UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Measuring ESG performance: A Text Mining Approach

Author(s): Emilien Caudron
Supervisor(s): Frédéric Vrins
Academic year 2021-2022
Dissertation for the master of
Business Engineering, Major in Financial Engineering
Daytime schedule

Acknowledgments

I would like to use this opportunity to express my gratitude and thanks to the various people who helped me undertake this research. First and foremost, I would like to thank my supervisor, Professor Fredéric Vrins, for its guidance and valuable advice during the whole progress of this thesis.

Then, this accomplishment would also not have been possible without the support of my friends all along my academic years and during the realisation of this work. I would like to express my gratitude to two of them in particular: Emma Gastoud and Mathilde Van Hijfte. Without your support in the last few months, I probably would not have been able to complete this work in time.

My last words will go to two people who are particularly dear to me: my parents. Thank you for always believing in me; for helping me; and for inspiring me.

TABLE OF CONTENTS

INTRODUCTION:

More than ever, environmental, social and governance issues are representing critical challenges for the modern society. Given the conclusions of the latest IPCC report on the impacts of climate change, it seems that besides being critical those are also becoming urgent to tackle. If mankind ever plans to reach its objective of protecting the environment while bringing prosperity, it is high time to act (UN, 2015). In this context, it appears that sustainable finance and ESG have an important role to play (Mann, 2021; Zesty, 2017). Indeed, those actors can support the financing of sustainable development by excluding investments in bad ESG companies while also finding and promoting the development of innovative and responsible business opportunities. Yet, to achieve those objectives, responsible investors are in dire need of clear, credible, accessible, and transparent measures of ESG performance. Currently, commonly used proxies for assessing the ESG performance of a company are ESG ratings. However, in their current format, those ratings do not meet the expectations of investors (low transparency, costly, biased, etc.). Hence, both investors and scholars have started to search for new and more appropriate ESG rating methodologies.

In this research, we propose to measure the ESG performance of a company by relying on the text mining analysis of its sustainability disclosure. More precisely, we are deriving ESG ratings from the number of occurrences of ESG terms in Annual and CSR reports. Of course, one cannot expect such a technique to provide a fully accurate indicator of ESG performance. Yet, in light of the importance of ESG described above and given the large number of firms investors have to analyze, this is a first step towards a more efficient and qualitative assessment of ESG performance. To achieve this, this research is divided in two main parts:

- Literature review: This first part should allow to contextualize the research and highlight the importance of the topic (role of ESG ratings, its shortcomings, etc.), but also to give the technical tools necessary to develop a score based on Text Mining (existing methodologies, tools, etc.).
- Empirical research: In this second part, we build and test the ability of several Text Mining models to measure ESG performance based on the number of occurrences of ESG terms in Annual and/or CSR reports. In particular, we test a simple word frequency sum, but also more complex classification models (Random Forest Classifiers, Support Vector machines, etc.). The quality of the models thus created is then judged based on a comparison against ratings from the agency Refinitiv. Based on this comparison, we will draw conclusions about the performance of our approach and about possible ways of improvement.

Part 1: LITERATURE REVIEW

Chapter 1: Measuring ESG performance

1. ESG Performance and ESG Ratings

1.1. Defining ESG performance

ESG performance refers to the efforts taken by a company to jointly address Environmental, Social and Governance (ESG) issues, i.e., reduce its impact in three of the main ethical finance pillars (Billio et al., 2021; Escrig-Olmedo et al., 2019; Semenova & Hassel, 2015). As for financial performance, assessing the ESG performance of a company is done by analyzing a series of metrics or ESG-focused KPI's. In the last decade, measuring the ESG performance of a company has become a common practice across the financial sector, especially since the explosion of Socially Responsible Investments (SRI). As a reminder, SRI, which remains a complex and blur notion to define, encapsulates a set of investing strategies aiming at integrating both financial and social information in the investment decision process (Boffo & Patalano, 2020; Salampasis, 2017). In recent years, this sustainable investing practice has become one of fastest-growing segment of the financial markets. Since 1995, and the first measurement of the SRI universe in the US, the number of sustainable assets under management has increased at an annual compound growth rate of 13.6% (US SIF, 2020).

According to multiple survey's, conducted across major actors of the financial sector, this growing interest for more responsible investments can be explained by two main drivers (Bernow al., 2017; Boffo & Patalano, 2020; Delevingne & Gründler, 2020). On the one hand, we observe that shareholders increasingly feel the need to find alignment between their investments and their values (addressing climate change, be more than financially performant, etc. (Bernow et al., 2017; Halbritter & Dorfleitner, 2015; Kachaner et al., 2020; Zesty, 2017)). For instance, under the pressure of "end-investors" and regulatory bodies, a growing number of institutional investors have started to look for investments in alignment with SDG's and to integrate ESG criteria in their decision-making process (Bernow et al., 2017; Boffo & Patalano, 2020). On the other hand, more and more investors are convinced that integrating ESG criteria in their decision-making process can support their search for better long-term financial value (BCG Global, n.d.; Gartenberg et al., 2018; Mirvis & Ryu, 2009). Note that though a lot of them seems to be convinced that this positive relationship between financial and non-financial

performance exists, we are still lacking clear empirical results to demonstrate this hypothesis (Eccles et al., 2014; Friede et al., 2015; Halbritter & Dorfleitner, 2015) (see Appendix II).

Yet, despite the uncertainty surrounding the ability of ESG stocks to outperform their conventional peers, ESG investing is still appealing to a lot of investors (for ethical or financial reasons) and is therefore in a booming trend. And according to many investments professional this boom is not about to stop (Bloomberg Intelligence, 2021). And it is for the best considering the gigantic financial gap humanity still has to fill to reach the SDG by the end of the decade. Yet, to support the development of sustainable investing, investors are in dire need of information and measures on both non-financial and financial performance (Crespi & Migliavacca, 2020). That is why in parallel to the SRI movement, another market has boomed in recent years, the market of ESG information and rating providers.

1.2. Measuring ESG performance

While there exist several ways of evaluating the ESG performance of a company, including in-house analysis, rankings, awards, reports from NGO's or consulting firms (Windolph, 2011), as of today, it seems that ratings are the preferred way of using ESG information (Halbritter & Dorfleitner, 2015; Scalet & Kelly, 2010; Wong & Petroy, 2020). The number of ESG ratings, and ESG rating providers, has therefore been in a growing trend since the early days of SRI, reaching the bar of 600 available ratings or scores in 2018 (Fish et al., 2019). Note however that in their paper "*Rating the Raters: Evaluating how ESG Rating Agencies Integrate Sustainability Principles*", Escrig-Olmedo et al. (2019) are observing that the ESG rating agency market has undergone a process of concentration and merger since the financial crisis, especially through the entrance of more traditional rating and information provider, e.g., Bloomberg, MSCI, S&P, etc.

According to Cash (2018) or Windolph (2011), the attraction of sustainable investors for ESG ratings can be partially explained by a form of 'rating addiction' proper to the financial world. Because it represents an easy-to-use, accessible, and generally independent source of information, ratings have been well-established tools in the capital markets (especially in the domain of evaluating credit worthiness). A situation which can explain a form of inertia among the investors that prefer to work with a format they trust and know.

However, in the case of ESG performance, the choice of ratings cannot be limited to this single factor. As described by several papers and reports, the term ESG remains a complex and volatile

notion that lacks standardization (Escrig-Olmedo et al., 2010; Mooij, 2017; Windolph, 2011). Indeed, ESG performance can be assessed using a very wide spectrum of KPIs or metrics, e.g., Refinitiv Rating includes more than 500 different measures (Refinitiv, 2021) while MSCI will be focusing on monitoring more than 3400 media sources daily (Moen, n.d.). Moreover, the choice of metrics to include in the assessment of a company can also vary according to the company or the sector that is evaluated, thereby creating an additional layer of complexity for the investor (Gyönyörová et al., 2021; Khan et al., 2016). Given this high level of complexity, it is not surprising that investors have chosen for ratings to integrate ESG information in their decision process. Under this format, the sustainable performance of a company can be summarized (maybe too much (LaBella et al., 2019)) in a simple and understandable way (Gyönyörová et al., 2021; Windolph, 2011). Besides, some investors highlight the fact that using scores facilitates the comparison process between various companies and investments (Ioannou & Serafeim, 2015; Windolph, 2011), i.e., it becomes easier to identify top performers in each industry.

Note that because investors have neither the time nor the resources to gather, process and analyze the various inputs required to build an ESG score, most of them will turn to external providers (Doyle, 2018; Wong & Petroy, 2020). Hence, ESG rating agencies are playing a crucial role in the promotion and growth of SRI: The more credible and qualitative information providers will get, the more money people will start to invest in sustainable companies (Escrig-Olmedo et al., 2010). Besides, financial investors are not the only consumers of ESG ratings, according to Saadaoui & Soobaroyen (2018) and Scalet & Kelly (2010) we should also include companies, wishing to measure their social impact, scholars, studying the ESG performance of companies, consumers, governments, etc. (see Appendix III). The impact of a more qualitative ESG rating market is therefore much more significant than a simple impact on investors.

1.3. Definition of ESG rating

To put it simply, an ESG rating, or score, should be seen as a tool that allows investment professionals to evaluate and compare the sustainability performance of various companies (López-Arceiz et al., 2020). The rating is built by evaluating each company "based on a comparative assessment of their quality, standard or performance on environmental, social or governance (ESG) issues." (Gyönyörová et al., 2021, p. 2; Wong & Petroy, 2020, p. 11), hence by using a set of metrics and KPI's that summarize non-financial performance. Besides, the literature also mentions that the process leading to the delivery of a rating generally includes

the following set of elements: An evaluation model – Data gathering and cleaning methodology – The definition of standards of assessment (Scalet & Kelly, 2010; Semenova & Hassel, 2015). In absence of a common evaluation framework, those three elements will be an important source of conflict among the different rating providers on which data to include in the model, the weight to give to each metrics, the source of information, the evaluation model, etc. (Doyle, 2018; LaBella et al., 2019; Mayor, 2019; Mooij, 2017; Wong & Petroy, 2020).

1.4. ESG rating providers – Who build the ratings?

An ESG rating provider consist of any organization that "rates or assesses corporations according to a standard of social and environmental performance that is at least in part based on non-financial data" (Scalet & Kelly, 2010, p. 70). This broad definition can therefore encapsulate a large spectrum of actors active in the ESG world, including the investors themselves. There actually exist investment professionals that decided to use their in-house capabilities to build custom ESG ratings. However, as it is stated in the *Rate the Raters* report from 2020, creating a reliable and qualitative sustainability rating for every company from one's investment universe require thousands of hours and a lot of manpower to collect, process and analyze the data (Wong & Petroy, 2020). Even for large organizations, it is therefore a common practice to delegate part or totality of this task to independent third parties (Amel-Zadeh & Serafeim, 2017; Saadaoui & Soobaroyen, 2018). Hence, the biggest producers of ESG ratings, and ESG information, are currently the so-called ESG rating agencies. Large ESG providers include MSCI, Sustainalytics, Bloomberg, Thomson Reuters (Refinitiv), and RobecoSAM. Also, traditional ratings agencies such as Moody's, S&P and Fitch now provide forms of ESG ratings (Boffo & Patalano, 2020).

2. How to measure ESG performance - ESG Rating methodologies

As explained in the last section, ESG ratings are aiming at approximating the corporate sustainability performance of various companies by evaluating the number and quality of the ESG practices they have implemented. Because of the high complexity and cost of the evaluation process, construction of ratings is usually conducted by independent private companies called ESG rating agencies. Contrary to financial ratings, there exist no common set of rules or consensus on the way those agencies are supposed to evaluate the ESG performance of a corporation. Each rating provider has thus developed its own customary assessment methodology for differentiation purposes, leading to regular divergences among agencies (Billio et al., 2021; Boffo & Patalano, 2020; Halbritter & Dorfleitner, 2015).

However, we can observe that agencies are often following similar steps, or structure, when measuring non-financial performance. By definition, a rating is obtained by collecting raw data from various sources, analyzing and structuring those inputs before finally assembling it into one single score (Boffo & Patalano, 2020; Saadaoui & Soobaroyen, 2018; Semenova & Hassel, 2015). Based on existing literature and an analysis of the documentations of several rating agencies, it can be concluded that ESG rating methodologies are likely to rely on the following process (Billio et al., 2021; Blank et al., 2016; Escrig-Olmedo et al., 2010; Halbritter & Dorfleitner, 2015; Moen, n.d.; Refinitiv, 2021; Saadaoui & Soobaroyen, 2018; Semenova & Hassel, 2015): Collecting and Assembling Raw Data – Defining the KPI's, measures of the different ESG criteria – Defining the importance to give to each metrics and criteria, i.e., define a materiality matrix

2.1. Data Sources

The first part of the ESG rating process consist in gathering the raw data from which analysts from Refinitiv, Bloomberg or MSCI will derive sustainability metrics. Rating agencies tend to make use of a very wide spectrum of sources, most of the time publicly available but also by directly contacting the evaluated company. According to Blank et al. (2016), which describe the full rating methodology implemented by Refinitiv, this step probably represents the most labor-intensive action of all the process as it generally requires lot of time and manpower to find, collect, verify and analyze those various sources. Though using independent sources of information are a good way to improve the credibility of agencies, it has been noted by Mooij (2017) that a majority of them are still relying at least partially on company-controlled sources of information. A company's rating is therefore often dependent on its degree of disclosure (LaBella et al., 2019). This phenomenon is further exacerbated by the fact that an information which is unavailable is often considered as a negative data hidden by the company and will therefore penalize the company's score (LaBella et al., 2019).

A non-exhaustive list of public sources used by agencies includes company official publications, i.e., annual reports but also CSR reports, media releases, governments and other third-party databases and reports, independent investigations, etc. (Escrig-Olmedo et al., 2019; Gyönyörová et al., 2021; Moen, n.d.; Mooij, 2017; Refinitiv, 2021; Scalet & Kelly, 2010; Semenova & Hassel, 2015). In general, the evaluated company will also be directly involved in the data collection process. This participation can come down to a simple verification or review of their score, or in most cases an on-hand involvement through the filling of a questionnaire,

contact with management or even on-site visits (Escrig-Olmedo et al., 2019; Gyönyörová et al., 2021; Mooij, 2017; Scalet & Kelly, 2010).

2.2. Evaluation model – Choice of Metrics and Criteria

The two following stages correspond to what Semenova & Hassel (2015) call the definition of an evaluation model: The definition of (1) the relevance and (2) the importance to give to ESG metrics and criteria. Whatever the rating approach, in the first step, the rating agency decides how the high-level categories, i.e., the Environmental, Social and Governance pillar, needs to be evaluated. Hence, it defines what is the set of positive criteria that allow to accurately assess each category, and what are the exclusionary practices that should have a negative impact on the rating (Boffo & Patalano, 2020; Escrig-Olmedo et al., 2010; Saadaoui & Soobaroyen, 2018). In this section, we will call this decision the definition of a criteria-mix, i.e., a combination of both positive and negative ESG criteria.

2.2.1. Negative Criteria – Exclusionary Approach

Most agencies consider a set of so-called negative criteria to assess the ESG performance of a company, or simply to keep investors and other users informed of its controversial activities. A negative criterion is therefore an indicator of exclusionary activities generally encapsulate those controversial activities or sectors in which a corporation could be involved such as Tobacco, Alcohol, etc. The objective of implementing those negative criteria is to allow for the agency to exclude, or strongly penalize, companies that are involved in sectors that are considered harmful for the environment, people's health, fair competition, etc. In most cases, the agency will limit its exclusionary approach to the gathering of information, leaving the final decision to their clients (Saadaoui & Soobaroyen, 2018). For instance, Refinitiv's database contains both a ESG score and a controversy score that is intended to help users identify companies active in unsustainable activities (Refinitiv, 2021).

Typical examples of controversial activities that are evaluated by most rating agencies have been studied by Escrig-Olmedo et al. (2010). It includes, among other, companies being active in the Tobacco or alcohol sectors, but also pornography, nuclear power, or military weapons. The full results gathered by their study for major rating agencies and indices can be seen on Appendix IV. According to a more recent studies, those criteria have not changed much over the last ten years (Boffo & Patalano, 2020; Saadaoui & Soobaroyen, 2018) though an increasing

number of agencies tend to adjust their ratings based on the degree of involvement rather than based on the simple presence of an exclusionary practice (Saadaoui & Soobaroyen, 2018).

2.2.2. Positive criteria – ESG themes

Regarding positive criteria, we can take the example of Refinitiv's Environmental pillar as a starting point. For this category, the agency believes relevant criteria should include *Resource Use*, *Emissions,* and *Innovation*. The *Emissions* criteria covers practices implemented by a company to address environmental issues linked to CO2 emissions, biodiversity, or waste management (Refinitiv, 2021). From a governance point of view, classical themes evaluated by rating agencies are the board diversity, the board independence, or the corporate ethics (Boffo & Patalano, 2020). However, the choice of criteria is not standardized across the ESG rating market each agency being free to define the criteria she considers as the most relevant to the ESG performance of a company. Some will rely solely on non-financial categories, but others will prefer to combine both financial and non-financial criteria to deliver measures of long-term value (Escrig-Olmedo et al., 2019; Scalet & Kelly, 2010)

As for negative criteria, scholars have studied the most common theme used by rating agencies in their evaluation model (see Appendix IV) (Escrig-Olmedo et al., 2010). Omnipresent criteria include for instance Environmental management, Human capital development or Community relations. Though some factors are common to almost all rating providers and indices (Escrig-Olmedo et al., 2010; Saadaoui & Soobaroyen, 2018), there is generally no uniformity on the choice and terminology of criteria to use. For example, both Bloomberg and MSCI are studying the pollution and waste management of companies. However, MSCI consider those as part of one single criterion while Bloomberg splits it in two (Boffo & Patalano, 2020)

Besides, it is noteworthy to mention that the criteria investigated by rating agencies tend to evolve with the regulatory and conceptual framework surrounding sustainability (Escrig-Olmedo et al., 2019; Saadaoui & Soobaroyen, 2018). In the field of environmental issues, Escrig-Olmedo et al. (2019) have observed that since the adoption and definition of the SDG's, reduction of inequalities and economic growth have become central pieces of the Social pillar as they are keys to sustainable development. This illustrates the fact that choices of both positive and negative criteria and their metrics are not coming out of nowhere but are generally based on international standards and frameworks on ESG whose definitions are building blocks of each agency's evaluation model (Escrig-Olmedo et al., 2010; Scalet & Kelly, 2010; Semenova & Hassel, 2015) (see details in Appendix IV).

2.3. Evaluation Model – Importance of the Metrics and Criteria

Even if two agencies would integrate the exact same set of criteria and metrics into their respective rating process, they could still deliver different scores for a same company. This is due to the last step of the ESG rating evaluation model, namely the weighting stage (Escrig-Olmedo et al., 2010, 2019; Madison & Schiehll, 2021; Saadaoui & Soobaroyen, 2018; Scalet & Kelly, 2010). Once an agency has gathered the required data for each of the relevant ESG metrics, a score will be generated by aggregating the different metrics into a set of criteria scores, and further aggregating those criteria scores into one single ESG rating (Moen, n.d.; Refinitiv, 2021). In both cases, the aggregation is done through a weighted sum where the weights given to each metric or criterion is dependent on the importance it has to the ESG performance of the company. This implies an important share of subjectivity which therefore leads to an important divergence among rating providers (Escrig-Olmedo et al., 2010, 2019; Scalet & Kelly, 2010). To be more precise, during this step the ESG rating agency defines a materiality matrix, a notion derived from financial accounting. To put it simply, an issue is said to be material if those are ESG factors that can have an impact on the financial value of the investment (LaBella et al., 2019; Madison & Schiehll, 2021; Valor & De la Cuesta, 2013).

Generally, rating agencies are hardly ever keen to disclose the entirety of their custom materiality map. According to Saadaoui & Soobaroyen (2018), rating providers are refusing to reveal their weights because they want to protect their intellectual property and thereby keep a potential competitive advantage. Therefore, it can be considered that the building of the materiality map is the core difference between the various rating processes of ESG rating agencies and therefore embodies their specific know-how in measuring ESG performance.

As for the choice of metrics and criteria, there exist international frameworks to guide rating agencies into the building of their map. As of today, the biggest contributor is the SASB which has developed a disclosure framework based on the materiality of ESG issues depending on the sector and industry of a company (Madison & Schiehll, 2021). Agencies will also tend to adapt their weights in function of the company, i.e., create evaluation models on a case by case, and the specific needs of the stakeholders (LaBella et al., 2019).

2.4. Criticism and recent trends

Although ESG rating agencies and their ratings are currently a cornerstone of the ESG ecosystem, they are not perfect and are often criticized by the investor's community (Doyle,

2018; Wong & Petroy, 2020). Furthermore, the question of the ESG scores quality has been the subject of numerous studies (Scalet & Kelly, 2010). From those studies it can be concluded that ESG ratings should be improved in the fields of standardization, transparency, credibility, independence, or biases (Escrig-Olmedo et al., 2019; Windolph, 2011).

2.4.1. Standardization

In the previous section, it has been mentioned that for most elements of the evaluation process, rating agencies could choose for a customary approach. Each agency defines its own sources of information, metrics, and materiality matrix. As a result, several papers have noted a high divergence between ESG ratings. For instance, Halbritter & Dorfleitner (2015) has found a correlation coefficient ranging between 0.2 and 0.6 for three major ratings that are Refinitiv, Bloomberg and MSCI. This high level of variability is often mentioned by investors as one of the major barriers to ESG investing (Doyle, 2018; Wong & Petroy, 2020). The resulting lack of comparability but also the possibility for companies to pick the most favorable rating are making the ESG scores almost obsolete for investors (Amel-Zadeh & Serafeim, 2017; Chatterji et al., 2016; Crespi & Migliavacca, 2020; Scalet & Kelly, 2010). In this context of high disagreement among ESG rating agencies, it is therefore of high importance for rating users to always deep dive into agency's methodologies to identify which are factors driving the rating, to make an educated decision (Boffo & Patalano, 2020; Halbritter & Dorfleitner, 2015).

The reasons behind this lack of standardization are straightforward. If ratings differ so widely it is because there are no common standards on how to define and measure sustainability or ESG (Escrig-Olmedo et al., 2010, 2019; LaBella et al., 2019; Mayor, 2019; Scalet & Kelly, 2010; Serafeim, 2021; Windolph, 2011). As stated by Berg et al. (2019), rating agencies do not have rules on what needs to be measured, on how it needs to be measured and on how it should be prioritized, as there currently exist plenty of available frameworks (SASB, GRI, IIRC, TCFD, etc.) that rating agencies can pick to design their own ESG rating process.

One of the keys to improving ESG ratings, and by extension one of the keys to the further development of the SRI movement, will therefore lie in the ability of the market to develop a unified ESG framework and standard. Note that this is a work in progress in the literature but also in the private sphere. For instance, the SASB and the GRI frameworks are currently in the process of developing guidelines on how to commonly use their respective frameworks ('SASB Standards & Other ESG Frameworks', n.d.). However, it is worth mentioning that in the opinion of some investors, it could be useful to keep a bit of variability across rating agencies (Wong

& Petroy, 2020). As sustainability remains a volatile notion, it would be surprising to be able to summarize it in one single score. Hence, relying on multiple ratings based on different visions of sustainability could give a better vision of what the true ESG performance of a company is.

2.4.2. Transparency

Another factor leading to difficulties for comparing ESG ratings is the lack of transparency (Escrig-Olmedo et al., 2019; Wong & Petroy, 2020). Most ESG rating providers will refuse to fully disclose their assessment methodology, i.e., their sources, metrics, and materiality matrix (Escrig-Olmedo et al., 2010, 2019; Windolph, 2011). For investors, this situation does not only raise issues in the field of comparability but also limits its ability to understand what the rating stands for (Semenova & Hassel, 2015). Without knowing what is exactly measured by the ESG score, an investor would not be able to determine whether a company is responsible according to its own principles. Increasing transparency is therefore a second way to improve the quality and readability of ESG ratings.

The reluctance of agencies to reveal their rating recipe is generally explained by their willingness to maintain their position in a competitive market situation. As mentioned in the previous section, in the ESG rating sector the evaluation model is the main source of differentiation for agencies (Saadaoui & Soobaroyen, 2018). It is therefore not surprising that most of the rating providers tend to protect their intellectual property. However, Serafeim (2021) notes that agencies would be more likely to agree on the sustainability performance of company if there was more disclosure, meaning that a higher transparency from the side of companies could also solve the standardization issue of the ESG rating market.

2.4.3. Credibility, Biases, and Independence

The question of credibility, biases, and independence in the ESG rating industry is fundamental as it can determine whether those ratings are reliable or not. If those cannot be trusted as good proxies of ESG performance, then sustainable investment would not be able favor sustainable development (Busch et al., 2016).

The credibility of ESG scores is already put in question by the issues of transparency and standardization. It is complicated for investors to trust the rating of a given company if every agency has a different opinion on the value it should have. Another aspect that tends to affect the credibility of ESG scores is the data availability. As mentioned by LaBella et al. (2019), a large share of ESG information finds its origin from public company documentation. ESG

rating agencies are therefore highly dependent on the willingness of assessed company to disclose relevant information. Yet, those sources of information are often unaudited and therefore less credible than independent sources (Doyle, 2018; Windolph, 2011).

Besides, investors are also criticizing ratings because of the presence of biases, making the rating less accurate (Doyle, 2018; Wong & Petroy, 2020). Typical biases includes (Doyle, 2018; Gyönyörová et al., 2021; Hawley, 2017; Mooij, 2017; Windolph, 2011): *Size bias* – As ESG scores are generally based on publicly available information, large corporations, which tend to have a higher degree of disclosure, are awarder better ESG rates. *Geography Bias* – As for the size bias, the geography bias can be explained by the degree of disclosure. Companies located in regions with high disclosure standards tend to have higher ESG scores. *Industry Bias* – ESG rating agencies usually develop a materiality map at an industry level. However, their vision of what is an industry is a simplification, which can by consequence lead to over- or underestimating the score of a company.

2.5. Conclusion – AI as a solution to ESG ratings critics

Given the numerous shortcomings listed above, it is no surprises that rating agencies, as well as investors, have started to investigate innovative ways of measuring ESG performance. Among the existing alternatives, investors believe that implementing Big Data and Artificial Intelligence in the evaluation process could deliver interesting returns especially in terms of biases (Hawley, 2017; Liberato Ferrao, 2019; Wong & Petroy, 2020). By combining machine learning and text mining techniques it seems that rating agencies might have the possibility to overcome some of their current challenges (credibility, transparency, biases, etc.). Though most providers have started to include those tools in their methodology, both research and practice are in their infancy. It is therefore interesting to look at how Text Mining can be concretely used in the world of sustainable finance.

Chapter 2: Text Mining and Sustainable Finance

1. Defining Text Mining – Set of Common Techniques

Since they were created by human beings, a large majority of the data available in our modern society are presented in an unstructured format (Shah et al., 2020), i.e., "information that is not arranged according to a pre-set data model or schema" (*What Is Unstructured Data?*, n.d.). Therefore, this valuable information cannot be used directly as an input for traditional computer programs. To generate value, or improve decision-making based on those contents, many companies, and researchers have thus started to invest in the development of methods and tools to (1) process large amounts of information, i.e., Big Data analysis, and (2) extract, or transform, unstructured information into computer-usable data, i.e., Text Mining techniques (Aaryan et al., 2020; Kumar & Ravi, 2016; Pejić Bach et al., 2019). Hence, interest for text mining has exploded in recent years which has been reflected by the development of a high number of applications for several sectors, including the financial markets (Aaryan et al., 2020).

1.1. Definition of Text Mining

Text mining and analytics are a branch of computer science that aims at extracting, processing, and analyzing data available under a textual format (Delen & Crossland, 2008; Miner et al., 2012). Starting from a semi-structured or unstructured data format, e.g., a news releases, a text mining algorithm will programmatically create high-quality information usable for classical statistical and data mining techniques (Aaryan et al., 2020; Kumar & Ravi, 2016).

The most basic version of text mining is the manual reading and analysis of a text by a human being (Aaryan et al., 2020; Miner et al., 2012). For instance, the writing of a paper abstract could be qualified of text mining and analysis: After the reading of a text, the analyst will extract the most relevant information and summarize its findings into an abstract. The machine-led version of text mining does basically the same, using data mining to extract knowledge from a given text. And thanks to the recent developments of artificial intelligence and machine learning automatic applications of text mining have appeared to exceed human capacities in terms of time- and cost-efficiency (Aaryan et al., 2020; Dumont, 2017; Pathan et al., 2020).

However, before being able to make any analysis using a computer, textual data needs to be converted to an exploitable format. Basically, a text mining tool works by going through two separates sub-operation (Kumar & Ravi, 2016; Liew et al., 2014; Pejić Bach et al., 2019).

Firstly, a Text Preprocessing step by which the raw data will be transformed to a structured format. Preprocessing techniques generally consist in converting the text into a word's frequency matrix, i.e., a vectorial representation of the text. The second step will then consist in the above-mentioned Knowledge Extraction process which corresponds to the application of the corpus of data mining tools to the preprocessed data (Dörre et al., 1999; Shahi et al., 2014).

1.2. Text Preprocessing – Common Techniques

Text preprocessing is the first step of any text mining or text analytics application. It consists of various transformations designed to overcome the many challenges of automatically processing textual data, i.e., large volumes of unstructured data. The main objective of pre-processing a text is always to convert raw data into a low dimensional structured format, usually an integer vector. As it is the first stage of any text mining tool, it has been seen that the quality of the results is dependent on the quality of the preprocessing (Uysal & Gunal, 2014).

1.1.1. Text Representation Model

The most basic form of structured representation of a text, and thereby the starting point of a large majority of text mining programs and applications, is to convert the text to a word frequency matrix, i.e., a vector where each element consists of the number of occurrences of a given word in the analyzed text. This vector representation of textual content is best known as the vector-space or bag-of-words model (Loughran & Mcdonald, 2016; Miner et al., 2012).

In this model, the data analyst assumes that the order of the words has no importance to the meaning of the text, i.e., most information is carried by the independent meaning of each word (Miner et al., 2012; Zhang et al., 2010). Algorithms using the bag-of-words model therefore assume that the semantic, the meaning of a word given the surrounding context, can be inferred from the syntax, the form, structure of a text. According to Miner et al. (2012), this hypothesis does not cause any issue when trying to classify or retrieve information from a text ("Though relying on syntactic information alone, these approaches work because documents that share many keywords are often on the same topic" (Miner et al., 2012, p. 45)) Yet, it can be problematic when dealing with application where semantic is more important. When semantics matters, other models such as word embeddings or BERT-models can offer better results in representing text into a structured format (Mikolov et al., 2013; Miner et al., 2012; Nugent et al., 2020).

1.1.2. Pre-processing steps

Representing a text document as a bag-of-words is achieved by following a set of transformation steps, generally associated to natural language processing techniques (Miner et al., 2012). The main objective of this pre-processing is to create a structured format of data. Yet it is also used to reduce the dimension of the text, hence, to handle the presence of noise in the raw data (Raghupathi et al., 2020). In a vector-space model, noise is mainly caused by the presence of characters and words often used in natural language but who don't carry any relevant information when taken separately, e.g., pronouns or punctuation. (A concrete illustration of pre-processing can be seen in Appendix VIII or on **Figure 1**).

The first of the pre-processing steps is called the *tokenization*, i.e., the transformation of text into a vector of tokens (Groß-Klußmann et al., 2019; Miner et al., 2012; Raghupathi et al., 2020). Tokens are units of text, i.e., words, sentences, or paragraphs, that will be considered as features of the text in a vector representation. A common technique of tokenization is to convert a continuous text into a vector of words by splitting it on white spaces and punctuation, i.e., your token delimiters (Miner et al., 2012).

After tokenizing your text, a stopping algorithm will generally be applied as a way to retrieve stop words from the text (Groß-Klußmann et al., 2019; Raghupathi et al., 2020). Stop words are commonly used terms that don't carry any relevant syntactic information. Typical examples of stop words in the English language are words such as *these, this, there,* etc. Removing those pronouns from the text allows for a reduction of the storage space and an increase in the speed of process (Miner et al., 2012).

Another important step when pre-processing unstructured data is called stemming, or lemmatization. To reduce the dimensionality of the vector-space model, most text mining applications will decide to consider words having a "common root" as being identical. Removing suffixes, prefixes, and plural forms, generally defined as stemming, is therefore a desirable transformation when working with textual data (Miner et al., 2012). One of the most popular stemmers (Groß-Klußmann et al., 2019) has been developed by Porter to efficiently remove suffixes (Porter, 2006).

After all the previous steps, and a potential lowercasing and a removal of punctuation (Groß-Klußmann et al., 2019; Raghupathi et al., 2020), the text should be ready for its last, and most

important, transformation: The "vectorization" of the text. Miner et al. (2012) lists three existing options for converting a set of tokens into a usable vector of data.

The binary and integer count vectors are similar techniques as they both assign an integer value to each token. In the binary vector this value is either 1 or 0 depending on the presence or not of the word in the text, while in the integer count, the value of the feature is always equal to the number of occurrences of the words in the text. The third option is called the term frequency-inverse document frequency, or TF-IDF, which constructs a float-value weighted vector. Behind the TF-IDF lies the intuition that words that are frequent in a document should only receive high weight if they are not common terms. Mathematically this is translated by weights which are both dependent on the term frequency, number of occurrences of a token in my text, and the inverse document frequency, 1 over the term frequency across all documents of my corpus (Miner et al., 2012; Raghupathi et al., 2020). Note that in Text Mining, the notion of frequency is an abuse of language since in reality it designates the number of occurrences of a word. Throughout this thesis, whenever we use the term frequency, we will in fact refer to the number of times a word is repeated in a text, thus following the terminology of Text Mining.

1.3. Knowledge Extraction – Common techniques

Once a text document, a news release or a tweet has been transformed into a bag-of-words, it becomes available for any application based on machine-led analysis. This corresponds to the second step of the text mining process: the Knowledge Extraction. Typical data mining tasks that can be applied on text includes, document classification, concept extraction, document clustering, web mining, information extraction, natural language processing (NLP), or sentiment analysis (Aaryan et al., 2020; Kumar & Ravi, 2016; Miner et al., 2012). Those different areas have each their own specificities, objectives, and models, but they are also often overlapping and used jointly (Miner et al., 2012). For instance, sentiment analysis can be used for classification purposes as you can use the sentiment score of a text to categorize it between positive and negative sentiment (Aaryan et al., 2020).

For purposes of brevity and clarity, we will limit our focus to two frequently used applications of text mining in finance, namely text classification and sentiment analysis.

1.3.1. Text Classification

Grouping and categorizing data based on a vector of features is one the of most common application of data mining and machine learning in general. When the data to classify are text

documents and the feature vector is a bag-of-words, then this specific classification task is called text classification (Delen & Crossland, 2008; Miner et al., 2012; Shahi et al., 2014). Typical tasks of text classification include document organization, text filtering, spam detection, or patent application categorization (Shahi et al., 2014).

The process of classifying a text is very similar to any classification task. Starting from a set of labelled feature vectors, e.g., vector-space model extracted from the text in the pre-processing stage, a classifier is selected based on accuracy to build the most efficient classification model (Aaryan et al., 2020; Miner et al., 2012). The available classification models are the same as for regular categorization tools. For instance, Logistic Regressions, Support Vector Machine and Naïve Bayes are often mentioned as being the best options for the classification of documents (Aaryan et al., 2020; Miner et al., 2012; Shahi et al., 2014). Yet, one is free to use any existing classifier (SVM, Neural Networks, Linear Discriminant Analysis, Latent Semantic Indexing, etc.) if this choice can help improve the performance of the model (Delen & Crossland, 2008; Groß-Klußmann et al., 2019; Miner et al., 2012).

However, given the high dimensionality of textual data, text classification will usually require an additional step to manage the curse of dimensionality, i.e., models overfitting the data because of a too large number of features and therefore having poor prediction capabilities (Kumar & Ravi, 2016; Miner et al., 2012). Therefore, before testing the performance of a classifier, an intermediate step of dimensionality reduction will generally be applied. This can consist of classical techniques of feature extraction such as Principal Component Analysis, Singular Value Decomposition, or Linear Discriminant analysis (Delen & Crossland, 2008; Groß-Klußmann et al., 2019), or be done through more sophisticated models specific to text analysis. For instance, topic models are popular ways of reducing the size of data when working with text (Groß-Klußmann et al., 2019). The basic assumption of topic models is that a text can be summarized as a combination of topics and not words. Therefore, to classify a text one should first aggregate the words into their respective topics, usually through a weighted sum, and then apply an appropriate classifier. The building of topics can be done manually by identifying the important subjects and terms of the document, but it will generally be done by applying automatic processes. A popular method in the world of text mining is the Latent Dirichlet Allocation model (LDA), a Bayesian model that identifies the various topics of a corpus of documents based on the used lexicon.

Another approach for reducing the dimensionality is the selection of the terms with the highest discriminatory power. There exist several feature selection methods in the field of text mining, including the ranking of words based on information gain, entropy, or on the results of a Chi square statistic (Groß-Klußmann et al., 2019; Kumar & Ravi, 2016; Ravi & Ravi, 2015).

1.3.2. Natural Language Processing and Sentiment Analysis

Sentiment analysis is one among the many text mining practices whose main objective is to extract or summarize the "meaning" of a text, i.e., its global message or topic (Delen & Crossland, 2008; Miner et al., 2012). To be more precise in sentiment analysis the data analyst tries to identify or measure the sentiment or opinion of the text. In general, the considered sentiment is a positive, or negative message, but this can be extended to any polarity detection in a text extract.

According to Cambria (2016), there can be two main objectives of applying sentiment analysis on textual data: Emotion recognition, "the extraction of a set of emotion labels" (Aaryan et al., 2020, p. 4), and Detection of polarity (Ravi & Ravi, 2015). Polarity detection corresponds to the measurement of the negative or positive tonality of a specific document (Dumont, 2017) and will generally be used as a first step to a global classification approach (Aaryan et al., 2020).

In most papers, the sentiment of a document is summarized by a measure computed based on the frequency of positive and negative terms in the text. For example, Groß-Klußmann et al. (2019) used the normalized difference of positive and negative words as the sentiment score of stock markets related tweets: $S = (\#\text{positive} - \#\text{negative})/(\#\text{positive} + \#\text{negative})$

Therefore, most of the time an important share of sentiment analysis will consist in defining what positive and negative terms concretely are. To address this issue, there exist two different approaches, and by extension two different approaches of sentiment analysis: Lexicon-Based and Machine Learning-Based (Aaryan et al., 2020; Ravi & Ravi, 2015).

Machine Learning-based sentiment analysis labels terms as positive or negative based on a classification model such as Naïve Bayes or Logistic Regression. The main advantage of this approach is to offer an automatic labelling of terms, i.e., it reduces human subjectivity to its lowest level (Aaryan et al., 2020; Ravi & Ravi, 2015). Yet, to build the model, one will need to find an appropriate training set for which words are already labeled as positive or negative (Al-

Natour & Turetken, 2020) or rely on an unsupervised classification model, e.g., a neural network (Ravi & Ravi, 2015).

An alternative approach to determine the polarity of a document is called lexicon-based sentiment analysis. This text mining method relies on dictionary, or lexicons, of positive and negative terms to perform the exact same process as Machine Learning based sentiment analysis (Aaryan et al., 2020; Baier et al., 2020; Groß-Klußmann et al., 2019; Loughran & Mcdonald, 2016; Miner et al., 2012). In this approach, the polarization is simply deducted from the presence (or absence) of the word in a list of positive (or negative) terms. Despite its simplicity, Taboada et al. (2011) have observed that this word list approach was still a relatively robust way of mining opinions. Provided that the analyst has an appropriate list of relevant polarity words, then using lexicon-based sentiment analysis has the advantage of combining simplicity, efficiency (even in large samples) and objectivity (Baier et al., 2020; Groß-Klußmann et al., 2019; Loughran & Mcdonald, 2016).

Finding an appropriate dictionary remains the biggest challenge of dictionary-based opinion mining, hence a certain number of papers have been focusing solely on the development of appropriate lexicons for different sectors or purposes (Baier et al., 2020; Lewis & Young, 2019; Loughran & Mcdonald, 2016). A very common dictionary in financial literature is the Harvard IV-4 list of positive and negative words used, among others, by Heston & Sinha (2017) to evaluate the relationship between positive Thomson-Reuters news articles and stocks returns. However, this lexicon was originally created for research in the field of opinion mining for sociology and psychology (Loughran & Mcdonald, 2016). For more finance related dictionaries, one could refer to the word list built by Loughran & Mcdonald (2011) based on their analysis of the relationship between 10-K fillings, stock returns, volatility, etc. (Lewis & Young, 2019; Loughran & Mcdonald, 2016).

2. Application of text mining in the world of finance

As of today, it is estimated that most of the financial information is reported under a textual format, e.g., reports directly published by the company, news releases, social media content, etc. (Dumont, 2017; Kuh, 2019; Lewis & Young, 2019; Shahi et al., 2014). As it is displayed under an unstructured format, the traditional way of processing such information, i.e., relying on an army of financial analysts, turns out to be time consuming, costly, and at risk of missing

key information or latent features (Aaryan et al., 2020; Lewis & Young, 2019; Shahi et al., 2014; Wei et al., 2019).

Finance is therefore one of the sectors that has been intensively investigating ways of using text mining and text analytics as a leverage for value (Aaryan et al., 2020; Dumont, 2017; Lewis & Young, 2019; Pejić Bach et al., 2019). The exponential growth of scientific papers but also FinTech's focusing on the development of financial applications of text mining is a good illustration of the growing interest of the financial community for those innovative technologies (Pejić Bach et al., 2019; Zavolokina et al., 2016). According to the existing literature on the topic, the specificities of text mining, i.e., the ability to summarize, or extract key information from a large corpus of data, are generally supporting decision-making in the area of stock predictions, reports and risk analysis, and more recently sustainable investing (Aaryan et al., 2020; Kumar & Ravi, 2016; Pejić Bach et al., 2019).

Regarding specific text mining techniques or methodologies, it seems that sentiment analysis and dictionary-based approach remain two of the most popular tools for the applications cited above (Aaryan et al., 2020; Groß-Klußmann et al., 2019; Kumar & Ravi, 2016; Pejić Bach et al., 2019). For instance, sentiment analysis can be used as an additional indicator of risks in the banking sector as investigated by Nopp & Hanbury (2015). In finance, sentiment analysis owes its success to proven capabilities for forecasting future financial performance (Hajek et al., 2014; Song et al., 2018). Another popular set of valuable techniques for financial applications, especially when analyzing trends in social media, includes topic and keywords extraction, for instance with Latent Dirichlet Allocation or other topic models (Pejić Bach et al., 2019). In both situations the focus is put on the ability of text mining to summarize a big data setting into a limited number of outputs.

2.1. Forecasting Prices

The ability of accurately predicting the future value of a stock, an exchange rate or a credit rating is obviously an asset for a financial institution. Hence, a lot of practitioners and scholars have started investigating whether text analytics could improve forecasting capabilities or not (Fisher et al., 2016). The intuition behind that research relies on the idea that the financial markets are, at least partially, reacting on information conveyed by company's reports, news releases, social media trends, etc., i.e., unstructured data (Pejić Bach et al., 2019). So, to forecast the next movement of the markets, one would have to quickly process this enormous amount

of textual information, hence the interest in text mining (Fisher et al., 2016; Kuh, 2019; Lewis & Young, 2019).

