0114
Gaussian NB	0.401	0.0103	0.3442	0.0172	0.3226	0.012
Logistic Regression	0.4834	0.0131	0.5166	0.0124	0.4385	0.024
Random Forest	0.5637	0.0302	0.5724	0.0282	0.4903	0.0507
SVM	0.5224	0.0382	0.5602	0.0329	0.479	0.0534
Average	0.482	0.0194	0.4915	0.0189	0.4222	0.0303

Table 9: Average and Variance of Key Metrics for Each Classifier

The fairly wide dispersion in accuracies and F1-scores suggests that the classifiers have delivered contrasting performances. To explore this further, we decided to compare their overall performance by aggregating the results from all datasets. This comparison is shown in **Figure 9**, where two boxplots illustrate the accuracy and Macro F-1 performance of each classifier.

Based only on the visuals, we observe that one model struggles particularly – the Gaussian NB - while the models with the best accuracy on average are the Random Forest and the SVM classifiers. Those models predict accurately more than one out of two companies. However, we also observe on both visuals that the most accurate models are also those with the most dispersed results. This suggests that the performance of these models were more impacted than the others when trained on certain datasets. We can also note that despite its low accuracy, the Multinomial NB classifier is the most stable classifier.

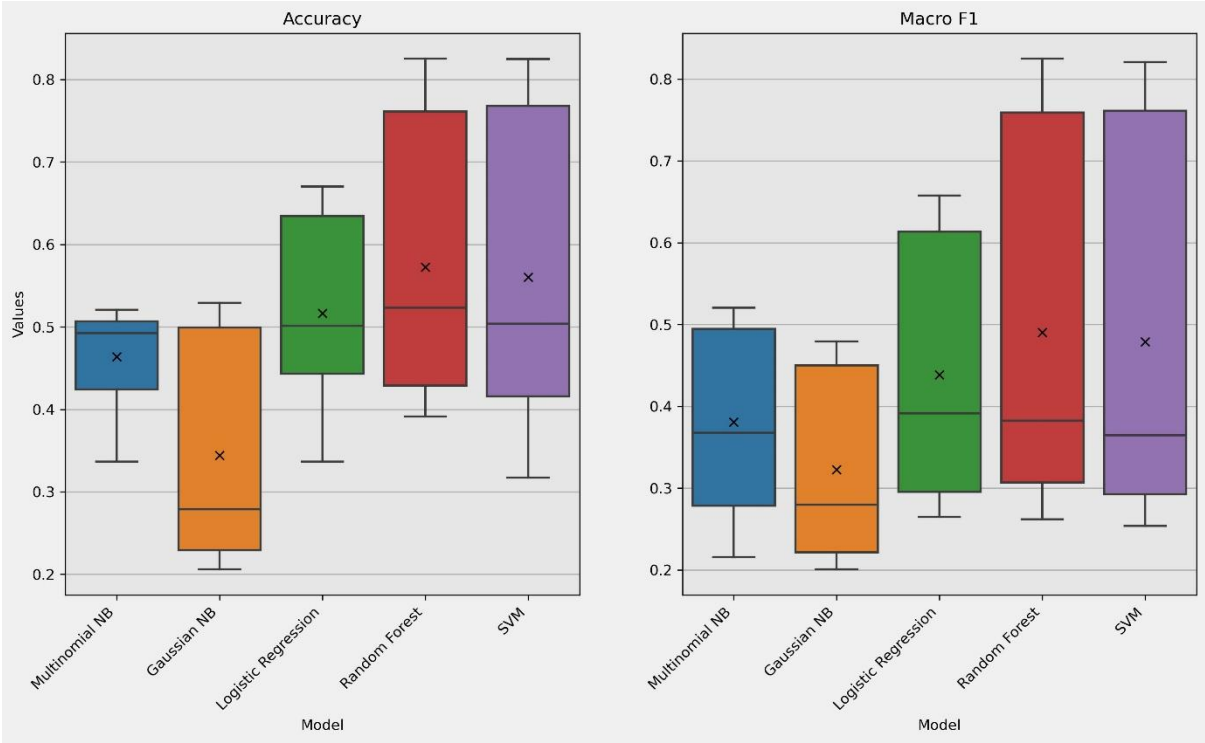

Figure 9: Boxplots of Accuracy and Macro F1-Score for Each Classifier

We observe on the right boxplot that the average Macro F1-scores follows a similar pattern to the average accuracies. However, the average Macro F1-scores are consistently lower than their corresponding accuracies. This may indicate, as previously mentioned, that the reasons of this issue are due to the struggling of the minority classes in the imbalanced datasets.

We will further investigate this issue by examining the overall performance obtained with the three dataset types: imbalanced, downsampled, and oversampled. As shown in the left part of **Figure 11**, the highest average accuracy (65.06%) is achieved with the oversampled datasets, which also deliver the best Macro F1-scores.

We also observe that, on average, the downsampled datasets show a higher accuracy than Macro F1-score.

Furthermore, we should note that the Macro F1-scores obtained with the original, and thus imbalanced datasets are significantly low. While the average accuracies and Macro F1-score are very similar across the other dataset types, we observe an important divergence between these metrics for the imbalanced datasets.

Lastly, we observe a significantly higher variance in the predictions made by classifiers trained on oversampled datasets. This suggests that the classification models exhibit highly inconsistent performance when trained on this type of data. We will explore this issue further in subsequent analyses.

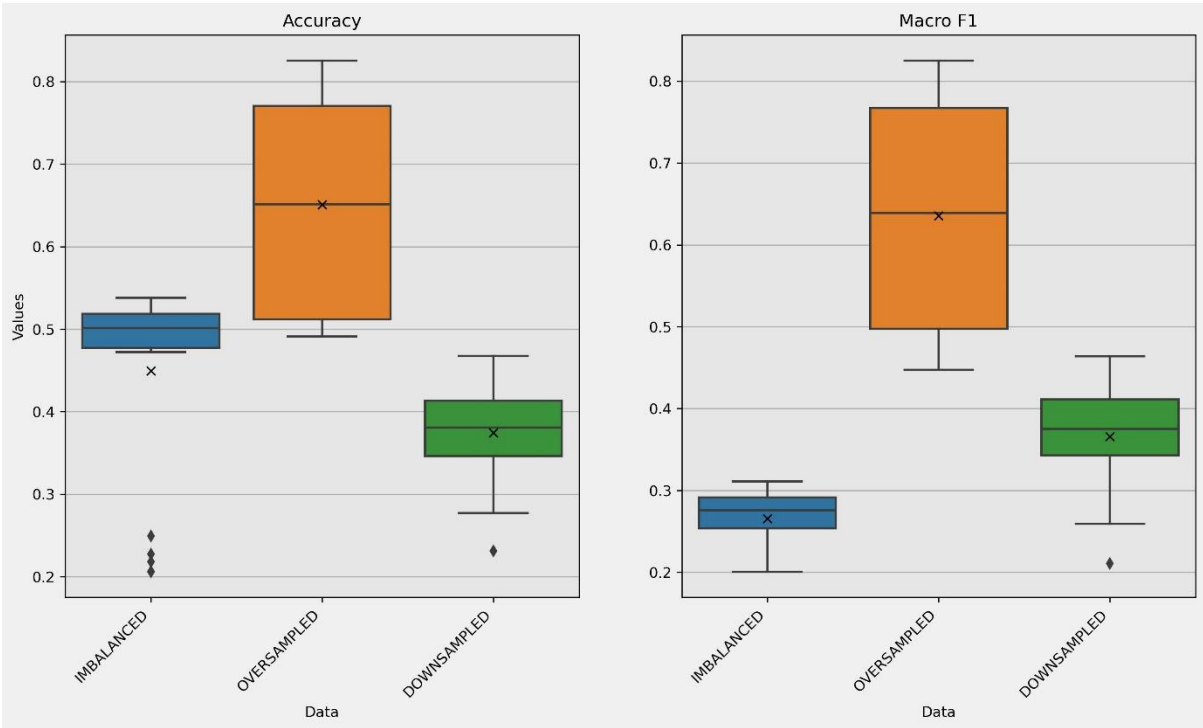

Figure 11: Boxplots of Accuracy and Macro F1-score Obtained with Different Sampling Approaches