Most of the forecasting models built in the literature are based on methods derived from sentiment analysis, both Lexicon- and Machine Learning-based (Aaryan et al., 2020; Groß-Klußmann et al., 2019; Kumar & Ravi, 2016; Pejić Bach et al., 2019). Their objective is therefore to automatically label a report, social media content, or news release in function of the impact of this same content on historic price movement. For instance, a tweet that shares the same features than a tweet that created an upward movement in the past should be categorized as positive (Groß-Klußmann et al., 2019). One concrete example of Lexicon-based sentiment analysis for stock prices prediction has been studied by Groß-Klußmann et al. (2019). Their main purpose was to evaluate the capacity of positive and negative tweets to forecast the daily variations of regional stock index futures. Based on a sample of more than 100 million finance-related tweets they constructed a daily sentiment score based on Harvard-IV 4 words list on the topic economy. They concluded that this Twitter sentiment measure could result in a profitable investment strategy on the stock market index futures, thereby illustrating the utility of text mining for prediction purposes. Another example of intra-day financial forecasting supported by text mining was conducted by Vijayan and Potey (2016, cited by Aaryan et al., 2020). In their study, a decision algorithm was used to identify positive and negative news headlines, i.e., they performed a Machine Learning-based sentiment analysis. The authors found out that stock prices could be more accurately forecast when combining a technical analysis with the sentiment measure of news headlines.

In terms of text mining algorithms and sources of information usually used for predicting price variations, Kumar & Ravi (2016) have observed that most of the reviewed papers had chosen for a Machine Learning approach of sentiment analysis with a preference for Naïve Bayes and Support Vector Machine as classifiers (based on a sample of 89 papers published between 2001 and 2016). One of the reasons for such observation being obviously the proven robustness of both methods in presence of big data settings (Miner et al., 2012). Besides, from the literature review, it seems that the preferred sources of qualitative information for this specific application are financial news and tweets.

2.2. *Fraud and Bankruptcy Detection*[1]

Fraud and Bankruptcy is a major concern for several actors active in the financial sector, e.g., companies, banks, regulators. (Gregor et al., 2020; Kumar & Ravi, 2016; Pejić Bach et al., 2019; Shirata et al., 2011). In traditional approaches, the detection of irregularities can be done based on the analysis of numerical values and financial ratios (Kumar & Ravi, 2016; Zavolokina et al., 2016). However, with the development of text mining and artificial intelligence, more and more papers have started to look for ways to integrate qualitative information in the fraud detection process.

In general, the fraud or bankruptcy detection consists in a classification task, i.e., the identification of specific fraudulent features by analyzing the data with a text mining algorithm (Pejić Bach et al., 2019). Therefore, most applications of text mining in this field are relying on classical text classification techniques with a preference for Support Vector Machine, Naïve Bayes, or Decision Trees (Kumar & Ravi, 2016; Pejić Bach et al., 2019). For instance, Appavu et al. (2009) used a Decision Tree classification algorithm to look for the presence of dangerous emails. Another application of Decision Tree and Text mining for fraud and bankruptcy detection was investigated by Shirata et al. (2011) who predicted the risk of bankruptcy based on the analysis of annual reports published by Japanese companies. Another interesting approach of detecting fraud was developed thanks to the analysis of internal employee communication by a text mining algorithm (Holton, 2009 cited by Aaryan et al., 2020). The author started from the assumption that unsatisfied employees represented a higher risk of fraud, hence determining the satisfaction through internal communication could help in the detection of irregularities. Therefore, they decided to classify those communication snippets based on a Naïve Bayes classifier.

As the three examples above suggest, the most used information sources for fraud and bankruptcy detection are company internal data, i.e., internal communication and published reports (Kumar & Ravi, 2016; Pejić Bach et al., 2019).

2.3. *Analysis of Company Reports*

Financial valuation, audit, ESG analysis are some of the many examples of tasks that are supported by the extraction of meaningful information from company disclosure (Aaryan et al.,

[1] The identification of fraud encapsulates a broad set of irregular situations, going from money-laundering to spam to creditworthiness.

2020; Baier et al., 2020; Fisher et al., 2016). Given the increasing number of company's reports and disclosure (Kuh, 2019; Shahi et al., 2014), automated ways of summarizing data represent an interesting opportunity for financial professionals to speed up their analysis (Aaryan et al., 2020; Lewis & Young, 2019; Pejić Bach et al., 2019). For instance, a simple dictionary-based keywords analysis can be used for extracting the most relevant, and sometime hidden, information of a company's financial statement and thereby supports the process of evaluating its future financial performance (Fisher et al., 2016).

In short, the main advantage of Text Analytics for this specific use case lies in its ability to summarize large amount of textual information (which represent an important share of the available information in reports (Song et al., 2018)). Typical applications of Text Mining will therefore try to extract meaningful, and material information from various reporting material by using topic models like the Latent Dirichlet Allocation algorithm, but also more simple natural language processing methods. For instance, highlighting the content of a report can be achieved by a simple analysis of the term-document matrix – looking for the presence of specific keywords, derive conclusion from the most frequent terms – or a sentiment analysis (Fisher et al., 2016; Lewis & Young, 2019). Some specific examples can be found in the context of ESG investing where Székely & vom Brocke (2017) and Goloshchapova et al. (2019) have used the Latent Dirichlet Allocation analysis to extract the most common ESG topics of public reports from companies. Another category of Text Mining applications based on automatic report analysis have the objective of studying the relationship between a company's disclosure and its financial performance (Aaryan et al., 2020; Fisher et al., 2016; Kloptchenko et al., 2004). For instance, Lee et al. (2018) studied the relationship between the presence of some textual patterns (text length, frequency of certain words, and tonality) in the annual reports and the sales performance of various American companies from the IT sector. Their conclusion was that both text length and frequency of certain words was associated to a higher performance but not the tonality.

3. Application in the context of ESG rating

According to TrueValue Labs (n.d.), a data driven ESG rating provider, Artificial Intelligence and Text Mining also have a high potential for applications in the context of ESG rating. In fact, it has been observed that over the last few years an increasing number of alternative ESG raters have come out of the ground, either as independent start up or FinTech's, or as a department of traditional rating providers. For instance, ESG scores delivered by Refinitiv,

formerly known as Thomson Reuters, are now partially determined by a state-of-the-art sentiment analysis model developed internally (Stichbury, 2020). More and more traditional rating providers are thus turning to at least partial automation of their ESG analysis, mainly in the hope of overcoming some of the many critics of the traditional ESG rating model.

Generally, we can observe two methods for measuring ESG performance based on Text Mining techniques (Azhar et al., 2019). Either the analysis of company's disclosure leads to the construction of alternative ratings based on the presence or absence of certain ESG related topics. Or the analysis of independent sources of information leads to the construction of a controversy or reputation index for each company. Usually based on a sentiment analysis, this score generally adds an extra layer of information on existing ESG ratings

3.1. Challenges of traditional ESG ratings

3.1.1. Dealing with a "superabundance" of data

Like any other segment of the financial sector, sustainable investing has recently entered in a period of "superabundance" of data (Kuh, 2019; Shahi et al., 2014). On the one hand, companies, willing to demonstrate their engagement towards sustainability, are multiplying reports and communications on CSR, while on the other hand external data, news releases and social media, has only increased in recent years (Aureli, 2017; Kuh, 2019; Shahi et al., 2014; Siew, 2015; Verbeeten et al., 2016). The use of Big Data and Artificial Intelligence has therefore gained some momentum in the ESG investing community, particularly with the recent outbreaks in the field of machine learning and natural language processing (Hughes et al., 2021). As for traditional finance, the biggest contribution of those techniques for alternative rating approaches relies on their ability to process unstructured data in a cost- and time-efficient way (Kuh, 2019). As a large amount of the above-mentioned information is presented in an unstructured format, Text Mining alternatives therefore logically appealed to most traditional rating agencies and investors wishing to improve the efficiency of their methodologies (Aaryan et al., 2020; Kuh, 2019; Modapothala et al., 2009; RavenPack, 2019; Shahi et al., 2014; S&P Global, 2020; Stichbury, 2020; Takuya et al., 2020).

In addition, rating providers are also valuing Text Analytics for its ability to efficiently extract meaningful and material information for large amount of textual data (Antoncic, 2020; Kuh, 2019). When dealing with Big Data, this second advantage can be of high importance given the

large presence of noise that might cause that more data does not always rhyme with better decision making (Bedenik & Barišić, 2019; Woods & Hollnagel, 2005).

3.1.2. Robustness, transparency, and comparability

Advantages of AI and NLP are multiple, but the most often mentioned regarding ESG ratings is that they can create more robust, more transparent and more comparable scores (Hughes et al., 2021; Kuh, 2019; Liberato Ferrao, 2019; Wong & Petroy, 2020).

As opposed to traditional rating providers, agencies using machine learning and AI argue that they can limit the risk of biases in their scores. By reducing the importance of human analysis, text mining can indeed reduce the risk of errors, increase the objectivity and thereby the quality of the ESG ratings (Kuh, 2019; Shahi et al., 2014; Takuya et al., 2020). Relying solely on automated solutions has also the advantage of potentially uncovering latent or hidden features from company disclosures (Antoncic, 2020; Kuh, 2019)

Besides, the ability of machine-led analysis to find and process an impressive quantity of data has also increased the capability of rating providers to access independent and trustworthy data, e.g., news releases and social media content (Antoncic, 2020; Kuh, 2019; Siew, 2015). Hence, by increasing the independence of rating providers from information disclosed by rated companies, they could drastically reduce the self-reporting bias, and thereby increase their credibility in the eyes of end-investors (Dumont, 2017; Kuh, 2019).

Finally, some authors argue that Text Mining can improve the transparency and comparability of ESG rating methodologies (Hughes et al., 2021; Kuh, 2019). However, as pointed out by Loughran & Mcdonald (2016) this can only be true if the used model is effectively disclosed by the rating provider, and that this model does not rely on a "black box" approach. For instance, Naïve Bayes classification, though very performant, can hardly ever be replicated, hence are lacking transparency and can hinder the level of comparability between ratings.

3.2. Alternative ESG Ratings – CSR reports

For ESG rating providers, a first way to use Text Mining for rating purposes consist in the automatic processing, analysis, and evaluation of company's official disclosure, e.g., annual and CSR reports, company's website, etc. Such scoring process is usually based on controlling for the presence of certain ESG issues in the various publications of a company (Azhar et al., 2019). Motivations for applying this method are essentially related to a gain in efficiency and

the reduction of human biases. Besides, when based on an appropriate Text mining model it can also serve the development of more transparent ESG ratings.

Yet, the quality of those alternative ratings is highly dependent on the relationship between ESG disclosure and ESG performance, i.e., whether textual information reported in company's documentation are representative of its real involvement towards sustainability. Some authors and practitioners therefore consider this method as a more efficient ways of doing ratings as usual, thereby forgetting to deal with another important shortcoming: Self-reporting bias (Antoncic, 2020; Azhar et al., 2019; Kuh, 2019).

3.2.1. CSR Disclosure and ESG performance

As previously mentioned, the key to the success of report based ESG ratings relies in the assumption that company disclosure is a proxy of ESG performance. Fortunately, given the increasing number of CSR reporting practices, this topic has been the object of many studies in the past, allowing us to get a better understanding of the relationship between these two elements.

According to Shahi et al. (2014, p. 6), disclosure, under its various format, is at least partially motivated by the willingness of a company to "communicate environmental performance, acknowledgement of environmental responsibility, gaining competitive edge, obtaining social approval and showing regulatory compliance". Therefore, as suggested by the economic theory, companies with higher ESG performance should also disclose more ESG information (Valor & De la Cuesta, 2013; Verbeeten et al., 2016). Yet, according to other theories, e.g., the stakeholder's model and legitimacy theory, corporate disclosure is rather a function of socio-political pressure from company's stakeholders. Therefore, poor ESG performers would tend to increase their disclosure to change the perception and reduce the pressure of stakeholders regarding their CSR performance (Aureli, 2017; Shahi et al., 2014; Verbeeten et al., 2016). Those theories are theoretically explaining the phenomenon of greenwashing, or more generally ESG washing. Greenwashing is a strategy implemented by companies that either hide or undermine the importance of an ESG problem, or exaggerate its performance on that issue (Siew, 2015). In this situation, we might thus have bad ESG performers mentioning ESG issues on a more often basis than their peers.

Now, from an empirical point of view, results on the relationship between CSR disclosure and CSR performance are also rather nuanced. For instance, Verbeeten et al. (2016) have observed

that though CSR reports contained relevant information about a company's CSR performance, this relevance was not uniform across all dimensions of sustainability. They found out that while Social disclosure was clearly linked to Social performance, this link was less obvious for the Environmental dimension. This observation was also made by Takuya et al. (2020) in their study on the quality of Text Mining created ESG ratings. Those past results could be explained by the importance taken by environmental and climate issues over the last few years, increasing the pressure on companies to be performant in that field, hence increasing the risk of greenwashing (Takuya et al., 2020; Verbeeten et al., 2016).

Besides, some studies have also tried to identify specific linguistic patterns that could help investors to discriminate good and bad ESG performers based solely on their disclosure (Li & Haque, 2019). A first clear correlation has been identified several times in the literature between the readability of reports and CSR performance. For example, Bacha & Ajina (2019) and Nazari et al. (2017) have shown that higher readability, e.g., sentence length and complexity of words, was associated with a better ESG performance. In a more in-depth study, Clarkson et al. (2020) have identified additional language features of good CSR performers that could help improve the quality of report based ESG ratings. Among others, they have observed that good performers tended to have a higher "level of disclosure", i.e., higher number of words and sentences, and more "advanced level of writing", i.e., longer words requiring a higher level of intellect.

3.2.2. Text Mining based ESG Ratings

The building of custom CSR scores based on sentiment analysis, content analysis or text classification of sustainability reports is becoming more and more common in the literature (Aaryan et al., 2020; Shahi et al., 2014; Song et al., 2018). To our knowledge, most papers focusing on that topic will assume that the number of mentions of an ESG-related topic in company's disclosure is a good proxy for the number of its CSR activities on that topic, hence a good proxy of its ESG performance (Aureli, 2017; Shahi et al., 2014; Takuya et al., 2020; Verbeeten et al., 2016). Note that those authors are still pointing out that this assumption could be seen as a limitation of their approach, and of Text Mining in general. Particularly, Aureli (2017) mentions the issue of the absence of context when dealing with words like "climate" which can be used with different tone or meaning than climate change.

Shahi et al. (2014) designed a score to evaluate the CSR performance of companies based on the completeness of their respective CSR reports. To determine the level of completeness of a

report their Text Mining model checks for the presence of Performance Indicators from the GRI 3.0 Content Index, i.e., a complete report should include as many items from the GRI reporting guidelines as possible. From a technical point of view, the presence of a specific item is here considered as a classification task and is therefore computed thanks to a Naïve Bayes algorithm (identified by the authors as the optimal classifier) applied to the different sections of the textual report. After the classification step, each company is then rated on a scale ranging from C, presence of at least 11 GRI performance indicators on 49, to A, 20 indicators on 49. Comparing their scores to completeness rates delivered by third parties, they found a correlation of 0.531, which was considered as moderate by the authors. They also discovered the presence of a disclosure bias in CSR reports with companies tending to disclose more intensively on Governance and Environmental issues than social.

Another Text Mining model for evaluating ESG performance was proposed by Takuya et al. (2020). Contrary to the previous paper, the authors are here considering that the number of ESG activities can be approximated by the frequency and specificity of ESG-related words in sustainability reports. A high ESG performer is therefore expected to have a higher number but also more precise or specific ESG terms, e.g., "Exhaust gas" instead of "Air pollution". Note that this approach can be considered as more compliant with the results of linguistic research on CSR reporting such as the one conducted by Bacha & Ajina (2019) or Clarkson et al. (2020). For the creation of the ESG glossary, the authors decided to rely (1) on a vector representation of words (as opposed to the bag-of-words model) using word2vec algorithm, (2) on a trained artificial neural networks algorithm to label words of the report and (3) on a word hierarchy for each of the E, S and G dimensions. Then, based on this glossary both a quantity score, the sum of log-frequencies of ESG related words in the report, and a specificity score, the average degree of divergence of the words used in the report compared to top words of the hierarchy (Environmental, Social and Governance), are generated. In terms of performance, they compared those scores to Refinitiv's ESG ratings and found a significant relationship between them, except for the Environmental dimension, suggesting a form of greenwashing.

The approach proposed in the previous papers, i.e., the use of ESG glossary or vocabulary for assessing the performance of a company, is used in several studies, mainly related to content analysis. Verbeeten et al. (2016) for instance defined the ESG performance of German listed companies based on the number of appearances of 32 specific GRI keywords and the presence or not of a separate CSR report. Similarly, Aureli (2017) suggested a Text Mining model based on a vocabulary derived from the most frequent terms of the GRI framework. In both cases,

words are labeled as being ESG-related based on a dictionary approach instead of a machine learning approach, an advantage in terms of interpretation, replicability, computing capabilities, but which can also be relatively robust if the glossary has been well defined.

Based on those papers, we can see that when analyzing CSR reports, measuring ESG performance can be done by relying on the frequency of ESG-related topics. This postulate is indeed supported by the fact that several authors managed to demonstrate a positive link between the frequency of ESG words and ESG performance. For instance, Takuya et al. (2020) with Refinitiv's ratings, but also Clarkson et al. (2020) that have found that among the key differences between good and bad ESG performers the level of disclosure, i.e., the number of CSR-related terms, was significant. But is this significance sufficient to create a performant approximation of an ESG rating?

Despite its obvious attractiveness and elegance, some theories, and existing studies are questioning this simplistic postulate. By showing the weakness of frequency scores on predicting the Environmental dimension, Verbeeten et al. (2016) and Takuya et al. (2020) are showing that, though useful, the frequency of ESG-related words or topics might not be sufficient to accurately evaluate the ESG performance of a company. Because of the existing risk of ESG-washing, we might thus need to consider additional factors to discriminate good and bad ESG performers. From the documentation of ESG rating agencies, we can observe that an ESG rating is not only the sum of the ESG activities of a company. Indeed, to evaluate ESG performance, rating providers will also rely on a materiality matrix, i.e., they will rely on a weighted sum of ESG activities (Escrig-Olmedo et al., 2010, 2019; Refinitiv, 2021). Therefore, to approximate ESG performance, a logical approach would be to combine the frequency of ESG words with other data mining techniques to identify material words, to find hidden relationships, etc. As experienced by Shahi et al. (2014) it can therefore be possible to use classification algorithms to measure ESG performance from the frequency of ESG terms (even though quality of this measure were found to be moderate in this case).

3.3. Alternative ESG Ratings – CSR News

The use of news releases or social media content as an alternative source of ESG information has also gained some momentum in the ESG rating community, especially with the improvement of Natural Language Processing and Artificial Intelligence (Azhar et al., 2019). Thanks to the ability of text analytics to both efficiently access and analyze the enormous amount of data published by external sources, and to manage their high level of noise, ESG

rating providers can eventually benefit from the many advantages of measuring ESG performance based on alternative sources of information (Kuh, 2019).

If the use of ESG news for ESG rating purposes has attracted the attention of ESG rating providers, it is because it can create more accurate and timely responsive scores (Guo et al., 2020; Sokolov et al., 2021). According to Antoncic (2020), by using external sources of ESG data, ESG rating agencies can reduce the risk of self-reporting bias in their scores, hence their accuracy. The intuition behind this argument is that the data presented in company's disclosure may not reflect the full spectrum of ESG issues relevant for its stakeholders (Sokolov et al., 2021). Relying on ESG News is therefore a potential solution against the risks of greenwashing and other manipulation by a bad ESG performer.

3.3.1. ESG Controversies score

First and foremost, it is important to note that ESG news are hardly ever used to create a complete and independent ESG rating. More generally, the automatic ESG analysis will generate an additional indicator that can be used in complement of more traditional ESG ratings. For instance, Refinitiv is known for having developed a controversy score based partially on a Text Analysis of the Thomson Reuters news database. This controversy score assesses the involvement of a company in controversy practices, but it does not replace the general ESG rating of the rating provider. As a matter of fact, it should rather be combined with it (Refinitiv, 2021). As wisely noted by Azhar et al. (2019, p. 11), this choice can be motivated by the fact that "the lack of firms' news associations with CSR should not indicate firms' lack of engagement in CSR, as news coverage is generally influenced by the principles of newsworthiness that includes the prominence of a particular firm, event or scenario".

Concretely, to derive an ESG indicator from an analysis of news releases or social media contents, one would have to (1) identify ESG related news about the company based on some keywords search (Guo et al., 2020) (2) define an evaluation model to score the company based on the information contained in those filtered articles. Sentiment analysis is a common technique in both the literature and practice to construct the news-based evaluation model. In other words, a good ESG performer should be associated to a higher number of positive ESG-related news.

For instance, to assess the prominence of corporate social responsibility in more than 500 Singaporean firms, Azhar et al. (2019) relied on the average sentiment measure of CSR related

articles retrieved from the GDELT database. The CSR prominence of a company was then estimated by the number of positive articles associated to the company. This specific scoring methodology appears to be quite common in the literature with the variation mainly on the specific sentiment analysis approach. Rating providers such as Refinitiv (Stichbury, 2020) or RavenPack (2019) chose for a machine learning-based sentiment analysis on their respective news database. Schmidt (2019) preferred to use Loughran & Mcdonald (2011) sentiment dictionary to calculate a sentiment index and evaluate the effect of ESG news on the performance of Dow Jones Industrial Average Constituents.

To improve the quality of the sentiment analysis, more and more scholars and practitioners are arguing that the sentiment score could be improved by taking some context into account. Therefore, one should prefer to represent the news using a text embedding model rather than the simple bag-of-words assumption (Guo et al., 2020; Stichbury, 2020). For instance, a popular approach is to use a context-based embedding model like a pre-trained Bidirectional Encoder Representation Transformer (BERT) model in the preprocessing approach.

Conclusion

The first step of this master thesis, which consisted in a review of the existing literature on Text Mining and ESG rating comes here to an end. At first, those more "theoretical" chapters consisted in a general explanation of what measuring ESG performance concretely was, what was its place and role in the financial sector, what were the existing indicators of ESG performance, and the methodologies used by ESG rating agencies. From this first set of observations, we concluded that ESG ratings, as the principal source of ESG information, had a crucial role to play in the further development of Sustainable Investing, and by extension in the actual implementation of sustainable development. Yet, we also found that to fulfil its information role, ratings would still need to overcome many shortcomings: Lack of credibility, lack of transparency, lack of standardization, presence of biases, etc.

Therefore, in a second time, we started asking ourselves whether artificial intelligence, and more specifically text mining, could support the creation of improved ESG scores. To answer this question, the second part of the literature review was required to give us the opportunity to have a better understanding of what text mining concretely was, and how its tools are applied in the world of sustainable finance. This allowed us to discover how a Text Mining based rating could improve the quality of ESG performance evaluation (less biased, independent source of

information, low cost, more credible, transparency, etc.) and how it could be built (ratings based on the frequency of ESG words (or topics) in Reports or on the sentiment of ESG-related News). However, we also found that research on those Text mining-based ratings was still in its infancy, with very few results on how accurately they could approximate ESG performance.

Therefore, in this research, we will try to partially fill that gap by investigating the performances of ESG ratings drawn by several Text Mining models from official company disclosure. In this way, we will seek to contribute to the knowledge on "*How to rate ESG performance based solely on a Text Mining analysis of a company's sustainability disclosure?*", and thereby contribute to the development of ESG ratings that combine advantages of Text Mining (high transparency, lower bias, higher credibility, and low cost) with the best possible approximation of ESG performance. Besides, by interrogating the possibility of building ESG ratings solely from publicly available information we are also questioning the transparency of companies and ESG rating providers.

Our starting point will be similar to the one of Aureli (2017), Shahi et al. (2014), Takuya et al. (2020) and Verbeeten et al. (2016), i.e., we are assuming that the frequency of ESG words in disclosure as a building block for our ratings. Yet, contrary to them, we will then combine the use of a more transparent ESG word labelling approach (a dictionary-based), with the testing of a bigger variety of Text Mining-based scoring models, and the study of performances for three different type disclosure formats as source of ESG words.

Part 2: EMPIRICAL RESEARCH

Chapter 1: Data & Methodology

1. Research Question and Hypothesis

To support the development of sustainable investing, and thereby the funding of sustainable development, investors need to have access to clear, credible and transparent measures of ESG performance. Unfortunately, in their traditional format, ESG ratings do not yet meet all these requirements. Therefore, both investors and scholars have started to search for new and more appropriate rating methodologies. Among those, text mining analysis of company's disclosure and/or ESG related news has recently gained some success. However, as this field of research remains in its infancy, empirical proof of the quality of those text mining-based ratings is rather unclear (Hummel et al., 2017). Therefore, through this empirical research, we will try to contribute to knowledge on that specific topic by addressing the following research question:

"How to rate ESG performance based solely on a Text Mining analysis of a company's sustainability disclosure?"

To fully answer this initial problematic, our empirical study will apply a comparative analysis on the prediction capabilities of a (non-exhaustive) set of Text Mining based "rating" models. To put it simply we are trying to see which of the existing Text Mining models could best approximate the ESG performance of companies based on the information contained in sustainability disclosure. In addition, we will also use this analysis to evaluate how accurate a Text Mining prediction of ESG performance can be in general.

To define our set of scoring models, we are relying on different hypothesis on the relationship between ESG performance and corporate disclosure derived from our literature review on ESG rating methodologies and Text Mining. This list will include simple word-frequency models, but also linear, non-linear, or rule-based algorithms from the general machine learning toolbox. Both the definition and the justification of each of those models can be found in subsection 3.4 of the methodology section. Regarding the ESG performance to predict and the information from sustainability disclosure, the various sources of data used to represent those variables (respectively noted $\boldsymbol{Y}$ and $\boldsymbol{X}$) are developed in the next section (see Data Sample). As for the way in which we decided to extract data from a company's disclosure, i.e., from textual documents, this methodology is developed further in this work (see Step 1 – Text Pre-processing).

2. Data Sample

In this section we describe the different datasets that were used to bring this empirical research to end, as well as the ways those data were collected.

2.1. ESG Ratings

To approximate the ESG performance of a company, i.e., the outcome variable ***Y*** of this empirical research (see Research Question and Hypothesis), it is a common practice in the literature to rely on the data provided by ESG rating agencies (Huber & Comstock, 2017). To cite some examples, Halbritter & Dorfleitner (2015) considered data from Thomson Reuters, Bloomberg and MSCI, to test the correlation between financial and non-financial performance, while Eccles et al. (2014) used similar ratings to define how high ESG performance could impact the processes and performance of a company, etc.

For this empirical study, we decided to rely specifically on the ESG ratings provided by Refinitiv, the ESG data provider from Thomson Reuters. This choice was motivated first by the availability of the database which can be accessed via the account of the Louvain School of Management to the Eikon platform. Then, another important factor that we had to consider was obviously the quality and credibility of the rating. Although Refinitiv is not among the preferred sources of information of sustainable investors (Mooij, 2017; Wong & Petroy, 2020), their ESG scores are nevertheless often cited and used in the literature for their quality (Blank et al., 2016; Clarkson et al., 2020; Halbritter & Dorfleitner, 2015; Semenova & Hassel, 2015; Takuya et al., 2020). For example, Clarkson et al. (2020) emphasizes their more comprehensive approach of the CSR performance evaluation process, while Semenova & Hassel (2015) points out that the rating agency claims to be a transparent, objective and audited information provider (Blank et al., 2016; Refinitiv, 2021). Given those various elements, Refinitiv's data seemed to be a reliable source of ESG information for the purpose of our research.

As we are only interested in the global ESG performance of a company, among the large amount of ESG data available on Eikon, this thesis only relied on the ESG scores, i.e., "the company's ESG performance based on verifiable reported data in the public domain" (Refinitiv, 2021, p. 6). Depending on the use case, those scores can be represented under a numerical or letter format (ranging from 0 to 100, from D- to A+, or from D to A) to allow for an easy and fair comparison of company's performances across and within industries and geographical regions (Blank et al., 2016). For this thesis, we have decided to rely on two out of the three different

formats (see **Table 1**) to compare the performance of our ratings models at different levels of granularity (though the three formats were downloaded).

Table 1: ESG Scores Format and Denomination

Name	Scores Range
Continuous - Y^c	$[0; 100]$
Letter-Grade - Y^g	$\{D-;\ D;\ D+;\ C-;\ C; C+;\ B-;\ B; B+; A-;\ A;\ A+\}$
Letter-Category - Y^{cat}	$\{D; C;\ B;\ A\}$

Note that as the ESG scores are updated on a yearly basis (frequency of publication of company disclosure), we collected ESG data for the four last available periods, i.e., from 2017 to 2020.

2.2. Sample of Ratings

As of today, Refinitiv's ESG rating universe covers the ESG performance of more than 9000 companies active in more than 50 industries and various geographical areas. For example, Refinitiv's database contains ESG information for all the companies listed on the MSCI world index, the S&P 500, or the FTSE 100 (Refinitiv, 2021). In an ideal world, our empirical study should have been based on the complete universe of ESG ratings available on Refinitiv. Unfortunately, because of limited storage and computational capabilities we decided to reduce our analysis to a smaller, yet representative, sample of companies.

Starting with the entire dataset of 9000 assessed companies, we reduced our sample a first time by narrowing the rating universe to companies listed on the MSCI world index only. This leaved us with a temporary list of ESG ratings for 1,540 large and mid-cap companies from 23 developed markets (MSCI, 2022). By keeping a certain level of variety of locations and industries, this first dataset ensured that the results of our study will not be specific to the disclosure characteristics (regulations, common practices, etc.) of one region or sector of activity, and could therefore be sufficiently generalized (Verbeeten et al., 2016). But, at the same time, by restricting ourselves to only large companies, operating in developed economies, we also ensured that the level of disclosure was sufficiently qualitative and harmonized, that we would have enough information available to build Text Mining based ratings. From this list, we then eliminated the companies for which an ESG score was not available for one of the four years under investigation (from 2017 to 2020). Finally, to comply with the storage and computing limitation we were facing, we applied a last reduction of the dataset by randomly

picking 367 constituents of the index (see list in Appendix VI). In this last operation, we were still careful to maintain a varied mix of companies in terms of location, industry, but also of ESG performance. Therefore, we have also added a sample of companies with lower ratings (see Appendix VI). Eventually, we ended up with a sample of 419 companies from 7 geographical areas and 44 industries. The exact repartition of location and industries are displayed in Appendix VII, while **Table 2** displays the number of companies per year together with the number of companies per ESG performance category.

Table 2: Number of Companies per ESG rating category and year (sample summary)

	2017	2018	2019	2020
A	154 (38,89%)	161 (39,08%)	163 (38,90%)	160 (38,28%)
B	164 (41,41%)	171 (41,50%)	176 (42,00%)	175 (41,87%)
C	67 (16,92%)	69 (16,75%)	69 (16,47%)	68 (16,43%)
D	11 (2,78%)	11 (2,67%)	11 (2,43%)	11 (2,66%)
Total	396	412	419	414

We can thus see that the selected sample is mainly composed of companies with relatively high ESG performance (more than 75% for every year). The difference in terms of size for samples of year 2020 and before is linked to the presence of Annual or CSR reports published in an unusable format (see **sections 2.3** and **3.3**).

For each of those companies, the full ESG score was downloaded from Refinitiv's database for the years 2017 to 2020. As previously explained, those ratings will be used as a proxy measure of their respective ESG performance, performance that our different rating models should try to predict based on the disclosure of the company.

2.3. Company Disclosure

Based on the literature review, we know that there are potentially two main sources of textual information that can be used to create a Text Mining based ESG rating: Company Disclosure and External information (news and social media). For this research, we have decided to focus solely on the company disclosure, as Clarkson et al. (2020), Shahi et al. (2014), Takuya et al. (2020), and Verbeeten et al. (2016) before us. Yet, a company's disclosure remains a rather broad term that encapsulates many different publications made by a company (webpages, articles, reports, codes of conduct, etc.) (Azhar et al., 2019; Refinitiv, 2021).

In this empirical research, we chose to only focus on the Annual and CSR reports[2] published by a company as an approximation of its sustainability disclosure, i.e., our explanatory variable ***X*****,** are made of the information from those two categories of report. This approach is relatively common across the literature on Text Mining analysis of ESG performance. Indeed, Takuya et al. (2020) relied on the information contained in CSR reports to build ESG quantity and specificity scores of Japanese firms. To justify this choice, they are citing results from a study by KPMG stating that a large majority of companies are using this format of report to disclose their ESG activities (Blasco & King, 2017). For their part, Baier et al. (2020) have considered annual reports of companies as being the most relevant source of information to build an ESG dictionary. This idea is shared by other authors like Giles & Murphy (2016) who are arguing that Annual Reports more than any other disclosure source are used by firms to publish information they consider as important. Besides, to evaluate the relationship between CSR performance and several CSR linguistic patterns, Clarkson et al. (2020) have relied on standalone CSR reports as a source of sustainability disclosure. We can see that both Annual and CSR reports can be used to link disclosure and ESG performance, yet we have no indication on which is the best source of information, neither do we know whether those are containing enough ESG information (Giles & Murphy, 2016; Unerman, 2000; Verbeeten et al., 2016). Therefore, the performance of our various scoring models will be firstly tested when deriving ratings from Annual reports (**A**). A second time when using CSR reports as explanatory variables (**C**). And a last time for a combination of the two formats (**A+C**). Thanks to this approach, we should also see which report type is the most relevant for the specific purpose of building Text Mining based ratings.

To collect our reports database, we visited the official websites of each of the companies in the sample presented above. There, we downloaded the Annual reports for fiscal year 2017, 2018, 2019 and 2020 under a PDF format[3]. While the reports for the year 2020 were available for every company, in several cases we had to deal with shortcomings for disclosure of other years (report only published every two years, report no longer available, etc.). This partly explains the difference in size between our 4 yearly samples (see **Table 2**).

When available, we also exported the CSR reports published by the different firms. We can see in **Table 3** that the percentage of firms publishing both reports is reasonable for most years with

[2] We include in the definition of CSR reports, sustainability reports, ESG reports, etc.
[3] Note that for some firms operating on a staggered accounting year, reports where then downloaded for the years 2018 to 2021

a peak at 86% for 2020. Note that when a firm does not publish any standalone sustainability report, we considered that the combination of Annual and CSR report for this company was simply equal to its annual report, i.e., $\boldsymbol{A + C = A}$. This means that they will automatically have a lower frequency of ESG related terms when using the data from Annual and CSR reports, hence they will automatically be penalized in their predicted ESG score. By relying on this approach, we will thus create rating models that tend to give higher rates to companies publishing CSR reports. In fact, we decided to keep this feature in our empirical research to be in line with what has been previously observed in the literature. Clarkson et al. (2020) noted that from a sample of 1835 standalone CSR reports 1325 (72.21%) were attributable to good ESG performers. In addition, among the results found by Verbeeten et al. (2016) in their study on the relationship between CSR disclosure, CSR performance and share prices of German Listed companies, we can also mention the evidence that publishing a standalone sustainability report is significantly associated to higher performance.

Table 3: Number of Companies with Annual vs. CSR vs. Annual and CSR reports for every year

	2017	2018	2019	2020
Annual X^A	396	412	419	414
CSR X^C	266 (67,17%)	305 (74,03%)	339 (80,91%)	360 (86,96%)
Annual-CSR X^{A+C}	396	412	419	414

3. Methodology

In this section, we describe the methodology implemented to obtain an answer to the question "*How to rate ESG performance based solely on a Text Mining analysis of a company's sustainability disclosure?"*. As previously explained, we suggest measuring ESG performance by relying on the number of occurrences (known as *frequency* in Text Mining) of ESG words in Annual and/or CSR reports. To be more precise, our study applied a comparative analysis on different Text Mining scoring models, i.e., we evaluated and compared the ability of several classification and prediction models to accurately measure the ESG performance of companies from our four years sample of ESG ratings by relying solely on the frequency of ESG terms in Annual and/or CSR reports. Of course, one cannot expect such a technique to provide a fully accurate indicator of ESG performance. Yet, considering the importance taken by ESG in the last few years and given the large number of firms responsible investors need to analyze, this is a first step towards more efficient and qualitative construction of ESG ratings.

3.1. General Overview

To achieve this comparative analysis, we needed to transform the textual information of Annual and CSR reports into a vectorial format usable by our rating models. So, like any other Text Analysis we began our empirical research with a step of Text pre-processing. The objective of this first operation was to represent each company's report by a vector summarizing the information they are containing. As previously explained, for our research the information we were interested in was the frequency of ESG-words. Moreover, a word was considered as ESG when it could be found in a list of 482 ESG terms from the literature. Hence, after the preprocessing a report was transformed into a vector of dimension 482, whose features are the number of occurrences of a specific ESG term in the report (see example in Appendix VIII).

Next, we entered the heart of this empirical research: measuring the ability of a set of models to create accurate ESG scores. First, we defined a list of models (and algorithms) that should generate a performant Text Mining based ESG score according to the literature. In this list, we considered both prediction and classification algorithms, i.e., models that can respectively predict ESG continuous[4] and letter-ratings. Then, to evaluate the performance of the considered scoring approaches we followed a commonly used train-test process. First, using the ESG words frequency vectors from 2020 as a validation set, we relied on a 10-fold cross validation to fit each model with its optimal parameters. Then, we used the full sample to test and compare the ability and stability of our optimal Text Mining models to measure ESG performance, and thereby provide an answer to the question asked by this paper. Note that the accuracy of each model is evaluated against the value of Refinitiv's ESG ratings under their category and grade formats (see **Table 1**), so, each model was fitted 2 times for each of the four years.

3.2. Notation

Before going into a more in-depth description of the different steps of the methodology, it is important to consider some of the notations that will be used in this chapter. Note that all the notations and their meanings are also listed in Appendix X.

- **Refinitiv's ESG rating of a company**: $y_i^{f,a}$

[4] For comparison purposes, the continuous ratings will then be converted to letter ratings by relying on Refinitiv's conversion rules (see Appendix V)

The index i refers to the company whose rating is given by the variable. It therefore takes as a value one of the names of the companies in the sample (see list in Appendix VI). If we name $Corp$ the set of company names, then we have that $i \in Corp$. Concerning the exponents, the letter f refers to the format of the rating y. As a reminder, Refinitiv scores can be presented in continuous (from 0 to 100, a set that we define as C), category (from D to A, a set of value that we define as Cat) or grade format (from D- to A+, a set that we define as G). Hence, f can take three different values that we define as c for continuous, g for grade, and cat for categories, i.e., $f \in \{c; g; cat\}$. Finally, the exponent a is a reference to the year for which the rating was delivered, hence we have $a \in \{2017; 2018; 2019; 2020\}$. For instance, in 2017, the company Boeing has received an ESG score of 83.09, which corresponds to a category A, and a grade A-. Therefore, from our notation we would have:

$$y_{Boeing\ Co}^{c,2017} = 83.09 \qquad y_{Boeing\ Co}^{g,2017} = A- \qquad y_{Boeing\ Co}^{cat,2017} = A$$

When we use the notation $\boldsymbol{Y^{f,a}}$, it simply refers to a column vector containing Refinitiv's rating from every company of the sample under format f for year a. Since the size of the sample is varying in function the year, the size of the vector is a function of a. For instance, for $a = 2017$, we have that $\boldsymbol{Y^{f,a}}$ with a dimension of (1×396) (see **Table *3***).

To symbolize a prediction of Refinitiv's ratings, we use the notation $\hat{y}_i^{\,r,f,a}$. The additional exponent r indicates from which type of report we derived the prediction of the rating. Recalling that we consider three sources of sustainability information to build ESG scores in this research (Annual (A), CSR (C) reports, or both (A+C)), we have that $r \in \{A; C; A + C\}$

- **ESG words Frequency vectors and matrix**: $\boldsymbol{w_i^{r,a}} - w_{i,j}^{\,r,a} - \boldsymbol{X^{r,a}}$

The variable $w_{i,j}^{\,r,a}$ refers to the frequency of the ESG word j in the report r published by company i in year a. The index j therefore refers to one of the ESG words from the list displayed on Appendix XV. Hence, it can take one of the 482 values of the set we call $Words$, i.e., $j \in Words$. The index i and the exponents a and r should be interpreted in the same way as for the previous variable. On the one hand, i is the name of the company who published the report in which word j was found, and $i \in Corp$. On the other hand, a refers to the year during which the report was published and $a \in \{2017; 2018; 2019; 2020\}$, and r is the type of report in which we found the frequency of word j and $r \in \{A; C; A + C\}$.

The vector $\boldsymbol{w_i}^{r,a}$ is a vector of dimension (1×482) that contains the frequencies of word j in the report r of company i in year a for each j. For instance, in 2017 Boeing used the word j = "approval", j = "approve", and j = "vulnerable" respectively 2 (3), 2 (8), and 1 (1) time in its CSR report (Annual Report). This resulted in the following vectors:

$$\boldsymbol{w}^{\boldsymbol{C,2017}}_{\textbf{Boeing Co}} = (2,\ \ 2,\ \ldots,\ \ 1) - \boldsymbol{w}^{\boldsymbol{A,2017}}_{\textbf{Boeing Co}} = (3,\ \ 8,\ \ldots,\ \ 1) - \boldsymbol{w}^{\boldsymbol{A+C,2017}}_{\textbf{Boeing Co}} = (5,\ \ 10,\ \ldots,\ \ 2)$$

When we use the notation $\boldsymbol{X^{r,a}}$ it refers to a matrix containing the vector $\boldsymbol{w_i}^{\boldsymbol{r,a}}$ for every company i:

$$\boldsymbol{X^{A,a}} = \begin{pmatrix} \boldsymbol{w^{A,a}_{Boeing}} \\ \vdots \\ \boldsymbol{w^{A,a}_{Aptiv}} \end{pmatrix} \qquad \boldsymbol{X^{C,a}} = \begin{pmatrix} \boldsymbol{w^{C,a}_{Beoing}} \\ \vdots \\ \boldsymbol{w^{C,a}_{Aptiv}} \end{pmatrix} \boldsymbol{X^{A+C,a}} = \begin{pmatrix} \boldsymbol{w^{A+C,a}_{Boeing}} \\ \vdots \\ \boldsymbol{w^{A+C,a}_{Aptiv}} \end{pmatrix}$$

3.3. Step 1 – Text Pre-processing

As it allows textual data to become usable by various statistical and data mining techniques, text pre-processing is an inevitable step of any Text Mining application (Miner et al., 2012). In fact, every research that has tried to study text mining tools has had to develop its own procedure, procedures from which we have drawn our inspiration for this thesis.

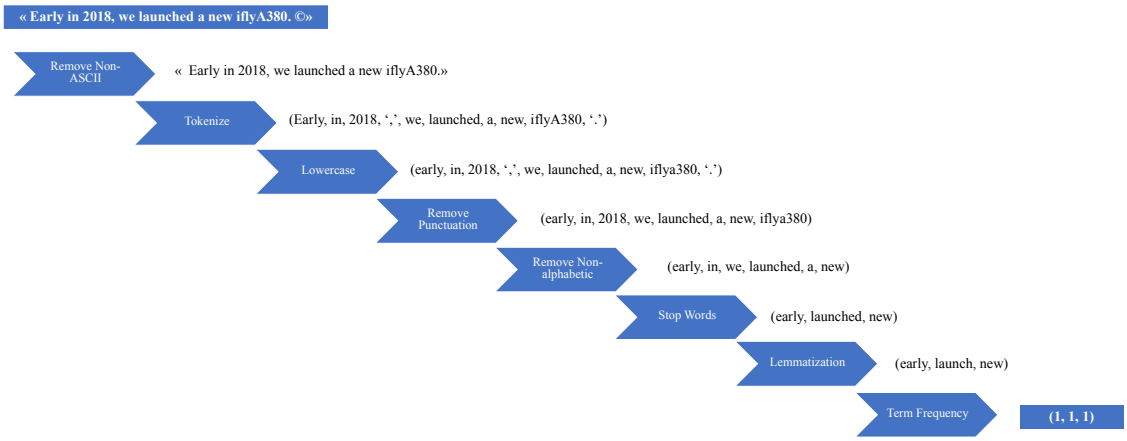
Concretely, in this research, the objective of the preprocessing is to transform a textual report published by a company into a vector of ESG-words frequency. Since we consider three different types of reports per company, after the preprocessing, each of them is represented by 3 vectors, one for the frequency of ESG words in the Annual report, one for the CSR report and one for the combination of both. Taking all the companies into account, we obtained three matrices whose rows are the word frequency vectors of each company: $\boldsymbol{X^{A,a}}$; $\boldsymbol{X^{C,a}}$; $\boldsymbol{X^{A+C,a}}$

To obtain this result we proceed in two distinct steps. First, we follow a classical pre-processing procedure to transform the report into a bag-of-words (vectorial representation). Then, as our approaches are based on one major postulate that the frequency of ESG words is linked to the ESG performance, we reduce the dimension of the bag-of-words to only keep the frequency of ESG words in the vectors. This second step is realized through the application of a dictionary-based approach as suggested by Clarkson et al. (2020), Verbeeten et al. (2016) or Baier et al. (2020)

3.3.1. Bag-of-words representation

The use of a bag-of-words representation is very common in Text Mining analysis of sustainability reports and sustainability performance (Clarkson et al., 2020; Raghupathi et al., 2020; Shahi et al., 2014; Takuya et al., 2020; Verbeeten et al., 2016). As explained during the literature review (see Chapter 2, section 1.1.1), this means that according to most authors on the topic, we can rely on the assumption that to evaluate ESG performance there is no need to consider context, that syntax is sufficient. This is of course a strong assumption which we will need to take into consideration when interpreting the results of this empirical research.