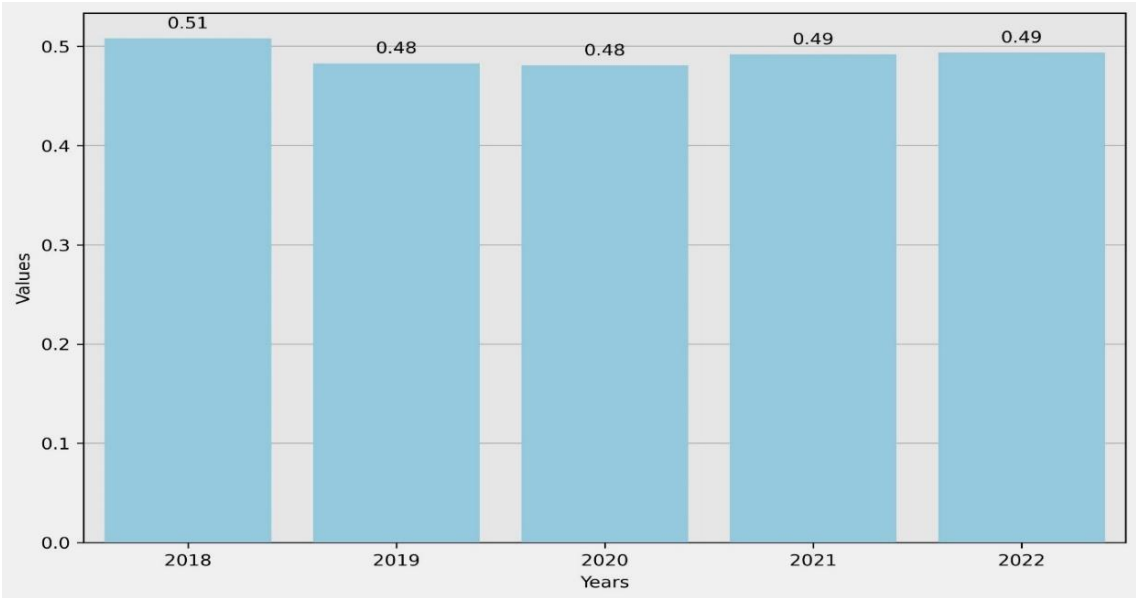

Figure 10: Bar Chart of The Average Accuracy per Year

Before concluding the presentation of the overall results, we compare the performance of the classification models based on the year of the datasets they were trained on. The bar chart in **Figure 10** illustrates the average performance for each year. We observe that prediction accuracy remains consistently low across the years, with only slight fluctuations. Furthermore, the numbers shown in **Table 8**: Average Accuracy for Each Classifier by Year show only minor changes in average accuracy over the years.

Lastly, we also provide some data over the quantity of ESG data that is contained in average each year in the datasets (see **Table 10**). We observe that, using Baier's (2020) ESG word list, the average number of ESG-related words in the 10-K filings increased steadily over the 2018–2022 period. This aligns with the findings discussed earlier in this thesis, indicating that ESG disclosure in the 10-Ks has expanded in recent years.

Year	2018	2019	2020	2021	2022
Number of Words	275,09	271,56	320,68	338,12	335,3

Table 10: Average Number of ESG Words in the 10-Ks

4.1.1. Imbalanced datasets

This section presents the performance of the various classifiers trained with the imbalanced, i.e. the original datasets for each of the surveyed year.

As we can see on the boxplot on the left of **Figure 11**, the accuracy values obtained by the classifiers with the imbalanced datasets are less dispersed than with the oversampled and downsampled datasets. However, we remark the presence of some outliers, i.e. extreme values, which can be explained in this case by the very poor performance delivered by one of our

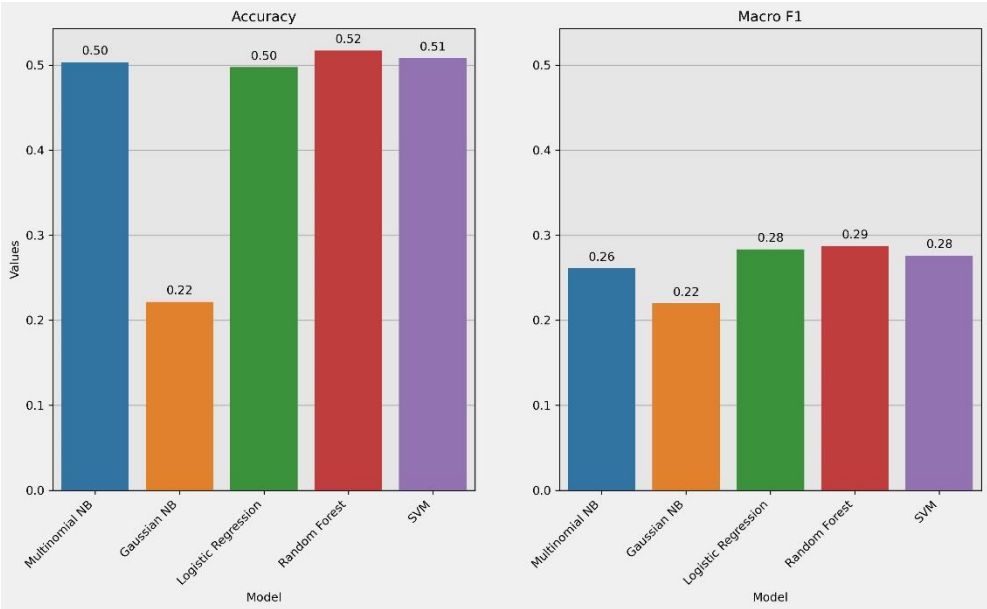

Figure 12: Bar Charts of Average Accuracy and Macro F1 for Each Classifier Trained on Imbalanced Datasets

prediction models. The model which performs worst seems to be the Gaussian NB (see **Figure 12**)

By comparing the two bar charts represented in **Figure 12**, we observe a significant difference between the average accuracy and F1-score when classifiers are trained on imbalanced datasets. This discrepancy, which is characteristic of imbalanced datasets, is even more apparent in the scatterplot in **Figure 13**. Across all models, accuracy consistently outperforms the Macro F1-Score, apart from the Gaussian Naive Bayes model, which performs poorly on both metrics.

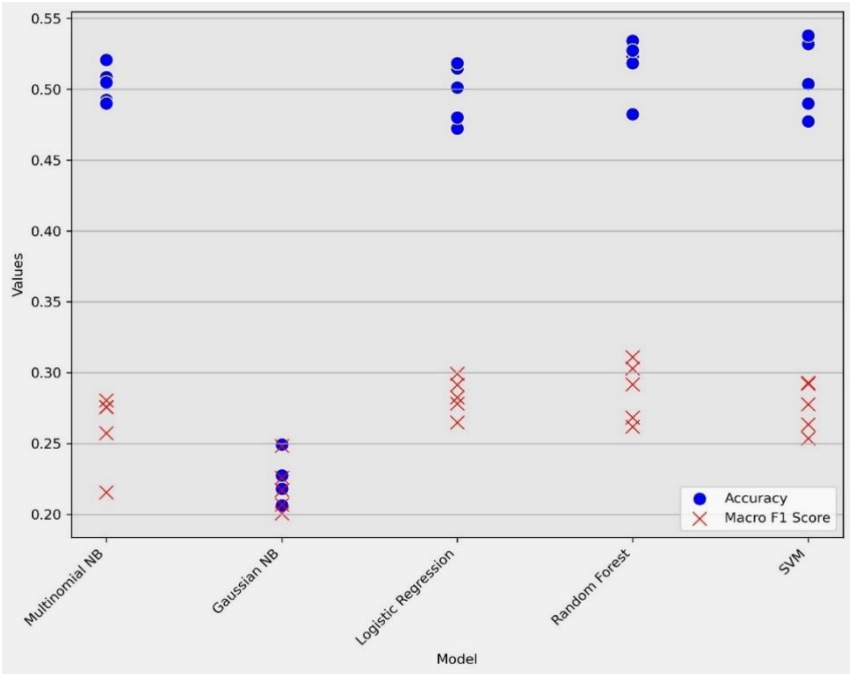

Figure 13:Scatterplot of Accuracy vs. Macro F1-Score Obtained for Classification Models Trained on Imbalanced Datasets

We illustrate the factors that lead to such situations by showing the classification report (**Table 11**:Classification Report of the Multinomial NB Classifier Trained on the 2018 Imbalanced Dataset.) and the confusion matrix (**Table 12**) of the Multinomial NB classifier trained on the 2018 imbalanced dataset. When we look narrowly, we realize that the model is unable to predict the minority classes.

It can be observed that no sample from the dataset is predicted as belonging to class A, and only 1 sample is predicted as class D. These are the minority classes, with only 23 and 98 samples in the entire dataset, respectively.

This example highlights why accuracy should be interpreted with caution when working with imbalanced datasets. At first glance, one might conclude that the model correctly predicts one out of two samples, but this is misleading.

	precision	recall	f1-score	support
A	0.00	0.00	0.00	23
B	0.38	0.12	0.18	90
C	0.50	0.97	0.66	191
D	1.00	0.01	0.02	98
Accuracy			0.49	402
Macro average	0.47	0.28	0.22	402
Weighted average	0.56	0.49	0.36	402

Table 11:Classification Report of the Multinomial NB Classifier Trained on the 2018 Imbalanced Dataset.

	Predicted A	Predicted B	Predicted C	Predicted D
Actual A	**0**	5	18	0
Actual B	0	11	79	0
Actual C	0	6	185	0
Actual D	0	7	90	**1**

Table 12:Confusion Matrix of the Multinomial NB Classifier Trained on the 2018 Imbalanced Dataset.

4.1.2. Downsampled datasets

We report now the results we obtained with the downsampled datasets. As explained previously, we decided to reduce our datasets to balance them. For this approach, it was achieved by randomly downsampling our datasets.

A key observation from **Figure 11***: Boxplots of Accuracy and Macro F1-score Obtained with Different Sampling Approaches* is that the classifiers trained on downsampled datasets achieve the lowest average accuracy (37.43%). However, we can observe that both the accuracies and Macro F1-score values obtained by the classifiers on these downsampled datasets are less

dispersed than with the other dataset types. these classifiers also exhibit the least variability in performance (variance of 0.0032).

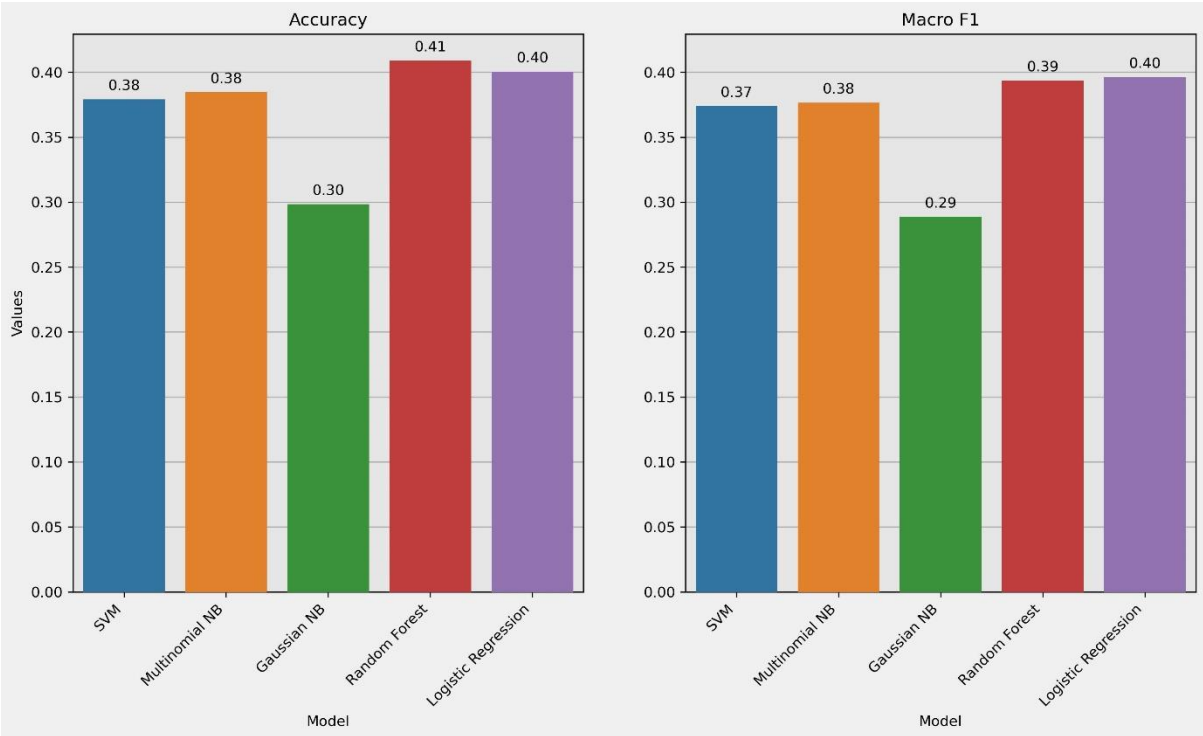

Figure 14: Bar Charts of Average Accuracy and Macro F1-Scores Obtained with Different Classifiers on Downsampled Datasets.

It is also worth noting that the accuracy and Macro F1-score values obtained with the downsampled datasets are very close. In fact, the average F1-score is 36.58%, which suggests that this resampling strategy has mitigated the issue we encountered with classifiers trained on imbalanced datasets. This holds true when comparing the accuracies and Macro F1-scores across the various classifiers (**Figure 14**). Indeed, we observe that both performance metrics—accuracy and Macro F1-score—are comparable across the various classification models. When trained on downsampled datasets, the best-performing models on average are the Random Forest classifier with 40.89% accuracy and the Logistic Regression model with 40.04%. Close behind are the Multinomial NB model, averaging 38.46%, and the SVM model, averaging 37.92%. The Gaussian NB classifier performs the worst, with an average accuracy of 29.85%.

We now take a look at the classification report and the confusion matrix of the same classification model we analyzed in the previous (see **Table 13** and **Table 14**). It can be seen directly on the confusion matrix that the model predicts samples of every class, compared to the model trained with imbalanced data which predicted no samples in the minority classes. Furthermore, the classification report shows that the model has now more lack identifying samples belonging the class that previously belonged to the majority (category C). It shows that the model trained with imbalanced data was biased by the overrepresentation of this class.

	Predicted A	Predicted B	Predicted C	Predicted D
Actual A	16	3	4	0
Actual B	7	9	4	3
Actual C	1	7	9	6
Actual D	6	0	9	8

Table 13:Confusion Matrix of the Multinomial NB Classifier Trained on the 2022 Downsampled Dataset.

	precision	recall	f1-score	support
A	0.53	0.70	0.60	23
B	0.47	0.39	0.43	23
C	0.35	0.39	0.37	23
D	0.47	0.35	0.40	23
Accuracy			0.46	92
Macro average	0.46	0.46	0.45	92
Weighted average	0.46	0.46	0.45	92

Table 14:Classification Report of the Multinomial NB Classifier Trained on the 2022 Downsampled Dataset.

Lastly, we take a look at some key performance metrics of the classification models over the surveyed years. We can see that the Multinomial NB, the Gaussian NB and the Logistic Regression classifiers deliver their best Macro F1-score when trained on the 2018 downsampled dataset. It must also be noted that all except one classifier (Gaussian NB) reach their worst F1-score when trained on the 2019 dataset.

To evaluate the stability of the classification models, we also calculated the standard deviation of the Macro F1-scores across the years 2018 to 2022. The results showed that the Random Forest had the lowest results – 1.39% - indicating minimal fluctuations in performance over the years. The more important fluctuation of the other classifiers suggests that they are more sensitive to the increase in size of the datasets.

We can also note that the precision and recall are very close to each other for each classifier when trained with the five different datasets (see **Table 15**), which suggests that the models are stable and have been well tuned.

		2018	2019	2020	2021	2022
MultinomialNB	Precision	0.4559	0.3367	0.3706	0.3726	0.3776
	Recall	0.4565	0.3365	0.3731	0.3649	0.3922
	Macro F1-score	0.4499	0.3363	0.3675	0.3540	0.3749
Gaussian NB	Precision	0.3609	0.2867	0.2090	0.2819	0.3495
	Recall	0.3587	0.2788	0.2313	0.2770	0.3464
	Macro F1-score	0.3514	0.2799	0.2106	0.2592	0.3427
Logistic Regression	Precision	0.4639	0.3304	0.3882	0.4174	0.3969
	Recall	0.4674	0.3365	0.3806	0.4189	0.3987
	Macro F1-score	0.4639	0.3307	0.3823	0.4127	0.3913
Random Forest	Precision	0.3801	0.3868	0.4131	0.3778	0.4219
	Recall	0.3913	0.4038	0.4328	0.3919	0.4248
	Macro F1-score	0.3824	0.3751	0.4160	0.3809	0.4127
SVM	Precision	0.4118	0.3170	0.3648	0.4205	0.3610
	Recall	0.4130	0.3173	0.3806	0.4189	0.3660
	Macro F1-score	0.4112	0.3167	0.3646	0.4186	0.3587

Table 15:Performance Metrics of Classification Models Trained on Downsampled Datasets

4.1.3. Oversampled Dataset

In this third section we report the results obtained with the various classification models trained on oversampled datasets. As reminder, the oversampling strategy consisted in creating artificial data samples to achieve balanced datasets. We achieved this task using SMOTE.

When we look back to the boxplots presented in the introduction of this chapter (see ***Figure 11)***, we see that the classification models deliver in average the best performance in terms of accuracy and F1-score when trained on oversampled datasets. In fact, they yield an average accuracy of 65.06% and a Macro F1-Score of 63.55%. However, it should also be remarked that the values obtained with those oversampled datasets are more spread than with both other dataset types. We report a standard deviation of 14.22% for the accuracy and 13.09% for the Macro F1-score. It suggests that the resampling strategy used to increase the size of the datasets impact very differently the performance of the classification models. We can also note that the values obtained for the accuracy and the Macro F1-scores are close to each other with the oversampled datasets, which is no surprise since those results were obtained with balanced datasets.

Figure 15 provides insights into the individual performance of the various classifiers. The SVM and the Random Forest classification models perform best on average. We report average accuracies of respectively 79.32% and 79.11%. We can see that the Multinomial NB classifier has now the worst average accuracy, shortly behind the Gaussian NB. Let also note that all but one classifier reports similar values in terms of average accuracy and Macro F1-score. In fact, when comparing the two bar charts we see that the Gaussian NB classifier reports different accuracy and Macro F1 values.

By looking at a classification report of this model (**Table 16**), we find that this gap between the two performances is due to the fact that the F1 score suffers from the high number of false positives and false negatives. In fact, the Macro F1-score penalizes the model for making incorrect predictions, while accuracy can remain relatively high if the model correctly identifies the majority of the instances. Furthermore, the classification report shows that the Gaussian NB classifier is better at predicting instances from the classes that were oversampled.