Figure 1: Fictional Example to illustrate the impact of every pre-processing sub steps on a sentence



To build our bag-of-words we have followed a set of common transformation steps using Python 3.10 NLTK library (see **Figure 1** and Appendix VIII). Starting from a report (Annual or CSR), we began with a simple removal of non-ASCII characters (Loughran et al., 2009). Then, as we were building bag-of-words, we logically tokenized the text on words using the *nltk.word_tokenize*( ) function (Clarkson et al., 2020; Miner et al., 2012; Takuya et al., 2020; Verbeeten et al., 2016). Then as suggested by Miner et al. (2012) we transformed each token to lowercase, removed punctuation and removed non-alphabetic characters to reduce both the dimensionality and noise of the data. Thereafter, we excluded stop words from the vector of tokens by using a list of English stop words provided by the NLTK library (Raghupathi et al., 2020). Finally, to further reduce the dimension of the token vectors we decided to rely on a lemmatization algorithm, (*nltk.stem.WordNetLemmatizer)* to group words with a similar root (Pejić Bach et al., 2019). Choosing for a lemmatization rather than a classical stemming algorithm was motivated by the finding that, for our dataset, it resulted in better and more coherent results. Eventually, the bag-of-words representation of a report was obtained by using

a term frequency approach, i.e., the value of the features was given by the number of occurrences of a token in the text.

As previously mentioned, this transformation is applied on the Annual and CSR reports (alone and combined) of each company for each of the four years. So, each company i is now represented by 12 vectors of words frequencies: $\boldsymbol{w_i}^{\boldsymbol{r,a}}, \quad \forall \boldsymbol{r, a}$

3.3.2. Bag-of-ESG-words representation

For the specific purpose of our research, we added an extra step to the preprocessing: The removal of non-ESG words. For this task, we relied on dictionary-based approaches rather than machine learning methods, thus distancing ourselves from the work of Takuya et al. (2020) or Shahi et al. (2014). As mentioned in the literature review, we can justify this choice by the fact that lexicon-based approaches have proven to be relevant methodologies offering simple, transparent, and robust models. However, those benefits are subject to the condition that the lexicon is correctly designed (Lewis & Young, 2019).

To comply with the above-mentioned constraint, we believed it was more appropriate to rely on an ESG dictionary already developed in the literature rather than creating our own. This dictionary, built by Baier et al. (2020), contains a list of 482 ESG terms divided into 10 ESG topics, and 40 sub-topics (see Appendix XV). It was constructed based on often used ESG terms in a sample of 10-K reports and proxy statements with the objective of helping future research to "quantitatively examine the extent of ESG reporting" (Baier et al., 2020, p. 1). An objective that corresponds quite well to the purpose of this research (quantifying the ESG reporting of a company to derive a score from it) making this lexicon a relevant choice for this paper.

Thanks to this dictionary, we had all the required tool to move from the simple word's frequency vector to an ESG words frequency vector, a vector of dimension (1×482). To put it simply, the bag-of-ESG-words representation is a vector whose feature's value is 0 unless the words it stands for is present in the initial bag-of-words and has a frequency higher than 0. If we denote the frequency of a given word j in the bag-of-words from report r of firm i in year a by $b_{i,j}^{r,a}$ we have that for all $j, i, r,$ and y:

$$w_{i,j}^{r,a} = \begin{cases} b_{i,j}^{r,a}, & b_{i,j}^{r,y} \exists \; and \; b_{i,j}^{r,y} > 0 \\ 0, & otherwise \end{cases}$$

A concrete example of this last step can be found in Appendix VIII.

3.4. Step 2 – Defining Scoring Models

After having looked at the pre-processing phase, we are describing here the Knowledge extraction stage of this empirical research. In other words, in the next section we are focusing on the definition of the different scoring models selected to derive Text Mining based ratings (from the pre-processed reports). For each model, a brief overview of their intuitions, main equations and parameters is given. Note that the list of the selected models was established based on the reviewed literature on Text Mining (hypothesis on the relationship between ESG performance and ESG disclosure, most often-cited classification algorithms in the literature, etc.) and on the specificities of our empirical research (high dimensional feature vectors, willingness to interpret results, etc.).

3.4.1. ESG-word frequency

Our first rating approach consists in using the frequency of ESG words as the ESG score of a company. Although at first sight simple, this initial model is also the most intuitive given the reviewed literature on Text Mining. It is indeed based on a very common assumption of the Text Mining literature which is to say that the occurrence of ESG terms approximates ESG activities and by extension ESG performance. Therefore, this has already been at least partially used by different authors to develop machine led ESG ratings (see section 3.2.2 Text Mining based ESG Ratings) (Shahi et al., 2014; Takuya et al., 2020).

Concretely, to build an ESG score this ESG-words frequency model takes an approach very similar to the one applied by Takuya et al. (2020) and their ESG quantity score: First, one must detect the frequency of ESG words in the CSR reports words frequency matrix. Then, one must sum the relevant frequencies to obtain a quantity score. Thanks to our pre-processing step, every company is already represented by a vector of frequency of ESG terms. Therefore, to create an ESG rating for firm i from its report r of year y we simply sum the features of vector $\boldsymbol{w_i^{r,a}}$:

$$\hat{y}_i^{r,c,a} = \sum_{\text{j}\in\text{Words}} w_{i,j}^{r,a}$$

At this point, the output does not yet correspond to an ESG score comparable to those of Refinitiv. The model still needs to bring them back to a range from 0 to 100[5], before converting them to letter-ratings using Refinitiv's conversion rule (see Appendix V). Note that this second

[5] By dividing each $\hat{y}_i^{r,y}$ by the predicted rating with the highest value for this year and report

conversion step is particularly important as it will allow us to compare this first approach to more sophisticated classification models in terms of accuracy, F1-score, precision, etc.

The main advantage of this rating approach relies obviously in its simplicity. Indeed, apart from the choice of ESG dictionary, the generated score does not rely on any parameter. Therefore, it is a scoring model that should be quite robust in our big data setting and thereby have stable out-of-sample prediction capabilities (Ben-Hur & Weston, 2010; Miner et al., 2012). In addition, this simplicity also ensures that the rating methodology remains highly transparent (no black box approach) which facilitates the interpretation and reproducibility of the results (Lewis & Young, 2019; Loughran & Mcdonald, 2016).

3.4.2. Multiple Linear Regression

For this second scoring model, we relied on a very common method in the world of statistics and data mining: the multiple linear regression. This time, to create our rating we are assuming that the ESG performance of a company can be obtained by a linear combination of its ESG word frequency vector:

$$y_i^{c,a} = \beta_0^{r,a} + \sum_{j \in \text{Words}} \beta_j^{r,a} \times w_{i,j}^{r,a} + \epsilon_i$$

Where ϵ_i is the error term and is assumed to follow a normal distribution $N(0,1)$.

To fit this model, one need to determine what this linear combination exactly is, i.e., one first need to estimate the value of the vector β $\big((1+p)\times 1\big)$. In the context of a multiple linear regression, the β's are chosen to minimize the squared prediction error on the training set. This estimator is therefore called the Ordinary Least Square (OLS) estimator and is obtained by solving the following optimization problem (for which we added a column of 1 to $\boldsymbol{X^{r,a}}$):

$$\widehat{\boldsymbol{\beta}}^{r,a} = arg\,\min_{\beta} \left|\left|\boldsymbol{Y^{c,a}} - \widehat{\boldsymbol{Y}}^{\boldsymbol{r,c,a}}\right|\right|_2^2 = \left|\left|\boldsymbol{Y^{c,a}} - \boldsymbol{X^{r,a}}\boldsymbol{\beta}^{\mathbf{r,a}}\right|\right|_2^2$$

$$\widehat{\boldsymbol{\beta}}^{r,y} = ((\boldsymbol{X^{r,a}})^T\, \boldsymbol{X^{r,a}})^{-1}(\boldsymbol{X^{r,a}})^T \boldsymbol{Y^{c,a}}$$

Though the multiple linear regression model is not the preferred method of data scientist when working with textual data (high risk of overfitting), we can still find several applications in a financial context with fairly positive results (Kumar & Ravi, 2016). In addition, from the documentation of ESG rating agencies, we can observe that an ESG rating is not only the sum of the ESG activities of a company. Indeed, to evaluate ESG performance, rating providers will

also rely on a materiality matrix, i.e., they will rely on a weighted sum of ESG activities (Escrig-Olmedo et al., 2010, 2019; Refinitiv, 2021). Therefore, to approximate ESG performance, it is a logical approach to consider that each ESG-related terms should be allocated a different weight in function of their contribution to the global performance of the company, i.e., to fit a linear regression. For our research, relying on a regression model has thus the advantage of building a continuous score which is closer to the one of Refinitiv's rating methodology.

Another important benefit of the linear regression is that this model keeps a relative simplicity in the interpretation of the results while adding the possibility to analyze more deeply the influence of each ESG words on the ESG performance of a company. Yet, the linear regression estimated by OLS has also one major drawback: Overfitting. As we are working with Big Data setting, it will be important to appropriately manage the curse of dimensionality (feature selection, cross-validation, etc.) to keep the predictive power of our model sufficiently high out-of-sample (Dougherty, 2013).

3.4.3. Naïve Bayes Classifier

The Naïve Bayes (NB) classification model is a linear statistical classifier that assigns individuals to their categories by relying on the Bayes theorem. The starting point of this classifier is the idea that an individual should be assigned to the class ω that is the most likely given its features X. In other words, the Naïve Bayes classifier assigns an individual to the class that maximizes the posterior probability of X:

$$\hat{\omega} = arg \max_{\omega \in \Omega} P\,(\omega|X)$$

Therefore, to make a classification one need to find a way to compute those different posterior probabilities and that is where the Bayes theorem comes in play. As a reminder, according to this theorem (which results from the definition of the conditional probability), the posterior probability is a function of (1) the probability of observing X given ω, the evidence, (2) the probability of ω, the prior, and (3) the probability of observing X, i.e., $P(X)$ (Dougherty, 2013):

$$P(\omega|X) = \frac{P(X|\omega) \times P(\omega)}{P(X)}$$

So, to classify a new observation, one must estimate the value of those three elements based on a training sample (a set of feature vectors of size n and their associated class ω) for every class. In fact, since $P(X)$ is independent from ω it has no influence on the classification rule presented earlier, and therefore does not require to be computed. Then, to facilitate the estimation of the

two remaining elements, the Naïve Bayes classifier is relying on a "naïve" hypothesis about the inputs. According to the classifier, the different features of the vector are independent, which allows us to reformulate the initial equation as follows:

$$P(\omega|X) \propto P(\omega) \times \prod_{i=1}^{n} P(x_i|\omega)$$

Now, how does the NB classification algorithm concretely estimate the prior and the evidence based on the training sample? To estimate the value of the prior $P(\omega)$, it will use the frequency of appearance of class ω in the training sample. For each $P(x_i|\omega)$, one will have to assume on the distribution of the variable x_i when the class is equal to ω, and use the training sample to estimate its parameters. For instance, one can rely on a Gaussian distribution with parameters μ and σ^2 given by the average and variance of the values taken by x_i when individuals are belonging to class ω. Once those different elements have been calculated, it only remains to rely on the simplified Bayes rule to calculate the posterior of each class and select the one that has the highest.

Although it is based on rather strong and naive assumptions, the naive Bayesian model remains one of the most used models for text classification. The authors particularly appreciate its ability to be robust and accurate even when dealing with high dimensional feature vectors. In the context of ESG rating, Shahi et al. (2014) compared the classification accuracy of several algorithms for the prediction of the completeness of CSR reports. They found that the Naïve Bayes outperformed all other models (Neural Networks, Decision Trees, and Decision tables) when fitted after a Correlation Feature Selection filter.

Nevertheless, this model does not allow for an easy interpretation and analysis of the results as it is more based on a Black Box approach. It is indeed not possible to identify what is the exact influence of each separate features on the final probability and so not easily reproductible (Loughran & Mcdonald, 2016)

3.4.4. Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) originally used for feature extraction has also shown interesting capabilities in classification tasks. For instance, Bekios-Calfa et al. (2011) compared the performance of a LDA, a SVM and, a boosting algorithm on a gender classification task. They found that the linear model had the advantage of combining high accuracy scores with simplicity and low computational requirements. Furthermore, they are showing that when data

and computational resources are limited, then the LDA model outperformed its non-linear peers. Equivalently, Mahmoudi & Duman (2015) investigated the classification ability of a LDA model for credit card fraud detection. Results from their study are again suggesting that this model can provide more accurate predictions, even in imbalanced data classification.

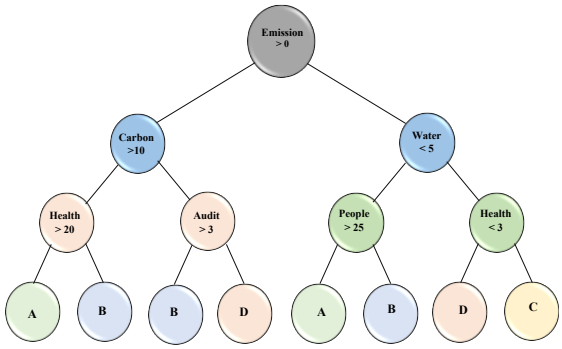
In practice, the classification through LDA is a process divided in two steps. First, the original high dimensional data is projected into a subspace of lower dimension (at most one dimension smaller than the number of classes). This linear projection is chosen to maximize the ratio Between Sum of Squares – Within-class Sum of Squares. This first transformation results in a new vector of feature $z_i^{r,y}$ given by a linear combination of the original data (in our case for each company i), new features that are supposed to most discriminate between classes (Dougherty, 2013). Then, based on this projected dataset, a classification rule is build based on the same process then for the NB classification (see above): Classification in the class which maximizes the posterior, posterior which is estimated based on the Bayes Theorem. Except that here the evidence is assumed to follow a multivariate Gaussian distribution with parameters μ_ω, the mean vector of the class ω in the sample, and Σ a variance-covariance matrix common to all classes (a pooled variance-covariance matrix) (Dougherty, 2013).

3.4.5. Random Forest (Regression or Classification)

The random forest (RF) model is an ensemble learning algorithm that generates its output (a continuous score or a class) by averaging the results of a given number of Decision Trees (DT) (Couronné et al., 2018; Shah et al., 2020). With this approach, we are therefore looking at the ability of a rule-based model to predict the ESG performance of a company. Rule-based approaches through Decision Trees or Random Forest are also common techniques for Text analysis and classification in the world of finance. For instance, Joshi et al. (2016) found that Random Forest classifiers could deliver better out-of-sample accuracy than Naïve Bayes for sentiment polarity detection. Random forest algorithms have also proven to be more efficient than other classical classifiers (SVM and logistic regression) for credit card fraud detection tasks, as explained by Bhattacharyya et al. (2011). According to Couronné et al. (2018) or Shah et al. (2020) this increased performance is due to the fact that averaging the predictions of several Decision Trees lowers the variance, and therefore stabilizes the out-of-sample predictions of this model.

Concretely, the RF model starts by randomly sampling n subsets from the original dataset (process of bootstrapping). On each of those samples a classification or regression task is achieved by fitting an independent DT model (Couronné et al., 2018). As a reminder, a DT model is a classification (or regression) tool that assigns individuals to their classes by building a hierarchy of rules. Each node of the Tree is thus a rule on one specific feature and is designed to split the dataset into increasingly homogenous subsets. For the RF algorithm used in this research, the DTs are built based on the *Classification And Regression Trees* (CART) algorithm, i.e., the rules are chosen to decrease the GINI impurity (see Appendix XII).

Figure 2: Fictional Example of Decision Tree for the ESG Category Rating Classification problem



So, to create a random forest, we apply a CART algorithm on the n sub-samples drawn from our initial dataset. Each model is creating a set of rules on the ESG-words frequency vectors to define the optimal rating of each company from the sub-sample. Then, by averaging the results of those n DT, we obtain a new model, a random forest, which is supposed to have a lower variance and potentially a better prediction power (Couronné et al., 2018).

As already mentioned, one of the great advantages of a RF algorithm is to create robust models with better out-of-sample capabilities. We can thus create a continuous score without the problems encountered by the linear regression in big data. However, this model is also by construction opaquer. It can therefore be problematic to use a RF in terms of interpretation of results and reproducibility.

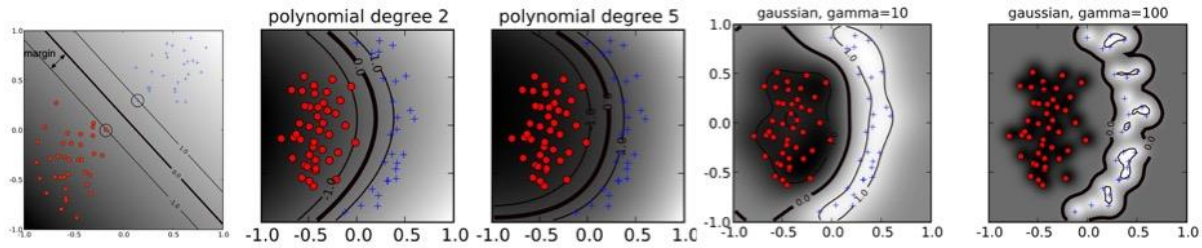
3.4.6. Support Vector Machine

Support Vector Machine (SVM) is with NB, one of the most often cited algorithm for Text classification. According to Kumar & Ravi (2016), SVM is even the most used algorithm in the financial literature on Text Mining for the period 2000-2016. This model owes its success to a proven performance, particularly in high dimensions (Ben-Hur & Weston, 2010; Miner et al.,

2012), but also, to its ability to manage both linear and non-linear data (Ben-Hur & Weston, 2010; Kumar & Ravi, 2016).

Concretely, SVM can help us discriminate among different categories of ESG ratings by building the hyperplane that best separate two classes (a new observation is then classified in function of its position with respect to the hyperplane). To find this optimal "separator", the algorithm relies on the notion of highest margin. The margin is equal to twice the Euclidean distance between the hyperplane and the nearest company from rating grade g ($y_i^{g,a} = +1$) and rating grade not g ($y_i^{g,a} = -1$), also known as the support vectors.

Figure 3: Illustration of the concepts of hyperplane, margin, and support vectors (the circled data points) for Linear (Ben-Hur & Weston, 2010 p. 5), Polynomial (p. 9) and Gaussian Kernels (p. 10)



For non-linear data, we can find this hyperplane by solving the following program (see details in Appendix XIII):

$$\min_{\boldsymbol{\alpha}} \sum_{i \in Corp} \alpha_i - \frac{1}{2} \sum_{i \in Corp} \sum_{k \in Corp} y_i^{g,a} y_k^{g,a} \alpha_i \alpha_k \left(\boldsymbol{w_i^{r,a}}\right)^T \boldsymbol{w_k^{r,a}}$$

Subject to $\sum_{i \in Corp} y_i^{g,a} \alpha_i = 0, 0 \leq \alpha_i \leq C$. To solve this SVM problem, one essentially need to compute the value of the dot product between $\left(\boldsymbol{w_i^{r,a}}\right)^T$ and $\boldsymbol{w_k^{r,a}}$. To do so, a powerful method is to rely on the Kernel trick, an alternative way of computing dot product by using kernel function (polynomial, gaussian, sigmoid, etc. in function of the type of data) (Ben-Hur & Weston, 2010).

Polynomial Kernel of degree d: $\left(\boldsymbol{w_i^{r,a}}\right)^T \boldsymbol{w_k^{r,a}} = K\left(\boldsymbol{w_i^{r,a}}; \boldsymbol{w_k^{r,a}}\right) = \left(\boldsymbol{\gamma}\left(\boldsymbol{w_i^{r,a}}\right)^T \boldsymbol{w_k^{r,a}} + \boldsymbol{cst}\right)^{\boldsymbol{d}}$

Gaussian Kernel of parameter γ: $\left(\boldsymbol{w_i^{r,a}}\right)^T \boldsymbol{w_k^{r,a}} = K\left(\boldsymbol{w_i^{r,a}}; \boldsymbol{w_k^{r,a}}\right) = \boldsymbol{e}^{-\boldsymbol{\gamma}\left(\left\|w_i^{r,a} - w_k^{r,a}\right\|\right)^2}$

As displayed by **Figure 3**, the parameters γ and the degree d of the polynomial kernel are influencing the flexibility of the SVM model to improve its prediction capabilities. To find the appropriate Kernel and parameters, the most common procedure is to apply a grid search on several parameters' values (Ben-Hur & Weston, 2010). Therefore, if this method can increase the accuracy of a classifier, it is also very important to note that it is computationally intensive to fit (many parameters, kernels, etc. to test and compare)

3.5. Step 3 – Analyzing Performance

In this section, we will look at the procedure implemented to assess the performance of the models described above, and thereby answer our research question. This performance analysis stage is divided into two steps, a very common procedure in comparative analysis Machine Learning and data mining tools (Bhattacharyya et al., 2011; Dougherty, 2013; Palmer & O'Connell, 2009; Shah et al., 2020): (1) Defining the optimal parameters of the rating models using cross-validation, and (2) Evaluating and comparing the performance on test samples.

3.5.1. Performance Metrics

To evaluate and compare the ability of Text Mining models to build accurate ESG ratings, we relied on a set of 7 different performance measures. Each measure is computed a first time on the different predictions of ESG Category ratings. Then, the process is repeated a second time for the more granular Grade ratings[6]. By comparing the accuracy of our scoring models for different level of granularities we have got the possibility to find a deeper insight on their ability to find the nuances of ESG performance.

3.5.1.1. Confusion Matrix analysis: Accuracy and Macro F1-score

To gauge the general performance of each scoring model we are relying on a set of threshold metrics that are commonly used for comparative analysis in the Machine Learning literature: the Accuracy score and the macro F1-score (Dougherty, 2013; Foody, 2009; Shah et al., 2020; Shahi et al., 2014). For the ESG ratings under letter formats $f \in \{g; cat\}$, the first performance indicator is quite intuitive as it is given by the percentage of correct predictions, i.e., the accuracy[7]:

$$\text{Accuracy}^{\text{f,r,a}} = \frac{1}{|Corp|} \sum_{\text{i}\in\text{Corp}} \mathbf{1}_{\{{y_i}^{f,a} = \hat{y}_i^{\,f,r,a}\}}$$

Where $\mathbf{1}_{\{{y_i}^{f,a} = \hat{y}_i^{\,f,r,a}\}}$ is the indicator who takes a value of 1 when the prediction of the rating under format f of company i based on report r is correct for year a (${y_i}^{f,a} = \hat{y}_i^{\,f,r,a}$) and 0 otherwise. This accuracy value has the advantage of offering a quick and easily interpretable vision of a classifier's performance. Yet when working on imbalanced datasets, relying solely on this this measure can generate misleading conclusions. For instance, for the year 2019, a

[6] For the ESG words frequency model, an additional format is also computed: Continuous ESG scores
[7] |Corp| refers to the number of company's name in the set of Corp

classification model on Category ratings could reach an accuracy of about 80% while completely ignoring class C and D. To complete our analysis, we therefore added an additional performance metrics: the macro F1-score. This measure is a composite indicator that summarizes both the precision (the percentage of companies from a class correctly assigned by the classifier) and the recall (the percentage of companies assigned to a class by the classifier that are actually belonging to this class) of classification models. For a given class g, the F1 score is obtained by computing the harmonic mean of both the precision and the recall (Dougherty, 2013; Foody, 2009; Shah et al., 2020; Shahi et al., 2014):

$$F1_g = 2 \times \frac{precision \times recall}{precision + recall}$$

For multi-label classification, the macro F1-score is given by the average of the F1-score of each class. Since an efficient classification model is supposed to combine both a high precision and a high recall, a macro F1-score is good when it has a value approaching 1.

3.5.1.2. Bias, Mean Absolute Error, and Correlation

To get a deeper understanding of the prediction capabilities of those models we wanted to get an insight on the average magnitude and sign of prediction errors. Traditionally, those insights can be obtained by the Mean Absolute Error and the Bias indicator. However, these are only applicable on continuous variables. So, before computing those indicators, we started by converting our letter-ratings (both Categories and Grade) into integer ratings (see Appendix X, e.g., a rating category A and D are becoming 4 and 1).

On the one hand, the bias indicator allowed us to see whether our predictions were over- or underestimating ESG ratings. On the other hand, the Mean Absolute error comes as an additional indicator of the precision of the different models:

$$\text{Bias}^{c,r,a} = \frac{\sum_{\text{i}\in\text{Corp}}\left(\hat{y}_i^{\,c,r,a} - y_i^{\,c,a}\right)}{|Corp|}$$

$$MAE^{c,r,a} = \frac{\sum_{\text{i}\in\text{Corp}}\left|\hat{y}_i^{\,c,r,a} - y_i^{\,c,a}\right|}{|Corp|}$$

For the ESG-words frequency model, the linear correlation is also computed to check for the strength of the relationship between the frequency of ESG words and the ESG performance:

$$\rho^{c,r,a} = \frac{\sum_{\text{i}\in\text{Corp}}[\,(y_i^{\,c,a} - \bar{y}^{c,a})(\hat{y}_i^{\,c,r,y} - \bar{\hat{y}}^{c,r,a})\,]}{\sqrt{\sum_{\text{i}\in\text{Corp}}(y_i^{\,c,a} - \bar{y}^{c,a})^2 \sum_{\text{i}\in\text{Corp}}(\hat{y}_i^{\,c,r,a} - \bar{\hat{y}}^{c,r,a})^2}}$$

3.5.1.3. Class Precision and Recall

Finally, we also calculated the macro precision and recall (average of all classes). A low precision indicates that this class is poorly identified by the scoring model, while a low recall indicates that the classifier tends to assign to many companies to a given class (Shah et al., 2020): $precision = \frac{TP}{TP+FP}$; $recall = \frac{TP}{TP+FN}$

Where TP is the number of companies correctly assigned to a class, FP is the number of companies assigned to a given class while not being part of it, and FN is the number of companies of a given class that were not assigned to it.

3.5.2. Fitting the optimal models

Before having the opportunity to test the performance of the models described above, we first needed to ensure that they had been fit with optimal parameters, i.e., parameters that should lead to the best out-of-sample performance (Dougherty, 2013). Otherwise, the results from the performance tests (step 2) might have underestimated the true prediction capabilities of our various approaches, and thereby distort the conclusions of this research. Therefore, here, we used the data from 2020 as a validation set to tune the different parameters.

In this research we considered two main categories of parameters, namely the number of features to keep in the model (or its complexity) and the choice of specific hyperparameters (see **Table 1**). The definition of the subset of features to keep corresponds to a Feature Selection task and its objective is to find a good balance between the bias and the variance of the model predictions (Kowshalya et al., 2019; Mohamad et al., 2021). This bias-variance tradeoff, also known as curse of dimensionality, comes from the observation that on the one hand reducing the bias is done by increasing complexity, while on the other hand increasing this complexity also increases the variance of the estimators. To be "optimal" in terms of prediction capabilities, a model should therefore be fitted on a certain subset of features selected to find the intersection between those two tendencies (Dougherty, 2013; Mohamad et al., 2021).

Table 4: List of parameters per rating model (a x indicates that this parameter is used by this specific model)

Model	Feature (p)	Distri (D)	Degree (d)	Gamma (γ)
Word-frequency				
MLR	x			
RF Regression	x			
LDA	x			
RF Classification	x			
NB	x	x		
SVM Polynomial	x		x	

SVM Gaussian	x			x

To select the optimal number of features, we relied on two techniques commonly used in Text Mining: Correlation Feature Selection and The Mutual information gain. Although there are other methods of feature selection, it has been proven that these two were able to significantly improve the performance of Text Mining models (Clarkson et al., 2020; Miner et al., 2012; Shiva Darshan & Jaidhar, 2018; Simeon & Hilderman, 2008). For instance, Shiva Darshan & Jaidhar (2018) compared the impact of several feature selection techniques on the classification of Portable Executable file as Malware or not. They found that a feature selection based on the Mutual Information gain could lead to the best accuracy. Another approach, the Correlation based feature selection, is also considered as being efficient for Text classification. Shahi et al. (2014) relied on this correlation factor to classify sections of CSR reports into various ESG topics arguing that this approach historically produced better results. Though there still exist many other alternatives to those metrics (e.g., Distinguishing Feature Selector, Categorical Proportional Difference, etc. (Shiva Darshan & Jaidhar, 2018)), in our research, we used the Pearson correlation coefficient (F-statistic from an Anova test) and Mutual information gain for the prediction of categorical ratings. Both indicators are measures of the strength of the relationship between two variables. In our case, the objective was to evaluate which were the ESG terms from our lexicon (our vector of features) that had a higher explanatory power, such that we could reduce the dimensionality of the model to relevant variables only.

Table 5: Grid of values visited by our algorithm for feature selection

Feature Selection – Grid Search									
Size	1	2	3	…	100	101	…	323	324

Concretely, to find the optimal number of features to keep, we wanted to evaluate the performance of models fitted with different number of features, each of those feature subsets containing the n features with the highest ANOVA F-statistic (or Mutual Information Gain). We tested a finite set of feature sizes using a grid-search with values ranging from 1 to 324. Then the grid-value that maximizes the performance of the model defined our optimal feature subset. Note that the performance of a model was evaluated based on a cross-validation approach (see below).

Besides feature selection, for some models (NB and SVM) we needed to define the value of their hyper-parameters. Again, for those we performed a grid-search as illustrated by **Table 6**.

Table 6: Grid of values for model specific parameters

Parameters Selection – Grid Search				
Degree	2	3	4	
γ	0.1	1	10	100
Distri	Gaussian	Bernoulli	Multinomial	

Both the feature selection and choice of parameters was made using a 10-fold cross validation approach. In other words, for various combinations of parameters, the performance of each model was computed using cross-validation, the optimal model being the one that maximizes the cross-validated performance. To obtain a cross-validated performance, we started by randomly splitting the training set into 10 subsets of equivalent length. Then iteratively, each of the subsets was used as a test sample for the model fitted on the 9 remaining subsets. At each iteration, an out-of-sample performance measure was computed (F1-score[8]). Finally, to obtain the cross-validated performance we took the average of those ten performance measures (Dougherty, 2013). Eventually, for each model we define 2 optimal combinations of parameters for each report type r, one that maximizes the cross-validated $F1^{cat,r,a}$ (the Accuracy on category ratings) and one that maximizes the cross-validated $F1^{g,r,a}$ (the accuracy on grade ratings) (see Appendix XIV for details on optimal parameters).

3.5.3. Testing the performance of optimal models

Once each of the models has been estimated with these optimal parameters, the last step of our evaluation process consists in the building of several performance measures. Simply, we use every of the different models defined above to make rating predictions for the year 2017 to 2020. Then, we computed the performance measures defined in the section Performance Metrics to evaluate their quality for every year and report type, measures that will serve as basics for comparison of models.

3.5.3.1. Comparing Models: Average, Standard Deviation and Statistical Tests

To facilitate the comparison between models, each of them is firstly characterized by the average and standard deviation of its prediction's accuracy, F1-score, bias, etc. By doing so, we allowed for a reduction of dimension and thereby an easier visualization and interpretation of the results. After this first step, we were able to draw a first conclusion on the general performance and stability of the different models. Yet, as explained by Demšar (2006) this

[8] Rather than the accuracy which is not a perfect performance indicator for training when working on imbalanced classes.

average comparison approach might not be the most robust way of comparing classifiers over multiple datasets, especially when the number of test sets is low. For instance, we could observe that the model with the highest average is also the one with the highest variance, hence it might not be the best performer. Therefore, as suggested by Dougherty (n.d.), we combined our average analysis with a set of statistical test designed to rank the performance of our scoring models in terms of accuracy, macro F1-score, bias and MAE. First, we applied a Kruskal-Wallis test, the non-parametric equivalent of the One-way ANOVA, to compare the mean accuracies of each model. Thanks to this test, we are questioning the null hypothesis that the mean accuracy of all our rating models is equivalent. Then, if the Kruskal-Wallis statistics leads to the conclusion that given our sample of accuracy results the null hypothesis must be rejected, we needed to conduct a post hoc test to compare the models mean pair-wise. In the context of classification, Dougherty (n.d.) recommends the use of Tukey's test. Finally, for the pairs for which a difference has been found (Tukey's test led to the rejection of the hypothesis that both models had an equivalent average), we can conclude that the model with the best average is indeed more performant. This series of tests is then repeated for the F1-score.

In the case of a disagreement between the F1-score and accuracy ranking (or the presence of no obvious difference) we perform a last set of tests to evaluate the hypothesis that model A has a different performance than model B given a set of average metrics (F1-score, Accuracy, Precision, Recall, MAE, and Bias). To do this we are relying on the Friedman test, the non-parametric equivalent of the Repeated measure ANOVA followed by the Nemenyi post hoc test to obtain a pairwise comparison (Demšar, 2006; Dougherty, 2013).

3.5.3.2. Comparing Disclosure type

The objective of this second set of performance analysis is to test for the presence of a difference between Text Mining based ratings drawn from different sources of ESG information. To do so, we started by identifying the source of disclosure used by the scoring model identified as the most performant in the previous analysis. This first result will serve as an initial answer to the question on the most relevant sources of ESG information (for Text Mining purposes). Then, to confirm this first result, we decided to compare the performance of models fitted on Annual reports to the one of models fitted on CSR, and on the combination of both. As for the general comparison of scoring approach, we perform this comparative analysis in two steps, an analysis of the averages and standard deviation followed by a ranking based on hypothesis tests.

Chapter 2: Analysis of the results

This second part is intended to give a general overview of the empirical research results to the reader. Hence, in the following pages, we will first describe the different features of our data sample, i.e., the distribution of ESG scores and sustainability disclosure. Thereafter, we are focusing on performance results given by various Text Mining based ratings, their accuracy, precision, etc. We begin with an analysis of the ESG-words frequency model, before deep-diving in the results of more complex machine-learning based models (linear, non-linear, and rule-based). Finally, we are concluding this part with a comparative analysis of the performance of various models based on different disclosure types.

1. Exploratory Analysis

1.1. Dependent variable – Refinitiv's ESG ratings

As previously mentioned, for this empirical research we used a sample of four years of Refinitiv's ESG ratings as a proxy for company's ESG performance. For the years 2017 to 2020, the samples were respectively including 396, 412, 419, and 414 ratings. In this section, we will briefly describe how those ratings are distributed, what are their main features and what it could imply for different Rating models and their results. As a reminder, we are testing the prediction ability of Text Mining models for ESG ratings under two distinct levels of granularity to evaluate to what extent they can predict nuances of ESG performance. So, our sample contains both a categorical variable with 4 distinct levels {D, C, B, A} and a second with 12 levels {D-, D, D+, C-, C, C+, B-, B, B+, A-, A, A+} (see Appendix X).

	Summary Statistics							
	Count	Mean	Std	Min	25%	50%	75%	Max
2017	396	66,68	18,90	0,38	56,72	70,59	81,27	94,35
2018	412	66,77	18,72	0,38	56,72	70,93	81,24	94,35
2019	419	66,72	18,71	0,38	56,68	70,74	81,28	94,35
2020	414	66,69	18,66	0,38	56,79	70,77	81,20	94,35

Table 7: Summary Statistics of Refinitiv's ESG scores (from 0 to 100) for the years 2017 to 2020

Table 7 offers some basic statistics regarding the distribution of the ESG ratings under a continuous format. Hence, we can note that for every year of our sample the average Refinitiv's score remains relatively steady with an average score of around 65 points.

A second element to highlight in this exploratory analysis is the value of the 1st quartile, a value that is displayed by **Table 7**. For all four years, the minimum 1st quartile score is equal to 56.68,

a score which corresponds to rating of category B-, i.e., 75 % of our sample is made of companies with a "good relative ESG performance and above- average degree of transparency in reporting material ESG data publicly" (Refinitiv, 2021, p. 7). This situation can be explained by the fact that our sample was built based on MSCI world index, with large-cap companies operating in developed market. Hence, a sample of companies that operate under higher regulatory and market pressure (but also with higher means) to disclose material information on sustainability which results in above-average ESG ratings. For instance, European companies (26% of the sample) who must comply with strict rules on disclosure have the highest average rating of all the regions in our sample, 74.91 (see Appendix VII).

Table 8: Repartition of ESG Grade Ratings for the full sample (A & A+C) (left), Comparison of the average ESG score of companies from sample C and sample A (right)

	2017	2018	2019	2020
A+	12	12	12	12
A	62	63	65	62
A-	80	86	86	86
B+	75	80	80	81
B	52	52	54	53
B-	37	39	42	41
C+	41	41	41	41
C	17	19	18	17
C-	9	9	10	10
D+	2	2	2	2
D	0	0	0	0
D-	9	9	9	9
Total	**396**	**412**	**419**	**414**

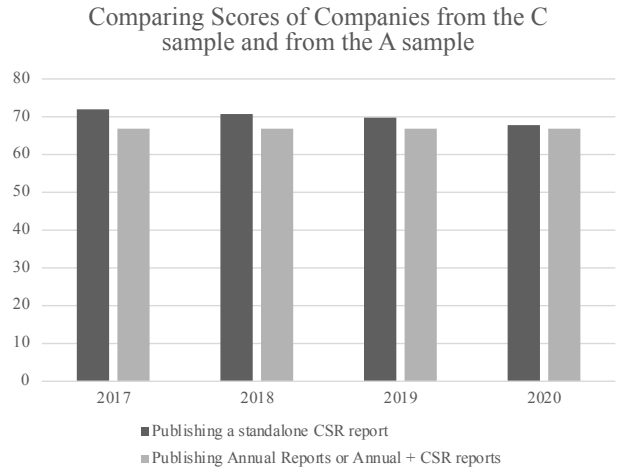


Table 8 gives a similar overview but for ratings taken in their categorical formats, i.e., the two letter scores. The observations are the same as above. Most of the companies have a rating higher than B with a maximum of 86 and 81 companies for categories A- and B+ in 2020. Besides, we can also note that over the period of 2017 to 2020 there is very few changes in the categories repartition. In fact, among the 396 companies already listed in the 2017 sample none have changed of rating group in 4 years. To be efficient, models should therefore be able to predict this relative stability in the ratings despite the publication of new reports with most likely some lexical differences. On this same figure, we can also note that on average, firms publishing an annual report tend to be better ESG performers. This further confirms the observation made by Clarkson et al. (2020) or Verbeeten et al. (2016) on the positive relationship between the publication of standalone CSR reports and ESG performance.

However, this relationship seems to be less and less clear over the years. In 2020, the difference between the two averages is reduced to 1.13 (4.12 less than in 2017).

1.2. Independent Variables – Company Disclosure

To predict ESG ratings, all the models of this research are relying on the frequency of ESG words used by a company in its official disclosure $X^{r,a}$. In this section, we are looking at some basic results and analysis that can be derived from those explanatory variables. From the initial list of ESG words mentioned, already mentioned in the methodology section, only 324 are used by at least one CSR or Annual reports from our sample (see Appendix XV). Therefore, we have a list of 324 explanatory variables of which 166 are related to Governance, 46 to Environment, and 112 to Social issues (see 3.3.2 Bag-of-ESG-words representation).

1.2.1. Statistical Overview

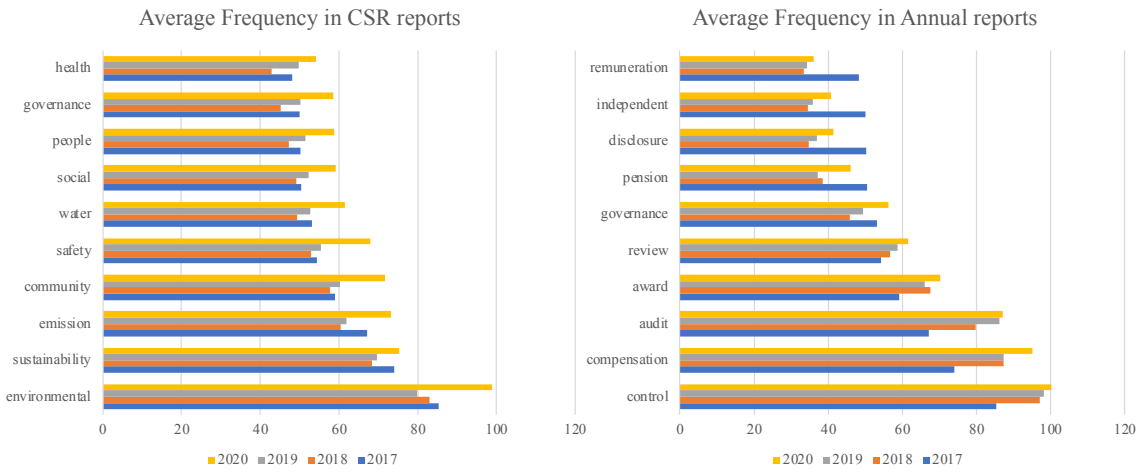


Figure 4: 10 most frequent ESG words in CSR (left) and Annual Report (right) for year 2017-2020

On average an ESG words appears 13.57 times in the disclosure of a company for the year 2020, 11.51 for 2019, 10.73 for 2018, and 9.78 for 2017 (see Appendix XVI). So, we observed that since 2017 companies have tended to use more ESG words in their various reports (Annual and CSR). **Figure 4** gives an overview of the ESG words that are the most frequent in our sample for the years 2017 to 2020. If the three generic words, Environmental, Social and Governance, are obviously three of the most frequent words in CSR reports we see nevertheless some more specific problems appearing in this list. We can for instance see Environmental and Social preoccupations linked to climate change (emission, climate, carbon) or health and safety (health, safety, people). Note that Annual Reports tend to be more governance oriented with the most frequent words being all associated to the G dimension.

We can also note that on average, a company will use a relatively wide lexicon, with 142 different ESG words in its Annual report and 151 in its CSR Reports. As displayed by **Table 9**, in general, the higher the ESG rating, the higher the richness of the lexicon. Yet, we can also note that companies with very bad ESG performance (rating of category D) do not follow this trend. Contrary to what could be expected, those firms are using ESG dictionaries that are as rich, especially in terms of Governance vocabulary, as the one of high performers.

Table 9: Average Length of ESG dictionary of Annual and CSR reports for different years and rating categories

	Annual Reports				*CSR Reports*			
Year	**2017**	**2018**	**2019**	**2020**	**2017**	**2018**	**2019**	**2020**
Global	**130,4**	**133,92**	**136,3**	**142**	**141,23**	**142,6**	**143,62**	**151,1**
A	146,16	148,73	152,22	159,99	152,29	155,34	155,95	168,08
B	122,59	128,21	129,01	132,45	133,86	135,78	138,36	144,54
C	110,10	111,26	115,43	122,43	102,61	107,96	113,21	123,44
D	149,73	148,09	147,82	153,09	145,50	139,43	133,67	147,78

Therefore, we can observe that **Table 9** further confirms the conclusions of past studies on that topic (Bacha & Ajina, 2019; Clarkson et al., 2020; Verbeeten et al., 2016). Clarkson et al. (2020) found that in a sample of 2398 CSR reports good ESG performers tended to use more rich and complex vocabularies (a more advanced level of writing) to describe their activities. Similarly, Bacha & Ajina (2019) and Nazari et al. (2017) identified the existence of a positive link between readability of CSR disclosure and ESG performance.