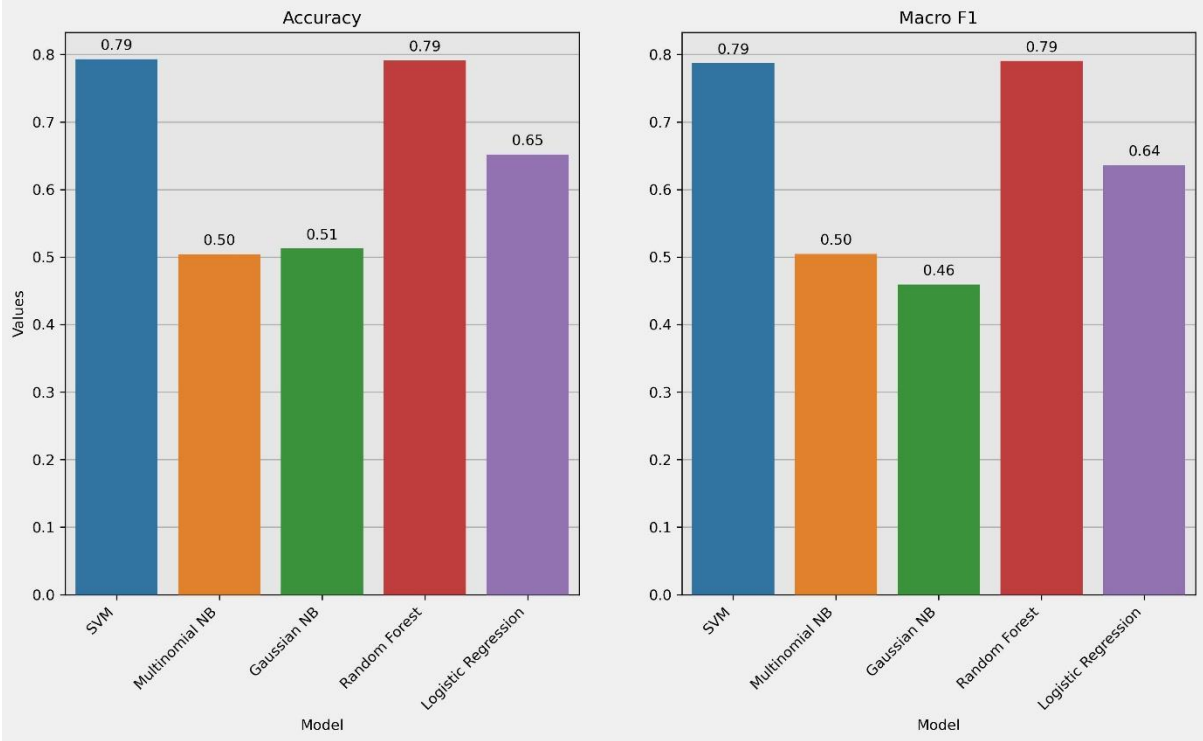

Figure 15:Bar Charts of Average Accuracy and Macro F1-Scores Obtained with Different Classification Models Trained on Oversampled Datasets.

	precision	recall	f1-score	support
A	0.55	0.85	0.67	181
B	0.54	0.12	0.19	181
C	0.49	0.30	0.37	182
D	0.48	0.78	0.59	181
accuracy			0.51	725
macro avg	0.51	0.51	0.46	725
weighted avg	0.51	0.51	0.46	725

Table 16 Classification Report of the Gaussian NB Classifier Trained on the 2020 Oversampled Dataset.

In **Table 17**:Performance Metrics of Classification Models Trained on Oversampled Datasets we report the performance metrics of the different classifiers over the years. The best Macro F1-score obtained is with a Random Forest classification model trained on the 2019, which is the largest oversampled dataset.

From the same table, we can see that the most variable model is the Random Forest classifier (we computed a standard deviation of 3.02%). We can also see that the accuracy of this model increases when the dataset size also does. The other models do not present such relationships.

		2018	2019	2020	2021	2022
Multinomial NB	Precision	0.5162	0.5042	0.4919	0.5223	0.4978
	Recall	0.5111	0.4987	0.4910	0.5191	0.5008
	Macro F1-score	0.5133	0.5002	0.4916	0.5204	0.4976
Gaussian NB	Precision	0.5298	0.5007	0.5147	0.5080	0.5114
	Recall	0.5242	0.4962	0.5117	0.5021	0.5293
	Macro F1-score	0.4795	0.4475	0.4567	0.4527	0.4609
Logistic Regression	Precision	0.6543	0.6099	0.6576	0.6382	0.6130
	Recall	0.6693	0.6327	0.6703	0.6511	0.6361
	Macro F1-score	0.6579	0.6119	0.6570	0.6388	0.6150
Random Forest	Precision	0.8115	0.8251	0.7973	0.7619	0.7523
	Recall	0.8105	0.8253	0.7972	0.7631	0.7594
	Macro F1-score	0.8106	0.8250	0.7974	0.7626	0.7555
SVM	Precision	0.8202	0.7867	0.8047	0.7653	0.7511
	Recall	0.8248	0.7946	0.8110	0.7702	0.7654
	Macro F1-score	0.8207	0.7867	0.8063	0.7670	0.7553

Table 17:Performance Metrics of Classification Models Trained on Oversampled Datasets

4.2. Micro Approach – ESG Rating Grades Prediction

In this section we aim to present the results we obtained for the micro approach, i.e. the prediction task at a more granular level. We begin by presenting the overall performance by using a boxplot which is presented in **Figure 16**. We observe that the overall performance of the models is very poor. Indeed, we observe an average accuracy of only 16.9% (see **Table 18**), which means that the classifiers were able to predict accurately the ESG rating grade of less than one company out of 6 companies.

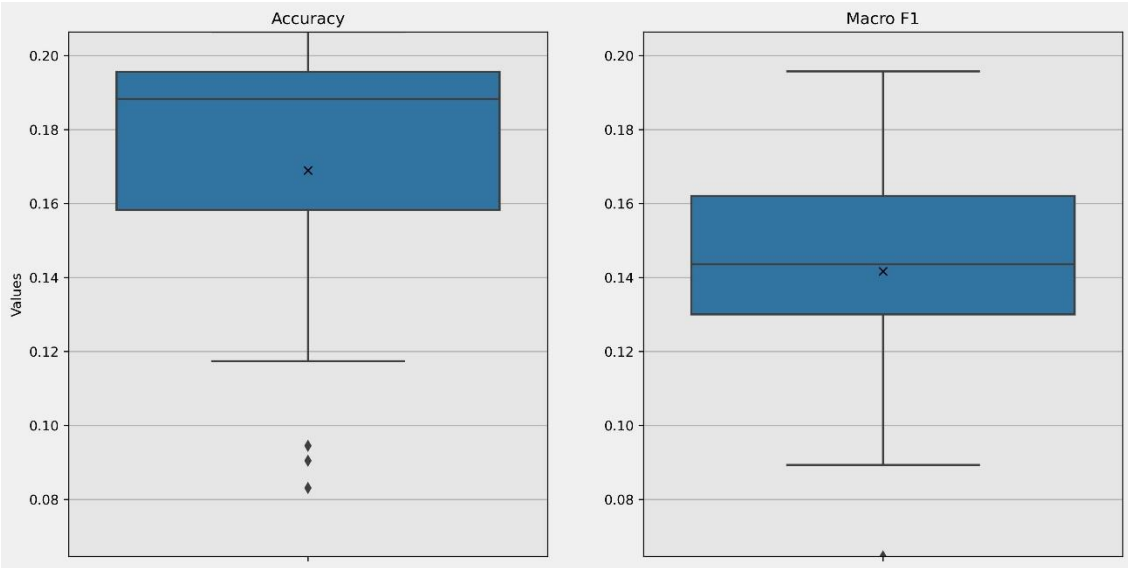

Figure 16: Boxplots of Overall Accuracy and Macro F1-Score

Furthermore, the boxplot reveals that the median accuracy is higher than the mean, suggesting that the majority of the observed accuracies are above 16.9%. This is explained by the presence of outliers at the lower end of the boxplot, which pull the average accuracy downwards.

We investigate the origin of the outliers on the left boxplot by plotting the distribution of each accuracy on a scatter plot (see **Figure 17**: Scatter Plot of The Accuracies for Each Classifier). We observe that the outliers belong to the Gaussian NB classifier, which obtains an average accuracy of only 10.17% and is with this number, by far the less accurate classifier.

The average performance of the other classifiers is similar and ranges between 18% and 20%. We find that the less worse classifier is the Logistic Regression classifier with an average accuracy of 19.20%, meaning that in the best case, this classifier accurately predict one out of five companies.

By looking at the boxplot on the right in (**Figure 16**), we see that the Macro F1-scores are dispersed at a lower level, which suggests the same issue that we encountered when we tried to accurately classify the companies in their ESG rating category.

	2018	2019	2020	2021	2022	AVERAGE
Multinomial NB	0.1741	0.1956	0.1583	0.1769	0.1932	**0.1796**
Gaussian NB	0.0945	0.0831	0.0905	0.1229	0.1174	**0.1017**
Logistic Regression	0.1965	0.2029	0.1608	0.2064	0.1932	**0.1920**
Random Forest	0.1990	0.1883	0.1583	0.1892	0.1883	**0.1846**
SVM	0.1766	0.1932	0.1658	0.2015	0.1980	**0.1870**
AVERAGE	**0.1682**	**0.1726**	**0.1467**	**0.1794**	**0.1780**	**0.1690**

Table 18: Average Accuracy of Each Classifier

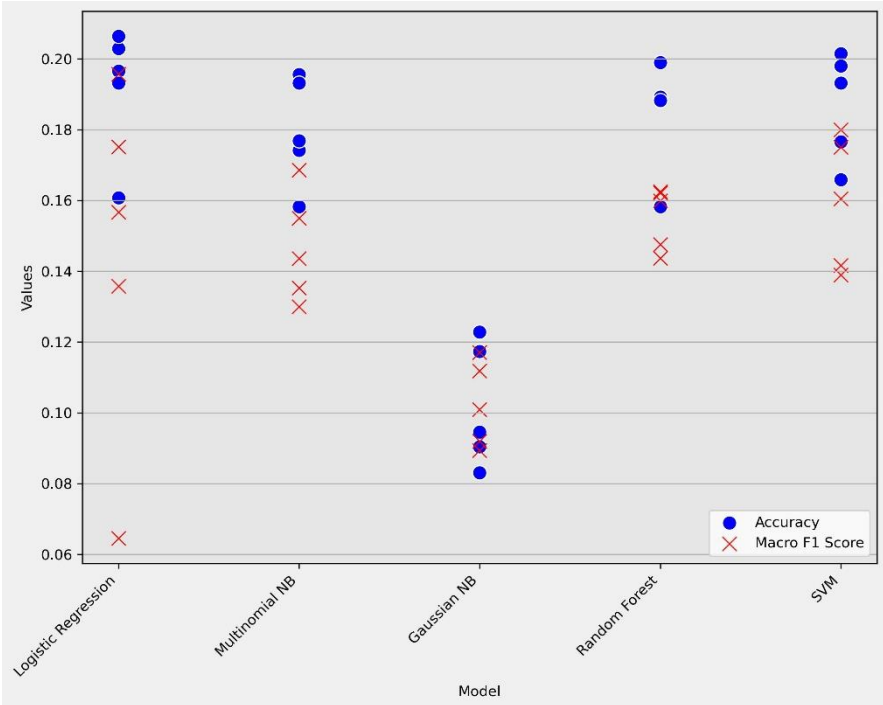

Figure 17: Scatter Plot of The Accuracies for Each Classifier

We investigate this by looking at the classification report (see **Table 19**) of the Multinomial NB classifier that was trained on the 2022 dataset. We observe that the accuracy is entirely driven by the majority grades. In fact, this is another perfect example of the class imbalance problem, which leads to a biased evaluation. It proves, again, that me must be careful when looking at the accuracy metric of classifiers trained on a severely imbalanced dataset.

Lastly, we take a look at the evolution of the average accuracy obtained over the years. We find, looking at the bottom line of **Table 18**, that the performance obtained is stable over the different years, except for the classifiers trained on the 2019 dataset, which shows a lower performance. We checked this finding by looking at the evolution of the average Macro-F1 score for each of the surveyed year (see **Table 20**). We can observe that the classifiers obtain a slightly higher average Macro F1-score by the end of the surveyed years, but the evolution is very small and we can thus not say that the performance has increased over the years.

	precision	recall	f1-score	support
A	0.00	0.00	0.00	6
A-	0.00	0.00	0.00	32
B	0.26	0.39	0.31	59
B+	0.21	0.26	0.23	50
B-	0.13	0.10	0.11	58
C	0.19	0.28	0.23	60
C+	0.16	0.17	0.17	58
C-	0.17	0.21	0.19	48
D	0.00	0.00	0.00	8
D+	0.00	0.00	0.00	29
D-	0.00	0.00	0.00	1
Accuracy			0.19	0.19
Macro average	0.10	0.13	0.11	409
Weighted Average	0.15	0.19	0.17	409

Table 19: Classification Report of The Multinomial NB Classifier Trained on The 2022 Dataset

	2018	2019	2020	2021	2022	AVERAGE
Multinomial NB	0.1436	0.1353	0.1300	0.1549	0.1686	0.1465
Gaussian NB	0.0919	0.0894	0.1009	0.1171	0.1119	0.1022
Logistic Regression	0.0646	0.1358	0.1567	0.1957	0.1752	0.1456
Random Forest	0.1621	0.1598	0.1436	0.1475	0.1625	0.1551
SVM	0.1417	0.1605	0.1390	0.1800	0.1751	0.1592
AVERAGE	0.1208	0.1362	0.1341	0.1590	0.1586	0.1417

Table 20: Average Macro F1-Score For Each Classifier and Each Year

5. Discussion

We will now review and discuss the results obtained in the previous chapter. Overall, we observe that using machine learning classification models to predict ESG ratings is a particularly challenging task, especially when dealing with imbalanced datasets. Although some models perform relatively better performance than others, the overall accuracy and F1-scores suggest that the predictive power of these models falls short for this task.

For clarity, we will first discuss the results of each approach individually before moving on to observations that apply to both approaches.

5.1. Macro Approach – ESG Rating Categories Prediction

5.1.1. Performance Overview

The macro approach, which aimed to classify the 10-K filings of companies in their correct ESG rating category was expected to be the less challenging task.

However, the overall performance of the classifiers in predicting ESG rating categories reveals significant limitations. On average, we observed an accuracy of 49.15%, indicating that the classifiers were only able to correctly predict the ESG rating categories for fewer than one out of two companies. This poor performance suggests that the models struggled to find the necessary patterns to accurately classify companies into different ESG categories. The issue is further highlighted by the overall average Macro F1-score of just 0.4222, clearly indicating that the classifiers particularly struggled with the minority classes in the datasets.

The high variability in both accuracy and Macro F1-scores also outlines the instability of the classifiers. While the Random Forest and SVM classifiers appeared to be the best models in terms of accuracy, they also showed a significant dispersion in results, indicating sensitivity to the features of specific datasets. On the other hand, the Multinomial NB classifier, although less accurate overall, was the most stable model. This trade-off between accuracy and stability should not be neglected when considering the application of these models in real-life scenarios.

5.1.2. Impact of Resampling Strategies

Training the classifiers on different types of datasets—imbalanced, oversampled, and undersampled—demonstrated a clear impact on model performance. The highest accuracy and Macro F1-scores were achieved with the oversampled datasets, suggesting that balancing the datasets through oversampling helped improve overall performance. However, this finding should be taken with a grain of salt. The increased variance in performance for oversampled datasets indicates that while some classifiers benefited from oversampling, other struggled. This could be due to overfitting, where certain models became too tailored to the training data and thus failed to generalize well to unseen samples.

The performance on downsampled datasets, though the lowest overall (with an average accuracy of 37.43%), exhibited less variability and more consistent results compared to the other datasets. This suggests that downsampling may have reduced some of the instability seen

with the other datasets, but at the expense of overall accuracy. The similar values of accuracy and F1-scores for downsampled datasets indicates that the models were better at handling class imbalances in this scenario.

5.1.3. Imbalanced Datasets and Minority Classes

As expected, classifiers trained on the original (and thus imbalanced) datasets highlighted significant performance issues, particularly with minority classes. We exemplified this with an illustrative example where the classifier failed to predict minority classes. It serves as an important reminder to interpret accuracy with caution when dealing with imbalanced datasets, as it can be misleading and overlook poor performance with minority classes.

5.2. Micro Approach – ESG Rating Grades Prediction

The micro approach, which aimed to predict ESG rating grades, was the most challenging among the two approaches. We obtained an overall average accuracy of 16.9%, with the best-performing model, Logistic Regression, reaching an average accuracy of 19.2%. This suggests that the classifiers struggled significantly at this level of analysis, indicating that they were largely unable to recognize meaningful patterns in the training data and had limited ability to generalize on unseen samples.

Like in the macro approach, accuracy in the micro approach was mainly driven by the classifier's ability to predict majority classes, while performance on minority classes remained poor. The boxplot analysis reveals that the median accuracy is higher than the mean, suggesting that the poor performance of certain outliers (specifically, the Gaussian Naive Bayes model) is dragging down the overall performance. The Macro F1-scores show a similar pattern, reinforcing the conclusion that models are struggling to identify the minority grades.

5.3. ESG Data Evolution Over Time

The steady increase in the number of ESG-related words in 10-K filings from 2018 to 2022 highlights the growing emphasis on ESG reporting. However, this increase in data did not translate into substantial improvements in model performance over time. The classifiers showed only minor changes in accuracy and Macro F1-scores across the years, indicating that the growing volume of ESG data alone was not sufficient to improve predictive accuracy.