1.2.2. Relationship between individual words and ESG ratings

After having reviewed some univariate statistics on the explanatory variables, we are now investigating the relationship between a given ESG word and the ESG rating under his continuous, category or grade format. By doing so, we should already get an idea of which variables have a higher predictive power, and thereby anticipate some potential results of the next section. **Figure 5** offers a visual representation of the linear relationship (the correlation coefficient) between the different column vectors of the ESG-words frequency matrix and Refinitiv's ESG continuous scores. It can be observed that regardless of the type of disclosure (CSR, Annual or both) or the year, correlation coefficients are relatively small, with value ranging from -0.2 and 0.3. So, on the one hand we found that for our sample of ratings, no ESG word of our dictionary had a sufficiently strong link with the outcome variable to deliver accurate prediction by itself. But on the other hand, we also show that an important number of

features do not have any predictive power at all (at least 25% of ESG words have a correlation $\rho \in [-0{,}1; 0{,}1\ ]$). Hence, this explains why an important reduction of dimensionality happened during the feature selection process (see Appendix XIV). A last important element uncovered by **Figure 5** is the presence of negative correlation in those graphs. In short, those negative relationships are suggesting that the presence of some ESG words is not a sign of good ESG performance. For instance, observing a higher frequency of "brother" in an Annual Report is not associated to better ESG performance ($\rho = -0.1751$ in 2020). For words like this one, we can therefore expect our models, e.g., the Multiple Linear regression, to give it a negative weight.

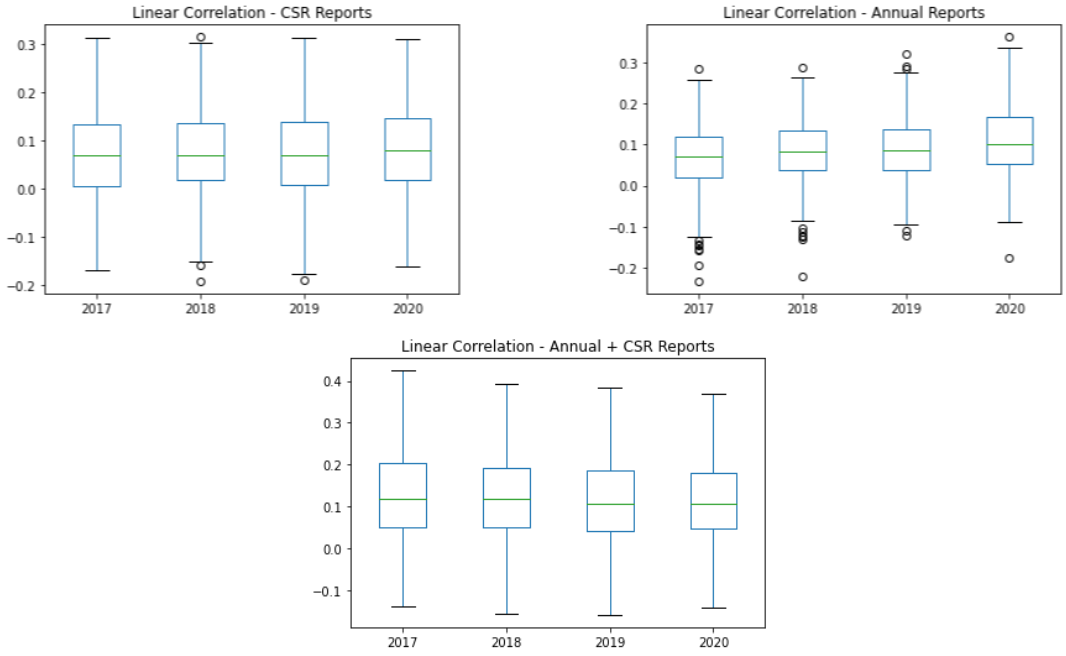


Figure 5: Boxplots of Linear Correlation between the frequency of ESG words (in Annual Reports $X^{A,y}$ (top left), CSR Reports $X^{C,y}$ (top right) and both $X^{A+C,y}$ (bottom)) and Refinitiv's ESG continuous scores $Y^{C,y}$ for year a 2017 to 2020

Besides, **Table 10** shows the ESG words with the most predictive power according to the linear correlation coefficient (maximum and minimum ρ). Thus, for the year 2020, we can see that "climate" or "pandemic" were most often used by good ESG performers. In the same spirit, Appendix XVI illustrates the 10 ESG words with the highest relationship to Refinitiv's rating based on Pearson Correlation and ANOVA tests, i.e., the criteria used for the feature selection. In both cases, those words with a high link to ESG performance should be included in the parameters of our different models.

Table 10: ESG Words with Maximum and Minimum Correlation to ESG ratings for every year and disclosure type

		CSR Reports			Annual Reports			Annual+CSR Reports	
Year		Word	Correlation		Word	Correlation		Word	Correlation
2017	min	alcohol:	-0,1701	min	brother:	-0,2326	min	alcohol:	-0,1402
	max	diversity:	0,3129	max	sustainable:	0,2826	max	diversity:	0,4247
2018	min	donates:	-0,1923	min	brother:	-0,2212	min	brother:	-0,1560
	max	transparency:	0,3143	max	transparency:	0,2873	max	diversity:	0,3917
2019	min	duly:	-0,1893	min	brother:	-0,1219	min	consent:	-0,1598
	max	climate:	0,3118	max	transparency:	0,3207	max	transparency:	0,3832
2020	min	epidemic:	-0,1616	min	brother:	-0,1751	min	epidemic:	-0,1427
	max	climate:	0,3094	max	pandemic	0,3614	max	climate:	0,3694

Table 10 and Appendix XVII we can also see that words with high explanatory power are not constant over time, except for Annual Reports. This suggests that the ESG lexicon of high performing companies is not constant over time and could therefore undermine the performance of Text Mining based scoring models.

2. Performance Analysis – Comparing Models

As explained in the methodology, to evaluate and compare the ability of Text Mining models to build accurate ESG ratings, we relied on a set of 7 different performance measures, each being computed for all four years, and for three different types of disclosure. Comparison of those metrics is made based on a combination of average analysis and hypothesis testing.

In this section, we are describing how each model performed on those metrics and statistical tests, and what this implies for our research question. We start by describing the performance of ESG-words frequency models whose results will serve as a form of benchmark for the following approaches. Then, we investigate the precision of a set of classification models in the objective of finding an appropriate ranking of Text mining based ESG ratings. Finally, we compare the relevance of the three different source of CSR disclosures for Text Mining purposes.

2.1. Word-frequency model

Table 11 reports the different performance results obtained when applying the ESG words-frequency model on company's disclosure from 2017 to 2020. As a reminder, the scores generated by this first scoring approach are obtained by summing the frequency of the 324 ESG-related terms in Annual and/or CSR reports.

In terms of general accuracy, those ESG ratings were found to be rather poor. For the year 2020, the prediction of ESG letter-categories happens to be correct for maximum 17.39% of the companies. Given the low value of the macro F1 indicator, it is suggested that nor the precision nor recall of this model could compensate for its bad accuracy performance. Furthermore, there is little to no difference between the models using the different types of disclosure. Still, we can highlight that the combination of Annual and CSR reports is apparently the most "appropriate" source of information for this rating approach as it delivers the best combination of average Accuracy and F1 score.

Table 11: General Accuracy metrics for the Word Frequency Model. From left to right: Accuracy and Macro F1 of the ESG Categories and Grade predictions

ESG words-Frequency												
Metric	Accuracy ($Y^{Cat,r,y}$ vs. $\hat{Y}^{Cat,r,y}$)			Macro F1 ($Y^{Cat,r,y}$ vs. $\hat{Y}^{Cat,r,y}$)			Accuracy ($Y^{g,r,y}$ vs. $\hat{Y}^{g,r,y}$)			Macro F1 ($Y^{g,r,y}$ vs. $\hat{Y}^{g,r,y}$)		
Year	**A**	**C**	**A+C**	**A**	**C**	**A+C**	**A**	**C**	**A+C**	**A**	**C**	**A+C**
2020	**15,46%**	**15,83%**	**17,39%**	**0,1605**	**0,1550**	**0,1691**	**4,83%**	**3,61%**	**3,86%**	**0,048309**	**0,036111**	**0,038647**
2019	20,53%	14,45%	18,14%	0,1757	0,1399	0,1759	6,92%	3,83%	6,21%	0,064439	0,044248	0,062053
2018	18,45%	15,74%	20,87%	0,1870	0,1299	0,1999	6,31%	3,28%	6,55%	0,067961	0,045902	0,065534
2017	22,22%	17,67%	22,22%	0,1815	0,1147	0,2137	4,29%	4,51%	6,82%	0,042929	0,033835	0,068182
Average	**17,50%**	**14,37%**	**19,66%**	**0,1761**	**0,1349**	**0,1897**	**5,59%**	**4,00%**	**5,86%**	**0,0608**	**0,0359**	**0,0579**
Variance	*0,022%*	*0,016%*	*0,052%*	*0,00013*	*0,00029*	*0,00043*	*1,217%*	*0,594%*	*1,354%*	*0,0110*	*0,0056*	*0,0129*

Unsurprisingly, **Table 11** reports results that are even worse when asking the model to provide more nuanced ratings. Accuracy rates (and F1 scores) are indeed dramatically dropping below the 6% (and 0.1) for prediction of letter-grades in 2020. As displayed in **Table 11**, performance results are showing that the word-frequency model has relatively stable performances with a variance that remains very low, particularly for category predictions. Therefore, it confirms that given the absence of any parameters, this scoring model has the advantage of not being keen to overfitting and can maintain a stable out-of-sample performance. Besides, this stability also suggests that the ESG vocabulary of companies is relatively stable over the period under investigation, even though the decrease in accuracy since 2017 might indicate an increasing similarity between the lexicons of Good and Bad performers.

Another set of interesting results are given by the bias and MAE measures (see **Table 12**). We can for instance see from the Bias that ESG categories and grades are under-estimated by this scoring model (around 2 rating classes below their actual performance on average) whatever the type of disclosure or year chosen. An issue that could be explained by a problem of too low intercept. However, in its current form, this suggests that the ESG words frequency model has

mainly trouble at identifying good ESG performers, i.e., those companies are differentiating themselves from others by other means than having a higher frequency of ESG terms.

Table 12: MAE, Correlation, Bias, Macro Precision and Recall of the ESG words Frequency model for Category (above) and Grade (below) Predictions

	Category														
	Annual					CSR					Annual & CSR				
	MAE	COR	BIAS	PRECISION	RECALL	MAE	COR	BIAS	PRECISION	RECALL	MAE	COR	BIAS	PRECISION	RECALL
2017	2,040	0,252	-2,030	0,292	0,195	2,030	0,202	-1,944	0,228	0,184	1,940	0,374	-2,006	0,319	0,246
2018	2,090	0,245	-2,079	0,315	0,202	2,080	0,261	-1,959	0,259	0,185	1,960	0,363	-2,043	0,303	0,260
2019	2,120	0,239	-2,083	0,293	0,201	2,080	0,260	-2,024	0,270	0,203	2,020	0,342	-2,077	0,296	0,216
2020	2,140	0,251	-2,039	0,265	0,200	2,040	0,343	-2,024	0,273	0,244	2,020	0,343	0,001	0,280	0,217
Average	**2,098**	**0,247**	**-2,058**	**0,291**	**0,200**	**2,058**	**0,266**	**-1,988**	**0,257**	**0,204**	**1,985**	**0,356**	**-1,531**	**0,299**	**0,235**
Variance	*0,00189*	*0,00004*	*0,00073*	*0,00042*	*0,00001*	*0,00069*	*0,00334*	*0,00178*	*0,00043*	*0,00079*	*0,00170*	*0,00024*	*1,04409*	*0,00027*	*0,00049*

	Grade														
	Annual					CSR					Annual & CSR				
	MAE	COR	BIAS	PRECISION	RECALL	MAE	COR	BIAS	PRECISION	RECALL	MAE	COR	BIAS	PRECISION	RECALL
2017	7,040	0,224	-6,992	0,067	0,072	6,990	0,185	-6,793	0,073	0,068	6,990	0,386	-6,943	0,131	0,082
2018	7,180	0,252	-7,170	0,136	0,079	7,170	0,217	-6,859	0,105	0,042	7,170	0,332	-7,071	0,119	0,128
2019	7,270	0,299	-7,242	0,129	0,114	7,240	0,278	-6,986	0,094	0,054	7,240	0,360	-7,167	0,117	0,081
2020	7,340	0,267	-7,164	0,107	0,069	7,160	0,341	-6,925	0,091	0,052	7,160	0,356	-7,142	0,073	0,087
Average	**7,208**	**0,260**	**-7,142**	**0,110**	**0,084**	**7,140**	**0,255**	**-6,891**	**0,091**	**0,054**	**7,140**	**0,358**	**-7,081**	**0,110**	**0,094**
Variance	*0,01676*	*0,00095*	*0,01121*	*0,00095*	*0,00044*	*0,01127*	*0,00474*	*0,00692*	*0,00018*	*0,00012*	*0,01127*	*0,00050*	*0,01011*	*0,00064*	*0,00050*

To sum up, this first set of results are suggesting that despite its obvious attract and simplicity, relying solely on the frequency of ESG words is not sufficient to measure the ESG performance of a firm. The main issue of this approach being that it tends to underestimate the performance of companies and is generally unable to evaluate the nuances of ESG performance. This observation is in line with findings made during the exploratory analysis. Nevertheless, the results do not indicate that there is no link between these two variables. Although weak, there is still a linear correlation between this Text Mining based rating and Refinitiv's one (average of minimum 0.247 for Annual Reports based ratings). Therefore, our study shows similar results to the one of Takuya et al. (2020) and Clarkson et al. (2020): Higher frequency of ESG terms in sustainability disclosure is associated to higher ESG performance (especially when extracted from a combination of Annual and CSR reports). Yet, those same results lead us to the conclusion that this relationship is not strong enough to consider the frequency of ESG words in sustainability disclosure as a good proxy for the ESG performance of a company.

2.2. Other Text Mining Models

In this subsection, we investigate how the predictions of different classification models performed in terms of accuracy, F1-score, bias, Mean Absolute Error, etc. compared to the ESG-words-frequency model and compared to each other. In the end we want to rely on those different metrics to build a ranking of Text Mining Based ratings, hence we want to find which approaches can be seen as optimal, and how precise they can be. For ease of comprehension, we are making a first set of analysis on the predictions of Category ratings before moving to the more granular Grade ratings predictions.

2.2.1. Category Ratings

Overall performances of scoring models on Category ratings are presented by the boxplots of Accuracies and F1-score depicted on **Figure 6, Figure 7, Figure 10,** and **Figure 11.** Results on other metrics can be found in the performance summary of **Table 13**. More precisely, it is the average and variance of the bias, MAE, macro precision and recall for the different models and disclosure type that can be seen in this table. Detailed versions (values of indicators for every test sample and every model) of this summary are listed in Appendix XVIII.

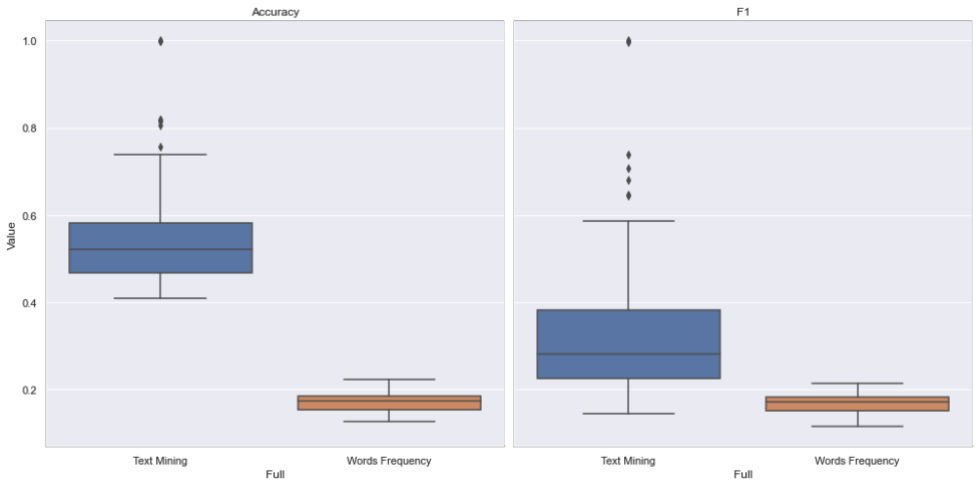


Figure 6: Boxplots of Accuracy and Macro F1 score of Text Mining models vs. ESG words Frequency

From those results, a first observation can be made regarding the performance of Text Mining based ratings in general. If we take the global average of results regardless of the scoring model or disclosure type, we can note that a rating built with Text Mining has a mean accuracy of 55.43% and a mean F1-score of 0.3414 (see **Table 19**). Therefore, from this point of view, it suggests that the relationship between ESG performance and ESG words frequency is better approximated by a linear, non-linear, or rule-based classification model than by a simple sum (see **Figure 6**). Even though we found that their average variance is higher, the smallest accuracy of any classification model is still higher than the best performance achieved by the ESG-words-frequency model on Category ratings (see **Figure 6** and **Figure 7**). Yet, this high variance results are also showing that not every Text Mining model can deliver satisfying results in terms of prediction of ESG ratings. Indeed, with an average accuracy of 55.43% and a standard deviation of 8% it can be concluded that generally those machine-led predictions were relatively poor and at best moderate. This conclusion is further confirmed by the analysis of the macro F1 score. With an average below the 0.35, it can be considered that most of our rating models are struggling in terms of precision (0.45 on average) and recall (0.38 on average).

Moreover, from the perspective of this measure, Text Mining based ratings are closer to the performance of the ESG words frequency model.

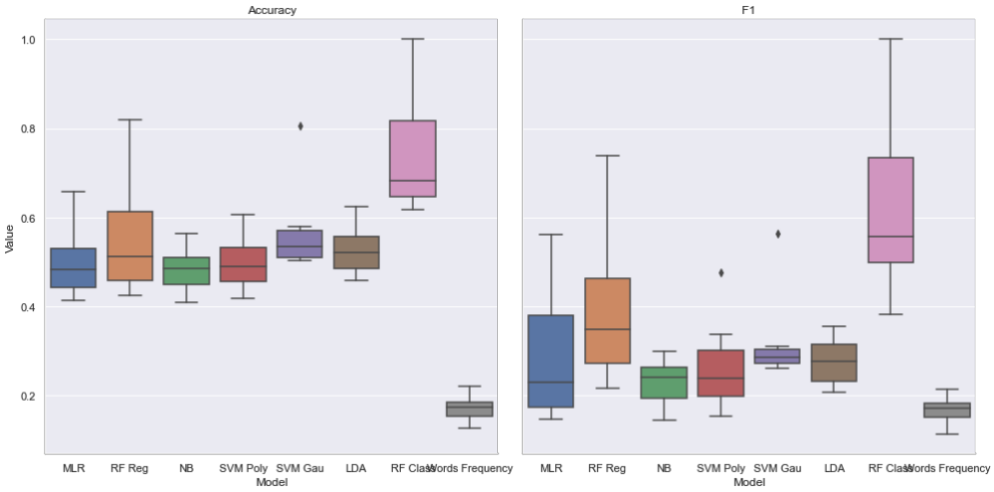


Figure 7: Comparison of different rating models performance in terms of Accuracy and Macro F1. Data from 2017 to 2020 for Annual, CSR and Annual & CSR reports

Nevertheless, from Figure 7 it can be observed that some Text Mining approaches have the potential to deliver highly accurate ESG ratings. For instance, both in terms of Accuracy and F1-score, it appears that the Random Forest Classier outperforms its peers, and this regardless of the type of reports from which it draws its predictions. On average it can reach 75% of accuracy (0.64 of F1-score) with a maximum of 75.66% for the Annual & CSR reports from 2019 (see **Table 20** and **Table 24**). Furthermore, it can also be observed from **Table 21** and **Table 25** that for 2020 the RF classifier delivered performance metrics near the perfection (around 100% accuracy and 1 of F1-score). Compared to out-of-sample performances, this represent a difference of at least 25% in accuracy which indicates a form of overfitting of the model, i.e., the words identified by RF as being discriminatory in 2020 are not equally discriminant in the other investigated periods. In fact, since we found this decrease in performance to happen for every model, it suggests that terms allowing for an efficient differentiation of ESG performance between companies are varying over time.

Behind this top performing model, differences between rating approaches are graphically more difficult to spot. Indeed, we can see from **Figure 7** that the 6 remaining models have values of accuracy and F1-score oscillating respectively between 40% and 80%, and between 0.05 and 0.75. Some models like the Random Forest Regression have on average a higher accuracy than the NB or MLR, but have at the same time a higher variance, while others like the SVM with Gaussian kernel tend to outperform their peers on average Accuracy but not on average F1.

To clarify this situation, it is therefore best to rely on the conclusions of the different statistical tests reported in **Table 29** and **Table 30**. Firstly, Kurksal-Wally's tests on accuracy suggest that we can reject the null hypothesis that there is no difference between the mean Accuracies of all our models on those metrics (Chi-Square = 56.584, $p < .05$, df = 7). Yet, the set of p-values derived from the pairwise Tukey's tests are indicating that from those models only the average performance in terms of Accuracy of the RF class and the ESG words frequency model are significantly different from the others (see **Table 29**). Hence, we can note that for the RF reg, MLR, SVM polynomial, SVM gaussian, NB, and LDA this first testing procedure did not lead to a clear and undebatable conclusion on Accuracy. Indeed, given our data the tests led to the conclusion that one could not reject the hypothesis that those models were similar in accuracy at a level of $\alpha = .05$. In terms of F1-score, once again the tests are suggesting that at least one pair of models are different in mean (Chi-Square = 51.776, $p < .05$, df = 7). We observe from Tukey's post hoc tests that it is the RF class model that has the highest performance, but for the other models, including ESG words frequency, we cannot say at a certainty level of .05 that they are different in average F1-score. Nevertheless, we observe that the RF reg has a significantly higher mean F1-score than the NB ($p < .05$) and Words frequency ($p < .05$).

To sum up, from the comparison of the different scoring approach on general performance metrics, we found that the RF classifier has significantly better performance than other classification tools. For the rest, on the one hand we observed that Words frequency had the worst performance in terms of accuracy but not in F1 score. On the other hand, it has been showed that all other models were significantly similar in performance, with a small advantage for the RF regression algorithm from the perspective of F1-scores. Note that both of those conclusions are valid for comparisons of models on F1-score and accuracy separately. When comparing models on all performance metrics combined, we must inspect the results from the Friedman and Niemeny posthoc tests (see **Table 31**). These are indicating that there is little to no difference between the various approaches when we compare them on the 7 metrics. Nervetheless, the Friedmann tests suggests that all models are not equal in average performance (Chi-Square = 31.952, $p < .05$, df = 6) but the Niemeny tests are returning p-values above the significance threshold of .05 for every pair of models except the RF class vs. Words Frequency, RF class vs. NB, RF reg vs. Words Frequency, and SVM Gaussian vs. Words Frequency. Hence, we have that the RF class, RF Reg and SVM with Gaussian kernel are significantly more performant at predicting ESG ratings than the two other above-mentioned models.

Regarding other performance metrics such as the Bias or the Mean Absolute Error, **Table 13** shows indeed that there are little variations among different rating approaches. Once again, the machine-led scoring approaches tend to underestimate company's ESG performance. This negative bias is however more contained then in the case of the ESG words-frequency model (less than a class below actual rating). Among the different models, it seems that the MLR and RF reg are more error prone than their peers (wrong by 1 category too low on average), while SVM with Gaussian Kernel, the LDA and the RF classifier have the less biased rating predictions with MAE of around 0,5 class error. Concerning linear correlation both RF models are showing better results with a reasonable coefficient slightly above 0.5.

Table 13: Average and Variance of MAE, Correlation, Bias, Precision, and Recall for each scoring model (Category ratings)

		Category (regardless of disclosure)								
Category		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
MAE	**Average**	0,93	1,00	0,55	0,53	0,49	0,52	0,28	0,61	1,63
	Variance	*0,00417*	*0,01482*	*0,00126*	*0,00376*	*0,00261*	*0,00200*	*0,03580*	*0,00920*	*0,68722*
COR	**Average**	0,44	0,67	0,20	0,25	0,31	0,27	0,65	0,40	0,34
	Variance	*0,00557*	*0,04411*	*0,00471*	*0,01794*	*0,00580*	*0,00831*	*0,05780*	*0,02061*	*0,01472*
BIAS	**Average**	-0,93	-1,00	-0,11	-0,13	0,07	-0,05	-0,03	-0,31	-1,53
	Variance	*0,00417*	*0,01504*	*0,00726*	*0,00687*	*0,00193*	*0,00836*	*0,00167*	*0,00647*	*1,04409*
PRECISION	**Average**	0,41	0,57	0,28	0,43	0,34	0,32	0,77	0,45	0,35
	Variance	*0,00473*	*0,04161*	*0,00203*	*0,01389*	*0,01017*	*0,00050*	*0,03253*	*0,01507*	*0,01848*
RECALL	**Average**	0,30	0,43	0,29	0,31	0,34	0,32	0,62	0,37	0,28
	Variance	*0,00007*	*0,02412*	*0,00051*	*0,00225*	*0,00211*	*0,00101*	*0,06405*	*0,01345*	*0,02051*

2.2.2. Grade Ratings

Regarding the performance of models on the generation of ESG grade ratings, results obtained during this research are reported jointly by **Figure 8**, **Figure 9**, **Figure 12**, **Figure 13** and **Table 14** (in Appendix XVIII for the detailed values). In a similar way as for the ESG words frequency model, **Figure 8** shows that the Accuracy and F1 score are strongly dropping compared to Category predictions. With an average accuracy of 33.48% and an average Macro F1 score of 0.2327, our results are thus suggesting that complex classification model are also failing to provide effective predictions of ESG performance at a more granular level (see **Table 32**). Yet, as for category ratings, the application of those models allowed for an improvement of both accuracy and F1-score compared to the simple sum provided by the ESG words-frequency rating. In terms of Bias and MAE, **Table 14** shows that on average the different rating approaches are more biased (-0.97) and are making errors of a bigger magnitude (2.09) than for category ratings. Yet this change remains reasonable given the increase in the number of classes (4 to 12). Besides, when inspecting the values by model, it can be observed that only the MLR and RF reg have seen a strong decrease in their Bias metric (see **Table 14**).

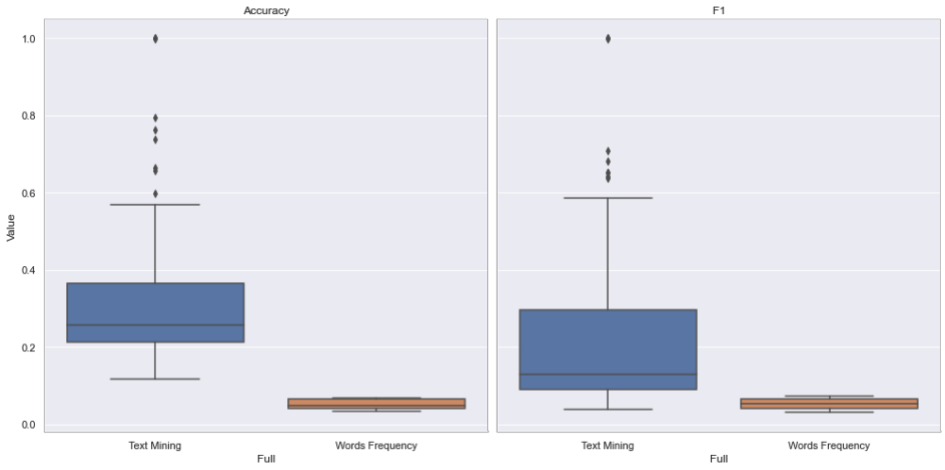


Figure 8: Boxplot of Accuracy and Macro F1 score of all Text Mining models vs. ESG words Frequency for Grade Ratings

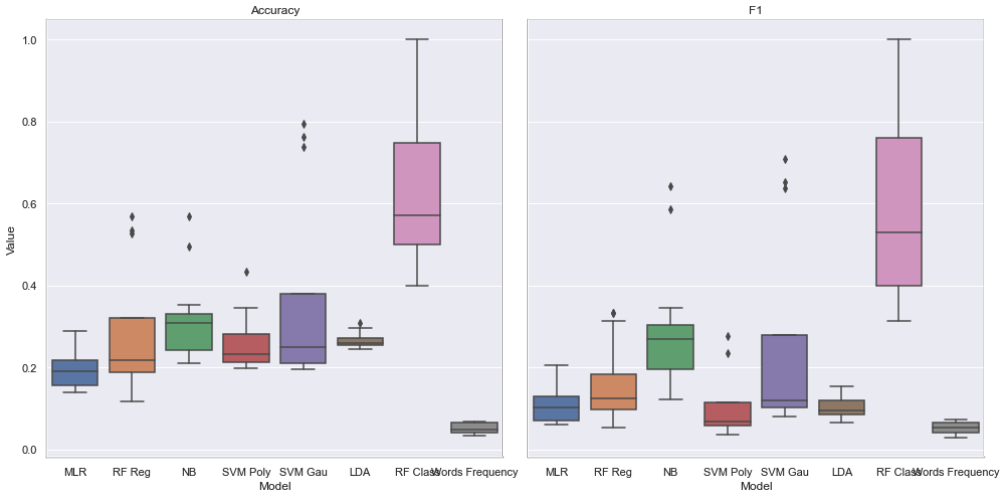


Figure 9: Comparison of different rating models performance in terms of Accuracy and Macro F1. Data from 2017 to 2020 for Annual, CSR and Annual & CSR reports

Furthermore, boxplots of **Figure 9** show that for both Categories and Grades ratings the performance hierarchy of Text Mining scoring models are similar. The RF class model has demonstrated to be the best performer with reasonable (though way lower than for categories) results in average accuracy (65%) and F1 score (0.61) (see **Table 21** and **Table 25**). Behind, it seems that the Naïve Bayes algorithm could have a better capacity at discriminating companies in terms of more nuanced ESG performance classes, especially given its F1 score (see Figure 9). Nevertheless, with most rating approach scoring between 20 and 40% of accuracy, those differences are remaining weak (see **Table 21**).

Statistical tests on the significance of those differences are coming as a confirmation of our initial observations. On the one hand, Kruskall-Wallis tests are confirming that there is at least on pair of models for which the mean Accuracy (Chi-Square = 59.728, $p < .05$, $df = 7$) and the mean F1-score (Chi-Square = 61.75, $p < .05$, $df = 7$) are significantly different (see **Table 42**

and **Table 43**). The subsequent Tukey's post hoc tests on the other hand are indicating a significant difference in accuracy only between RF class and every other model, and between Words Frequency and every other model except MLR. Yet, results in terms of F1-score are leading to other conclusions. The RF classifier is still found to be better on average than all other options, but Words Frequency is only less performant than the NB, the RF class, and the SVM with Gaussian Kernel. In addition, we found that the Naïve Bayes was significantly better in mean F1-score than the SVM with Polynomial kernel, the LDA, and the MLR. Therefore, we can conclude that behind the RF classifier, it is the Naïve Bayes model that should be preferred for delivering precise predictions of ESG grade ratings.

Furthermore, Friedmann and Niemeny tests reported on **Table 44** are showing that when considering every performance metric, differences exist (Chi-Square = 28.762, $p < .05$, $df = 6$) but are only significant for the pairs Word Frequency vs. RF class ($p < .05$), Words Frequency vs. NB ($p < .05$), Words Frequency vs. SVM Gaussian ($p < .05$), and MLR vs. RF class ($p < .05$). For remaining models, given the pairwise Niemeny statistics we fail to reject the Null hypothesis which suggests that most Text Mining based models are equally performant in terms of average MAE, Bias, Precision, Recall, F1-score, and Accuracy. The global average of those metrics can therefore serve as a good indicator of the general ability of Text Mining models at predicting ESG rating in their Grade format.

Table 14: Average and Variance of MAE, Correlation, Bias, Precision, and Recall for each scoring model (Grade ratings)

		Grade (regardless of disclosure)								
Grade		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
MAE	**Average**	3,77	3,79	1,89	1,51	1,37	1,50	0,78	2,09	7,16
	Variance	*0,05132*	*0,06742*	*0,10670*	*0,01347*	*0,28228*	*0,00031*	*0,30872*	*0,11860*	*0,01179*
COR	**Average**	0,45	0,69	0,39	0,15	0,33	0,26	0,69	0,42	0,29
	Variance	*0,00771*	*0,03839*	*0,01208*	*0,03353*	*0,09250*	*0,00312*	*0,05581*	*0,03474*	*0,00086*
BIAS	**Average**	-3,77	-3,86	-0,88	0,47	0,45	0,76	0,05	-0,97	-7,08
	Variance	*0,05097*	*0,07673*	*0,04212*	*0,02596*	*0,00946*	*0,02470*	*0,00350*	*0,03335*	*0,01011*
PRECISION	**Average**	0,15	0,18	0,33	0,24	0,30	0,11	0,75	0,30	0,10
	Variance	*0,00042*	*0,01040*	*0,01348*	*0,03116*	*0,09187*	*0,00061*	*0,04317*	*0,02730*	*0,00024*
RECALL	**Average**	0,14	0,19	0,37	0,14	0,26	0,14	0,58	0,26	0,08
	Variance	*0,00046*	*0,00984*	*0,01871*	*0,00232*	*0,06172*	*0,00079*	*0,08755*	*0,02591*	*0,00005*

2.3. Conclusion

Before moving on to the next part of this performance analysis, this short sub-section is intended to briefly summarize the findings made on the performance of the different ESG scoring models we have tested:

- The frequency of ESG words in Annual, CSR or even both reports is not sufficiently accurate and precise to serve as a proxy for ESG ratings. One potential driver of this finding was the presence of an important negative bias in the predictions of both Category and Grade ratings (that could however be linked to a too low intercept).
- We observed a certain stability in the ESG words frequency model performances, thereby suggesting the stability of the ESG lexicon used by companies.
- Classification models can improve the performances of the ESG words frequency model by deriving ratings from a subset of ESG terms on which various weights could be given.
- On average this improvement is not sufficient for the creation of a good proxy. Indeed, we found that in general Text Mining model could moderately approximate ESG Category rates (accuracy of 55.43%), but completely failed to predict more granular ESG performances.
- From a set of statistical tests, we observed that most of the considered rating models had equivalent performances, except for the Random Forest classifier which was able to reach encouraging levels of Accuracy and F1-score but is also highly variable.

3. Performance Analysis – Comparing Disclosure Formats

In the next pages, we are investigating the role of disclosure format on the performance of Text Mining based ESG ratings. In other words, we are trying to identify whether the frequency of ESG words in Annual reports has led to better prediction accuracies, or F1-score than the one of CSR reports or of a combination of both. Empirical results on this question are reported by **Figure 10, Figure 11, Figure 12**, and **Figure 13** (tables of detailed values can be found in Appendix XVIII). As for the model comparison, we are splitting our analysis in two subsections, one for the predictions of Category and the other for Grade ratings, and we are comparing performances in terms of average, variances, and statistical differences between metrics.

3.1. Category Ratings

From **Figure 10** it can be observed that there are no clear graphical differences in Accuracy and Macro F1-score between models fitted on Annual, CSR or Annual & CSR reports. Still, we can note that in terms of Category predictions, best general performances are reached by the RF classifier, and more particularly when built from Annual & CSR reports (Accuracy: 76.88% – F1-score: 0.6898). This suggest that combining both types of reports constitute the most appropriate source of ESG information for building most accurate Text Mining based ratings.

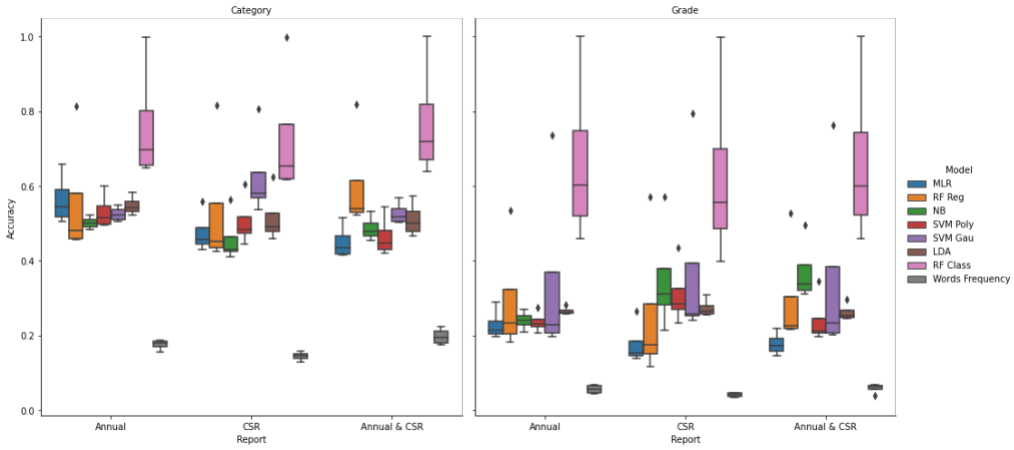


Figure 10: Comparison of different rating models performance in terms of Accuracy and Macro F1 for different type of disclosure. Data from 2017 to 2020

However, Tables of average accuracies and F1-score from Appendix XVIII (section 1.2 part 1) show that differences in both average Accuracy and F1-score between the three versions of the RF classifier are relatively thin ($\Delta_{max}^{Accuracy} = 5.05\% - \Delta_{max}^{F1} = 0.1602$). Moreover, accuracies and F1-scores are less variable for the Annual and particularly for the RF classifier on CSR reports. In addition, graphically we found that Annual & CSR reports did not always produce optimal performance results compared to other disclosure type for other classifiers. For instance, CSR reports are on average more appropriate for fitting accurate SVM with Gaussian kernel, while Annual reports are best suited for highly accurate MLR models (**Figure 10**).

To evaluate if the choice of disclosure type had any significant impact on the average Accuracy and F1-score of each model, we applied a set of Kruskal-Wallis and Tukey's test (1 per scoring model). Statistics and p-values resulting from those tests are depicted on **Table 49** and **Table 50**. Those results of Kruskal-Wallis tests are suggesting that disclosure type does not make any significant difference for Accuracy of RF reg ($p > .05$), NB ($p > .05$), LDA ($p > .05$), SVM with Polynomial kernel ($p > .05$), and RF class ($p > .05$). For the remaining three models, a significant difference has been found by Tukey's test between Annal and Annual & CSR reports for MLR (A > A+C at .05 level), CSR and Annual reports for SVM poly (A > C at .05 level), and between Annual and CSR (A > C at .05 level), as well as between CSR and Annual & CSR reports (C > A+C at .05 level). Concerning F1-scores, tests led to the conclusions that we failed to reject the hypothesis that averages were equal for the RF reg, NB, LDA, SVM with Gaussian kernel, and RF classifier. In the other cases, Annual reports were found to better than CSR for MLR and Words Frequency, while Annual report were also found to be more appropriate than Annual & CSR reports for SVM with polynomial kernel and Words Frequency. If in most cases,

all three disclosure formats have generated similar performances, those results are still indicating that Annual reports have outperformed the two others more regularly.

Figure 11 shows the boxplots of accuracies and F1 scores of models when those metrics are grouped by disclosure types rather than by model. It displays more clearly that there is, from a general point of view, little to no difference between models fitted with Annual, CSR or both reports in terms of F1-score and Accuracy. Still, it can be noted that Annual reports-based models have had on average higher and more stable Accuracies than their peers. Results from the Kruskal-Wallis tests reported on **Table 46** are suggesting that this difference is indeed significant at a 5% level (Chi-Square = 6.7181, $p < .05$, df = 3). However, this difference did not appear to be relevant for the performances in terms of F1-score (Chi-Square = 1.8647, $p = .39$, df = 3). From the perspective of other performance metrics (MAE, bias, precision, etc.), average values of metrics per disclosure type are not showing any major difference (see **Table 45**). To confirm this initial analysis, a Friedmann test was performed on those averages. Contrary to what was expected this test suggested that there was a significant difference between at least one pair of report format (Chi-Square = 6.0, $p = .0498$, df = 2). Statistics from Nemenyi's post hoc tests are indicating that given our data Annual reports-based models are on average better performer than their CSR reports peers ($p = .0427$).

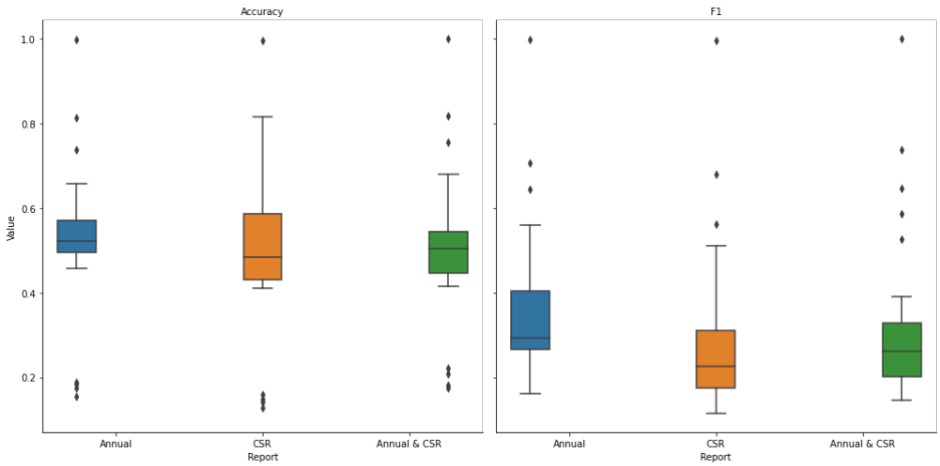


Figure 11: Boxplot of Accuracy and Macro F1 score of Text Mining Models given the used Disclosure Type for Category Ratings

3.2. Grade Ratings

Like previous section, results on Grade ratings are reported by boxplots of Accuracy and F1 scores of every models separately (see **Figure 12**) and of models grouped by disclosure type (see **Figure 13**). Combined with detailed performance metrics displayed in **Table 51**, it can be observed that results for Grade ratings are leading to similar conclusions than Category ratings.

Graphically, **Figure 11** does not provide a clear indication on whether a specific type of report may be more appropriate to predict ESG ratings with Text Mining tools. RF classifier is found to be more accurate when fitted on Annual & CSR reports, yet this does no longer hold when looking at the average performances in F1-scores. To confirm those visual results, we apply once again a set of Kruskal-Wallis tests. Those tests are suggesting that only the NB (Chi-Square = 5.6538, p = .0592, df = 3) is significantly different in terms of accuracy (see **Table 55**). For both models, Tukey's tests do not yield to any indication on which pair of disclosure type was different. Concerning F1-score, we have that MLR is better when fitted with Annual reports rather than with the two other options. A second significant difference was found between Annual reports and CSR reports (A > C at .05 level), and between CSR reports and Annual & CSR reports (A+C > C at .05 level) for Words Frequency (see **Table 56**). As for category ratings, we have thus that Annual reports are delivering better average performances on a more regular basis than other disclosure type. However, it is here limited to two out of 7 scoring models.

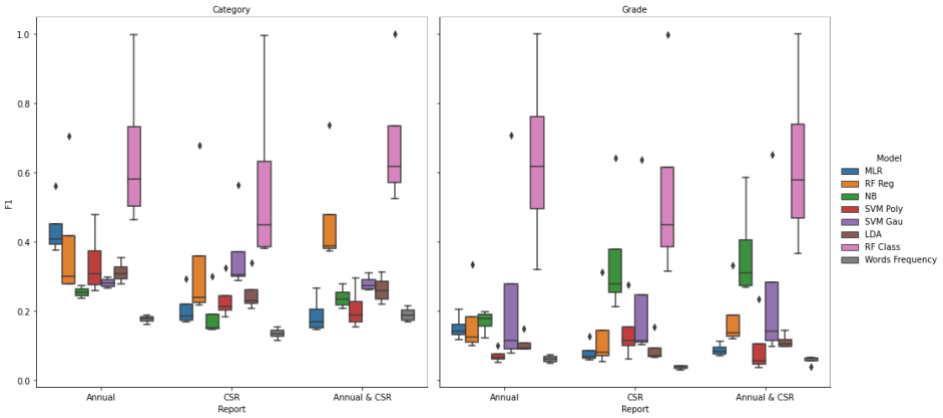


Figure 12: Comparison of different rating models performance in terms of Accuracy and Macro F1 for different type of disclosure. Data from 2017 to 2020 (Grades)

When comparing Accuracy and F1-score of models grouped by type of report, it can be seen in **Table 51** that differences are even smaller than for category ratings ($\Delta_{max}^{Accuracy} = 1.15\%$ – $\Delta_{max}^{F1} = 0.0308$). Very small values of Kruskal-Wallis statistics are confirming the intuition that we cannot reject the hypothesis that the respective impacts of Annual, CSR, and Annual & CSR reports on performance are on average equivalent (Accuracy: Chi-Square = .6688, p > .05, df = 3 – F1-score: Chi-Square = .161, p > .05, df = 3) (see **Table 52** and **Table 53**).