6. Conclusion

6.1. Summary

The aim of this thesis was to explore whether the quantity of ESG-related information disclosed by U.S. listed companies in their 10-K filings could be utilized to predict their ratings using machine learning classifiers. To evaluate the effectiveness of these classifiers, two distinct approaches were employed: one focused on predicting ESG categories, and the other on predicting ESG rating grades. Our findings demonstrate that predicting ESG ratings using machine learning classifiers is a challenging task. On average, the classifiers successfully predicted one out of every two ESG categories. However, the more detailed task of predicting ESG grades resulted in poorer outcomes, with less than one out of six ESG rating grades being predicted accurately.

Despite these challenges, the findings of this thesis provide several key insights into the application of machine learning for ESG classification. Firstly, the results emphasize the importance of data quality in training machine learning models for this task. We observed that the classifiers struggled significantly with certain datasets, particularly when trained on the original, imbalanced dataset. While downsampling helped improve model stability, it did so at the cost of reduced accuracy in ESG rating predictions. Furthermore, oversampling the datasets using the SMOTE technique emerged as the least unfavorable approach to addressing class imbalance. However, uncertainty remains as to whether machine learning classifiers trained on oversampled datasets would generalize effectively to unseen ESG ratings, given the heightened risk of overfitting.

However, it should not be the end of the prediction of ESG using 10-K filings, since we found, as already suggested by the literature, an increased quantity of ESG data.

6.2. Sources of Improvement

To enhance the accuracy of ESG rating predictions, the first and most crucial lever to consider is the training data itself. A more balanced dataset is essential to avoid bias and provide a more accurate response to the research question posed. Achieving this, however, is difficult, as the distribution of ESG ratings provided by Refinitiv is heavily skewed towards middle-range classes.

We could also implement a feature selection process to reduce the number of features, which would facilitate quicker and more efficient pattern recognition by the machine learning classifiers.

Additionally, given that the machine learning classifiers performed poorly on downsampled datasets, another area of improvement could involve modifying other variables within the experiment. For instance, one potential lever could be to use a different ESG word list than Baier’s. While companies do disclose material information in their 10-K filings, the materiality of this information varies depending on the specific context of each company and its industry. Baier's list is not industry-specific and may not capture the ESG information of all companies accurately. A more tailored approach could involve using industry-specific word lists, such as those developed by Rouen et al. (2024), to better reflect the nuances of ESG disclosures.

Another potential improvement lies in the application of natural language processing. In the current experiment, we operated under the assumption that the quantity of ESG-related terms in a company's disclosures could approximate its ESG performance. However, the methodology relied solely on the frequency of ESG-related words in the documents. Integrating more advanced natural language processing techniques would allow for deeper insights into the textual data, enabling the extraction of more nuanced ESG information and, consequently, potentially improving the accuracy of the predictions.

Appendix : ESG Word List

clean	assessments	motivated	
environmental	audit	motivates	bargaining
epa	audited	motivating	eeo
sustainability	auditing	motivation	fairness
climate	auditor	recruit	fla
warming	auditors	recruiting	harassment
biofuel	audits	recruitment	injury
biofuels	control	retain	labor
green	controls	retainer	overtime
renewable	coso	retainers	ruggie
solar	detect	retaining	sick
stewardship	detected	retention	wage
wind	detecting	talent	wages
atmosphere	detection	talented	workplace
emission	evaluate	talents	bisexual
emissions	evaluated	brother	diversity
emit	evaluates	clicking	ethnic
ghg	evaluating	conflict	ethnically
ghgs	evaluation	conflicts	ethnicities
greenhouse	evaluations	family	ethnicity
agriculture	examination	grandchildren	female
deforestation	examinations	grandparent	females
pesticide	examine	grandparents	gay
pesticides	examined	inform	gays
wetlands	examines	insider	gender
zoning	examining	insiders	genders
biodiversity	irs	inspector	homosexual
species	oversee	inspectors	immigration
wilderness	overseeing	interlocks	lesbian
wildlife	oversees	nephews	lesbians
freshwater	oversight	nieces	lgbt
groundwater	review	posting	minorities
water	reviewed	relatives	minority
cleaner	reviewing	siblings	ms
cleanup	reviews	sister	race
coal	rotation	son	racial
contamination	test	spousal	religion
fossil	tested	spouse	religious
resource	testing	spouses	sex
air	tests	stepchildren	transgender
carbon	treadway	stepparents	woman
nitrogen	backgrounds	transparency	women

pollution	independence	transparent	occupational
superfund	leadership	visit	safe
biphenyls	nomination	visiting	safely
hazardous	nominations	visits	safety
householding	nominee	webpage	ilo
pollutants	nominees	website	labour
printing	perspectives	announce	eicc
recycle	qualifications	announced	children
recycling	refreshment	announcement	epidemic
toxic	skill	announcements	health
waste	skills	announces	healthy
wastes	succession	announcing	ill
weee	tenure	communicate	illness
climate change	vacancies	communicated	pandemic
conservation	vacancy	communicates	childbirth
environmentally	appreciation	communicating	drug
footprint	award	erm	medicaid
global warming	awarded	fairly	medicare
pollutant	awarding	integrity	medicine
recycled	awards	liaison	medicines
sustainable	bonus	presentation	hiv
sustainably	bonuses	presentations	alcohol
align	cd	sustainable	drinking
aligned	compensate	asc	bugs
aligning	compensated	disclose	conformance
alignment	compensates	disclosed	defects
aligns	compensating	discloses	fda
bylaw	compensation	disclosing	inspection
bylaws	eip	disclosure	inspections
charter	iso	disclosures	minerals
charters	isos	fasb	standardization
culture	payout	gaap	warranty
death	payouts	objectivity	endowment
duly	pension	press	endowments
independent	prsu	sarbanes	people
parents	prsus	engagement	philanthropic
cobc	recoupment	engagements	philanthropy
ethic	remuneration	feedback	socially
ethical	reward	hotline	societal
ethically	rewarding	investor	society
ethics	rewards	invite	welfare
honesty	rsu	invited	charitable
bribery	rsus	mail	charities
corrupt	salaries	mailed	charity
corruption	salary	mailing	donate

crimes	severance	mailings	donated
embezzlement	vest	notice	donates
grassroots	vested	relations	donating
influence	vesting	stakeholder	donation
influences	vests	stakeholders	donations
influencing	ballot	compact	donors
lobbied	ballots	ungc	foundation
lobbies	cast	citizen	foundations
lobby	consent	citizens	gift
lobbying	elect	csr	gifts
lobbyist	elected	disabilities	nonprofit
lobbyists	electing	disability	poverty
whistleblower	election	disabled	courses
compliance	elections	human	educate
conduct	elects	nations	educated
conformity	nominate	social	educates
governance	nominated	un	educating
misconduct	plurality	veteran	education
parachute	proponent	veterans	educational
parachutes	proponents	vulnerable	learning
perquisites	proposal	dignity	mentoring
plane	proposals	discriminate	scholarships
planes	proxies	discriminated	teach
poison	quorum	discriminating	teacher
retirement	vote	discrimination	teachers
approval	voted	equality	teaching
approvals	votes	freedom	training
approve	voting	humanity	employ
approved	attract	nondiscrimination	employment
approves	attracting	sexual	headcount
approving	attracts	communities	hire
assess	incentive	community	hired
assessed	incentives	expression	hires
assesses	interview	marriage	hiring
assessing	interviews	privacy	staffing
assessment	motivate	peace	unemployment

Bibliography

Abiteboul, S. (1996). Querying semi-structured data. In F. Afrati & P. Kolaitis (Eds.), Database Theory — ICDT '97. ICDT 1997. Lecture Notes in Computer Science (Vol. 1186). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-62222-5_33

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919. https://arxiv.org/abs/1707.02919

Amel-Zadeh, A., & Serafeim, G. (2018). Why and how investors use ESG information: Evidence from a global survey. Financial Analysts Journal, 74(3), 87-103. https://doi.org/10.2139/ssrn.2925310

Antolín-López, R., Delgado-Ceballos, J., & Montiel, I. (2016). Deconstructing corporate sustainability: A comparison of different stakeholder metrics. Journal of Cleaner Production, 136(Part A), 5-17. https://doi.org/10.1016/j.jclepro.2016.01.111

Aureli, S. (2017). A comparison of content analysis usage and text mining in CSR corporate disclosure. International Journal of Digital Accounting Research, 17, 1-22.

Avetisyan, E., & Hockerts, K. (2017). The consolidation of the ESG rating industry as an enactment of institutional retrogression. Business Strategy and the Environment, 26(7), 844-857. https://doi.org/10.1002/bse.1919

Barkemeyer, R., Revelli, C., & Douaud, A. (2023). Selection bias in ESG controversies as a risk for sustainable investors. Journal of Cleaner Production, 405, 137035.