Yet, other performance metrics are showing some differences. For instance, from **Table 51** it can be seen that on average Annual reports model are less biased but have a higher MAE. Besides, Annual & CSR reports lead to predictions with a higher correlation and a higher recall

than with other form of disclosure. Furthermore, Friedmann test on all average metrics reported in **Table 54** (Chi-Square = 7.1428, p = .0281, df = 2) is suggesting that there is indeed a significant difference between the three-disclosure format. Yet, this difference is only concerning Annual vs. Annual & CSR report (p = .0206).

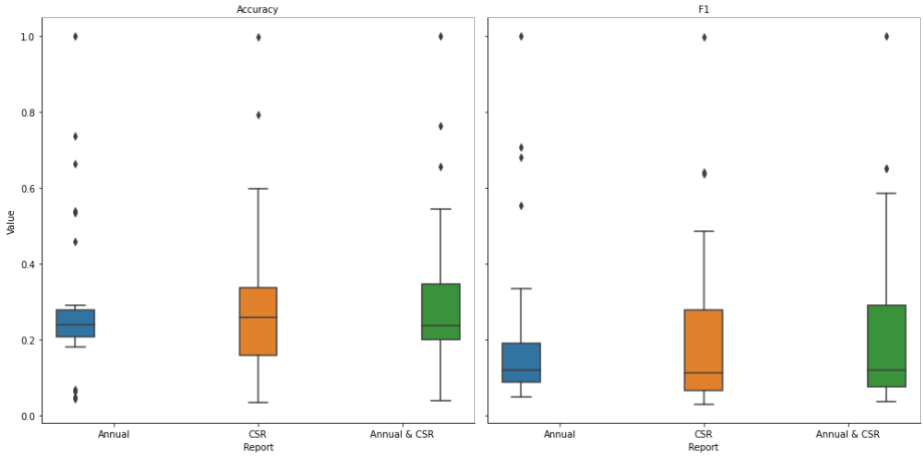


Figure 13: Boxplot of Accuracy and Macro F1 score of Text Mining Models given the used Disclosure Type for Grade Ratings

Chapter 3: Discussion and Implications

With the aim of contributing to the knowledge on AI and Text Mining in the field of ESG ratings, the question *"How to rate ESG performance based solely on a Text Mining analysis of a company's sustainability disclosure?"* has played the role of a unifying thread throughout this all Thesis. Having collected a set of results on the performance of several Text Mining-based ratings and sources of sustainability disclosure, we will use this section to eventually sketch an answer to the research question. By discussing the implication of the collected results considering the reviewed literature, we will draw conclusions on the general accuracy of Text Mining based ESG ratings, from their different scoring approaches to their different sources of sustainable information.

1.1.Frequency of ESG words and ESG performance

Several authors having shown the significance of the relationship between the frequency of ESG words and ESG performance (Clarkson et al., 2020; Takuya et al., 2020; Verbeeten et al., 2016), a first postulate to measure this performance based on Text Mining was to simply rely on the frequency of ESG words in the company's report. During the empirical research, we found that this approach, though theoretically attractive, was struggling at predicting Refinitiv's ESG ratings. Furthermore, we observed that this poor performance was mainly due to the underestimation of the companies with high ESG performances. Although the presence of a correlation confirmed the results of Clarkson et al. (2020) or Takuya et al. (2020) on the presence of a significant relationship between the two variables, the poor results in terms of accuracy (around 20% for category and 5% for grade) are indicating that this model does not capture enough information to deliver an appropriate differentiation between good and bad ESG performers.

To explain this poor accuracy, it should therefore be considered that ESG performers, from very bad to very good, use an equivalent number of ESG labelled words in their Annual and in their CSR reports. Therefore, this observation may call into question the quality of these two disclosure formats as a source of ESG information for investors. Indeed, this suggests that to measure ESG performance relying solely on Annual and CSR reports is not sufficient, hence it is necessary to look for additional, potentially non-public, sources of information as suggested by Giles & Murphy (2016), Unerman (2000), or Verbeeten et al. (2016). Furthermore, those poor results can also be justified by the presence of ESG-washing, i.e., bad ESG performers

unjustifiably increasing their disclosure on ESG topics in these two public documents (Aureli, 2017; Gyönyörová et al., 2021; Walter, 2020). This issue is concretely translated by the presence of ESG words that are negatively correlated to Refinitiv scores. The shortcoming of the ESG-words frequency model would therefore support the hypothesis of legitimacy theories as suggested by Verbeeten et al. (2016) and Takuya et al. (2020) for the Environmental dimension, or shown by Li & Haque (2019) for CSR disclosure in general.

Another approach for explaining these results is to take a more technical point view. The poor performances can indeed be seen as the proof that the bag-of-words assumption does not hold in the ESG rating context (Miner et al., 2012). In other words, we might have demonstrated here that to evaluate the ESG performance of a company, context surrounding a word does have an influence. In other words, the bad accuracy could have found its sources in the fact that good and bad ESG performers use the same amount of ESG words but not, for instance, with the same sentiment. This specific limitation had already been pointed out by Aureli (2017), Giles & Murphy (2016) or Verbeeten et al. (2016) as a limitations of their studies. This study therefore highlights the interest that can have word embeddings and other semantic models to measure ESG performance with Text Mining.

1.2.Text Classification Algorithms and ESG performance

A first objective of our empirical research was to investigate to what extent different Text Classification algorithms could accurately approximate ESG ratings. Thereby, we were hoping to find how to best combine the claimed advantages of Text Mining based ratings with appropriate measures of ESG performance. We found that among the 7 considered rating approaches, only one could reach “interesting” accuracy levels from the frequency of ESG words in Annual, or CSR, or a combination of the two. In general, we observed that Text Mining were able to moderately (average accuracy around 50%) measure ESG performance of companies when ESG performance is modelized by a 4-class categorization, from very bad to very good. Furthermore, we have seen that the best way to measure ESG performances into this format was to rely on a Random Forest classification algorithm which showed a reasonable level of accuracy, average of 75%, with a peak at almost 100% for ratings of 2020. When ESG performance is defined by a higher number of categories (12), we found that most of the Text Mining were unable to correctly measure this performance. For this task, the Random Forest classifier remained the best option with a lower average accuracy, but still with a peak at almost 100% in 2020.

In short, to measure ESG performance with a Text Mining model, we found that the optimal approach was to rely on a Random Forest classifier. Yet, even with this optimal algorithm, Refinitiv's ratings could not be "perfectly" approximated for the year 2017 to 2020, especially when ESG performance was characterized by more nuanced ratings. To explain this shortcoming, we can once again take the point of view of Giles & Murphy (2016), Unerman (2000), and Verbeeten et al. (2016) to conclude that the information contained in "traditional" disclosure support is not sufficient to evaluate the ESG involvement of a company. Thus, from the side of the reporting company, it indicates a problem of transparency on ESG issues in those essential documents (main source of information for the investor (Giles & Murphy, 2016)), either the company uses other means to communicate, or information is hidden or exaggerated (thus falling into the problematic of ESG-washing described previously). Moreover, the fact of having to resort to other documents than public reports to approximate Refinitiv's ESG scores highlights the problems of transparency that surround the ESG rating market (Gyönyörová et al., 2021; Walter, 2020). If in the best situation we can only explain 67% of the ratings, then it means that Rating agencies have access to other sources of information, sources that can't be verified by investors (Doyle, 2018; Windolph, 2011). However, this point must be put into perspective as we did not collect data from every public source presented by Refinitiv (Refinitiv, 2021). In addition, the ESG rating provider is known for implementing a sector specific approach (Madison & Schiehll, 2021) while we assigned ESG ratings without taking that dimension into account.

As for the ESG-words frequency model, a more technical interpretation of the results can be made here. Though, an improvement of accuracy shows that some extra hidden features can be extracted from the frequency of ESG terms (Miner et al., 2012), this improvement remains insufficient to perfectly measure ESG performance. Once again, this brings out the limitations of the bag-of-words hypothesis and the potential necessity to add other elements, such as context, to build proper ESG ratings from Text Mining. Another point of interpretation concerns the stability of the different Text Mining based ratings. As previously mentioned, the Random Forest algorithm was able to reach near to 100% accuracy in 2020, before dropping to less accurate results on other year sample. This decrease in performance suggested an instability in the most discriminant ESG terms. Hence, the ESG lexicon used by companies in their sustainability disclosure tends to vary from year to year. This observation partly shows the limits of a static dictionary-based approach to identify ESG terms (Lewis & Young, 2019). Relying on a machine learning approach combined with a word embedding model (such as the

one implemented by Takuya et al. (2020)) might decrease the sensitivity of the model to those changes in lexicons.

1.3.Sustainability disclosure and ESG performance

A last element that needs to be discussed concerns the influence of the source of sustainability information on performance of Text Mining scoring models. We found that generally there was no significant difference between relying on Annual, CSR or both reports for building Text Mining based ESG ratings. Yet, we have observed that Annual reports could outperform their peers on a more regular basis. This could be explained by the analysis of Giles & Murphy (2016) or Guthrie & Abeysekera (2006) who claimed that those reports were the document on which companies usually rely to disclose material information. Knowing from the literature review that materiality is a key point of any ESG rating methodology (Escrig-Olmedo et al., 2010), it could justify this slightly better performance. Since Annual reports are containing less immaterial ESG words it allows for a reduction of the noise in the textual data, hence it allows for a better approximation of Refinitiv's ratings.

CONCLUSION

On the 24th of September 2019, the UN secretary-general Antonio Guterres called on all actors of society for a decade of action (Guterres, 2019). A decade during which governments, companies, and people will have to further mobilize to enable humanity to comply with the agenda set by the Sustainable Development Goals. Yet, since concretely implementing actions requires an important financial effort from the private sector (an effort estimated at $3 to $5 trn of private capital a year) (Crédit Suisse, 2020; KPMG UK, 2019), the achievement of Sustainable development will most likely pass through the development of a more Sustainable finance and that is where ESG ratings can play an important role (Joselow, 2021; Mann, 2021; Zesty, 2017). Currently, ESG ratings represent one of the cornerstones of Sustainable finance, with investors using them to allocate their capital, companies to position themselves, and scientists to approximate ESG performance of companies (see ESG Performance and ESG Ratings). Though essential to the development of this sector, those scores are not without criticism: Time consuming, Lack of transparency, lack of standardization, biased, etc. As elaborated during the literature review, an increasing number of scholars and practitioners from the ESG sector are therefore claiming that AI and Text Mining are a solution to those specific issues. Feeding on this statement, throughout this thesis we attempted to address the following research question: "*How to rate ESG performance based solely on a Text Mining analysis of a company's sustainability disclosure?"*.

Based on a literature review on Text Mining and its applications in Sustainable finance, we came up with a simple methodology to rate ESG performance. Relying on the observation that the frequency of ESG terms in Annual and/or CSR reports is associated to the ESG performance of a company (Clarkson et al., 2020; Takuya et al., 2020), we used different classification and prediction algorithms to build an ESG score from those frequencies. To evaluate the quality of this scoring methodology, we then compared predictions of ESG scores to official ESG ratings from Refinitiv for a sample of 414 companies and four year (2017-2020). The quality was assessed based on several metrics including accuracy, F1-score, bias, etc.

After conducting the empirical research, we were able to show the limitations of this methodology and its main assumption for our sample. Among the different investigated models, only one managed to reach high accuracies and F1-score, the Random Forest classifier. And even in this "optimal" situation, the quality of the predictions remained moderate. An observation which suggested that, though a share of the information on ESG performance could be captured by the frequency of ESG terms, this share remains insufficient to perfectly

approximate ESG ratings. To explain this poor performance, several reasons were investigated including the questioning of the ESG rating market (transparency of the rating agencies and companies, and presence of greenwashing in company's reports) or the potential "technical" shortcomings of our Text Mining approach (bag-of-words representation, relying solely on Annual and CSR reports, etc.). In addition, we found that relying on Annual reports, or CSR reports, or a combination of both had few influences on the quality of ESG scores predictions, with a little advantage to the Annual Reports. This last result indicated that this specific category of reports was best suited for the construction of ESG rating based on our Text Mining approach.

While our study has highlighted the limitations of the proposed rating methodology (potentially due to its simplicity), it is nevertheless a first step towards the more efficient (in terms of time but also quality) construction of ESG ratings. It could serve as a benchmark paving the way for future research. In particular, the discussion of the results highlighted several possibilities for improvement to be investigated. For instance, it could be interesting to see the impact of applying the same methodology to different sources of sustainability disclosure. Instead of simply relying on Annual and CSR reports, one could also collect information from company's website, social media accounts, etc. and see whether it improves the model's performances. Another avenue to explore would be the impact of external sources of information on our ESG score quality, a study that could highlight to what extent greenwashing is present in Annual and CSR reports. Finally, future research could focus on the questioning of the bag-of-words assumption for the construction of Text Mining based ESG ratings. It could be useful to see whether the implementation of a word embedding model to represent ESG words can improve the quality of the predictions, thereby showing the importance of semantic for that specific task.

Before closing this paper, it is still important to recall that all our results are subject to certain limitations specific to our research. First, the ability of our methodology to measure ESG performance was judged by comparing our predictions to a single ESG rating, that of Refinitiv. To confirm our results, we would need to repeat the research on ratings from other providers (MSCI, Bloomberg, etc.). The, we were only able to test the performance of our different scoring models over 4 years. To be able to confirm our results in a more robust way, we would need to obtain more data on each model's accuracy, F1-score, etc. Finally, we chose to rely on a specific ESG dictionary to label words from Annual and CSR reports. Given that this dictionary may not be optimal, it would be interesting to perform this research either on a different dictionary (still to be created) or by using a machine learning based approach.

BIBLIOGRAPHY

Aaryan, G., Vinya, D., Abubakar, K. H., & Manan, S. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation*, *6*(1). http://dx.doi.org/10.1186/s40854-020-00205-1

Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. *International Journal of Information Management*, *54*, 102132. https://doi.org/10.1016/j.ijinfomgt.2020.102132

Amel-Zadeh, A., & Serafeim, G. (2017). *Why and How Investors Use ESG Information: Evidence from a Global Survey* (SSRN Scholarly Paper ID 2925310). Social Science Research Network. https://doi.org/10.2139/ssrn.2925310

Antoncic, M. (2020). Uncovering hidden signals for sustainable investing using Big Data: Artificial intelligence, machine learning and natural language processing. *Journal of Risk Management in Financial Institutions*. https://hstalks.com/article/5506/uncovering-hidden-signals-for-sustainable-investin/

Appavu, S., Rajaram, R., Muthupandian, M., Athiappan, G., & Kashmeera, K. S. (2009). Data mining based intelligent analysis of threatening e-mail. *Knowledge-Based Systems*, *22*(5), 392–393. https://doi.org/10.1016/j.knosys.2009.02.002

Aureli, S. (2017). A comparison of content analysis usage and text mining in CSR corporate disclosure. *International Journal of Digital Accounting Research*, *17*, 1–32. http://dx.doi.org/10.4192/1577-8517-v17_1

Azhar, N. A., Pan, G., Seow, P.-S., Koh, A., & Tay, W.-Y. (2019). *Text Analytics Approach to Examining Corporate Social Responsibility* (SSRN Scholarly Paper No. 3392876). Social Science Research Network. https://papers.ssrn.com/abstract=3392876

Bacha, S., & Ajina, A. (2019). CSR performance and annual report readability: Evidence from France. *Corporate Governance: The International Journal of Business in Society*, *20*(2), 201–215. https://doi.org/10.1108/CG-02-2019-0060

Baier, P., Berninger, M., & Kiesel, F. (2020). Environmental, social and governance reporting in annual reports: A textual analysis. *Financial Markets, Institutions & Instruments*, *29*(3), 93–118. https://doi.org/10.1111/fmii.12132

BCG Global. (n.d.). *Sustainable Finance*. BCG Global. Retrieved 17 March 2022, from https://www.bcg.com/capabilities/social-impact-sustainability/how-sustainable-finance-is-shifting-future-of-investing

Bedenik, N., & Barišić, P. (2019). *Nonfinancial Reporting: Theoretical and Empirical Evidence*. https://doi.org/10.5772/intechopen.87159

Bekios-Calfa, J., Buenaposada, J. M., & Baumela, L. (2011). Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(4), 858–864. https://doi.org/10.1109/TPAMI.2010.208

Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. *Methods in*

Molecular Biology (Clifton, N.J.), *609*(Journal Article), 223–239. https://doi.org/10.1007/978-1-60327-241-4_13

Berg, F., Kölbel, J. F., & Rigobon, R. (2019). *Aggregate Confusion: The Divergence of ESG Ratings* (SSRN Scholarly Paper No. 3438533). Social Science Research Network. https://doi.org/10.2139/ssrn.3438533

Bernow, S., Klempner, B., & Magnin, C. (2017, October 25). *From 'why' to 'why not': Sustainable investing as the new normal | McKinsey*. https://www.mckinsey.com/industries/private-equity-and-principal-investors/our-insights/from-why-to-why-not-sustainable-investing-as-the-new-normal

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008

Billio, M., Costola, M., Hristova, I., Latino, C., & Pelizzon, L. (2021). Inside the ESG ratings: (Dis)agreement and performance. *Corporate Social Responsibility and Environmental Management*, *28*(5), 1426–1445. https://doi.org/10.1002/csr.2177

Blank, H., Sgambati, G., & Truelson, Z. (2016). Best Practices in ESG Investing. *The Journal of Investing*, *25*(2), 103–112. https://doi.org/10.3905/joi.2016.25.2.103

Blasco, J. L., & King, A. (2017). *The KPMG Survey of Corporate Responsibility Reporting 2017* (p. 9). KPMG Global. https://www.business-humanrights.org/en/latest-news/the-road-ahead-the-kpmg-survey-of-corporate-responsibility-reporting-2017/

Bloomberg Intelligence. (2021, February 23). ESG assets may hit $53 trillion by 2025, a third of global AUM. *Bloomberg Professional Services*. https://www.bloomberg.com/professional/blog/esg-assets-may-hit-53-trillion-by-2025-a-third-of-global-aum/

Boffo, R., & Patalano, R. (2020). *ESG Investing: Practices, Progress and Challenges* (p. 88). OECD Paris. https://www.oecd.org/finance/ESG-Investing-Practices-Progress-Challenges.pdf

Busch, T., Bauer, R., & Orlitzky, M. (2016). Sustainable Development and Financial Markets: Old Paths and New Avenues. *Business & Society*, *55*(3), 303–329. https://doi.org/10.1177/0007650315570701

Cambria, E. (2016). Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems*, *31*(2), 102–107. https://doi.org/10.1109/MIS.2016.31

Capelle-Blancard, G., & Monjon, S. (2012). Trends in the literature on socially responsible investment: Looking for the keys under the lamppost. *Business Ethics: A European Review*, *21*(3), 239–250. https://doi.org/10.1111/j.1467-8608.2012.01658.x

Cash, D. (2018). Sustainable finance ratings as the latest symptom of 'rating addiction'. *Journal of Sustainable Finance & Investment*, *8*, 1–17. https://doi.org/10.1080/20430795.2018.1437996

Chatterji, A. K., Durand, R., Levine, D. I., & Touboul, S. (2016). Do ratings of firms converge?

Implications for managers, investors and strategy researchers. *Strategic Management Journal*, *37*(8), 1597–1614. https://doi.org/10.1002/smj.2407

Chelawat, H., & Trivedi, I. (2013). Ethical Finance: Trends and Emerging Issues for Research. *International Journal of Business Ethics in Developing Economies*, *2*, 34–42. https://doi.org/10.26643/think-india.v16i2.7819

Clarkson, P. M., Ponn, J., Richardson, G. D., Rudzicz, F., Tsang, A., & Wang, J. (2020). A Textual Analysis of US Corporate Social Responsibility Reports. *Abacus*, *56*(1), 3–34. https://doi.org/10.1111/abac.12182

Clementino, E., & Perkins, R. (2021). How Do Companies Respond to Environmental, Social and Governance (ESG) ratings? Evidence from Italy. *Journal of Business Ethics*, *171*(2), 379–397. https://doi.org/10.1007/s10551-020-04441-4

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*, 270. https://doi.org/10.1186/s12859-018-2264-5

Crédit Suisse. (2020, September 24). *SDGs put impact at the heart of sustainable investing*. Credit Suisse. https://www.credit-suisse.com/about-us-news/en/articles/news-and-expertise/sdgs-put-impact-at-the-heart-of-sustainable-investing-202009.html

Crespi, F., & Migliavacca, M. (2020). The Determinants of ESG Rating in the Financial Industry: The Same Old Story or a Different Tale? *Sustainability*, *12*(16), 6398. https://doi.org/10.3390/su12166398

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, *34*(3), 1707–1720. https://doi.org/10.1016/j.eswa.2007.01.035

Delevingne, L., & Gründler, A. ,. Kane, S. ,. &. Koller, T. (2020). *The ESG premium: New perspectives on value and performance*.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*(Journal Article), 1–30.

Dörre, J., Gerstl, P., & Seiffert, R. (1999). Text mining: Finding nuggets in mountains of textual data. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 398–401. https://doi.org/10.1145/312129.312299

Dougherty, G. (2013). *Pattern Recognition and Classification*. Springer. https://doi.org/10.1007/978-1-4614-5323-9

Doyle, T. (2018). *Ratings that don't rate: The subjective world of ESG rating agencies*. ACCF. https://accfcorpgov.org/wp-content/uploads/2018/07/ACCF_RatingsESGReport.pdf

Dumont, N. H. R. (2017). Sentiment Analysis & Natural Language: Processing Techniques for Capital Markets & Disclosure. *The Corporate Governance Advisor*, *25*(6), 16–23.

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The Impact of Corporate Sustainability on Organizational Processes and Performance. *Management Science*, *60*(11), 2835–2857.

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

Escrig-Olmedo, E., Muñoz-Torres, M., & Fernandez-Izquierdo, M. (2010). Socially responsible investing: Sustainability indices, ESG rating and information provider agencies. *International Journal of Sustainable Economy*, *2*, 442–461. https://doi.org/10.1504/IJSE.2010.035490

Fink, L. (2018, January 17). A Sense of Purpose. *The Harvard Law School Forum on Corporate Governance*. https://corpgov.law.harvard.edu/2018/01/17/a-sense-of-purpose/

Fish, A. J. M., Kim, D. H., & Venkatraman, S. (2019). *The ESG Sacrifice* (SSRN Scholarly Paper No. 3488475). Social Science Research Network. https://doi.org/10.2139/ssrn.3488475

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157–214. https://doi.org/10.1002/isaf.1386

Foody, G. M. (2009). Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, *113*(8), 1658–1663. https://doi.org/10.1016/j.rse.2009.03.014

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

Gartenberg, C. M., Prat, A., & Serafeim, G. (2018). *Corporate Purpose and Financial Performance* (SSRN Scholarly Paper ID 2840005). Social Science Research Network. https://doi.org/10.2139/ssrn.2840005

Giles, O., & Murphy, D. (2016). SLAPPed: The relationship between SLAPP suits and changed ESG reporting by firms. *Sustainability Accounting, Management and Policy Journal*, *7*(1), 44–79. https://doi.org/10.1108/SAMPJ-12-2014-0084

Goloshchapova, I., Poon, S.-H., Pritchard, M., & Reed, P. (2019). Corporate social responsibility reports: Topic analysis and big data approach. *The European Journal of Finance*, *25*(17), 1637–1654. https://doi.org/10.1080/1351847X.2019.1572637

Gregor, D., externe, L. vers un site, fenêtre, celui-ci s'ouvrira dans une nouvelle, Kreuzer, C., & Christian, S. (2020). ESG controversies and controversial ESG: About silent saints and small sinners. *Journal of Asset Management*, *21*(5), 393–412. http://dx.doi.org/10.1057/s41260-020-00178-x

Groß-Klußmann, A., König, S., & Ebner, M. (2019). Buzzwords build momentum: Global financial Twitter sentiment and the aggregate stock market. *Expert Systems with Applications*, *136*, 171–186. https://doi.org/10.1016/j.eswa.2019.06.027

Guo, T., Jamet, N., Betrix, V., Piquet, L.-A., & Hauptmann, E. (2020). *ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction*.

https://doi.org/10.48550/arXiv.2005.02527

Guterres, A. (2019, September 24). *Remarks to High-Level Political Forum on Sustainable Development | United Nations Secretary-General*. https://www.un.org/sg/en/content/sg/speeches/2019-09-24/remarks-high-level-political-sustainable-development-forum

Guthrie, J., & Abeysekera, I. (2006). Content analysis of social, environmental reporting: What is new? *Journal of Human Resource Costing & Accounting*, *10*(2), 114–126. https://doi.org/10.1108/14013380610703120

Gyönyörová, L., Stachoň, M., & Stašek, D. (2021). ESG ratings: Relevant information or misleading clue? Evidence from the S&P Global 1200. *Journal of Sustainable Finance & Investment*, *0*(0), 1–35. https://doi.org/10.1080/20430795.2021.1922062

Hajek, P., Olej, V., & Myskova, R. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, *20*(4). https://doi.org/10.3846/20294913.2014.979456

Halbritter, G., & Dorfleitner, G. (2015). The wages of social responsibility — where are they? A critical review of ESG investing. *Review of Financial Economics*, *26*, 25–35. https://doi.org/10.1016/j.rfe.2015.03.004

Hawley, J. (2017). *ESG Ratings and Rankings All Over the Map* (p. 18). TrueValue Labs.

Heston, S. L., & Sinha, N. R. (2017). News vs. Sentiment: Predicting Stock Returns from News Stories. *Financial Analysts Journal*, *73*(3), 67–83. https://doi.org/10.2469/faj.v73.n3.3

Huber, B. M., & Comstock, M. (2017, July 27). ESG Reports and Ratings: What They Are, Why They Matter. *The Harvard Law School Forum on Corporate Governance*. https://corpgov.law.harvard.edu/2017/07/27/esg-reports-and-ratings-what-they-are-why-they-matter/

Hughes, A., Urban, M. A., & Wójcik, D. (2021). Alternative ESG Ratings: How Technological Innovation Is Reshaping Sustainable Investment. *Sustainability*, *13*(6), 3551. https://doi.org/10.3390/su13063551

Hummel, K., Mittelbach-Hoermanseder, S., Cho, C. H., & Matten, D. (2017). *Implicit Versus Explicit Corporate Social Responsibility Disclosure: A Textual Analysis*. https://doi.org/10.2139/SSRN.3090976

Ioannou, I., & Serafeim, G. (2015). *The Impact of Corporate Social Responsibility on Investment Recommendations: Analysts' Perceptions and Shifting Institutional Logics* (SSRN Scholarly Paper ID 1507874). Social Science Research Network. https://doi.org/10.2139/ssrn.1507874

Joselow, M. (2021, December 7). *John Kerry is counting on the private sector to help solve climate change—The Washington Post*. https://www.washingtonpost.com/politics/2021/12/07/john-kerry-is-counting-private-sector-help-solve-climate-change/

Joshi, K., H. N, B., & Rao, J. (2016). Stock Trend Prediction Using News Sentiment Analysis. *International Journal of Computer Science & Information Technology*, *8*(3), 67–76. https://doi.org/10.5121/ijcsit.2016.8306

Kachaner, N., Nielsen, J., Portafaix, A., & Rodzko, F. (2020, July 15). *The Pandemic Is Heightening Environmental Awareness*. BCG Global. https://www.bcg.com/publications/2020/pandemic-is-heightening-environmental-awareness

Khan, M., Serafeim, G., & Yoon, A. (2016). *Corporate Sustainability: First Evidence on Materiality* (SSRN Scholarly Paper No. 2575912). Social Science Research Network. https://doi.org/10.2139/ssrn.2575912

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., & Visa, A. (2004). Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance and Management*, *12*(1), 29–41.

Kowshalya, A. M., Madhumathi, R., & Gopika, N. (2019). Correlation Based Feature Selection Algorithms for Varying Datasets of Different Dimensionality. *Wireless Personal Communications*, *108*(3), 1977–1993. https://doi.org/10.1007/s11277-019-06504-w

KPMG UK. (2019). *The numbers that are changing the world—Revealing the growing appetite for responsible investing*. https://assets.kpmg/content/dam/kpmg/ie/pdf/2019/10/ie-numbers-that-are-changing-the-world.pdf

Kuh, T. (2019). *ESG Research in the Information Age: AI, Unstructured Data and the Future of ESG Investing*. TrueValue Labs. https://cdn2.hubspot.net/hubfs/4137330/White%20Papers/WP_ESGResearch_InfoAge.pdf?__hssc=16054825.1.1578655262760&__hstc=16054825.bac067346c4aadc8a5c43f15f7d0ae07.1578655262760.1578655262760.1578655262760.1&__hsfp=1603390744&hsCtaTracking=60c937d4--47dd-4a22-bdfc-2d575f9cd566%7C1a3c1c13-d812--4521--811b-e5f05cf1f141

Kumar, B. S., & Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, *114*, 128–147. https://doi.org/10.1016/j.knosys.2016.10.003

LaBella, M. J., Sullivan, L., Russell, J., & Novikov, D. A. (2019). *The Devil is in the Details: The Divergence in ESG Data and Implications for Sustainable Investing*. https://www.semanticscholar.org/paper/The-Devil-is-in-the-Details%3A-The-Divergence-in-ESG-LaBella-Sullivan/4c649ab05d826959294c9b3538d82c96ab1694b0

Lee, B., Park, J.-H., Kwon, L., Moon, Y.-H., Shin, Y., Kim, G., & Kim, H. (2018). About relationship between business text patterns and financial performance in corporate data. *Journal of Open Innovation: Technology, Market, and Complexity*, *4*(1), 3. https://doi.org/10.1186/s40852-018-0080-9

Lewis, C., & Young, S. (2019). Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, *49*(5), 587–615. https://doi.org/10.1080/00014788.2019.1611730

Li, Z., & Haque, S. (2019). Corporate social responsibility employment narratives: A linguistic analysis. *Accounting, Auditing & Accountability Journal*, *32*(6), 1690–1713.

https://doi.org/10.1108/AAAJ-10-2016-2753

Liberato Ferrao, M. (2019). *ESG Rating Agencies: What is the Path towards a Bright Future?*

Liew, W. T., Adhitya, A., & Srinivasan, R. (2014). Sustainability trends in the process industries: A text mining-based analysis. *Computers in Industry*, *65*(3), 393–400. https://doi.org/10.1016/j.compind.2014.01.004

López-Arceiz, F. J., Del Río, C., & Bellostas, A. J. (2020). Sustainability performance indicators: Definition, interaction, and influence of contextual characteristics. *Corporate Social Responsibility and Environmental Management*, *27*(6), 2615–2630. https://doi.org/10.1002/csr.1986

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

Loughran, T., & Mcdonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Loughran, T., McDonald, B., & Yun, H. (2009). A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports. *Journal of Business Ethics*, *89*, 39–49.

Madison, N., & Schiehll, E. (2021). The Effect of Financial Materiality on ESG Performance Assessment. *Sustainability*, *13*(7), 1–21.

Mann, M. (2021). *The New Climate War* (Scribe Publications). Scribe Publications. https://www.scribd.com/book/492203184/The-New-Climate-War-the-fight-to-take-back-our-planet

Mayor, T. (2019, August 26). *Why ESG ratings vary so widely (and what you can do about it)*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/why-esg-ratings-vary-so-widely-and-what-you-can-do-about-it

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Miner, G. D., Elder, J., IV, Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Elsevier Science & Technology. http://ebookcentral.proquest.com/lib/uclouvainbe/detail.action?docID=842198

Mirvis, P., & Ryu, K. (2009). *How Virtue Creates Value for Business and Society: Investigating the value of environmental, social and governance activities*. https://ccc.bc.edu/content/ccc/reports/how-virtue-creates-value-for-business-and-society--investigating.html

Modapothala, J. R., Issac, B., & Jayamani, E. (2009). *Appraising the Corporate Sustainability Reports—Text Mining and Multi-Discriminatory Analysis*. 489–494. https://doi.org/10.1007/978-90-481-9112-3_83

Moen, E. (n.d.). *Integrating MSCI ESG Ratings into investment decision making can help identify risks and opportunities that may not be captured by conventional*. 5.

Mohamad, M., Selamat, A., Krejcar, O., Crespo, R. G., Herrera-Viedma, E., & Fujita, H. (2021). Enhancing Big Data Feature Selection Using a Hybrid Correlation-Based Feature Selection. *Electronics*, *10*(23), 2984. https://doi.org/10.3390/electronics10232984

Mooij, S. (2017). *The ESG Rating and Ranking Industry; Vice or Virtue in the Adoption of Responsible Investment?* https://doi.org/10.13140/RG.2.2.33379.76323

MSCI. (2022). *MSCI World Index (USD)*. MSCI. https://www.msci.com/documents/10199/149ed7bc-316e-4b4c-8ea4-43fcb5bd6523

Nazari, J. A., Hrazdil, K., & Mahmoudian, F. (2017). Assessing social and environmental performance through narrative complexity in CSR reports. *Journal of Contemporary Accounting & Economics*, *13*(2), 166–178. https://doi.org/10.1016/j.jcae.2017.05.002

Nopp, C., & Hanbury, A. (2015). Detecting Risks in the Banking System by Sentiment Analysis. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 591–600. https://doi.org/10.18653/v1/D15-1071

Nugent, T., Stelea, N., & Leidner, J. L. (2020). *Detecting ESG topics using domain-specific language models and data augmentation approaches*. https://doi.org/10.48550/arXiv.2010.08319

Palmer, P. B., & O'Connell, D. G. (2009). Regression Analysis for Prediction: Understanding the Process. *Cardiopulmonary Physical Therapy Journal*, *20*(3), 23–26.

Pathan, M., Patel, N., Yagnik, H., & Shah, M. (2020). Artificial cognition for applications in smart agriculture: A comprehensive review. *Artificial Intelligence in Agriculture*, *4*, 81–95. https://doi.org/10.1016/j.aiia.2020.06.001

Pejić Bach, M., Krstić, Ž., Seljan, S., & Turulja, L. (2019). Text Mining for Big Data Analysis in Financial Sector: A Literature Review. *Sustainability*, *11*(5), 1277. https://doi.org/10.3390/su11051277

Porter, M. F. (2006). An algorithm for suffix stripping. *Program*, *40*(3), 211–218. https://doi.org/10.1108/00330330610681286

Raghupathi, V., Ren, J., & Raghupathi, W. (2020). Identifying Corporate Sustainability Issues by Analyzing Shareholder Resolutions: A Machine-Learning Text Analytics Approach. *Sustainability*, *12*(11), 4753. https://doi.org/10.3390/su12114753

RavenPack. (2019). *Generate Alpha by Enhancing MSCI ESG Ratings with News Sentiment*. RavenPack. https://www.ravenpack.com/research/msci-esg-ratings-sentiment

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, *89*, 14–46. https://doi.org/10.1016/j.knosys.2015.06.015

Refinitiv. (2021). *ENVIRONMENTAL, SOCIAL AND GOVERNANCE SCORES FROM REFINITIV*.

https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf

Saadaoui, K., & Soobaroyen, T. (2018). An analysis of the methodologies adopted by CSR rating agencies. *Sustainability Accounting, Management and Policy Journal*, *9*(1), 43–62. http://dx.doi.org/10.1108/SAMPJ-06-2016-0031

Salampasis, D. (2017). Leveraging robo-advisors to fill the gap within the SRI marketplace. *Journal of Innovation Management*, *5*(3), 6–13. https://doi.org/10.24840/2183-0606_005.003_0002

SASB Standards & Other ESG Frameworks. (n.d.). *SASB*. Retrieved 11 April 2022, from https://www.sasb.org/about/sasb-and-other-esg-frameworks/

Scalet, S., & Kelly, T. F. (2010). CSR Rating Agencies: What is Their Global Impact? *Journal of Business Ethics*, *94*(1), 69–88.

Schmidt, A. (2019). *Sustainable News – A Sentiment Analysis of the Effect of ESG Information on Stock Prices* (SSRN Scholarly Paper No. 3809657). Social Science Research Network. https://doi.org/10.2139/ssrn.3809657

Semenova, N., & Hassel, L. G. (2015). On the Validity of Environmental Performance Metrics. *Journal of Business Ethics*, *132*(2), 249–258.

Serafeim, G. (2021). *ESG: Hyperboles and Reality* (SSRN Scholarly Paper ID 3966695). Social Science Research Network. https://doi.org/10.2139/ssrn.3966695

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1), 12. https://doi.org/10.1007/s41133-020-00032-0

Shahi, A. M., Issac, B., & Modapothala, J. R. (2014). Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications*, *13*(01), 1450006. https://doi.org/10.1142/S1469026814500060

Shirata, C. Y., Takeuchi, H., Ogino, S., & Watanabe, H. (2011). Extracting Key Phrases as Predictors of Corporate Bankruptcy: Empirical Analysis of Annual Reports by Text Mining. *Journal of Emerging Technologies in Accounting*, *8*(1), 31–44. https://doi.org/10.2308/jeta-10182

Shiva Darshan, S. L., & Jaidhar, C. D. (2018). Performance Evaluation of Filter-based Feature Selection Techniques in Classifying Portable Executable Files. *Procedia Computer Science*, *125*, 346–356. https://doi.org/10.1016/j.procs.2017.12.046

Siew, R. Y. J. (2015). A review of corporate sustainability reporting tools (SRTs). *Journal of Environmental Management*, *164*, 180–195. https://doi.org/10.1016/j.jenvman.2015.09.010

Simeon, M., & Hilderman, R. (2008). *Categorical Proportional Difference: A Feature Selection Method for Text Categorization. 87*, 201–208.

Sokolov, A., Mostovoy, J., Ding, J., & Seco, L. (2021). Building Machine Learning Systems for Automated ESG Scoring. *The Journal of Impact and ESG Investing*. https://doi.org/10.3905/jesg.2021.1.010

Song, Y., Wang, H., & Zhu, M. (2018). Sustainable strategy for corporate governance based on the sentiment analysis of financial reports with CSR. *Financial Innovation*, *4*(1), 2. https://doi.org/10.1186/s40854-018-0086-0

S&P Global. (2020, April 25). *How can AI help ESG investing?* https://www.spglobal.com/en/research-insights/articles/how-can-ai-help-esg-investing

Stichbury, J. (2020, March 20). *Next-level NLP and potential ESG controversies*. Refinitiv Perspectives. https://www.refinitiv.com/perspectives/ai-digitalization/next-level-nlp-and-potential-esg-controversies/

Székely, N., & vom Brocke, J. (2017). What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLoS ONE*, *12*(4), e0174807. https://doi.org/10.1371/journal.pone.0174807

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Takuya, K., externe, L. vers un site, fenêtre, celui-ci s'ouvrira dans une nouvelle, & Masatoshi, N. (2020). A Text Mining Model to Evaluate Firms' ESG Activities: An Application for Japanese Firms. *Asia - Pacific Financial Markets*, *27*(4), 621–632. http://dx.doi.org/10.1007/s10690-020-09309-1

UN. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development | Department of Economic and Social Affairs*. https://sdgs.un.org/2030agenda

Unerman, J. (2000). Methodological issues—Reflections on quantification in corporate social reporting content analysis. *Accounting, Auditing & Accountability Journal*, *13*(5), 667–680.

US SIF. (2020). *Report on US Sustainable and Impact Investing Trends 2020*. US SIF. https://www.ussif.org/files/US%20SIF%20Trends%20Report%202020%20Executive%20Summary.pdf

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Valor, C., & De la Cuesta, M. (2013). Evaluation of the environmental, social and governance information disclosed by Spanish listed companies. *Social Responsibility Journal*, *9*. https://doi.org/10.1108/SRJ-08-2011-0065

Verbeeten, F. H. M., Gamerschlag, R., & Möller, K. (2016). Are CSR disclosures relevant for investors? Empirical evidence from Germany. *Management Decision*, *54*(6), 1359–1382. https://doi.org/10.1108/MD-08-2015-0345

Walter, I. (2020). *Sense and Nonsense in ESG Ratings* (SSRN Scholarly Paper No. 3696718).

Social Science Research Network. https://papers.ssrn.com/abstract=3696718

Wei, L., Li, G., Zhu, X., & Li, J. (2019). Discovering bank risk factors from financial statements based on a new semi-supervised text mining algorithm. *Accounting & Finance*, *59*(3), 1519–1552. https://doi.org/10.1111/acfi.12453

What Is Unstructured Data? (n.d.). MongoDB. Retrieved 12 April 2022, from https://www.mongodb.com/unstructured-data

Windolph, S. (2011). Assessing Corporate Sustainability Through Ratings: Challenges and Their Causes. *Journal of Environmental Sustainability*, *1*. https://doi.org/10.14448/jes.01.0005

Wong, C., & Petroy, E. (2020). *Rate the Raters 2020:Investor Survey and Interview Results*. SustainAbility. https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sustainability-ratetheraters2020-report.pdf

Woods, D. D., & Hollnagel, E. (2005). *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. Routledge & CRC Press. https://www.routledge.com/Joint-Cognitive-Systems-Foundations-of-Cognitive-Systems-Engineering/Hollnagel-Woods/p/book/9780367864200

Zavolokina, L., Dolata, M., & Schwabe, G. (2016). The FinTech phenomenon: Antecedents of financial innovation perceived by the popular press. *Financial Innovation*, *2*, 16. https://doi.org/10.1186/s40854-016-0036-7

Zesty, W. design and website development by S. (2017, November 21). *The role of financial services in helping to meet the UN Sustainable Development Goals*. Financial Services Thought Gallery. https://eyfinancialservicesthoughtgallery.ie/un-sustainable-development-goals/

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1), 43–52. https://doi.org/10.1007/s13042-010-0001-0

APPENDICES

I. The success of ESG investing

Undoubtfully, sustainable investing has become one of fastest-growing segment of the financial markets. Since 1995, and the first measurement of the SRI universe in the US, the number of sustainable assets under management has increased at an annual compound growth rate of 13.6%. Currently, this is representing 16.6 trillion of dollars' worth of assets in the US (US SIF, 2020). When we take a worldwide perspective, those numbers rise to just under 31 trillion dollars, which represented more than 25% of the global asset market in 2018 (Bernow et al., 2017; KPMG UK, 2019; US SIF, 2020).

According to multiple survey's conducted across major actors of the financial sector, this growing interest for more responsible investments can be explained by two main drivers (Bernow et al., 2017; Boffo & Patalano, 2020; Delevingne & Gründler, 2020). On the one hand, we observe that shareholders increasingly feel the need to find alignment between investments and their values. On the other hand, more and more investors are convinced that integrating ESG criteria in their decision-making process can support their search for better long-term financial value.

i. Alignment with the values

Over the last few years, we have observed a growing awareness surrounding the topics of climate change, and more generally sustainability, across all actors of societies. As of today, we can for example count more than 9000 companies trying to align their strategy with the SGDs, but also an increasing number consumers that are considering sustainability and the climate to be the key obstacles to overcome before the end of the decade (70% according to a 2020 BCG survey) (Kachaner et al., 2020; Zesty, 2017).

Obviously, this shift in mentality has also reached the financial markets with an increasing number of investors that are convinced that companies should be more than simply financially performant (Halbritter & Dorfleitner, 2015). This recent trend is among others embodied by the advent of a new demographic of investors: Millennials and Woman. Investors who openly confess that their main concern is currently that their investments have a good ESG impact (Bernow et al., 2017; Zesty, 2017). With this pressure from the "end-investors" and from various other stakeholders (including regulatory bodies), institutional investors have also

increasingly started to look for investments in alignment with SDG's and to integrate ESG criteria in their decision-making process (Bernow et al., 2017; Boffo & Patalano, 2020).

ii. Search for better long-term returns

"Society is demanding that companies, both public and private, serve a social purpose. To prosper over time, every company must not only deliver financial performance, but also show how it makes a positive contribution to society." (Fink, 2018). This famous 2018 quote from Blackrock's CEO illustrates another recent trend across the financial sector: Investors believe non-financial performance can create value. So investors might not only choose for responsible investments because it is ethical, but because responsible companies, i.e., companies with a better non-financial performance, could have a better financial performance (Bernow et al., 2017; Delevingne & Gründler, 2020; Halbritter & Dorfleitner, 2015; KPMG UK, 2019; Mirvis & Ryu, 2009). In fact, as of today most investment professionals and executives are convinced that ESG initiatives will certainly have an interesting pay-off at a long-term horizon (Delevingne & Gründler, 2020).

When asked to explain this positive link, CFO's, investors or ESG professionals generally mention reputation, employee attraction and motivation as key performance drivers (BCG Global, n.d.; Boffo & Patalano, 2020; Delevingne & Gründler, 2020; Gartenberg et al., 2018; Mirvis & Ryu, 2009). Besides, it is also believed that implementing ESG programs can develop an innovative mindset and bring a gain in efficiency, thereby opening the path to resilient competitive advantages and increased financial performance (Boffo & Patalano, 2020; Mirvis & Ryu, 2009). Finally, ESG investing is considered to allow for a better risk management, and therefore improve financial returns (Bernow et al., 2017; Boffo & Patalano, 2020; Mirvis & Ryu, 2009).

II. ESG and Financial Performance

Though a lot of investors seems to be convinced that there exists a positive relationship between financial and non-financial performance (because of the several performance drivers that sustainability programs can unleash), we are still lacking clear empirical results to demonstrate this hypothesis. Over the years, several scientific papers have focused on testing for the presence of this positive correlation, but unfortunately their results are far from unanimous.

According to Friede et al. (2015), more than 2000 empirical studies have been written about this topic, and among those, 90% concluded that ESG performance had a positive influence on corporate performance. But, when looking at the performance of responsible investment strategies (SRI funds), another research conducted by Capelle-Blancard & Monjon (2012) showed that a large majority of the reviewed papers did not find any difference between SRI funds and regular funds in terms of performance.

Even when considering individual stocks instead of investment strategies it is complicated to find unambiguous results. For instance, Eccles et al. (2014), have demonstrated that ESG performant companies, i.e. companies with a high ESG rating according to Robecco SAM and Asset4, tended to outperform bad performers when taking a long-term horizon (they tracked the performance of 180 US companies over 18 years). Similar results were found by studies with equivalent methodologies but focusing on one aspect of sustainability, e.g., impact of ESG on the employee performance that impacted stock performance by Gartenberg et al. (2018). But in another recent paper, by studying the full ESG data universe of both Asset4, Bloomberg and KLD, Halbritter & Dorfleitner (2015) have come to the conclusion that you cannot expect to achieve abnormal return from trading ESG high performer long and bad performers low. According to this paper, the main reason why past studies found a different result is because they are biased by the choice of rating provider, and study sample.

This last point illustrates the importance that ESG ratings also play in the scientific literature. Because of the complexity of assessing ESG performance (evolving concept, no common reporting standards, etc.), scientists are generally relying on proxy, i.e., ESG ratings, to evaluate companies performance (Mayor, 2019; Mooij, 2017; Windolph, 2011). It is therefore of high interest to keep investigating ways of improving availability and quality of those ratings (Chelawat & Trivedi, 2013).

III. Consumers of ESG ratings

If ESG ratings have gained such visibility and recognition over the last decades, it is largely due to an increasing demand for ESG information from financial stakeholders. However, the consumers of ESG scores also includes companies, wishing to measure their social impact, scholars, studying the performance of ESG top scorer, consumers, governments, etc. (Saadaoui & Soobaroyen, 2018; Scalet & Kelly, 2010).

1.1.1. Investors

Firstly, it is obvious many of the users of sustainability scores are the broad spectrum of investors considering themselves as responsible. Based on a 2019 OECD report (Boffo & Patalano, 2020), this includes both ESG investors, i.e., investors integrating ESG metrics in their decision process to improve the long-term value of their portfolios, and more socially oriented investors applying strategies like Social or Impact Investing, i.e., strategies that aim at jointly reaching high social and financial return. The distinction between those three different ways of investing responsibly can be seen on figure 3, which displays the full spectrum of social and financial investing. Besides, note that the investors using ESG ratings includes a broad set of investment professionals, e.g., asset managers, institutional investors and smaller end-investors from both the private and public sector (Boffo & Patalano, 2020).

Concerning the way investors are making use of ESG information, it is worth mentioning that as of today integrating non-financial metrics into investment decisions "remains much more an art than a science" (Doyle, 2018, p. 7). There is currently no common and standardized way of using ESG information (Amel-Zadeh & Serafeim, 2017; Wong & Petroy, 2020). For instance, the Rate the raters report issued by SustainAbility noted divergences between large and smaller investors, where the first mentioned tended to be more keen to use ratings only as a starting point for their ESG analysis, relying then on their internal expertise and manpower to improve and adapt the results to their specific situation and purposes (Wong & Petroy, 2020). In addition, this same report highlights the fact that most investors will rely on more than one ESG rating to make investment decisions.

Despite this variety of possible usage and methodologies across investors, in general we can observe that most are using ESG ratings to define where they should allocate their capital to comply with their sustainability objectives. In a survey of 413 investment professionals conducted by Amel-Zadeh & Serafeim (2017), it has been highlighted that ratings were

generally used to define the boarders of an investors investment universe. In other words, by considering ESG scores as proxy for ESG performance, investors will screen the capital market and exclude bad performers from their portfolios, i.e., apply a negative screening strategy (Halbritter & Dorfleitner, 2015; López-Arceiz et al., 2020; Mooij, 2017). Currently, a smaller number of investors are using ratings for positive screening purposes or active ownership, i.e., engaging in the management of firms, but it is expected to increase in importance in the near future as more and more investors are becoming convinced that investing in 'responsible' companies can generate abnormal return (Amel-Zadeh & Serafeim, 2017). We can also mention that at least 1/3 of surveyed investors are integrating ESG rating into their valuation model.

1.1.2. Academic Research

As previously stated, ESG ratings are not exclusively reserved to financial professionals and can also support various actors in their respective projects. For instance, sustainability ratings happen to have a relative influence in the field of academic research. Scholars willing to study the ESG performance of companies, e.g., the relationship between Corporate Sustainable Performance and Financial performance, tended to rely on ESG scores provided by rating agencies as proxies (Crespi & Migliavacca, 2020). For example, Halbritter & Dorfleitner (2015) relied on ESG data from three different rating providers, namely Thomson Reuters, Bloomberg and MSCI, to evaluate the correlation between financial and non-financial performance, while other authors like Eccles et al. (2014) used similar ratings to define how high ESG performance could impact the processes and performance of a company.

The multiplication and the growing availability of ESG scores has therefore supported and facilitated the development of knowledge around the benefits of integrating into business and financial decision-processes, thereby further improving the attractivity of sustainable investments and companies. However, it goes without saying that this can only be beneficial if the ratings are valid and credible in the eyes of society, otherwise we would have the put an important number of studies into questions. It is therefore essential to keep investigating ways of improving the quality of ESG ratings and address the numerous shortcomings they are still facing.

1.1.3. Companies

Another important category of ESG ratings users to consider is the rated companies themselves. It is mentioned in several papers that many corporations tended to rely on ratings to evaluate their global sustainability impact and adjust their actions in consequences (Blank et al., 2016;

Halbritter & Dorfleitner, 2015). For instance, Gyönyörová et al. (2021) cites a set of studies showing how the performance in terms of toxic pollution of poor ESG scorers was improving more importantly than their high-scoring peers. However, it is worth mentioning that the impact of ratings on companies is not homogenous as it will depend on "managers' beliefs regarding the material benefits of adjusting to and scoring well on ESG ratings and their alignment with corporate strategy" (Clementino & Perkins, 2021, p. 1).

Some authors argue that ESG ratings would rather be used by firms as learning tools (López-Arceiz et al., 2020; Scalet & Kelly, 2010). The idea here is to consider that ratings can be seen as "a forum for conveying best practices, reduce cumulative negative events across the business system" (Scalet & Kelly, 2010, p. 79). In other words, corporations can use ESG ratings to identify good practices and processes in terms of sustainability among competitors of its sectors.

IV. Evaluation model – Choice of metrics and criteria

Figure 14: Most common Positive (right) and Negative (left) Criteria used by ESG rating agencies in their custom ESG rating methodologies (Escrig-Olmedo et al., 2010, pp. 450–451)

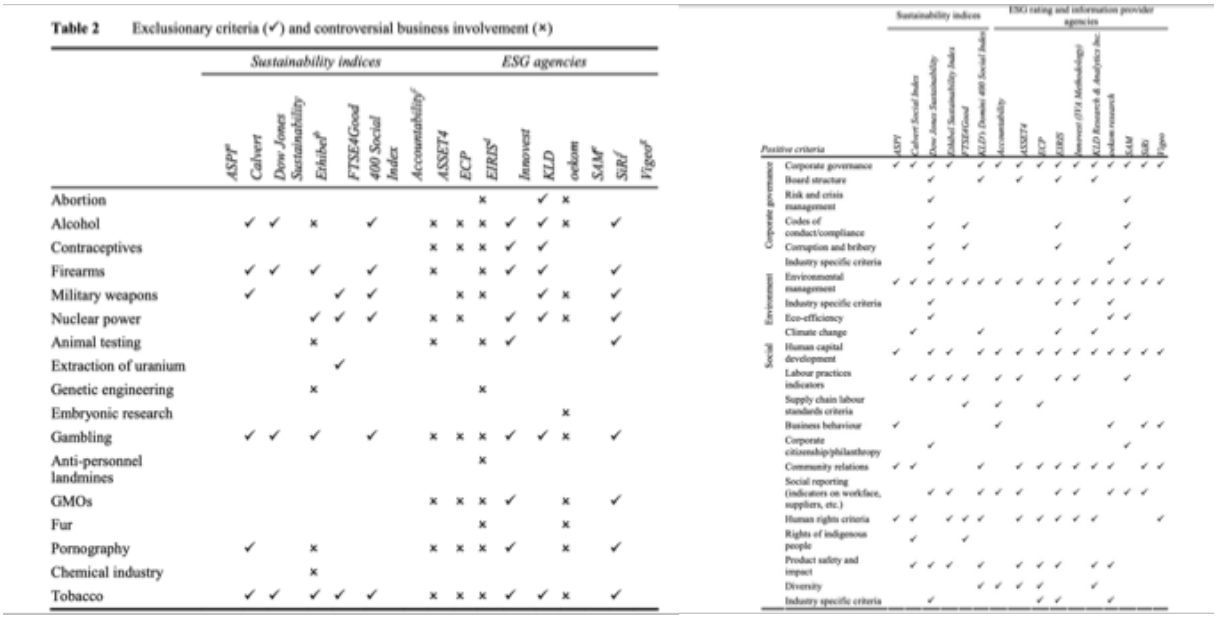


Table 2 Exclusionary criteria (✓) and controversial business involvement (×)

	Sustainability indices							*ESG agencies*							
	ASPI[a]	*Calvert*	*Dow Jones Sustainability*	*Ethibel*[b]	*FTSE4Good*	*400 Social Index*	*Accountability*[c]	*ASSET4*	*ECP*	*EIRIS*[d]	*Innovest*	*KLD*	*oekom*	*SAM*[e]	*SiRi/ Vigeo*[f]
Abortion										×		✓	×		
Alcohol		✓	✓	×		✓		×	×	×	✓	✓	×		✓
Contraceptives								×	×	×	✓	✓			
Firearms		✓	✓	✓		✓		×		×	✓	✓			✓
Military weapons		✓			✓	✓			×	×		✓	×		✓
Nuclear power				✓	✓	✓		×	×		✓	✓	×		✓
Animal testing				×				×		×	✓				✓
Extraction of uranium					✓										
Genetic engineering				×						×					
Embryonic research													×		
Gambling		✓	✓	✓		✓		×	×	×	✓	✓	×		✓
Anti-personnel landmines										×					
GMOs								×	×	×	✓		×		✓
Fur										×			×		
Pornography		✓		×				×	×	×	✓		×		✓
Chemical industry				×											
Tobacco		✓	✓	✓	✓	✓		×	×	×	✓	✓	×		✓

Once every criterion, and potential sub-criterion, has been defined, agencies still need to complete their evaluation model by identifying what metrics is relevant to assess the performance of the company in terms of *Resource Use*, *Emissions*, *Board Diversity*, etc. once they are aggregated (Boffo & Patalano, 2020). As there is no unique framework or methodology, the choice of metrics will diverge from agency to agency in function of the chosen criteria, the data sources and the audiences they are willing to reach (Crespi & Migliavacca, 2020; Escrig-Olmedo et al., 2010; Scalet & Kelly, 2010)

(1) International standards and frameworks:

Both the criteria mix and the choice of metrics, are usually decisions taken independently by the agency itself. Unfortunately, this decision is hardly ever justified or at least explained in the providers documentation (Escrig-Olmedo et al., 2019; LaBella et al., 2019; Saadaoui & Soobaroyen, 2018). However, several papers have noted that most agencies derived their evaluation model both from a series of international standards on environmental management and from corporate social responsibility conventions (Escrig-Olmedo et al., 2010, 2019; Saadaoui & Soobaroyen, 2018; Semenova & Hassel, 2015). There are multiple existing standards, or frameworks, that are wishing to offer guidance to investors and rating agencies in their willingness to evaluate non-financial performance. Most of those standards are disclosure frameworks and are therefore not fundamentally intended to create ratings. But as it defines a

set of rules and guidelines on what and how information should be made public by companies, it clearly defines what factors are relevant or not to assess the sustainability of a company (Boffo & Patalano, 2020). In addition, since those standards are used by the assessed company to define the ESG metrics they agree to disclose, the available ESG data is conditional to the frameworks used by the company.

Among the various existing standard setters, the Sustainability Accounting Standard Board (SASB), the International Integrated Reporting Council (IIRC) and the Global Reporting Initiative (GRI) are the most used. Aside from those general, or global, disclosure frameworks, over the years, initiatives have also emerged to define disclosure standards for specific ESG issues, e.g., the Taskforce on Climate-related Financial Disclosure (TCFD) has become a popular reporting standard in the domain of climate risks (Boffo & Patalano, 2020; Madison & Schiehll, 2021).

(2) Stakeholders Model approach

A second commonality between agencies when choosing criteria and metrics is that they take their decisions based on an adapted Stakeholders Model approach (Escrig-Olmedo et al., 2010, 2019; Saadaoui & Soobaroyen, 2018; Semenova & Hassel, 2015). To put it simply, they will try to assess ESG issues that are relevant for the needs and expectations of the different stakeholders of the evaluated company. Saadaoui & Soobaroyen (2018) have therefore noted that some rating providers tended to take a client-led approach, hence tried to involve the end-user as much as possible to create a score adapted to its needs. For instance, they mentioned that they are reporting the controversial activities but let the client decide whether a company's involvement in such activity should be exclusionary or not.

V. Refinitiv Rating Methodology:

In terms of methodology, to establish these scores, analysts of Refinitiv are relying on public information published by the companies (Annual Reports, CSR Reports, Company website and SEC fillings) or external stakeholders (NGO websites and News Sources[9]). From those various sources, more than 500 ESG metrics are then manually extracted and grouped into 10 different ESG sub-category scores. Category scores are computed on an industry basis (values of the metrics are compared to the values of companies belonging to the same industry according to Refinitiv's classification) by using a percentile rank scoring methodology, i.e.:

$$C = \frac{n°\ of\ companies\ with\ a\ worse\ value\ +\ 0.5\ \times\ n°\ of\ companies\ with\ same\ value}{n°\ of\ companies\ with\ a\ value}$$

Finally, the ESG score of the company i is obtained by a weighted sum of the different category scores C_j^i, where the weightings matrix is the custom materiality matrix developed by Refinitiv. As for the category scoring, the materiality matrix w_j^k will vary in function of the considered industry k (Blank et al., 2016; Huber & Comstock, 2017; Refinitiv, 2021).

$$\mathrm{ESG^i} = \sum_{j=1}^{10} w_j^k \times C_j^i$$

This score has a value between 0 and 100% which can then be converted to a letter-rating using the conversion rules of **Figure 15**.

Figure 15: Refinitiv's Scores conversion table (Refinitiv, 2021, p. 7)

Score range	Grade	Description
0.0 <= score <= 0.083333	D -	'D' score indicates poor relative ESG performance and insufficient degree of transparency in reporting material ESG data publicly.
0.083333 < score <= 0.166666	D	
0.166666 < score <= 0.250000	D +	
0.250000 < score <= 0.333333	C -	'C' score indicates satisfactory relative ESG performance and moderate degree of transparency in reporting material ESG data publicly.
0.333333 < score <= 0.416666	C	
0.416666 < score <= 0.500000	C +	
0.500000 < score <= 0.583333	B -	'B' score indicates good relative ESG performance and above-average degree of transparency in reporting material ESG data publicly.
0.583333 < score <= 0.666666	B	
0.666666 < score <= 0.750000	B +	
0.750000 < score <= 0.833333	A -	'A' score indicates excellent relative ESG performance and high degree of transparency in reporting material ESG data publicly.
0.833333 < score <= 0.916666	A	
0.916666 < score <= 1	A +	

[9] Note that News sources are rather used for the building of the ESG controversy score than for the "regular" ESG score (Refinitiv, 2021)

VI. Sample of Companies

Figure 16: List of Companies from the MSCI World Index which are included in the Data Sample

List of companies from the MSCI world Index						
Boeing Co	Raytheon Technologies Corp	Airbus SE	L3harris Technologies Inc	Safran SA	BAE Systems PLC	
Rolls-Royce Holdings PLC	Toyota Motor Corp	General Motors Co	Daimler AG	Ford Motor Co	Volkswagen AG	Tesla Inc
Cummins Inc	Aptiv PLC	Ferrari NV	Bridgestone Corp	Denso Corp	BNP Paribas SA	JPMorgan Chase & Co
Bank of America Corp	Wells Fargo & Co	HSBC Holdings PLC	Citigroup Inc	US Bancorp	Banco Santander SA	Mitsubishi UFJ Financial Group Inc
Bank of Montreal	Royal Bank of Canada	Toronto-Dominion Bank	Commonwealth Bank of Australia	Bank of Nova Scotia	Australia and New Zealand Banking Group Lt	Sumitomo Mitsui Trust Holdings Inc
ING Groep NV	Capital One Financial Corp	Truist Financial Corp	Barclays PLC	Credit Suisse Group AG	Standard Chartered PLC	Oversea-Chinese Banking Corporation Lt
Nordea Bank Abp	Deutsche Bank AG	Regions Financial Corp	Orix Corp	Coca-Cola Co	PepsiCo Inc	Diageo PLC
Pernod Ricard SA	Constellation Brands Inc	Heineken NV	Monster Beverage Corp	Linde PLC	BASF SE	Ecolab Inc
Air Products and Chemicals Inc	Shin-Etsu Chemical Co Ltd	L'Air Liquide Societe Anonyme pour l'Etude et l'Exploitation des Procedes Georges Claude	Koninklijke DSM NV	LyondellBasell Industries NV	Givaudan SA	Akzo Nobel NV
Henkel AG & Co KGaA	Cisco Systems Inc	Telefonaktiebolaget LM Ericsson	Nokia Oyj	Motorola Solutions Inc	Calix Inc	ADT Inc
Apple Inc	Sony Group Corp	Sony Group Corp_2	HP Inc	Hewlett Packard Enterprise Co	TE Connectivity Ltd	Caterpillar Inc
Caterpillar Inc_2	Keyence Corp	Deere & Co	Schneider Electric SE	Emerson Electric Co	Abb Ltd	Johnson Controls International PLC
Eaton Corporation PLC	Fanuc Corp	Daikin Industries Ltd	Nidec Corp	Murata Manufacturing Co Ltd	Parker-Hannifin Corp	Fortive Corp
AMETEK Inc	Rockwell Automation Inc	Kyocera Corp	Kyocera Corp_2	Volvo AB	Sandvik AB	
Kubota Corp	Ball Corp	Amcor PLC	Amazon.com Inc	TJX Companies Inc	Target Corp	Dollar General Corp
Wesfarmers Ltd	Dollar Tree Inc	Mitsubishi Corp	Itochu Corp	Suncor Energy Inc	Exxon Mobil Corp	Chevron Corp
Royal Dutch Shell PLC	BP PLC	TotalEnergies SE	TC Energy Corp	Enbridge Inc	Schlumberger NV	EOG Resources Inc
Walmart Inc	Walgreens Boots Alliance Inc	Sysco Corp	Koninklijke Ahold Delhaize NV	Tesco PLC	Woolworths Group Ltd	Mckesson Corp
Nextera Energy Inc	Iberdrola SA	Enel SpA	American Electric Power Company Inc	Exelon Corp	Consolidated Edison Inc	Xcel Energy Inc
Orsted A/S	Edison International	Union Pacific Corp	United Parcel Service Inc	CSX Corp	Norfolk Southern Corp	FedEx Corp
Canadian Pacific Railway Ltd	Canadian National Railway Co	Deutsche Post AG	Abbott Laboratories	Medtronic PLC	Danaher Corp	Becton Dickinson and Co
Intuitive Surgical Inc	Koninklijke Philips NV	Hoya Corp	Agilent Technologies Inc	Veeva Systems Inc	Smith & Nephew PLC	Cerner Corp
UnitedHealth Group Inc	CVS Health Corp	Fresenius SE & Co KGaA	Nestle SA	Philip Morris International Inc	British American Tobacco PLC	Danone SA
General Mills Inc	Japan Tobacco Inc	Imperial Brands PLC	Kraft Heinz Co	Archer-Daniels-Midland Co	Tyson Foods Inc	McDonald's Corp
Starbucks Corp	Booking Holdings Inc	Marriott International Inc	Yum! Brands Inc	Oriental Land Co Ltd	Hilton Worldwide Holdings Inc	Expedia Group Inc
Las Vegas Sands Corp	Royal Caribbean Cruises Ltd	Chipotle Mexican Grill Inc	Restaurant Brands International Inc	Intel Corp	Texas Instruments Inc	Broadcom Inc
ASML Holding NV	NVIDIA Corp	Qualcomm Inc	Micron Technology Inc	Advanced Micro Devices Inc	Infineon Technologies AG	
NXP Semiconductors NV	Tokyo Electron Ltd	KLA Corp	Microchip Technology Inc	Ferguson PLC	Assa Abloy AB	Geberit AG
Nintendo Co Ltd	CK Hutchison Holdings Ltd	CME Group Inc	Goldman Sachs Group Inc	BlackRock Inc	Morgan Stanley	Intercontinental Exchange Inc
Charles Schwab Corp	Brookfield Asset Management Inc	UBS Group AG	Hong Kong Exchanges and Clearing L	Macquarie Group Ltd	London Stock Exchange Group PLC	Deutsche Boerse AG
Northern Trust Corp	Sempra Energy	National Grid PLC	E.ON SE	PPL Corp	Ameren Corp	AIA Group Ltd
Allianz SE	Zurich Insurance Group AG	Marsh & McLennan Companies Inc	AXA SA	American International Group Inc	Progressive Corp	Aon PLC
MetLife Inc	Prudential Financial Inc	Tokio Marine Holdings Inc	Sun Life Financial Inc	Assicurazioni Generali SpA	Hartford Financial Services Group Inc	Allstate Corp
Sampo plc	Walt Disney Co	Comcast Corp	ViacomCBS Inc	Vivendi SE	Hong Kong and China Gas Co Ltd	BHP Group PLC
Rio Tinto PLC	Glencore PLC	Barrick Gold Corp	Newmont Corporation	Anglo American PLC	Franco-Nevada Corp	Canon Inc
Mettler-Toledo International Inc	Alimentation Couche-Tard Inc	Canadian Natural Resources Ltd	Pioneer Natural Resources Co	Repsol SA	Equinor ASA	Woodside Petroleum Ltd
Kinder Morgan Inc	Williams Companies Inc	Pembina Pipeline Corp	Halliburton Co	UPM-Kymmene Oyj	East Japan Railway Co	Central Japan Railway Co
Procter & Gamble Co	Colgate-Palmolive Co	Reckitt Benckiser Group PLC	Kimberly-Clark Corp	Estee Lauder Companies Inc	L'Oreal SA	Shiseido Co Ltd
Clorox Co	Johnson & Johnson	Novartis AG	Roche Holding AG	Amgen Inc	AstraZeneca PLC	GlaxoSmithKline PLC
Abbvie Inc	Novo Nordisk A/S	Bristol-Myers Squibb Co	CSL Ltd	Bayer AG	Takeda Pharmaceutical Co Ltd	
Zoetis Inc	Biogen Inc	Pfizer Inc	Chugai Pharmaceutical Co Ltd	Fiserv Inc	S&P Global Inc	Waste Management Inc
Relx PLC	Recruit Holdings Co Ltd	Moody's Corp	Square Inc	Experian PLC	Cintas Corp	Waste Connections Inc
Republic Services Inc	Wolters Kluwer NV	Vonovia SE	Mitsubishi Estate Co Ltd	CBRE Group Inc	Mitsui Fudosan Co Ltd	American Tower Corp
Simon Property Group Inc	Prologis Inc	Public Storage	Welltower Inc	Equity Residential	Avalonbay Communities Inc	Ventas Inc
Digital Realty Trust Inc	Realty Income Corp	Link Real Estate Investment Trust	Boston Properties Inc	Meta Platforms Inc	Alphabet Inc	Visa Inc
Mastercard Inc	Adobe Inc	Oracle Corp	Netflix Inc	International Business Machines Corp	PayPal Holdings Inc	Accenture PLC
SAP SE	Salesforce.Com Inc	Activision Blizzard Inc	Roper Technologies Inc	Amadeus IT Group SA	Cognizant Technology Solutions Corp	Shopify Inc
Autodesk Inc	eBay Inc	Twitter Inc	Mercadolibre Inc	Electronic Arts Inc	Verisign Inc	Cadence Design Systems Inc
Dassault Systemes SE	CGI Inc	Capgemini SE	ANSYS Inc	CDW Corp	Twilio Inc	AT&T Inc
Verizon Communications Inc	SoftBank Group Corp	Deutsche Telekom AG	KDDI Corp	Vodafone Group PLC	Telefonica SA	Nippon Telegraph and Telephone Corp
T-Mobile US Inc	Orange SA	SoftBank Corp	Rogers Communications Inc	BT Group PLC	Home Depot Inc	Lowe's Companies Inc
Kering SA	Ross Stores Inc	Fast Retailing Co Ltd	H & M Hennes & Mauritz AB	LVMH Moet Hennessy Louis Vuitton S	Nike Inc	Adidas AG
Compagnie Financiere Richemont SA	EssilorLuxottica SA	VF Corp	Lululemon Athletica Inc	Hermes International SCA	Transurban Group	American Water Works Company Inc

Figure 17: List of Companies out of the MSCI World Index which are included in the Data Sample

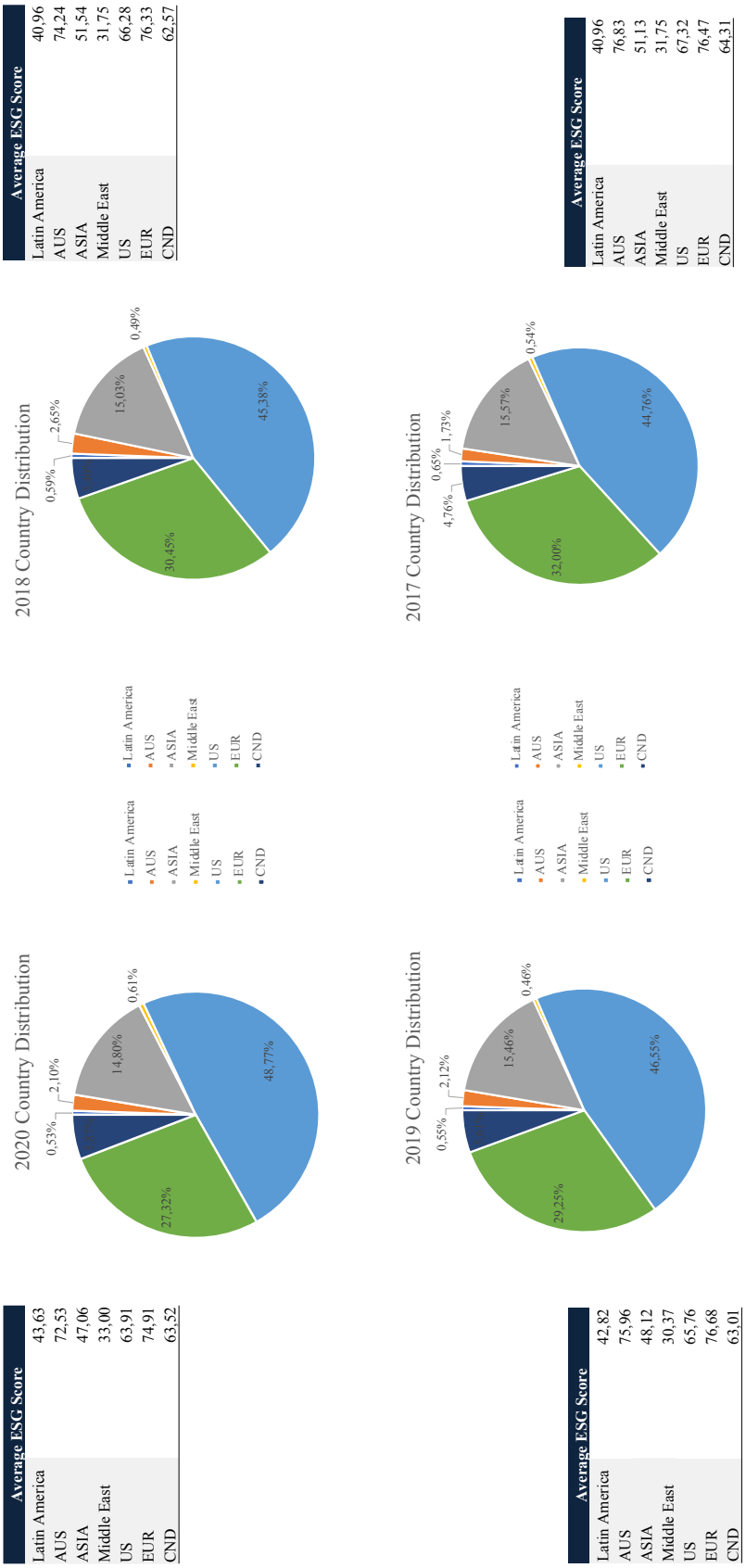
List of companies not from the MSCI world Index						
Lumentum Holdings Inc	Viavi Solutions Inc	Goertek Inc	Shanghai Pudong Development Bank Co Ltd	First Republic Bank	State Bank of India	
SVB Financial Group	Ping An Bank Co Ltd	Signature Bank	Al Rajhi Banking & Investment Corporation SJSC	Industrial Bank Co Ltd	Axis Bank Ltd	China Resources Beer Holdings Co Ltd
Wuliangye Yibin Co Ltd	Western Digital Corp	Xiaomi Corp	Synaptics Inc	Banzl plc	Shimano Inc	Square Enix Holdings Co Ltd
Pool Corp	Take-Two Interactive Software Inc	Merida Industry Co Ltd	Avis Budget Group Inc	United Airlines Holdings Inc	Ryanair Holdings PLC	
Sixt SE	Seazen Holdings Co Ltd	Henderson Land Development Co Ltd	Colliers International Group Inc	Zillow Group Inc	Fastighets AB Balder	Hongkong Land Holdings Ltd
Azrieli Group Ltd	Sumitomo Realty & Development Co Ltd	Emaar Properties PJSC	New World Development Co Ltd	Ceridian HCM Holding Inc	Zoom Video Communications Inc	Fortinet Inc
CoStar Group Inc	Spotify Technology SA	Alibaba Group Holding Ltd	Atlassian Corporation PLC	Ubisoft Entertainment SA	Japan Airport Terminal Co Ltd	Flughafen Zuerich AG

VII. Sample Distribution – Country and Sector

Figure 18: Number of companies per Industry present in the Samples (based on Refinitiv's nomenclature(Refinitiv, 2021))

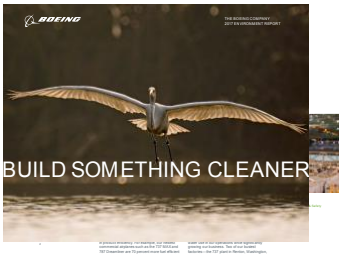
List of Industries (Refinitiv's Categorization)				
	2017	**2018**	**2019**	**2020**
Aerospace and defence	7	7	7	7
Automobiles and auto parts	12	12	12	11
Banking services	36	36	37	36
Beverages	8	8	9	9
Chemicals	10	11	11	11
Communications and networking	8	8	8	8
Computers, phones and household electronics	8	8	9	9
Construction materials	20	20	21	23
Containers and packaging	2	2	2	2
Diversified Retail	6	6	6	6
Diversified industrial goods wholesalers	3	3	3	3
Oil and gas	12	14	13	13
Oil and gas related equipment and services	7	7	7	7
Food and drug retailing	7	7	7	7
Electric utilities and IPPs	9	9	9	9
Freight and logistics services	8	9	8	8
Healthcare equipment and supplies	11	11	11	11
Healthcare providers and services	2	3	3	3
Food and tobacco	10	9	11	10
Hotels and entertainment services	12	12	12	12
Semiconductors and semiconductor equipment	13	13	13	13
Machinery, tools, heavy vehicles, trains and ships	3	3	3	3
Leisure products	6	6	6	6
Investment holding companies	1	1	1	1
Investment banking and investment services	13	13	14	13
Multiline utilities	4	6	5	5
Insurance	14	15	16	16
Media and publishing	2	5	4	4
Natural gas utilities	1	1	1	1
Metals and mining	7	7	7	7
Office equipment	2	2	2	2
Paper and forest products	1	1	1	1
Passenger transportation services	6	6	6	6
Personal and household products and services	8	8	8	8
Pharmaceuticals	15	17	17	16
Professional and commercial services	12	12	12	12
Real estate operations	14	14	14	14
Residential and commercial REITs	11	12	12	12
Software and IT services	37	38	39	38
Telecommunications services	11	12	14	13
Specialty retailers	6	6	6	6
Textiles and apparel	7	8	8	8
Transport infrastructure	3	3	3	3
Water and related utilities	1	1	1	1
Total	**396**	**412**	**419**	**414**

Figure 19: Pie Chart of the number of Company and Table Average ESG scores per Geographical Region in each year sample



VIII. Text Pre-processing – Illustration

Example of ESG words frequency vectors created by the pre-processing: Boeing Co CSR report 2017 and Airbus SE Annual Report 2017



Pre-processing

approve *vulnerable*

□ = (2, 2, 0, …, 5, 2, 0, 1)

Vector of ESG words frequency, dimension (1×482)

The vector displayed above indicates that the word « approve » is repeated twice in Boeing 2017 CSR reports, while « vulnerable » is only used once

Pre-processing

approve *vulnerable*

□ = (25, 43, 1, …, 32, 13, 0, 0)

Vector of ESG words frequency, dimension (1×482)

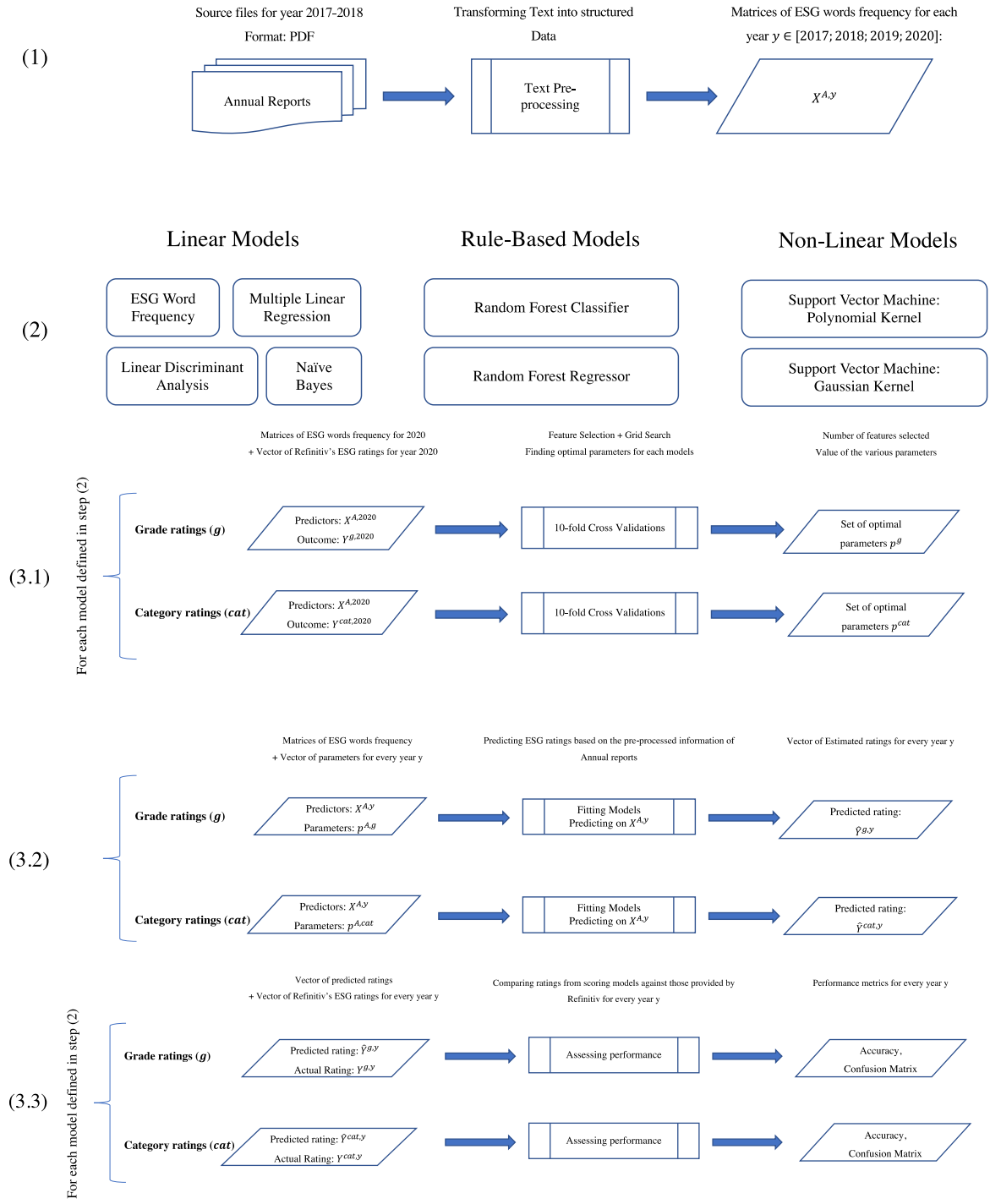
Similarly, the value of the second feature of the vector indicates that the word « approve » is repeated 43 times in Airbus 2017 Annual report, while « vulnerable » is never used.

Figure 20: Illustration of the pre-processing step for Boeing CSR report (above) and Airbus Annual report (below) of 2017

IX. Methodology – General overview

Evaluating performance of Text Mining Based Ratings – Methodology

Illustration for Annual Reports



Methodology

(1) Pre-processing text from Annual Reports
(2) Defining a set of scoring Models
(3.1) Finding the values of the optimal parameters
(3.2) Making out-of-sample predictions
(3.3) Assessing the prediction performance

Figure 21: Schema of the research methodology for sustainability disclosure approximated by Annual Report

X. Variables and Parameters

Name	Definition
Different years of the sample:	$a \in \{2017; 2018; 2019; 2020\}$
Different Types of reports:	$r \in \{A;\ C;\ A + C\}$
Different Format of ratings:	$f \in \{c;\ g;\ cat\}$
Set of Category Ratings	$\mathrm{C} = \{\mathrm{D}; \mathrm{C}; \mathrm{B}; \mathrm{A}\} = \{1; 2; 3; 4\}$
Set of Grade Ratings	$\mathrm{G} = \{\mathrm{D}-; \mathrm{D}; \mathrm{D}+; \mathrm{C}-; \mathrm{C}; \mathrm{C}+; \mathrm{B}-; \mathrm{B}; \mathrm{B}+; \mathrm{A}-; \mathrm{A}; \mathrm{A}+\} = \{1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12\}$
Number of companies per sample for year a and report r	$n^{r,a} \in \{396; 412; 419; 414\}$
Matrix of ESG words frequency for year a and report r	$X^{r,a}\ (n^{r,a} \times \mathrm{p})$
Vector of ESG words frequency for year a and report r of company i	$\boldsymbol{w_i^{r,a}}\ (1 \times p)$
Number of ESG words	$p = 482\ \forall r, y$ and i
Frequency of ESG word j in report r of company i for year a	$w_{i,j}^{r,a} \in \boldsymbol{R}^+$
Vector of Refinitiv's ratings under format f for year a and report r	$Y^{f,a}(n^{r,a} \times 1)$

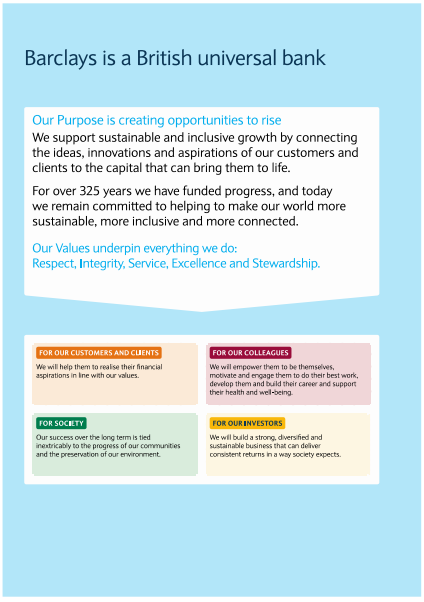
Refinitiv's rating under continuous format of company i for year y and report r	$y_i{}^{c,a} \in [0;100]$
Refinitiv's rating under grade format of company i for year y and report r	$y_i{}^{g,a} \in \mathrm{G}$
Refinitiv's rating under category format of company i for year y and report r	$y_i{}^{g,a} \in \mathrm{C}$
Prediction of variable λ	$\hat{\lambda}$

XI. Text Pre-processing – Example

To illustrate how we are proposing to transform a report into a vector of ESG-words frequency, we are displaying here the different pre-processing steps on the 1st page of the ESG report from Barclays PLC in 2019.

As explained in section 3.3, a first step of our text pre-processing methodology consists in the removal of non-ASCII elements.

Step 1: Remove Non-ASCII Characters



Barclays is a British universal bank
Our Purpose is creating opportunities to rise
We support sustainable and inclusive growth by connecting the ideas, innovations and aspirations of our customers and clients to the capital that can bring them to life.
For over 325 years we have funded progress, and today we remain committed to helping to make our world more sustainable, more inclusive and more connected.
Our Values underpin everything we do:
Respect, Integrity, Service, Excellence and Stewardship.

FOR OUR CUSTOMERS AND CLIENTS
We will help them to realise their financial aspirations in line with our values.

FOR SOCIETY
Our success over the long term is tied

inextricably to the progress of our communities and the preservation of our environment.

FOR OUR COLLEAGUES
We will empower them to be themselves, motivate and engage them to do their best work, develop them and build their career and support their health and well-being.

FOR OUR INVESTORS
We will build a strong, diversified and sustainable business that can deliver consistent returns in a way society expects.