Bell, M. (2021, March 9). Why ESG performance is growing in importance for investors. EY. Retrieved August 11, 2024, from https://www.ey.com/en_gl/insights/assurance/why-esg-performance-is-growing-in-importance-for-investors

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

Billio, M., Costola, M., Hristova, I., Latino, C., & Pelizzon, L. (2020). Inside the ESG ratings: (Dis)agreement and performance. SAFE Working Paper No. 284. https://doi.org/10.2139/ssrn.3674502

Bloomberg Intelligence. (2023, June 14). Global ESG assets predicted to hit $40 trillion by 2030 despite challenging environment: Bloomberg Intelligence. Bloomberg. Retrieved from https://www.bloomberg.com/company/press/global-esg-assets-predicted-to-hit-40-trillion-by-2030-despite-challenging-environment-forecasts-bloomberg-intelligence/

Boffo, R., & Patalano, R. (2020). ESG Investing: Practices, Progress and Challenges. OECD.

Bose, S. (2020). Evolution of ESG reporting frameworks. In Sustainable Investing: Revolutions in Theory and Practice (pp. 17-34). Springer. https://doi.org/10.1007/978-3-030-55613-6_2

Bouyé, E., Klingebiel, D., & Ruiz, M. (2021). Environmental, social, and governance investing: A primer for central banks' reserve managers. https://doi.org/10.1596/36285

Boyle, E. J., Higgins, M. M., & Rhee, S. G. (1997). Stock market reaction to ethical initiatives of defense contractors: Theory and evidence. Critical Perspectives on Accounting, 8(6), 541–561. https://doi.org/10.1006/cpac.1997.0124

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010950718922

Brown, H., de Jong, W. M., & Levy, D. (2009). Building institutions based on information disclosure: Lessons from GRI's sustainability reporting. Journal of Cleaner Production, 17(6), 571-580. https://doi.org/10.1016/j.jclepro.2008.12.009

Bruder, Benjamin and Cheikh, Yazid and Deixonne, Florent and Zheng, Ban, Integration of ESG in Asset Allocation (October 25, 2019). Available at SSRN: https://ssrn.com/abstract=3473874 or http://dx.doi.org/10.2139/ssrn.3473874

Byrne, D. (n.d.). What's the difference between ESG reporting standards and frameworks? The Corporate Governance Institute. Retrieved from https://www.thecorporategovernanceinstitute.com/insights/guides/whats-the-difference-between-esg-reporting-standards-and-frameworks/

Caplan, L., Griswold, J. S., & Jarvis, W. F. (2013). From SRI to ESG: The changing world of responsible investing. Commonfund Institute.

Castellanos, A., Parra, C., & Tremblay, M. (2015). Corporate social responsibility reports: Understanding topics via text mining.

Caudron, E., & Vrins, F. (2022). Measuring ESG Performance: A Text Mining Approach. Louvain School of Management, Université catholique de Louvain. CFA Institute,[Online] available: https://www. cfainstitute. org/-/media/documents/protected/esg-candidate/pdf/2021-Chapter3. pdf [2022 August 20].

Cervantes, J., García-Lamont, F., Rodríguez, L., & Lopez-Chau, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

CFA Institute. (n.d.). ESG investing. CFA Institute. Retrieved from https://www.cfainstitute.org/en/rpc-overview/esg-investing

Chatterji, A. K., Levine, D. I., & Toffel, M. W. (2009). How well do social ratings actually measure corporate social responsibility? Journal of Economics & Management Strategy, 18(1), 125-169. https://doi.org/10.1111/j.1530-9134.2009.00210.x

Chatterji, A., Durand, R., Levine, D. I., & Touboul, S. (2014). Do ratings of firms converge? Implications for managers, investors, and strategy researchers. Strategic Management Journal. (Forthcoming). HEC Paris Research Paper No. SPE-2015-1076. https://doi.org/10.2139/ssrn.2514361

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, M., von Behren, R., & Mussalli, G. (2021). The unreasonable attractiveness of more ESG data. SSRN. https://doi.org/10.2139/ssrn.3883651

Chowdhury, G. G. (2003). Natural language processing. Annual Review of Information Science and Technology, 37(1), 51-89. https://doi.org/10.1002/aris.1440370103

Christensen, D. M., Serafeim, G., & Sikochi, A. (2021). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. The Accounting Review. https://doi.org/10.2308/TAR-2019-0506

Corporate Finance Institute. (n.d.). Environmental, Social, and Governance (ESG) course. Corporate Finance Institute. Retrieved from https://corporatefinanceinstitute.com/course/environmental-social-governance/

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Council of the European Union. (2024). Proposal for a regulation of the European Parliament and the Council on the transparency and integrity of Environmental, Social and Governance (ESG) rating activities, and amending Regulation (EU) 2019/2088 (Document No. 6255/24).

D'Amato, V., D'Ecclesia, R., & Levantesi, S. (2022). ESG score prediction through random forest algorithm. Computational Management Science, 19(2), 347-373. https://doi.org/10.1007/s10287-021-00412-3

Deloitte. (2024). Earning trust with investors through better sustainability data. Deloitte Global. Retrieved from https://www.deloitte.com/global/en/issues/climate/earning-trust-with-investors-through-better-sustainability-data.html

Deutsche Wealth. (n.d.). Corporate governance (G) in ESG. Deutsche Wealth. Retrieved August 11, 2024, from https://www.deutschewealth.com/en/our-capabilities/esg/what-is-esg-investing-wealth-management/corporate-governance-g-in-esg-governance.html

Dimson, E., Marsh, P., & Staunton, M. (2020). Historical returns, long-term perspectives, and new insights for today's market. The Journal of Portfolio Management, 47(1), 75-87. https://doi.org/10.3905/jpm.2020.1.175

Drempetic, S., Klein, C., & Zwergel, B. (2020). The influence of firm size on the ESG score: Corporate sustainability ratings under review. Journal of Business Ethics, 167(2), 333-360. https://doi.org/10.1007/s10551-019-04164-1

Eccles, R., Lee, L.-E., & Stroehle, J. (2019). The social origins of ESG?: An analysis of Innovest and KLD. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3318225

Escrig-Olmedo, E., Fernández-Izquierdo, M. Á., Ferrero-Ferrero, I., Rivera-Lirio, J. M., & Muñoz-Torres, M. J. (2019). Rating the Raters: Evaluating how ESG Rating Agencies Integrate Sustainability Principles. Sustainability, 11(3), 915. https://doi.org/10.3390/su11030915

ESG Today. (2024). SEC defends its climate disclosure rule in court. ESG Today. Retrieved from https://www.esgtoday.com/sec-defends-its-climate-disclosure-rule-in-court/

European Parliament, & Council of the EU. (2014). Directive 2014/95/EU of the European Parliament and of the Council of22 October 2014 amending directive 2013/34/EU as regards disclosure of non-financial and diversity information bycertain large undertakings and groups. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014L0095

Fatemi, A., Glaum, M., & Kaiser, S. (2018). ESG performance and firm value: The moderating role of disclosure. Global finance journal, 38, 45-64.

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. Systems Science & Control Engineering, 2(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), 210-233. https://doi.org/10.1080/20430795.2015.1118917

García, F., González-Bueno, J., Guijarro, F., & Oliver, J. (2020). Forecasting the environmental, social, and governance rating of firms by using corporate financial performance variables: A rough set approach. Sustainability, 12(8), 3324. https://doi.org/10.3390/su12083324

Gez, M., Anagnosti, E., & Pullins, T. (2022, July 16). ESG disclosure trends in SEC filings. Harvard Law School Forum on Corporate Governance. Retrieved from https://corpgov.law.harvard.edu/2022/07/16/esg-disclosure-trends-in-sec-filings/

Giglio, S., Maggiori, M., Stroebel, J., Tan, Z., Utkus, S. P., & Xu, X. (2023). Four facts about ESG beliefs and investor portfolios. Wharton Pension Research Council Working Paper No. 2023-04. https://doi.org/10.2139/ssrn.4415012

Gupta, V., Koller, T., & Stumpner, P. (2021). Reports of corporates' demise have been greatly exaggerated. McKinsey & Company.

Haber, S., Kepler, J. D., Larcker, D. F., Seru, A., & Tayan, B. (2022). ESG investing: What shareholders do fund managers represent? Rock Center for Corporate Governance at Stanford University Working Paper. Stanford University Graduate School of Business Research Paper No. 4267270.

Hayes, A. (2024, June 27). Russell 3000 index: Stocks and limitations. Investopedia. Retrieved from https://www.investopedia.com/terms/r/russell_3000.asp

Henisz, W., Koller, T., & Nuttall, R. (2019). Five ways that ESG creates value. McKinsey Quarterly, 4, 1-12.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. Journal for Language Technology and Computational Linguistics, 20(1), 19-62.

Hughey, C., & Sulkowski, A. (2012). More disclosure = better CSR reputation? An examination of CSR reputation leaders and laggards in the global oil and gas industry. Journal of Academy of Business and Economics, 12(1), 24-34.

Hummel, K., & Jobst, D. (2024). An overview of corporate sustainability reporting legislation in the European Union. Accounting in Europe. (Forthcoming). https://doi.org/10.2139/ssrn.4004598

Ignatov, K. (2020). When ESG talks: ESG tone of annual reports and its significance to stock markets. SSRN. https://doi.org/10.2139/ssrn.3715427

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), Machine Learning: ECML-98. ECML 1998.