Then, we proceed to a tokenization of the resulting text. As displayed below, this consist in the creation of a vector whose features are obtained by splitting the original text on white spaces. In the schema presented below, we combined both the tokenization and the lowercasing step.

Step 2: Tokenization

Barclays is a British universal bank
Our Purpose is creating opportunities to rise
We support sustainable and inclusive growth by connecting the ideas, innovations and aspirations of our customers and clients to the capital that can bring them to life.
For over 325 years we have funded progress, and today we remain committed to helping to make our world more sustainable, more inclusive and more connected.
Our Values underpin everything we do:
Respect, Integrity, Service, Excellence and Stewardship.

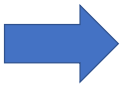
FOR OUR CUSTOMERS AND CLIENTS
We will help them to realise their financial aspirations in line with our values.

FOR SOCIETY
Our success over the long term is tied

inextricably to the progress of our communities and the preservation of our environment.

FOR OUR COLLEAGUES
We will empower them to be themselves, motivate and engage them to do their best work, develop them and build their career and support their health and well-being.

FOR OUR INVESTORS
We will build a strong, diversified and sustainable business that can deliver consistent returns in a way society expects.



['barclays', 'is', 'a', 'british', 'universal', 'bank', 'our', 'purpose', 'is', 'creating', 'opportunities', 'to', 'rise', 'we', 'support', 'sustainable', 'and', 'inclusive', 'growth', 'by', 'connecting', 'the', 'ideas', ',', 'innovations', 'and', 'aspirations', 'of', 'our', 'customers', 'and', 'clients', 'to', 'the', 'capital', 'that', 'can', 'bring', 'them', 'to', 'life', '.', 'for', 'over', '325', 'years', 'we', 'have', 'funded', 'progress', ',', 'and', 'today', 'we', 'remain', 'committed', 'to', 'helping', 'to', 'make', 'our', 'world', 'more', 'sustainable', ',', 'more', 'inclusive', 'and', 'more', 'connected', '.', 'our', 'values', 'underpin', 'everything', 'we', 'do', ':', 'respect', ',', 'integrity', ',', 'service', ',', 'excellence', 'and', 'stewardship', '.', 'for', 'our', 'customers', 'and', 'clients', 'we', 'will', 'help', 'them', 'to', 'realise', 'their', 'financial', 'aspirations', 'in', 'line', 'with', 'our', 'values', '.', 'for', 'society', '\x07', 'ur', 'success', 'over', 'the', 'long', 'term', 'is', 'tied', 'o', 'inextricably', 'to', 'the', 'progress', 'of', 'our', 'communities', 'and', 'the', 'preservation', 'of', 'our', 'environment', '.', 'for', 'our', 'colleagues', 'we', 'will', 'empower', 'them', 'to', 'be', 'themselves', ',', 'motivate', 'and', 'engage', 'them', 'to', 'do', 'their', 'best', 'work', ',', 'develop', 'them', 'and', 'build', 'their', 'career', 'and', 'support', 'their', 'health', 'and', 'well-being', '.', 'for', 'our', 'investors', 'we', 'will', 'build', 'a', 'strong', ',', 'diversified', 'and', 'sustainable', 'business', 'that', 'can', 'deliver', 'consistent', 'returns', 'in', 'a', 'way', 'society', 'expects', '.']

Thereafter, we remove the punctuation (highlighted in yellow in the schema below) from the vector of tokens.

Step 3: Removing Punctuation

['barclays', 'is', 'a', 'british', 'universal', 'bank', 'our', 'purpose', 'is', 'creating', 'opportunities', 'to', 'rise', 'we', 'support', 'sustainable', 'and', 'inclusive', 'growth', 'by', 'connecting', 'the', 'ideas', ',', 'innovations', 'and', 'aspirations', 'of', 'our', 'customers', 'and', 'clients', 'to', 'the', 'capital', 'that', 'can', 'bring', 'them', 'to', 'life', ',', 'for', 'over', '325', 'years', 'we', 'have', 'funded', 'progress', ',', 'and', 'today', 'we', 'remain', 'committed', 'to', 'helping', 'to', 'make', 'our', 'world', 'more', 'sustainable', ',', 'more', 'inclusive', 'and', 'more', 'connected', '.', 'our', 'values', 'underpin', 'everything', 'we', 'do', ':', 'respect', ',', 'integrity', ',', 'service', ',', 'excellence', 'and', 'stewardship', '.', 'for', 'our', 'customers', 'and', 'clients', 'we', 'will', 'help', 'them', 'to', 'realise', 'their', 'financial', 'aspirations', 'in', 'line', 'with', 'our', 'values', ',', 'for', 'society', '\x07', 'ur', 'success', 'over', 'the', 'long', 'term', 'is', 'tied', 'o', 'inextricably', 'to', 'the', 'progress', 'of', 'our', 'communities', 'and', 'the', 'preservation', 'of', 'our', 'environment', ',', 'for', 'our', 'colleagues', 'we', 'will', 'empower', 'them', 'to', 'be', 'themselves', ',', 'motivate', 'and', 'engage', 'them', 'to', 'do', 'their', 'best', 'work', ',', 'develop', 'them', 'and', 'build', 'their', 'career', 'and', 'support', 'their', 'health', 'and', 'well-being', ',', 'for', 'our', 'investors', 'we', 'will', 'build', 'a', 'strong', ',', 'diversified', 'and', 'sustainable', 'business', 'that', 'can', 'deliver', 'consistent', 'returns', 'in', 'a', 'way', 'society', 'expects', '.']

['barclays', 'is', 'a', 'british', 'universal', 'bank', 'our', 'purpose', 'is', 'creating', 'opportunities', 'to', 'rise', 'we', 'support', 'sustainable', 'and', 'inclusive', 'growth', 'by', 'connecting', 'the', 'ideas', '', 'innovations', 'and', 'aspirations', 'of', 'our', 'customers', 'and', 'clients', 'to', 'the', 'capital', 'that', 'can', 'bring', 'them', 'to', 'life', '', 'for', 'over', '325', 'years', 'we', 'have', 'funded', 'progress', '', 'and', 'today', 'we', 'remain', 'committed', 'to', 'helping', 'to', 'make', 'our', 'world', 'more', 'sustainable', '', 'more', 'inclusive', 'and', 'more', 'connected', '', 'our', 'values', 'underpin', 'everything', 'we', 'do', '', 'respect', '', 'integrity', '', 'service', '', 'excellence', 'and', 'stewardship', '', 'for', 'our', 'customers', 'and', 'clients', 'we', 'will', 'help', 'them', 'to', 'realise', 'their', 'financial', 'aspirations', 'in', 'line', 'with', 'our', 'values', '', 'for', 'society', '\x07', 'ur', 'success', 'over', 'the', 'long', 'term', 'is', 'tied', 'o', 'inextricably', 'to', 'the', 'progress', 'of', 'our', 'communities', 'and', 'the', 'preservation', 'of', 'our', 'environment', '', 'for', 'our', 'colleagues', 'we', 'will', 'empower', 'them', 'to', 'be', 'themselves', '', 'motivate', 'and', 'engage', 'them', 'to', 'do', 'their', 'best', 'work', '', 'develop', 'them', 'and', 'build', 'their', 'career', 'and', 'support', 'their', 'health', 'and', 'wellbeing', '', 'for', 'our', 'investors', 'we', 'will', 'build', 'a', 'strong', '', 'diversified', 'and', 'sustainable', 'business', 'that', 'can', 'deliver', 'consistent', 'returns', 'in', 'a', 'way', 'society', 'expects', '']

Next, we are deleting figures and other non-alphabetical characters (highlighted in yellow)

Step 4: Removing Non-alphabetical characters

['barclays', 'is', 'a', 'british', 'universal', 'bank', 'our', 'purpose', 'is', 'creating', 'opportunities', 'to', 'rise', 'we', 'support', 'sustainable', 'and', 'inclusive', 'growth', 'by', 'connecting', 'the', 'ideas', '', 'innovations', 'and', 'aspirations', 'of', 'our', 'customers', 'and', 'clients', 'to', 'the', 'capital', 'that', 'can', 'bring', 'them', 'to', 'life', '', 'for', 'over', '325', 'years', 'we', 'have', 'funded', 'progress', '', 'and', 'today', 'we', 'remain', 'committed', 'to', 'helping', 'to', 'make', 'our', 'world', 'more', 'sustainable', '', 'more', 'inclusive', 'and', 'more', 'connected', '', 'our', 'values', 'underpin', 'everything', 'we', 'do', '', 'respect', '', 'integrity', '', 'service', '', 'excellence', 'and', 'stewardship', '', 'for', 'our', 'customers', 'and', 'clients', 'we', 'will', 'help', 'them', 'to', 'realise', 'their', 'financial', 'aspirations', 'in', 'line', 'with', 'our', 'values', '', 'for', 'society', '\x07', 'ur', 'success', 'over', 'the', 'long', 'term', 'is', 'tied', 'o', 'inextricably', 'to', 'the', 'progress', 'of', 'our', 'communities', 'and', 'the', 'preservation', 'of', 'our', 'environment', '', 'for', 'our', 'colleagues', 'we', 'will', 'empower', 'them', 'to', 'be', 'themselves', '', 'motivate', 'and', 'engage', 'them', 'to', 'do', 'their', 'best', 'work', '', 'develop', 'them', 'and', 'build', 'their', 'career', 'and', 'support', 'their', 'health', 'and', 'wellbeing', '', 'for', 'our', 'investors', 'we', 'will', 'build', 'a', 'strong', '', 'diversified', 'and', 'sustainable', 'business', 'that', 'can', 'deliver', 'consistent', 'returns', 'in', 'a', 'way', 'society', 'expects', '']

['barclays', 'is', 'a', 'british', 'universal', 'bank', 'our', 'purpose', 'is', 'creating', 'opportunities', 'to', 'rise', 'we', 'support', 'sustainable', 'and', 'inclusive', 'growth', 'by', 'connecting', 'the', 'ideas', 'innovations', 'and', 'aspirations', 'of', 'our', 'customers', 'and', 'clients', 'to', 'the', 'capital', 'that', 'can', 'bring', 'them', 'to', 'life', 'for', 'over', 'years', 'we', 'have', 'funded', 'progress', 'and', 'today', 'we', 'remain', 'committed', 'to', 'helping', 'to', 'make', 'our', 'world', 'more', 'sustainable', 'more', 'inclusive', 'and', 'more', 'connected', 'our', 'values', 'underpin', 'everything', 'we', 'do', 'respect', 'integrity', 'service', 'excellence', 'and', 'stewardship', 'for', 'our', 'customers', 'and', 'clients', 'we', 'will', 'help', 'them', 'to', 'realise', 'their', 'financial', 'aspirations', 'in', 'line', 'with', 'our', 'values', 'for', 'society', 'ur', 'success', 'over', 'the', 'long', 'term', 'is', 'tied', 'o', 'inextricably', 'to', 'the', 'progress', 'of', 'our', 'communities', 'and', 'the', 'preservation', 'of', 'our', 'environment', 'for', 'our', 'colleagues', 'we', 'will', 'empower', 'them', 'to', 'be', 'themselves', 'motivate', 'and', 'engage', 'them', 'to', 'do', 'their', 'best', 'work', 'develop', 'them', 'and', 'build', 'their', 'career', 'and', 'support', 'their', 'health', 'and', 'wellbeing', 'for', 'our', 'investors', 'we', 'will', 'build', 'a', 'strong', 'diversified', 'and', 'sustainable', 'business', 'that', 'can', 'deliver', 'consistent', 'returns', 'in', 'a', 'way', 'society', 'expects']

The biggest reduction of dimension happens when removing the stop words from the remaining tokens.

Step 5: Removing Stop Words

['barclays', 'is', 'a', 'british', 'universal', 'bank', 'our', 'purpose', 'is', 'creating', 'opportunities', 'to', 'rise', 'we', 'support', 'sustainable', 'and', 'inclusive', 'growth', 'by', 'connecting', 'the', 'ideas', 'innovations', 'and', 'aspirations', 'of', 'our', 'customers', 'and', 'clients', 'to', 'the', 'capital', 'that', 'can', 'bring', 'them', 'to', 'life', 'for', 'over', 'years', 'we', 'have', 'funded', 'progress', 'and', 'today', 'we', 'remain', 'committed', 'to', 'helping', 'to', 'make', 'our', 'world', 'more', 'sustainable', 'more', 'inclusive', 'and', 'more', 'connected', 'our', 'values', 'underpin', 'everything', 'we', 'do', 'respect', 'integrity', 'service', 'excellence', 'and', 'stewardship', 'for', 'our', 'customers', 'and', 'clients', 'we', 'will', 'help', 'them', 'to', 'realise', 'their', 'financial', 'aspirations', 'in', 'line', 'with', 'our', 'values', 'for', 'society', 'ur', 'success', 'over', 'the', 'long', 'term', 'is', 'tied', 'o', 'inextricably', 'to', 'the', 'progress', 'of', 'our', 'communities', 'and', 'the', 'preservation', 'of', 'our', 'environment', 'for', 'our', 'colleagues', 'we', 'will', 'empower', 'them', 'to', 'be', 'themselves', 'motivate', 'and', 'engage', 'them', 'to', 'do', 'their', 'best', 'work', 'develop', 'them', 'and', 'build', 'their', 'career', 'and', 'support', 'their', 'health', 'and', 'wellbeing', 'for', 'our', 'investors', 'we', 'will', 'build', 'a', 'strong', 'diversified', 'and', 'sustainable', 'business', 'that', 'can', 'deliver', 'consistent', 'returns', 'in', 'a', 'way', 'society', 'expects']

['barclays', 'british', 'universal', 'bank', 'purpose', 'creating', 'opportunities', 'rise', 'support', 'sustainable', 'inclusive', 'growth', 'connecting', 'ideas', 'innovations', 'aspirations', 'customers', 'clients', 'capital', 'bring', 'life', 'years', 'funded', 'progress', 'today', 'remain', 'committed', 'helping', 'make', 'world', 'sustainable', 'inclusive', 'connected', 'values', 'underpin', 'everything', 'respect', 'integrity', 'service', 'excellence', 'stewardship', 'customers', 'clients', 'help', 'realise', 'financial', 'aspirations', 'line', 'values', 'society', 'ur', 'success', 'long', 'term', 'tied', 'inextricably', 'progress', 'communities', 'preservation', 'environment', 'colleagues', 'empower', 'motivate', 'engage', 'best', 'work', 'develop', 'build', 'career', 'support', 'health', 'wellbeing', 'investors', 'build', 'strong', 'diversified', 'sustainable', 'business', 'deliver', 'consistent', 'returns', 'way', 'society', 'expects']

Then, words are transformed back to their root form based on a lemmatization algorithm.

Step 6: Lemmatization

['barclays', 'british', 'universal', 'bank', 'purpose', 'creating', 'opportunities', 'rise', 'support', 'sustainable', 'inclusive', 'growth', 'connecting', 'ideas', 'innovations', 'aspirations', 'customers', 'clients', 'capital', 'bring', 'life', 'years', 'funded', 'progress', 'today', 'remain', 'committed', 'helping', 'make', 'world', 'sustainable', 'inclusive', 'connected', 'values', 'underpin', 'everything', 'respect', 'integrity', 'service', 'excellence', 'stewardship', 'customers', 'clients', 'help', 'realise', 'financial', 'aspirations', 'line', 'values', 'society', 'ur', 'success', 'long', 'term', 'tied', 'inextricably', 'progress', 'communities', 'preservation', 'environment', 'colleagues', 'empower', 'motivate', 'engage', 'best', 'work', 'develop', 'build', 'career', 'support', 'health', 'wellbeing', 'investors', 'build', 'strong', 'diversified', 'sustainable', 'business', 'deliver', 'consistent', 'returns', 'way', 'society', 'expects']

['barclays', 'british', 'universal', 'bank', 'purpose', 'create', 'opportunity', 'rise', 'support', 'sustainable', 'inclusive', 'growth', 'connect', 'idea', 'innovation', 'aspiration', 'customer', 'client', 'capital', 'bring', 'life', 'year', 'fund', 'progress', 'today', 'remain', 'commit', 'help', 'make', 'world', 'sustainable', 'inclusive', 'connect', 'value', 'underpin', 'everything', 'respect', 'integrity', 'service', 'excellence', 'stewardship', 'customer', 'client', 'help', 'realise', 'financial', 'aspiration', 'line', 'value', 'society', 'ur', 'success', 'long', 'term', 'tie', 'inextricably', 'progress', 'community', 'preservation', 'environment', 'colleague', 'empower', 'motivate', 'engage', 'best', 'work', 'develop', 'build', 'career', 'support', 'health', 'wellbeing', 'investor', 'build', 'strong', 'diversified', 'sustainable', 'business', 'deliver', 'consistent', 'return', 'way', 'society', 'expect']

This vector is eventually transformed into a vector of frequency. For instance, we can see that the 9th word from the vector, "support", is present twice on this page.

Step 7: Words frequency vector

['barclays', 'british', 'universal', 'bank', 'purpose', 'create', 'opportunity', 'rise', 'support', 'sustainable', 'inclusive', 'growth', 'connect', 'idea', 'innovation', 'aspiration', 'customer', 'client', 'capital', 'bring', 'life', 'year', 'fund', 'progress', 'today', 'remain', 'commit', 'help', 'make', 'world', 'sustainable', 'inclusive', 'connect', 'value', 'underpin', 'everything', 'respect', 'integrity', 'service', 'excellence', 'stewardship', 'customer', 'client', 'help', 'realise', 'financial', 'aspiration', 'line', 'value', 'society', 'ur', 'success', 'long', 'term', 'tie', 'inextricably', 'progress', 'community', 'preservation', 'environment', 'colleague', 'empower', 'motivate', 'engage', 'best', 'work', 'develop', 'build', 'career', 'support', 'health', 'wellbeing', 'investor', 'build', 'strong', 'diversified', 'sustainable', 'business', 'deliver', 'consistent', 'return', 'way', 'society', 'expect']

[1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 2, 1, 2, 1, 1, 2, 2, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

The last step consists in the creation of the bag-of-ESG-words. In the page, we identified 8 words as being ESG. For those, the feature of the ESG word frequency vector will be given by the bag-of-words, while the 474 others will have a value of 0.

Step 8: Bag-of-ESG-words

['barclays', 'british', 'universal', 'bank', 'purpose', 'create', 'opportunity', 'rise', 'support', 'sustainable', 'inclusive', 'growth', 'connect', 'idea', 'innovation', 'aspiration', 'customer', 'client', 'capital', 'bring', 'life', 'year', 'fund', 'progress', 'today', 'remain', 'commit', 'help', 'make', 'world', 'sustainable', 'inclusive', 'connect', 'value', 'underpin', 'everything', 'respect', 'integrity', 'service', 'excellence', 'stewardship', 'customer', 'client', 'help', 'realise', 'financial', 'aspiration', 'line', 'value', 'society', 'ur', 'success', 'long', 'term', 'tie', 'inextricably', 'progress', 'community', 'preservation', 'environment', 'colleague', 'empower', 'motivate', 'engage', 'best', 'work', 'develop', 'build', 'career', 'support', 'health', 'wellbeing', 'investor', 'build', 'strong', 'diversified', 'sustainable', 'business', 'deliver', 'consistent', 'return', 'way', 'society', 'expect']

+

[1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 2, 1, 2, 1, 1, 2, 2, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

XII. GINI Impurity

(Dougherty, 2013, p. 31) defines the GINI index, also called Gini Impurity, as the "expected error rate if the class label is selected randomly from the class distribution present". In Decision Trees algorithm, this is mathematically translated by the following equation:

$$GINI(i) = 1 - \sum_{j} P(\omega_j|i)^2$$

Where $P(\omega_j|i)$ is the probability of the class j at given node i. When the CART algorithm has to define a new rule to further split one node into two subsets, it will use this GINI measure to assess the homogeneity resulting from the split:

$$Global\ GINI = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Where k is the number of nodes created by the split, n is the number of items at the node before the split, and n_i the number of individuals at the node i created by the split. Therefore, to assess homogeneity, the CART algorithm looks at the remaining impurity at each node created by a split and choose the one that minimizes it.

XIII. SVM Model – Full development

Concretely, SVM can help us discriminate among different categories of ESG ratings by building the hyperplane that best separate two classes. To find this optimal "separator", the algorithm relies on the notion of highest margin. In our case, the margin is equal to twice the Euclidean distance between the hyperplane and the nearest company from rating grade g (${y_i}^{g,r,y} = +1$) and rating grade not g (${y_i}^{g,r,y} = -1$), or the support vectors.

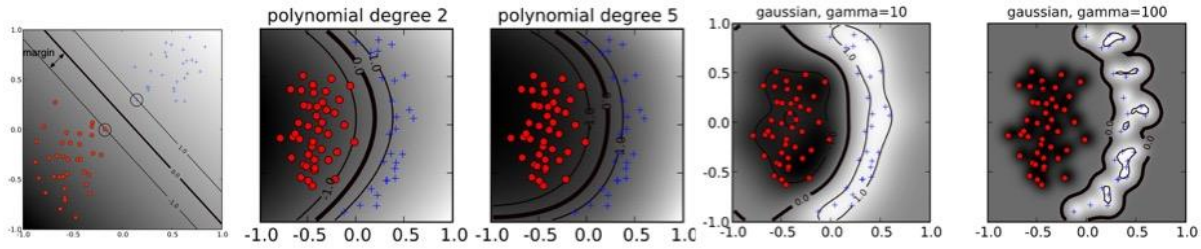


Figure 22: Illustration of the concepts of hyperplane, margin, and support vectors (the circled data points) for Linear (Ben-Hur & Weston, 2010 p. 5), Polynomial (p. 9) and Gaussian Kernels (p. 10)

For linear data, we can find the parameters of the hyperplane by solving the following quadratic program:

$$\min_{\boldsymbol{\beta}} \frac{||\boldsymbol{\beta}||^2}{2}$$

Subject to ${y_i}^{g,r,y}\left(\boldsymbol{\beta}^T \boldsymbol{w_i^{r,y}} + \beta_0\right) \geq 1, \forall i$

To put it simply, we try to find the parameters $\boldsymbol{\beta}$ that will maximize the margin between two classes under the constraint that no observation can be found in the area situated between the hyperplane + the margin and the hyperplane – the margin (see **Error! Reference source not found.**). However, when dealing with non-linear data, this optimization problem does not always lead to the optimal classification. Indeed, if data is not linearly separable, then accepting that some points are misclassified can still increase the margin. Therefore, the initial optimization program needs to be modified by relaxing the constraint: ${y_i}^{g,r,y}\left(\boldsymbol{\beta}^T w_{ij}^{r,y} + \beta_0\right) \geq 1 - \xi_i, \forall i$. ξ_i is a slack variable whose value will determine whether a point can be misclassified or not. If $\xi_i \geq 1$ then we accept that the company i is misclassified. Knowing this, we can define $\sum_{i=1}^{n^{r,y}} \xi_i$ as the bound on the number of accepted misclassifications and reformulate the optimization as the *soft margin SVM* problem (Ben-Hur & Weston, 2010):

$$\min_{\boldsymbol{\beta}} \frac{||\boldsymbol{\beta}||^2}{2} + C\sum_{i=1}^{n^{r,y}} \xi_i$$

Subject to ${y_i}^{g,r,y}\left(\boldsymbol{\beta}^T \boldsymbol{w_i^{r,y}} + \beta_0\right) \geq 1 - \xi_i, \forall i;\ \xi_i \geq 0$

Using Lagrange Multiplier this can also be given by:

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n^{r,y}} \alpha_i - \frac{1}{2}\sum_{i=1}^{n^{r,y}}\sum_{k=1}^{n^{r,y}} {y_i}^{g,r,y} {y_k}^{g,r,y}\, \alpha_i\, \alpha_k \left(\boldsymbol{w_i^{r,y}}\right)^{\boldsymbol{T}} \boldsymbol{w_k^{r,y}}$$

Subject to $\sum_{i=1}^{n^{r,y}} {y_i}^{g,r,y}\alpha_i = 0, 0 \leq \alpha_i \leq C$

To solve this SVM problem, one essentially need to compute the value of the dot product between $\left(\boldsymbol{w_i^{r,y}}\right)^{\boldsymbol{T}}$ and $\boldsymbol{w_k^{r,y}}$. To do so, a powerful method is to rely on the Kernel trick, an alternative way of computing dot product by using kernel function (polynomial, gaussian, sigmoid, etc. in function of the type of data) (Ben-Hur & Weston, 2010).

Polynomial Kernel of degree d: $\left(\boldsymbol{w_i^{r,y}}\right)^{\boldsymbol{T}} \boldsymbol{w_k^{r,y}} = K\left(\boldsymbol{w_i^{r,y}}; \boldsymbol{w_k^{r,y}}\right) = \left(\boldsymbol{\gamma}\left(\boldsymbol{w_i^{r,y}}\right)^{\boldsymbol{T}} \boldsymbol{w_k^{r,y}} + \boldsymbol{cst}\right)^{\boldsymbol{d}}$

Gaussian Kernel of parameter γ: $\left(\boldsymbol{w_i^{r,y}}\right)^{\boldsymbol{T}} \boldsymbol{w_k^{r,y}} = K\left(\boldsymbol{w_i^{r,y}}; \boldsymbol{w_k^{r,y}}\right) = \boldsymbol{e}^{-\boldsymbol{\gamma}\left(\left|\left|w_i^{r,y} - w_k^{r,y}\right|\right|\right)^2}$

As displayed by **Figure 3**, the parameters γ and the degree d of the polynomial kernel are influencing the flexibility of the SVM model to improve its prediction capabilities. To find the appropriate Kernel and parameters, the most common procedure is to apply a grid search on several parameters values (Ben-Hur & Weston, 2010). Therefore, if this method can increase the accuracy of a classifier, it is also very important to note that it is computationally intensive to fit (many parameters, kernels, etc. to test and compare)

XIV. Optimal parameters – Cross-validation

Optimal Number of Features						
	CSR		**Annual**		**A+C**	
	Category	*Grade*	*Category*	*Grade*	*Category*	*Grade*
MLR	15	19	124	107	4	33
RF Reg	19	312	61	65	203	187
NB	1	260	1	19	1	223
SVM poly	4	42	17	4	3	6
SVM Gau	2	2	1	2	1	2
LDA	6	5	6	2	2	281
RF Class	295	176	56	83	2	268

Table 15: Optimal Number of features (based on 10-fold cross-validation) for each model, rating format, and disclosure type

Optimal Parameters						
	CSR		**Annual**		**A+C**	
	Category	*Grade*	*Category*	*Grade*	*Category*	*Grade*
Distri	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian
Degree	5	5	5	5	5	5
Gamma	0,1	10	0,1	10	0,1	10

Table 16: Optimal value of model specific parameters (based on 10-fold cross-validation) for each rating format, and disclosure type

XV. List of ESG words – Features of the different models

List of ESG words					
approval	approve	approves	assess	assessment	audit
auditor	control	coso	detect	detection	evaluate
evaluates	evaluation	examination	examine	examines	irs
oversee	oversees	oversight	review	rotation	test
treadway	independence	leadership	nomination	nominee	refreshment
skill	succession	tenure	vacancy	appreciation	award
bonus	cd	compensate	compensates	compensation	eip
iso	isos	payout	payouts	pension	prsu
prsus	recoupment	remuneration	reward	rsu	rsus
salary	severance	vest	vested	ballot	cast
consent	elect	election	nominate	plurality	proponent
proposal	quorum	vote	voting	brother	conflict
family	grandparent	inform	insider	inspector	posting
sister	son	spousal	spouse	transparency	transparent
visit	webpage	website	attract	attracts	incentive
interview	motivate	motivates	motivation	recruit	recruiting
recruitment	retain	retainer	retention	talent	talented
compliance	conduct	conformity	governance	misconduct	parachute
plane	poison	retirement	align	alignment	aligns
bylaw	charter	culture	death	duly	independent
cobc	ethic	ethical	ethically	honesty	bribery
corrupt	corruption	embezzlement	grassroots	influence	lobby
lobbying	lobbyist	whistleblower	announce	announcement	announces
communicate	communicates	erm	fairly	integrity	liaison
presentation	sustainable	asc	disclose	discloses	disclosure
fasb	gaap	objectivity	press	sarbanes	engagement
feedback	hotline	investor	invite	mail	mailing
notice	stakeholder	compact	ungc	biofuels	biofuel
green	renewable	solar	stewardship	wind	emission
ghg	ghgs	greenhouse	atmosphere	emit	biodiversity
wilderness	wildlife	freshwater	groundwater	water	air
carbon	nitrogen	pollution	superfund	biphenyls	hazardous
householding	printing	recycling	toxic	waste	weee
recycle	climate	agriculture	deforestation	pesticide	cleaner
cleanup	coal	contamination	fossil	resource	clean
environmental	epa	sustainability	childbirth	drug	medicaid
medicare	medicine	hiv	alcohol	drinking	conformance
fda	inspection	standardization	warranty	epidemic	health
healthy	ill	illness	pandemic	community	expression
marriage	privacy	peace	dignity	discriminate	discrimination
equality	freedom	humanity	nondiscrimination	sexual	bisexual
diversity	ethnic	ethnically	ethnicity	female	gay
gender	homosexual	immigration	lesbian	lgbt	minority
ms	race	racial	religion	religious	sex
transgender	woman	occupational	safe	safely	safety
ilo	labour	eicc	bargaining	eeo	fairness
fla	harassment	injury	labor	overtime	ruggie
sick	wage	workplace	charitable	charity	donate
donates	donation	foundation	gift	nonprofit	poverty
educate	educates	education	educational	learning	teach
teacher	teaching	training	employ	employment	headcount
hire	unemployment	endowment	people	philanthropic	philanthropy
socially	societal	society	welfare	citizen	csr
disability	disabled	human	social	un	veteran

Table 17: List of ESG words found in the different Annual and CSR reports of our sample (based on the ESG dictionary of Baier et al. (2020))

XVI. Average Frequency of ESG words in Annual and CSR reports

	2017	2018	2019	2020
Average Rep	9,78	10,73	11,51	13,57
Min	3316	1550	2063	1778
Max	0	0	0	0

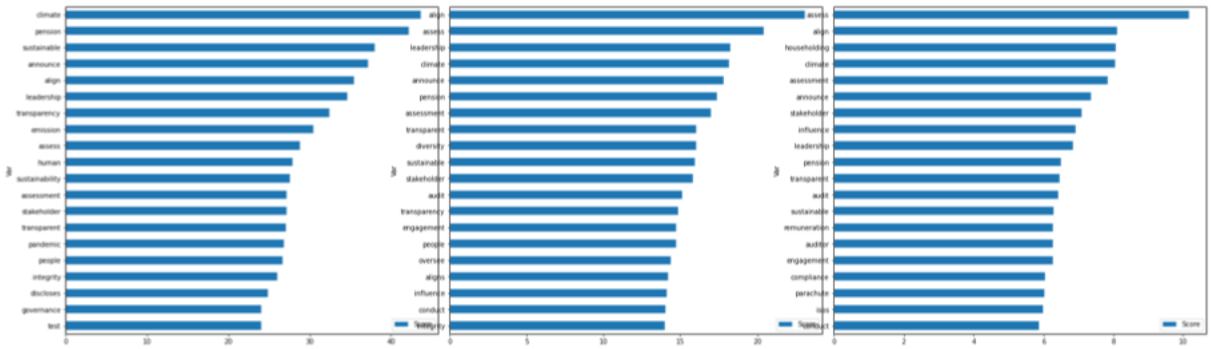
Table 18: Average, Minimum, and Maximum frequency of an ESG word in the Annual and CSR reports of companies (2017-2020)

XVII. List of most Relevant Features

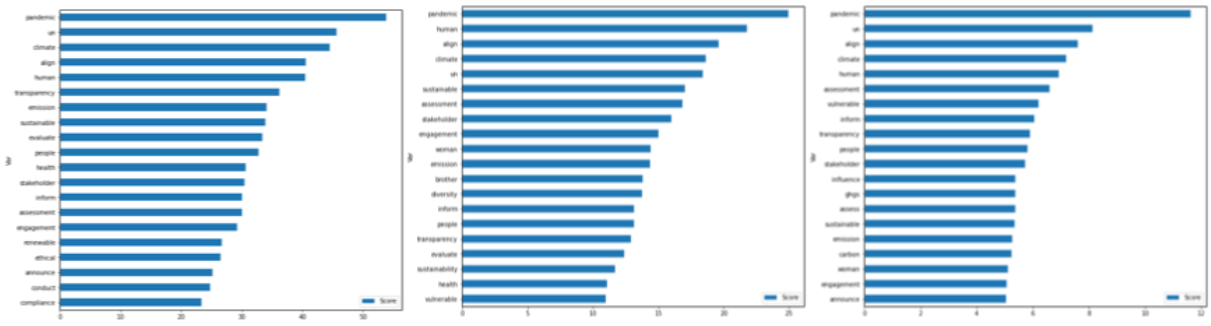
For each of the reports type and year we can define a set of features that have the highest explanatory power by relying on the statistic of a Correlation Test between an explanatory variable and the continuous ratings $Y^{c,r,y}$, or by relying on the statistic of an ANOVA test between the same variable and both categorical rating $Y^{cat,r,y}$ and $Y^{g,r,y}$. In this Appendix, we display graphically the 10 most relevant features according to (from left to right) Correlation test on $Y^{c,r,y}$, ANOVA test on $Y^{cat,r,y}$ and ANOVA test on $Y^{g,r,y}$, for each year y and report r.

2020

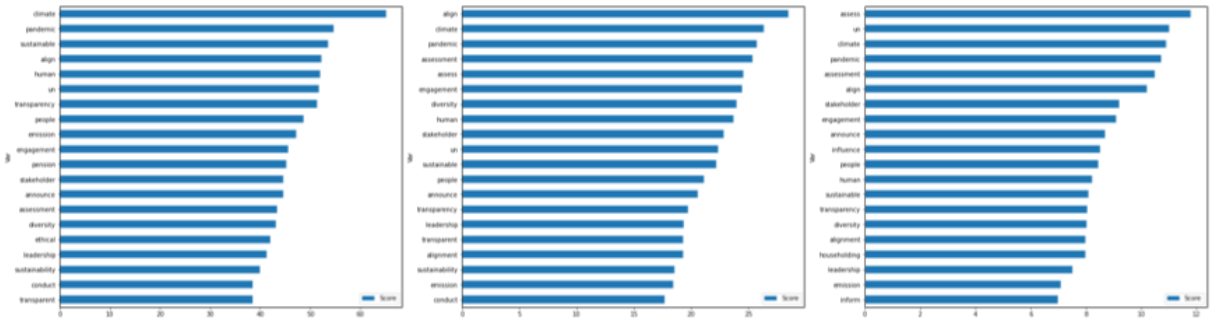
Annual Reports



CSR Reports

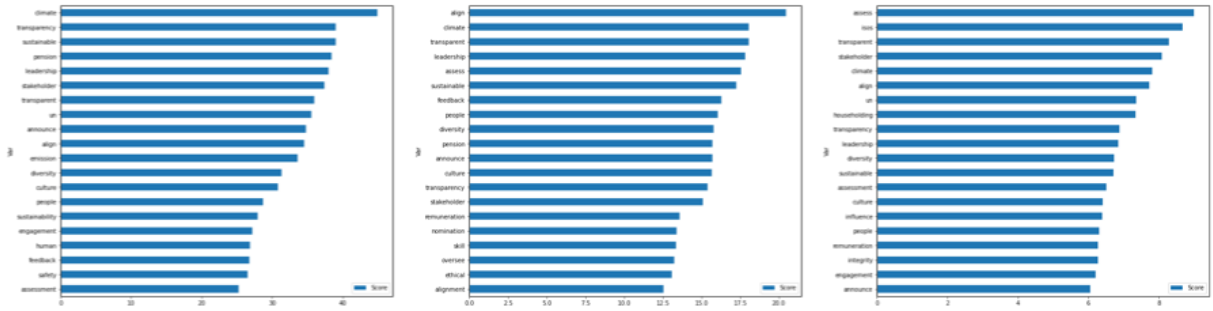


Annual + CSR Reports

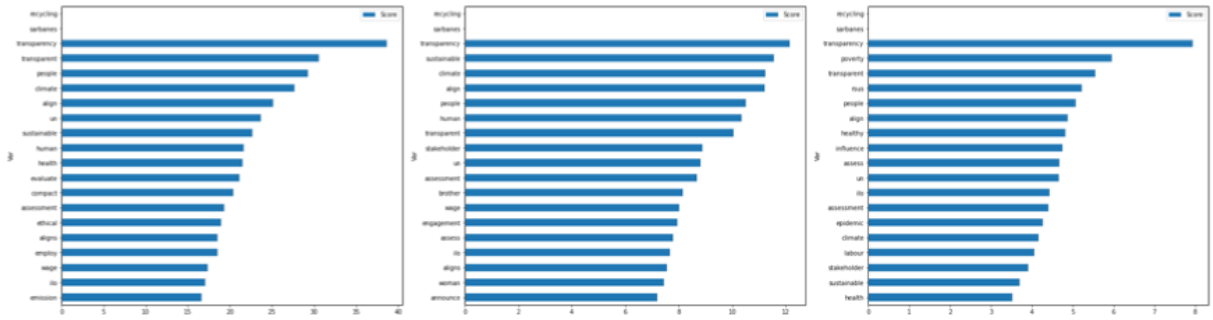


2019

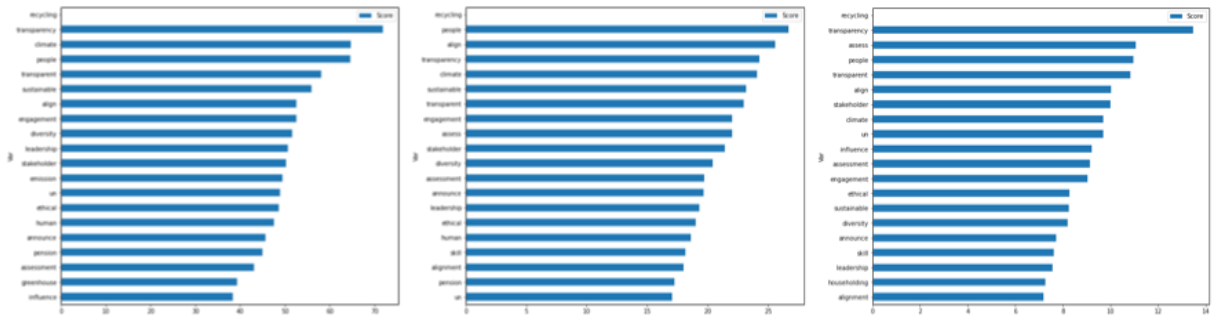
Annual Reports



CSR Reports

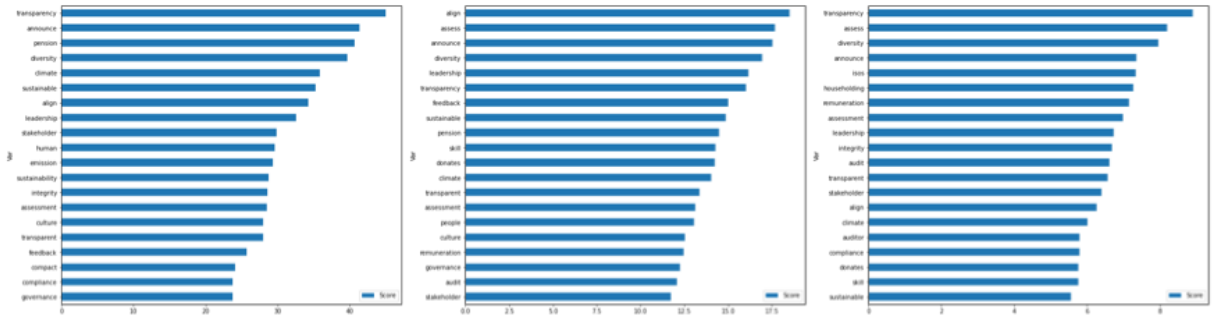


Annual + CSR Reports

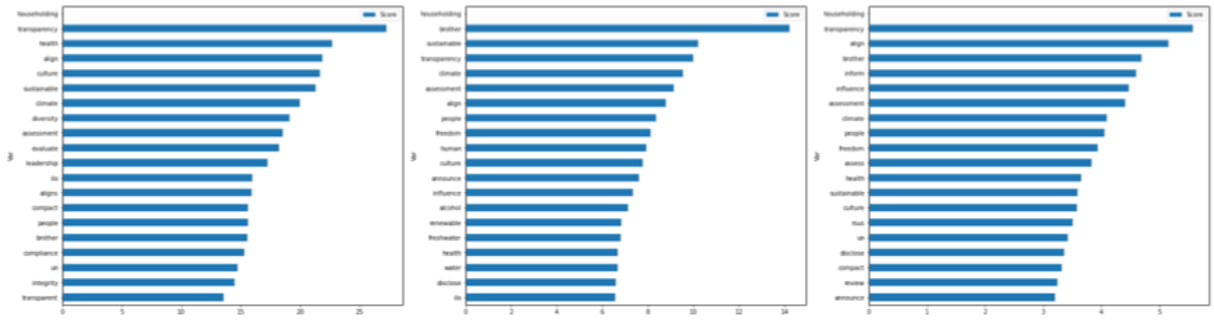


2018

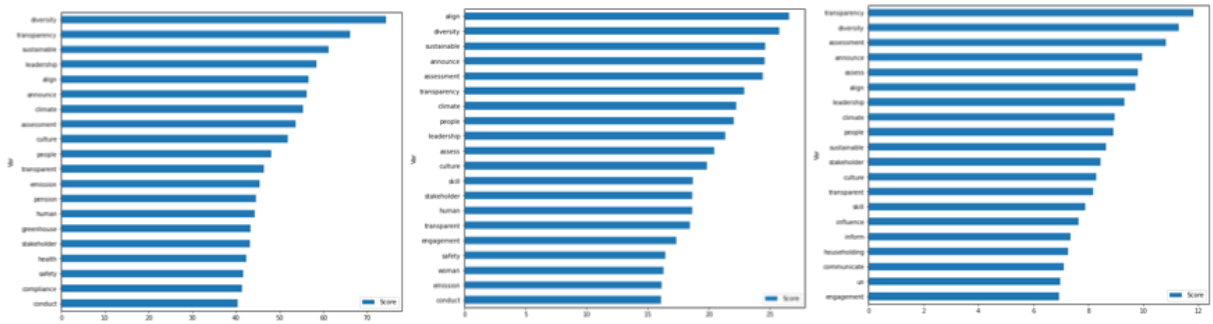
Annual Reports



CSR Reports

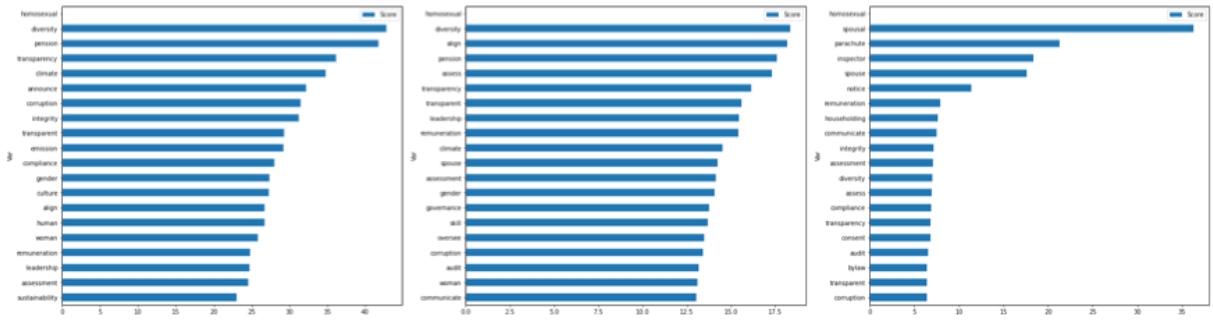


Annual + CSR Reports

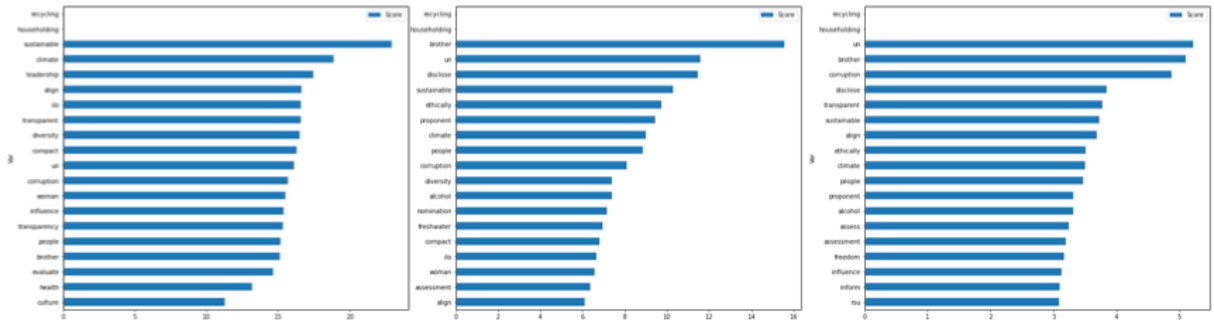


2017

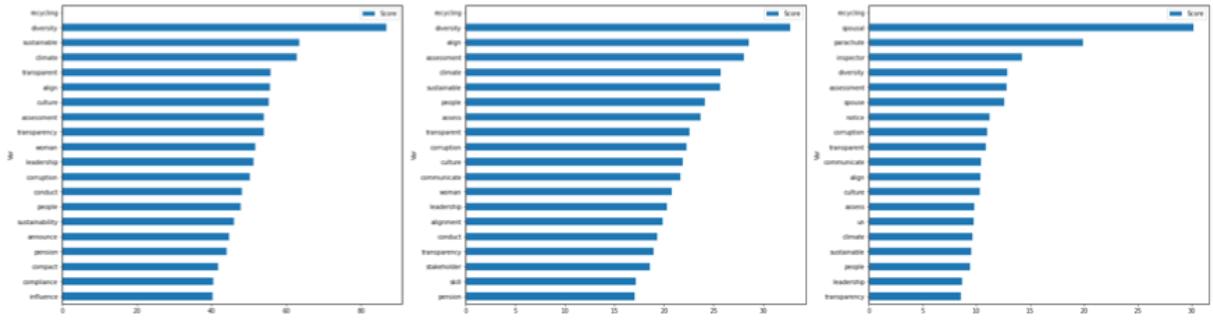
Annual Reports



CSR Reports



Annual + CSR Reports



XVIII. Results

Part 1: Category Ratings

1. Average and Variance of Metrics (2017-2020)

1.1. General Performance of Text Mining Models (ESG Words Frequency excluded)

Table 19: Average Performance of all Text Mining Based ratings

General Performance of Text Mining Based Ratings								
		MAE	COR	ACC	F1	BIAS	PRECISION	RECALL
Category	Average	**0,6132**	**0,3980**	**55,43%**	**0,3414**	**-0,3103**	**0,4463**	**0,3766**
	Variance	*0,0065*	*0,0161*	*0,71%*	*0,0106*	*0,0054*	*0,0096*	*0,0072*
	Max	1,1400	1,0000	1,0000	1,0000	0,1750	1,0000	1,0000
	Min	0,0000	0,0000	0,4098	0,1453	-1,1439	0,1024	0,2500