Lecture Notes in Computer Science (Vol. 1398). Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0026683

Jusoh, S., & Alfawareh, H. M. (2012). Techniques, applications and challenging issues in text mining. International Journal of Computer Science Issues (IJCSI), 9(6), 431.

Kay, I., Brindisi, C., & Martin, B. (2020, September 14). The stakeholder model and ESG. Harvard Law School Forum on Corporate Governance. Retrieved from https://corpgov.law.harvard.edu/2020/09/14/the-stakeholder-model-and-esg/

Kotsantonis, S., & Serafeim, G. (2019). Four things no one will tell you about ESG data. Journal of Applied Corporate Finance, 31(2), 50-58. https://doi.org/10.1111/jacf.12346

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190. https://doi.org/10.1007/s10462-007-9052-3

Krantz, T. (2024). The history of ESG: A journey towards sustainable investing. IBM Blog. Retrieved from https://www.ibm.com/blog/environmental-social-and-governance-history/

Kräussl, R., Oladiran, T., & Stefanova, D. (2024). A review on ESG investing: Investors' expectations, beliefs, and perceptions. Journal of Economic Surveys, 38(2), 476-502.

La Torre, M., Sabelfeld, L., Blomkvist, M., & Dumay, J. (2020). Rebuilding trust: Sustainability and non-financial reporting and the European Union regulation. Meditari Accountancy Research, 28(5), 701-725. https://doi.org/10.1108/MEDAR-06-2020-0914

Larcker, D. F., Pomorski, L., Tayan, B., & Watts, E. M. (2022). ESG ratings: A compass without direction. Rock Center for Corporate Governance at Stanford University Working Paper. (Forthcoming). Available at SSRN: https://ssrn.com/abstract=4179647

Longadge, R., & Dongre, S. (2013). Class imbalance problem in data mining: Review. International Journal of Computer Science and Network, 2(1).

Mandas, M., Lahmar, O., Piras, L., & De Lisa, R. (2023). ESG in the financial industry: What matters for rating analysts? Research in International Business and Finance, 66, 102045. https://doi.org/10.1016/j.ribaf.2023.102045

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge: Cambridge University Press.

Mazzacurati, J. (2021). ESG ratings: Status and key issues ahead. ESMA Report on Trends, Risks, and Vulnerabilities No. 1, 2021. European Securities and Markets Authority. Retrieved from

https://www.esma.europa.eu/press-news/esma-news/esma-calls-legislative-action-esg-ratings-and-assessment-tools

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. AAAI Conference on Artificial Intelligence.

McWilliams, A., & Siegel, D. (2000). Corporate social responsibility and financial performance: correlation or misspecification?. Strategic management journal, 21(5), 603-609.

Miner, G. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Principles for Responsible Investment. (2024). 2024 Annual Report.

PwC. (2022). ESG-focused institutional investment seen soaring 84% to US$33.9 trillion in 2026, making up 21.5% of assets under management: PwC report. PwC. Retrieved from https://www.pwc.com/id/en/media-centre/press-release/2022/english/esg-focused-institutional-investment-seen-soaring-84-to-usd-33-9-trillion-in-2026-making-up-21-5-percent-of-assets-under-management-pwc-report.html

Renneboog, L., Ter Horst, J., & Zhang, C. (2008). Socially responsible investments: Institutional aspects, performance, and investor behavior. Journal of Banking & Finance, 32(9), 1723-1742. https://doi.org/10.1016/j.jbankfin.2007.12.039

Riedl, A., & Smeets, P. (2017). Why do investors hold socially responsible mutual funds?. The Journal of Finance, 72(6), 2505-2550.

Roselle, P. (2016). The evolution of integrating ESG analysis into wealth management decisions. Journal of Applied Corporate Finance, 28(2), 75-79. https://doi.org/10.1111/jacf.12178

Rouen, E., Sachdeva, K., & Yoon, A. (2024). Sustainability meets substance: Evaluating ESG reports in the context of 10-Ks and firm performance. SSRN. https://doi.org/10.2139/ssrn.4227934

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620. https://doi.org/10.1145/361219.361220

Scalet, S., & Kelly, T. F. (2010). CSR rating agencies: What is their global impact? Journal of Business Ethics, 94(1), 69–88. https://doi.org/10.1007/s10551-009-0250-6

Senadheera, S. S., Gregory, R., Rinklebe, J., Farrukh, M., Rhee, J. H., Ok, Y. S., & You, S. (2022). The development of research on environmental, social, and governance (ESG): A bibliometric

analysis. Sustainable Environment, 8(1), 2125869. https://doi.org/10.1080/27658511.2022.2125869

Serafeim, G., & Grewal, J. (2016). ESG Metrics: Reshaping Capitalism? Harvard Business School Technical Note No. 116-037.

Stubbs, W., & Rogers, P. (2013). Lifting the veil on environment-social-governance rating methods. Social Responsibility Journal, 9(4), 622-640. https://doi.org/10.1108/SRJ-03-2012-0035

SustainAbility. (2020). Rate the raters 2020 investor survey and interview results. SustainAbility. Retrieved from https://www.sustainability.com/sustainability-report-2020/impact-with-clients/esg-and-sustainable-finance/rate-the-raters-2020-investor-survey-and-interview-results/

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: Techniques, applications and issues. International Journal of Advanced Computer Science and Applications, 7(11), 414-418.

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

Taylor, J., Vithayathil, J., & Yim, D. (2018). Are corporate social responsibility (CSR) initiatives such as sustainable development and environmental policies value-enhancing or window dressing? Corporate Social Responsibility and Environmental Management, 25(6), 971–980. https://doi.org/10.1002/csr.1513

The SustainAbility Institute by ERM. (2023). ESG ratings at a crossroads: Rate the Raters 2023. Retrieved from https://www.sustainability.com/sustainability-report-2023

U.S. Securities and Exchange Commission. (2024, March 15). SEC announces updates to climate disclosure regulations. Retrieved August 11, 2024, from https://www.sec.gov/newsroom/press-releases/2024-31

United Nations Global Compact. (2004). Who Cares Wins: Connecting Financial Markets to a Changing World. United Nations.

Vance, S. (1975). Are socially responsible corporations good investment risks? Managerial Review, 64, 18-24.

Vartiak, L. (2016). CSR reporting of companies on a global scale. Procedia Economics and Finance, 39, 176-183. https://doi.org/10.1016/S2212-5671(16)30276-3

Vijayarani, S., Ilamathi, J., & Nithya, M. (2015). Preprocessing techniques for text mining - An overview. International Journal of Computer Science & Communication Networks, 5(1), 7-16.

Windolph, S. E. (2011). Assessing corporate sustainability through ratings: Challenges and their causes. Journal of Environmental Sustainability, 1(1), Article 5. https://doi.org/10.14448/jes.01.0005

Wong, C., & Petroy, E. (2020). Rate the Raters 2020: Investor survey and interview results. The SustainAbility Institute by ERM. Retrieved from https://sustainability.com/our-work/reports/rate-raters-2020-investor-survey-and-interview-results/

World Economic Forum. (2022, July 13). The "G" in ESG: 3 ways to not miss it. World Economic Forum. Retrieved from https://www.weforum.org/agenda/2022/07/the-g-in-esg-3-ways-to-not-miss-it/

Yoon, Joohye & Han, Sujin & Lee, Yongseok & Hwang, Hyesun. (2023). Text Mining Analysis of ESG Management Reports in South Korea: Comparison With Sustainable Development Goals. SAGE Open. 13. 10.1177/21582440231202896.

Zhang, Q., & Segall, R. S. (2008). Web mining: A survey of current research, techniques, and software. International Journal of Information Technology & Decision Making, 7(4), 683-720.

Executive summary

This master's thesis[2] explores the use of ESG-related textual information from 10-K filings of U.S.-listed companies to predict their ESG ratings using machine learning classifiers. As ESG ratings play an increasingly important role in responsible investing, this study provides an alternative approach to traditional ESG rating methods by leveraging publicly available corporate disclosures.

Two approaches were used: a macro approach to predict broader ESG rating categories and a micro approach to predict more detailed ESG grades. Various machine learning models, including Support Vector Machines, Logistic Regression, and Random Forests, were trained on processed 10-K data. The results show that while machine learning models managed an accuracy of around 49% for ESG categories, predicting specific ESG grades proved more challenging, with accuracy dropping to 16.9%. This highlights the complexity of the task, especially due to the imbalanced nature of the dataset.

Key insights include the importance of data quality in training models and the challenges of using imbalanced datasets, where models struggled with minority classes. Oversampling using the SMOTE technique showed promise in improving performance, though the risk of overfitting remains a concern. Despite these challenges, the study demonstrates the potential of using machine learning for ESG rating predictions as ESG disclosures in 10-K filings increase.

Overall, this thesis contributes to the growing field of machine learning-based ESG analysis and underscores the need for further refinement of models and techniques. It offers a foundation for future research into more reliable methods for ESG rating prediction.

[2] Number of words = 19814

Note on the use of GenAI: After completing the writing of this dissertation, I tasked ChatGPT-4 to review certain paragraphs with the goal of enhancing clarity and ensuring the overall coherence of the thesis.