1.2. Accuracy and Macro F1-score by Rating Model

Table 20: Average Accuracy and Macro F1-score of Every Text Mining Models (regardless of Disclosure Type; for Annual Reports; for CSR reports; for both)

		Category (regardless of disclosure)								
Category		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
ACCURACY	**Average**	0,50	0,57	0,48	0,50	0,56	0,52	0,75	0,55	0,30
	Variance	*0,00323*	*0,02803*	*0,00158*	*0,00331*	*0,00298*	*0,00234*	*0,02831*	*0,00997*	*0,06179*
F1	**Average**	0,28	0,40	0,23	0,26	0,31	0,28	0,64	0,34	0,25
	Variance	*0,00422*	*0,04126*	*0,00155*	*0,00544*	*0,00300*	*0,00183*	*0,06120*	*0,01693*	*0,01472*

		Annual Category								
Category		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
ACCURACY	**Average**	0,56	0,56	0,50	0,53	0,52	0,55	0,76	0,57	0,18
	Variance	*0,00469*	*0,02951*	*0,00029*	*0,00230*	*0,00039*	*0,00067*	*0,02658*	*0,00920*	*0,00022*
F1	**Average**	0,44	0,40	0,25	0,34	0,28	0,31	0,66	0,38	0,18
	Variance	*0,00700*	*0,04308*	*0,00026*	*0,00968*	*0,00020*	*0,00104*	*0,05756*	*0,01697*	*0,00004*

		CSR Category								
Category		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
ACCURACY	**Average**	0,47	0,54	0,46	0,50	0,63	0,52	0,73	0,55	0,14
	Variance	*0,00330*	*0,03512*	*0,00503*	*0,00491*	*0,01483*	*0,00551*	*0,03276*	*0,01450*	*0,00016*
F1	**Average**	0,21	0,34	0,19	0,23	0,37	0,25	0,57	0,31	0,13
	Variance	*0,00333*	*0,05037*	*0,00560*	*0,00385*	*0,01744*	*0,00360*	*0,08498*	*0,02417*	*0,00334*

		Annual & CSR Category								
Category		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
ACCURACY	**Average**	0,45	0,60	0,49	0,46	0,53	0,51	0,77	0,54	0,20
	Variance	*0,00212*	*0,02055*	*0,00112*	*0,00307*	*0,00091*	*0,00218*	*0,02614*	*0,00801*	*0,00052*
F1	**Average**	0,19	0,47	0,24	0,21	0,28	0,26	0,69	0,33	0,19
	Variance	*0,00298*	*0,03155*	*0,00097*	*0,00397*	*0,00054*	*0,00167*	*0,04523*	*0,01242*	*0,00024*

1.3. Other Metrics by Rating Model: MAE, Bias, Precision, Recall

		Category (regardless of disclosure)								
Category		**MLR**	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
MAE	**Average**	0,93	1,00	0,55	0,53	0,49	0,52	0,28	0,61	1,63
	Variance	*0,00417*	*0,01482*	*0,00126*	*0,00376*	*0,00261*	*0,00200*	*0,03580*	*0,00920*	*0,68722*
COR	**Average**	0,44	0,67	0,20	0,25	0,31	0,27	0,65	0,40	0,34
	Variance	*0,00557*	*0,04411*	*0,00471*	*0,01794*	*0,00580*	*0,00831*	*0,05780*	*0,02061*	*0,01472*
BIAS	**Average**	-0,93	-1,00	-0,11	-0,13	0,07	-0,05	-0,03	-0,31	-1,53
	Variance	*0,00417*	*0,01504*	*0,00726*	*0,00687*	*0,00193*	*0,00836*	*0,00167*	*0,00647*	*1,04409*
PRECISION	**Average**	0,41	0,57	0,28	0,43	0,34	0,32	0,77	0,45	0,35
	Variance	*0,00473*	*0,04161*	*0,00203*	*0,01389*	*0,01017*	*0,00050*	*0,03253*	*0,01507*	*0,01848*
RECALL	**Average**	0,30	0,43	0,29	0,31	0,34	0,32	0,62	0,37	0,28
	Variance	*0,00007*	*0,02412*	*0,00051*	*0,00225*	*0,00211*	*0,00101*	*0,06405*	*0,01345*	*0,02051*

		Annual Category								
Category		**MLR**	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
MAE	**Average**	0,91	1,02	0,54	0,51	0,52	0,51	0,27	0,61	2,10
	Variance	*0,00273*	*0,01489*	*0,00017*	*0,00409*	*0,00033*	*0,00089*	*0,03427*	*0,00820*	*0,00189*
COR	**Average**	0,54	0,69	0,25	0,31	0,26	0,28	0,67	0,43	0,25
	Variance	*0,01516*	*0,03972*	*0,00025*	*0,02412*	*0,00070*	*0,00445*	*0,05109*	*0,01936*	*0,00004*
BIAS	**Average**	-0,91	-1,02	-0,03	-0,06	0,06	0,02	0,02	-0,27	-2,10
	Variance	*0,00279*	*0,01512*	*0,00114*	*0,00068*	*0,00087*	*0,00129*	*0,00037*	*0,00318*	*0,00171*
PRECISION	**Average**	0,60	0,58	0,31	0,53	0,29	0,35	0,78	0,49	0,29
	Variance	*0,02084*	*0,04109*	*0,00003*	*0,05851*	*0,00003*	*0,00187*	*0,03887*	*0,02303*	*0,00042*
RECALL	**Average**	0,42	0,42	0,30	0,36	0,32	0,35	0,63	0,40	0,20
	Variance	*0,00374*	*0,02566*	*0,00009*	*0,00462*	*0,00012*	*0,00056*	*0,06226*	*0,01386*	*0,00001*

		CSR Category								
Category		**MLR**	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
MAE	**Average**	0,93	0,96	0,57	0,52	0,43	0,52	0,29	0,60	2,06
	Variance	*0,00583*	*0,01257*	*0,00429*	*0,00480*	*0,01456*	*0,00443*	*0,03929*	*0,01225*	*0,00069*
COR	**Average**	0,35	0,55	0,10	0,25	0,36	0,22	0,57	0,34	0,27
	Variance	*0,00992*	*0,08094*	*0,02540*	*0,01595*	*0,03522*	*0,02075*	*0,08686*	*0,03929*	*0,00334*
BIAS	**Average**	-0,93	-0,95	-0,25	-0,21	0,12	-0,15	-0,05	-0,35	-2,06
	Variance	*0,00594*	*0,01255*	*0,03285*	*0,02206*	*0,00265*	*0,02897*	*0,00208*	*0,01530*	*0,00073*
PRECISION	**Average**	0,33	0,56	0,22	0,34	0,44	0,31	0,71	0,41	0,26
	Variance	*0,00037*	*0,04792*	*0,01812*	*0,00031*	*0,08862*	*0,00042*	*0,06537*	*0,03159*	*0,00043*
RECALL	**Average**	0,28	0,38	0,27	0,30	0,37	0,30	0,56	0,35	0,20
	Variance	*0,00107*	*0,02834*	*0,00166*	*0,00168*	*0,01252*	*0,00212*	*0,08646*	*0,01912*	*0,00079*

		Annual & CSR Category								
Category		**MLR**	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
MAE	**Average**	0,96	1,02	0,54	0,57	0,51	0,52	0,27	0,63	1,99
	Variance	*0,00443*	*0,01807*	*0,00096*	*0,00277*	*0,00056*	*0,00169*	*0,03429*	*0,00897*	*0,00170*
COR	**Average**	0,41	0,76	0,26	0,20	0,31	0,31	0,70	0,42	0,36
	Variance	*0,00024*	*0,02204*	*0,00222*	*0,01631*	*0,00130*	*0,00487*	*0,04062*	*0,01252*	*0,00024*
BIAS	**Average**	-0,96	-1,02	-0,06	-0,11	0,03	-0,01	-0,06	-0,31	-1,99
	Variance	*0,00431*	*0,01849*	*0,00184*	*0,00560*	*0,00321*	*0,00520*	*0,00600*	*0,00638*	*0,00178*
PRECISION	**Average**	0,28	0,59	0,31	0,41	0,30	0,31	0,83	0,43	0,30
	Variance	*0,01483*	*0,03731*	*0,00002*	*0,01332*	*0,00005*	*0,00007*	*0,01437*	*0,01142*	*0,00027*
RECALL	**Average**	0,27	0,48	0,29	0,28	0,32	0,31	0,68	0,38	0,23
	Variance	*0,00078*	*0,01903*	*0,00039*	*0,00124*	*0,00033*	*0,00082*	*0,04688*	*0,00992*	*0,00049*

2. Detailed Performance metrics

2.1. MAE

Categories									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,98	1,09	0,58	0,58	0,51	0,56	0,41	0,671	2,00
2018	0,98	1,05	0,56	0,56	0,51	0,54	0,39	0,66	2,04
2019	0,94	1,03	0,56	0,54	0,51	0,52	0,31	0,63	2,07
2020	**0,84**	**0,82**	**0,50**	**0,44**	**0,41**	**0,45**	**0**	**0,49**	**0,38**
Average	**0,93**	**1,00**	**0,55**	**0,53**	**0,49**	**0,52**	**0,28**	**0,61**	**1,63**
Variance	*0,00417*	*0,01482*	*0,00126*	*0,00376*	*0,00261*	*0,00200*	*0,03580*	*0,00652*	*0,68722*

Annual Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,95	1,09	0,55	0,55	0,54	0,54	0,4	0,66	2,04
2018	0,95	1,09	0,54	0,56	0,53	0,52	0,38	0,65	2,09
2019	0,9	1,07	0,53	0,52	0,51	0,5	0,3	0,62	2,12
2020	**0,84**	**0,84**	**0,52**	**0,42**	**0,5**	**0,47**	**0**	**0,51**	**2,14**
Average	**0,91**	**1,02**	**0,54**	**0,51**	**0,52**	**0,51**	**0,27**	**0,61**	**2,10**
Variance	*0,00273*	*0,01489*	*0,00017*	*0,00409*	*0,00033*	*0,00089*	*0,03427*	*0,00461*	*0,00189*

CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,98	1,04	0,61	0,58	0,47	0,57	0,41	0,66571429	2,03
2018	0,98	0,99	0,59	0,54	0,47	0,54	0,42	0,65	2,08
2019	0,95	1	0,6	0,54	0,52	0,54	0,34	0,64	2,08
2020	**0,82**	**0,79**	**0,47**	**0,42**	**0,25**	**0,42**	**0**	**0,45**	**2,04**
Average	**0,93**	**0,96**	**0,57**	**0,52**	**0,43**	**0,52**	**0,29**	**0,60**	**2,06**
Variance	*0,00583*	*0,01257*	*0,00429*	*0,00480*	*0,01456*	*0,00443*	*0,03929*	*0,00997*	*0,00069*

Annual & CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	1	1,14	0,57	0,61	0,53	0,56	0,41	0,68857143	1,94
2018	1	1,08	0,56	0,59	0,53	0,55	0,37	0,67	1,96
2019	0,97	1,03	0,54	0,57	0,51	0,51	0,29	0,63	2,02
2020	**0,86**	**0,83**	**0,5**	**0,49**	**0,48**	**0,47**	**0**	**0,52**	**2,02**
Average	**0,96**	**1,02**	**0,54**	**0,57**	**0,51**	**0,52**	**0,27**	**0,63**	**1,99**
Variance	*0,00443*	*0,01807*	*0,00096*	*0,00277*	*0,00056*	*0,00169*	*0,03429*	*0,00577*	*0,00170*

2.2. *Correlation*

Categories									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,39	0,52	0,15	0,15	0,26	0,18	0,48	0,304	0,28
2018	0,38	0,53	0,17	0,19	0,27	0,24	0,49	0,33	0,29
2019	0,43	0,64	0,18	0,23	0,28	0,27	0,62	0,38	0,28
2020	**0,54**	**0,97**	**0,30**	**0,45**	**0,42**	**0,40**	**1,00**	**0,58**	**0,52**
Average	**0,44**	**0,67**	**0,20**	**0,25**	**0,31**	**0,27**	**0,65**	**0,40**	**0,34**
Variance	*0,00557*	*0,04411*	*0,00471*	*0,01794*	*0,00580*	*0,00831*	*0,05780*	*0,01614*	*0,01472*

Annual Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,484	0,533	0,230	0,211	0,223	0,210	0,497	0,341	0,2524
2018	0,434	0,573	0,243	0,197	0,255	0,251	0,533	0,36	0,2447
2019	0,544	0,669	0,257	0,291	0,273	0,285	0,647	0,42	0,2390
2020	**0,717**	**0,973**	**0,266**	**0,532**	**0,284**	**0,367**	**0,992**	**0,59**	**0,2508**
Average	**0,54**	**0,69**	**0,25**	**0,31**	**0,26**	**0,28**	**0,67**	**0,43**	**0,25**
Variance	*0,01516*	*0,03972*	*0,00025*	*0,02412*	*0,00070*	*0,00445*	*0,05109*	*0,01306*	*0,00004*

CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,27	0,39	0,00	0,15	0,24	0,08	0,38	0,22	0,20
2018	0,29	0,34	0,05	0,22	0,30	0,21	0,37	0,26	0,26
2019	0,34	0,51	0,00	0,20	0,26	0,19	0,55	0,29	0,26
2020	**0,49**	**0,97**	**0,33**	**0,44**	**0,64**	**0,42**	**1,00**	**0,61**	**0,34**
Average	**0,35**	**0,55**	**0,10**	**0,25**	**0,36**	**0,22**	**0,57**	**0,34**	**0,27**
Variance	*0,00992*	*0,08094*	*0,02540*	*0,01595*	*0,03522*	*0,02075*	*0,08686*	*0,03300*	*0,00334*

Annual & CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,43	0,64	0,21	0,07	0,31	0,25	0,57	0,36	0,37
2018	0,40	0,68	0,22	0,15	0,26	0,27	0,58	0,37	0,36
2019	0,40	0,76	0,29	0,21	0,32	0,34	0,66	0,43	0,34
2020	**0,42**	**0,97**	**0,31**	**0,37**	**0,34**	**0,40**	**1,00**	**0,55**	**0,34**
Average	**0,41**	**0,76**	**0,26**	**0,20**	**0,31**	**0,31**	**0,70**	**0,42**	**0,36**
Variance	*0,00024*	*0,02204*	*0,00222*	*0,01631*	*0,00130*	*0,00487*	*0,04062*	*0,00758*	*0,00024*

2.3. Accuracy

Table 21: Accuracy per year, on average, and variance for every model regardless of disclosure type (2017-2020)

Categories									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	45,43%	47,16%	44,89%	45,26%	52,97%	48,28%	63,48%	49,64%	17,81%
2018	45,80%	47,25%	46,36%	47,21%	53,16%	50,07%	65,24%	50,73%	17,89%
2019	49,33%	50,54%	47,53%	49,16%	53,30%	52,16%	72,71%	53,53%	16,77%
2020	**57,66%**	**81,65%**	**53,90%**	**58,27%**	**64,05%**	**59,32%**	**99,83%**	**67,81%**	**67,20%**
Average	**49,55%**	**56,65%**	**48,17%**	**49,98%**	**55,87%**	**52,46%**	**75,31%**	**55,43%**	**29,92%**
Variance	*0,323%*	*2,803%*	*0,158%*	*0,331%*	*0,298%*	*0,234%*	*2,831%*	*0,709%*	*6,179%*

Table 22: Accuracy per year, on average, and variance for every model when fitted on Annual Reports (2017-2020)

Annual Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	52,02%	45,96%	48,23%	49,49%	50,51%	52,27%	64,90%	51,91%	18,43%
2018	50,49%	45,63%	49,27%	50,00%	51,21%	53,40%	65,78%	52,25%	18,69%
2019	56,80%	50,36%	50,60%	52,74%	53,22%	55,13%	73,75%	56,09%	17,42%
2020	**65,70%**	**81,40%**	**52,17%**	**59,90%**	**54,83%**	**58,21%**	**99,76%**	**67,43%**	**15,46%**
Average	**56,25%**	**55,84%**	**50,07%**	**53,04%**	**52,44%**	**54,75%**	**76,05%**	**56,92%**	**17,50%**
Variance	*0,469%*	*2,951%*	*0,029%*	*0,230%*	*0,039%*	*0,067%*	*2,658%*	*0,526%*	*0,022%*

Table 23: Accuracy per year, on average, and variance for every model when fitted on CSR Reports (2017-2020)

CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	42,86%	42,48%	40,98%	44,36%	57,89%	45,86%	61,65%	48,01%	12,78%
2018	44,92%	43,93%	42,95%	48,20%	58,03%	48,52%	61,97%	49,79%	14,10%
2019	46,31%	46,61%	43,07%	48,67%	53,69%	49,56%	68,73%	50,95%	14,75%
2020	**55,83%**	**81,67%**	**56,39%**	**60,56%**	**80,56%**	**62,50%**	**99,72%**	**71,03%**	**15,83%**
Average	**47,48%**	**53,67%**	**45,85%**	**50,45%**	**62,54%**	**51,61%**	**73,02%**	**54,95%**	**14,37%**
Variance	*0,330%*	*3,512%*	*0,503%*	*0,491%*	*1,483%*	*0,551%*	*3,276%*	*1,165%*	*0,016%*

Table 24: Accuracy per year, on average, and variance for every model when fitted on Annual & CSR Reports (2017-2020)

Annual & CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	41,41%	53,03%	45,45%	41,92%	50,51%	46,72%	63,89%	48,99%	22,22%
2018	41,99%	52,18%	46,84%	43,45%	50,24%	48,30%	67,96%	50,14%	20,87%
2019	44,87%	54,65%	48,93%	46,06%	52,98%	51,79%	75,66%	53,56%	18,14%
2020	**51,45%**	**81,88%**	**53,14%**	**54,35%**	**56,76%**	**57,25%**	**100,00%**	**64,98%**	**17,39%**
Average	**44,93%**	**60,44%**	**48,59%**	**46,44%**	**52,62%**	**51,01%**	**76,88%**	**54,42%**	**19,66%**
Variance	*0,212%*	*2,055%*	*0,112%*	*0,307%*	*0,091%*	*0,218%*	*2,614%*	*0,533%*	*0,052%*

2.4. Macro F1-score

Table 25: F1-score per year, on average, and variance for every model regardless of disclosure type (2017-2020)

Categories									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,24	0,30	0,20	0,21	0,28	0,24	0,46	0,273	0,17
2018	0,23	0,29	0,21	0,21	0,28	0,26	0,49	0,28	0,17
2019	0,27	0,32	0,22	0,25	0,29	0,27	0,60	0,32	0,16
2020	**0,37**	**0,708**	**0,28**	**0,37**	**0,39**	**0,34**	**1,00**	**0,49**	**0,48**
Average	**0,2777**	**0,4038**	**0,2266**	**0,2594**	**0,3088**	**0,2754**	**0,6382**	**0,3414**	**0,2460**
Variance	*0,0042*	*0,0413*	*0,0015*	*0,0054*	*0,0030*	*0,0018*	*0,0612*	*0,0106*	*0,0239*

Table 26: F1-score per year, on average, and variance for every model when fitted on Annual Reports (2017-2020)

Annual Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,3989	0,2799	0,2368	0,2800	0,2667	0,2795	0,4642	0,3151	0,1815
2018	0,3748	0,2782	0,2475	0,2592	0,2732	0,3172	0,5173	0,3239	0,1870
2019	0,4161	0,3216	0,2592	0,3376	0,2864	0,2976	0,6453	0,3663	0,1757
2020	**0,5604**	**0,7064**	**0,2744**	**0,4775**	**0,2982**	**0,3547**	**0,9974**	**0,5241**	**0,1605**
Average	**0,4375**	**0,3965**	**0,2545**	**0,3385**	**0,2811**	**0,3122**	**0,6560**	**0,3824**	**0,1762**
Variance	*0,0070*	*0,0431*	*0,0003*	*0,0097*	*0,0002*	*0,0010*	*0,0576*	*0,0094*	*0,0001*

Table 27: F1-score per year, on average, and variance for every model when fitted on CSR Reports (2017-2020)

CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1686	0,2263	0,1453	0,1831	0,3033	0,2074	0,3862	0,2314	0,1147
2018	0,1754	0,2166	0,1533	0,2090	0,3070	0,2260	0,3813	0,2384	0,1299
2019	0,1948	0,2520	0,1505	0,2179	0,2876	0,2336	0,5109	0,2639	0,1399
2020	**0,2928**	**0,6795**	**0,2992**	**0,3238**	**0,5629**	**0,3403**	**0,9967**	**0,4993**	**0,1550**
Average	**0,2079**	**0,3436**	**0,1871**	**0,2334**	**0,3652**	**0,2518**	**0,5688**	**0,3082**	**0,1349**
Variance	*0,0033*	*0,0504*	*0,0056*	*0,0038*	*0,0174*	*0,0036*	*0,0850*	*0,0164*	*0,0003*

Table 28: F1-score per year, on average, and variance for every model when fitted on Annual & CSR Reports (2017-2020)

Annual & CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1464	0,3835	0,2072	0,1534	0,2630	0,2206	0,5258	0,2714	0,2137
2018	0,1533	0,3739	0,2220	0,1724	0,2616	0,2401	0,5860	0,2870	0,1999
2019	0,1852	0,3906	0,2451	0,2032	0,2847	0,2752	0,6472	0,3187	0,1759
2020	**0,2655**	**0,7376**	**0,2787**	**0,2955**	**0,3112**	**0,3132**	**1,0000**	**0,4574**	**0,1691**
Average	**0,1876**	**0,4714**	**0,2383**	**0,2061**	**0,2801**	**0,2623**	**0,6898**	**0,3336**	**0,1897**
Variance	*0,0030*	*0,0315*	*0,0010*	*0,0040*	*0,0005*	*0,0017*	*0,0452*	*0,0072*	*0,0004*

2.5. *Bias*

Categories									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	-0,98	-1,09	-0,18	-0,19	0,02	-0,13	-0,08	-0,376	-2,01
2018	-0,98	-1,05	-0,15	-0,17	0,06	-0,09	-0,04	-0,35	-2,04
2019	-0,94	-1,03	-0,12	-0,13	0,08	-0,04	0,00	-0,31	-2,08
2020	**-0,84**	**-0,82**	**0,01**	**-0,01**	**0,13**	**0,08**	**0,00**	**-0,21**	**0,00**
Average	**-0,9330**	**-0,9989**	**-0,1122**	**-0,1262**	**0,0720**	**-0,0454**	**-0,0288**	**-0,3103**	**-1,5313**
Variance	*0,0042*	*0,0150*	*0,0073*	*0,0069*	*0,0019*	*0,0084*	*0,0017*	*0,0054*	*1,0441*

Annual Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	-0,9495	-1,0909	-0,0606	-0,0783	0,0328	-0,0126	0,0101	-0,3070	-2,0429
2018	-0,9539	-1,0898	-0,0461	-0,0752	0,0413	0,0049	0,0291	-0,2985	-2,0922
2019	-0,9021	-1,0692	-0,0239	-0,0644	0,0597	0,0310	0,0382	-0,2758	-2,1241
2020	**-0,8406**	**-0,8382**	**0,0169**	**-0,0217**	**0,0990**	**0,0700**	**-0,0048**	**-0,2170**	**-2,1353**
Average	**-0,9115**	**-1,0220**	**-0,0284**	**-0,0599**	**0,0582**	**0,0233**	**0,0181**	**-0,2746**	**-2,0986**
Variance	*0,0028*	*0,0151*	*0,0011*	*0,0007*	*0,0009*	*0,0013*	*0,0004*	*0,0016*	*0,0017*

CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	-0,9774	-1,0376	-0,3872	-0,3459	0,0526	-0,2857	-0,0977	-0,4398	-2,0301
2018	-0,9770	-0,9934	-0,3311	-0,2787	0,1311	-0,2230	-0,0721	-0,3920	-2,0787
2019	-0,9528	-0,9971	-0,2950	-0,2183	0,1386	-0,1770	-0,0236	-0,3607	-2,0826
2020	**-0,8167**	**-0,7889**	**0,0167**	**-0,0028**	**0,1750**	**0,1000**	**0,0028**	**-0,1877**	**-2,0389**
Average	**-0,9310**	**-0,9542**	**-0,2492**	**-0,2114**	**0,1244**	**-0,1464**	**-0,0477**	**-0,3451**	**-2,0576**
Variance	*0,0059*	*0,0126*	*0,0328*	*0,0221*	*0,0027*	*0,0290*	*0,0021*	*0,0121*	*0,0007*

Annual & CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	-1,0000	-1,1439	-0,0985	-0,1591	-0,0152	-0,0859	-0,1641	-0,3810	-1,9444
2018	-0,9951	-1,0777	-0,0850	-0,1505	-0,0049	-0,0558	-0,0631	-0,3474	-1,9587
2019	-0,9714	-1,0310	-0,0501	-0,1217	0,0453	0,0143	0,0000	-0,3021	-2,0239
2020	**-0,8599**	**-0,8285**	**-0,0024**	**0,0024**	**0,1087**	**0,0749**	**0,0000**	**-0,2150**	**-2,0242**
Average	**-0,9566**	**-1,0203**	**-0,0590**	**-0,1072**	**0,0335**	**-0,0131**	**-0,0568**	**-0,3114**	**-1,9878**
Variance	*0,0043*	*0,0185*	*0,0018*	*0,0056*	*0,0032*	*0,0052*	*0,0060*	*0,0052*	*0,0018*

2.6. *Macro Precision*

Categories									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequer**
2017	0,32	0,46	0,24	0,36	0,30	0,31	0,59	0,37	0,28
2018	0,39	0,45	0,33	0,35	0,29	0,32	0,67	0,40	0,29
2019	0,43	0,51	0,24	0,40	0,29	0,31	0,83	0,43	0,29
2020	**0,48**	**0,88**	**0,31**	**0,60**	**0,49**	**0,35**	**1,00**	**0,59**	**0,56**
Average	**0,4058**	**0,5734**	**0,2796**	**0,4263**	**0,3437**	**0,3224**	**0,7730**	**0,4463**	**0,3541**
Variance	*0,0047*	*0,0416*	*0,0020*	*0,0139*	*0,0102*	*0,0005*	*0,0325*	*0,0096*	*0,0185*

Annual Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,5378	0,4652	0,3125	0,3860	0,2854	0,3148	0,5218	0,4034	0,2923
2018	0,4617	0,4554	0,3090	0,3233	0,2886	0,3525	0,7833	0,4248	0,3150
2019	0,6042	0,5140	0,3051	0,5395	0,2956	0,3173	0,8364	0,4875	0,2933
2020	**0,7988**	**0,8804**	**0,3005**	**0,8647**	**0,2958**	**0,4077**	**0,9964**	**0,6492**	**0,2649**
Average	**0,6007**	**0,5787**	**0,3068**	**0,5284**	**0,2913**	**0,3481**	**0,7845**	**0,4912**	**0,2914**
Variance	*0,0208*	*0,0411*	*0,0000*	*0,0585*	*0,0000*	*0,0019*	*0,0389*	*0,0124*	*0,0004*

CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,3189	0,4180	0,1024	0,3332	0,2942	0,2837	0,5224	0,3247	0,2277
2018	0,3591	0,4089	0,3569	0,3628	0,2921	0,3037	0,4631	0,3638	0,2591
2019	0,3173	0,5215	0,1077	0,3258	0,2730	0,3029	0,8398	0,3840	0,2702
2020	**0,3325**	**0,8752**	**0,3158**	**0,3558**	**0,8815**	**0,3333**	**0,9984**	**0,5846**	**0,2726**
Average	**0,3319**	**0,5559**	**0,2207**	**0,3444**	**0,4352**	**0,3059**	**0,7059**	**0,4143**	**0,2574**
Variance	*0,0004*	*0,0479*	*0,0181*	*0,0003*	*0,0886*	*0,0004*	*0,0654*	*0,0135*	*0,0004*

Annual & CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1035	0,4955	0,3094	0,3541	0,3127	0,3186	0,7356	0,3756	0,3190
2018	0,3543	0,4825	0,3094	0,3558	0,2979	0,3078	0,7576	0,4093	0,3033
2019	0,3581	0,4889	0,3082	0,3355	0,3002	0,3049	0,8209	0,4167	0,2959
2020	**0,3228**	**0,8751**	**0,3188**	**0,5786**	**0,3075**	**0,3217**	**1,0000**	**0,5321**	**0,2797**
Average	**0,2847**	**0,5855**	**0,3114**	**0,4060**	**0,3046**	**0,3133**	**0,8286**	**0,4334**	**0,2995**
Variance	*0,0148*	*0,0373*	*0,0000*	*0,0133*	*0,0000*	*0,0001*	*0,0144*	*0,0046*	*0,0003*

2.7. *Macro Recall*

Categories									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,30	0,35	0,27	0,28	0,31	0,29	0,46	0,325	0,21
2018	0,30	0,34	0,28	0,28	0,32	0,31	0,48	0,33	0,22
2019	0,31	0,36	0,28	0,31	0,32	0,31	0,56	0,35	0,21
2020	**0,38**	**0,66**	**0,32**	**0,38**	**0,41**	**0,37**	**1,00**	**0,50**	**0,50**
Average	**0,30**	**0,4267**	**0,2900**	**0,3133**	**0,3386**	**0,3195**	**0,6241**	**0,3766**	**0,2817**
Variance	*0,00007*	*0,0241*	*0,0005*	*0,0022*	*0,0021*	*0,0010*	*0,0641*	*0,0072*	*0,0205*

Annual Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,3920	0,3312	0,2944	0,3235	0,3101	0,3210	0,4577	0,3471	0,1952
2018	0,3773	0,3257	0,3004	0,3053	0,3139	0,3487	0,4885	0,3514	0,2015
2019	0,4037	0,3611	0,3065	0,3644	0,3245	0,3359	0,5926	0,3841	0,2012
2020	**0,5115**	**0,6582**	**0,3161**	**0,4577**	**0,3345**	**0,3766**	**0,9984**	**0,5218**	**0,2003**
Average	**0,4211**	**0,4190**	**0,3043**	**0,3627**	**0,3207**	**0,3456**	**0,6343**	**0,4011**	**0,1995**
Variance	*0,0037*	*0,0257*	*0,0001*	*0,0046*	*0,0001*	*0,0006*	*0,0623*	*0,0068*	*0,0000*

CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,2590	0,3031	0,2500	0,2665	0,3209	0,2720	0,3836	0,2936	0,1837
2018	0,2623	0,2796	0,2518	0,2799	0,3256	0,2804	0,3827	0,2946	0,1845
2019	0,2680	0,3053	0,2500	0,2813	0,3066	0,2860	0,4727	0,3100	0,2034
2020	**0,3280**	**0,6319**	**0,3322**	**0,3568**	**0,5409**	**0,3708**	**0,9950**	**0,5079**	**0,2437**
Average	**0,2793**	**0,3800**	**0,2710**	**0,2961**	**0,3735**	**0,3023**	**0,5585**	**0,3515**	**0,2038**
Variance	*0,0011*	*0,0283*	*0,0017*	*0,0017*	*0,0125*	*0,0021*	*0,0865*	*0,0109*	*0,0008*

Annual & CSR Category									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,2500	0,4268	0,2765	0,2532	0,3095	0,2846	0,5343	0,3336	0,2462
2018	0,2531	0,4015	0,2847	0,2624	0,3072	0,2943	0,5649	0,3383	0,2601
2019	0,2684	0,4089	0,2957	0,2763	0,3228	0,3149	0,6187	0,3580	0,2156
2020	**0,3107**	**0,6875**	**0,3217**	**0,3317**	**0,3468**	**0,3492**	**1,0000**	**0,4782**	**0,2165**
Average	**0,2706**	**0,4812**	**0,2946**	**0,2809**	**0,3216**	**0,3107**	**0,6795**	**0,3770**	**0,2346**
Variance	*0,0008*	*0,0190*	*0,0004*	*0,0012*	*0,0003*	*0,0008*	*0,0469*	*0,0047*	*0,0005*

3. Hypothesis Tests

3.1. Kruskal-Wallis and Tukey's Tests

3.1.1. Average Accuracy

Table 29: Results from the Kurskal Wallis Test on Mean Accuracy of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	df	p-value
Mean Accuracy: Words Frequency vs. MLR vs. RF Reg vs. … vs. RF Class	56.584	7	7.232e-10

Pair Comparison using Tukey's Test

	MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class
MLR	-	-	-	-	-	-	-
RF Reg	0,5631	-	-	-	-	-	-
NB	0,9000	0,3398	-	-	-	-	-
SVM Poly	0,9000	0,6280	0,9000	-	-	-	-
SVM Gau	0,6833	0,9000	0,4681	0,7483	-	-	-
LDA	0,9000	0,9000	0,9000	0,9000	0,9000	-	-
RF Class	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	-
Words Frequency	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***

3.1.2. Average F1-score

Table 30: Results from the Kurskal Wallis Test on Mean F1-score of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	df	p-value
Mean F1: Words Frequency vs. MLR vs. RF Reg vs. … vs. RF Class	51.776	7	6.463e-09

Pair Comparison using Tukey's Test

	MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class
MLR	-	-	-	-	-	-	-
RF Reg	0,2421	-	-	-	-	-	-
NB	0,9000	***0,0213***	-	-	-	-	-
SVM Poly	0,9000	0,1143	0,9000	-	-	-	-
SVM Gau	0,9000	0,5898	0,7340	0,9000	-	-	-
LDA	0,9000	0,2228	0,9000	0,9000	0,9000	-	-
RF Class	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	-
Words Frequ	0,4068	***0,0010***	0,9000	0,6187	0,1275	0,4341	***0,0010***

3.2. Friedmann and Nemenyi's Tests

Table 31: Results from the Friedmann Test comparing the Mean F1-score, Accuracy, MAE, Bias, Correlation, Precision, and Recall of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Nemenyi's test (below).

Friedmann rank sum Test

	Kruskall-Wallis Chi Squared	df	p-value
Words Frequency vs. MLR vs. RF Reg vs. … vs. RF Class	31.952	6	4.145e-05

Pair Comparison using Nemenyi's Test

	MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class
MLR	-	-	-	-	-	-	-
RF Reg	0,9000	-	-	-	-	-	-
NB	0,8314	0,1910	-	-	-	-	-
SVM Poly	0,9000	0,8314	0,9000	-	-	-	-
SVM Gau	0,8967	0,9000	0,1145	0,7010	-	-	-
LDA	0,9000	0,9000	0,7662	0,9000	0,9000	-	-
RF Class	0,1490	0,7662	***0,0014***	0,0640	0,8967	0,1910	-
Words Frequency	0,4352	***0,0335***	0,9000	0,6357	***0,0165***	0,3636	***0,0010***

Part 2: Grade Ratings

1. Average and Variance of Metrics (2017-2020)

1.1. General Performance of Text Mining Models (ESG Words Frequency excluded)

Table 32: Average Performance of all Text Mining Based Grade ratings

General Performance of Text Mining Based Ratings								
		MAE	COR	ACC	F1	BIAS	PRECISION	RECALL
Grade	Average	**2,0885**	**0,4229**	**33,48%**	**0,2327**	**-0,9674**	**0,2957**	**0,2583**
	Variance	*0,0804*	*0,0269*	*1,61%*	*0,0168*	*0,0191*	*0,0172*	*0,0154*
	Max	4,1300	1,0000	1,0000	1,0000	1,0652	1,0000	1,0000
	Min	0,0000	-0,0555	0,1165	0,0370	-4,3056	0,0461	0,0812

1.2. Accuracy and Macro F1-score by Rating Model

Table 33: Average Accuracy and Macro F1-score of Every Text Mining Models (regardless of Disclosure Type; for Annual Reports; for CSR reports; for both)

	Grade (regardless of disclosure)									
Grade		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
ACCURACY	**Average**	0,19	0,28	0,32	0,26	0,36	0,27	0,65	0,33	0,05
	Variance	*0,00184*	*0,03020*	*0,00752*	*0,00376*	*0,07154*	*0,00040*	*0,06006*	*0,02504*	*0,00007*
F1	**Average**	0,11	0,16	0,30	0,10	0,25	0,10	0,61	0,23	0,05
	Variance	*0,00081*	*0,01217*	*0,01481*	*0,00470*	*0,07667*	*0,00092*	*0,08074*	*0,02726*	*0,00006*

	Annual Grade									
Grade		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
ACCURACY	**Average**	0,23	0,29	0,24	0,24	0,35	0,27	0,67	0,33	0,06
	Variance	*0,00176*	*0,02621*	*0,00064*	*0,00082*	*0,06789*	*0,00014*	*0,05678*	*0,02203*	*0,00015*
F1	**Average**	0,15	0,17	0,17	0,07	0,25	0,11	0,64	0,22	0,06
	Variance	*0,00139*	*0,01204*	*0,00113*	*0,00043*	*0,09231*	*0,00081*	*0,08024*	*0,02691*	*0,00012*

	CSR Grade									
Grade		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
ACCURACY	**Average**	0,18	0,26	0,35	0,31	0,39	0,27	0,63	0,34	0,04
	Variance	*0,00341*	*0,04376*	*0,02326*	*0,00749*	*0,07370*	*0,00061*	*0,06756*	*0,03140*	*0,00004*
F1	**Average**	0,08	0,13	0,35	0,14	0,24	0,09	0,55	0,23	0,04
	Variance	*0,00100*	*0,01477*	*0,03819*	*0,00866*	*0,06939*	*0,00174*	*0,09320*	*0,03242*	*0,00003*

	Annual & CSR Grade									
Grade		MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class	Average	Words Frequency
ACCURACY	**Average**	0,18	0,30	0,37	0,24	0,36	0,26	0,66	0,34	0,06
	Variance	*0,00094*	*0,02326*	*0,00723*	*0,00490*	*0,07380*	*0,00059*	*0,05640*	*0,02387*	*0,00018*
F1	**Average**	0,09	0,18	0,37	0,10	0,26	0,11	0,63	0,25	0,06
	Variance	*0,00036*	*0,01027*	*0,02205*	*0,00867*	*0,06999*	*0,00045*	*0,07414*	*0,02656*	*0,00017*

1.3. Other Metrics by Rating Model: MAE, Bias, Precision, Recall

		Grade (regardless of disclosure)								
Grade		**MLR**	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
MAE	**Average**	3,77	3,79	1,89	1,51	1,37	1,50	0,78	2,09	7,16
	Variance	*0,05132*	*0,06742*	*0,10670*	*0,01347*	*0,28228*	*0,00031*	*0,30872*	*0,11860*	*0,01179*
COR	**Average**	0,45	0,69	0,39	0,15	0,33	0,26	0,69	0,42	0,29
	Variance	*0,00771*	*0,03839*	*0,01208*	*0,03353*	*0,09250*	*0,00312*	*0,05581*	*0,03474*	*0,00086*
BIAS	**Average**	-3,77	-3,86	-0,88	0,47	0,45	0,76	0,05	-0,97	-7,08
	Variance	*0,05097*	*0,07673*	*0,04212*	*0,02596*	*0,00946*	*0,02470*	*0,00350*	*0,03335*	*0,01011*
PRECISION	**Average**	0,15	0,18	0,33	0,24	0,30	0,11	0,75	0,30	0,10
	Variance	*0,00042*	*0,01040*	*0,01348*	*0,03116*	*0,09187*	*0,00061*	*0,04317*	*0,02730*	*0,00024*
RECALL	**Average**	0,14	0,19	0,37	0,14	0,26	0,14	0,58	0,26	0,08
	Variance	*0,00046*	*0,00984*	*0,01871*	*0,00232*	*0,06172*	*0,00079*	*0,08755*	*0,02591*	*0,00005*

		Annual Grade								
Grade		**MLR**	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
MAE	**Average**	3,63	3,79	2,26	1,65	1,59	1,65	0,74	2,19	7,21
	Variance	*0,01287*	*0,04453*	*0,03749*	*0,00330*	*0,34800*	*0,00049*	*0,29207*	*0,10554*	*0,01676*
COR	**Average**	0,52	0,69	0,13	0,20	0,28	0,21	0,70	0,39	0,26
	Variance	*0,01289*	*0,03776*	*0,00553*	*0,00682*	*0,09438*	*0,00175*	*0,05641*	*0,03079*	*0,00095*
BIAS	**Average**	-3,63	-3,79	-0,59	0,62	0,61	0,91	0,20	-0,81	-7,21
	Variance	*0,01255*	*0,04409*	*0,05598*	*0,00273*	*0,01361*	*0,00077*	*0,01940*	*0,02130*	*0,01618*
PRECISION	**Average**	0,19	0,23	0,26	0,16	0,33	0,10	0,77	0,29	0,11
	Variance	*0,00216*	*0,00747*	*0,00429*	*0,00299*	*0,13489*	*0,00069*	*0,04086*	*0,02762*	*0,00095*
RECALL	**Average**	0,16	0,19	0,26	0,11	0,25	0,14	0,61	0,25	0,08
	Variance	*0,00086*	*0,01040*	*0,00216*	*0,00023*	*0,06965*	*0,00080*	*0,08453*	*0,02409*	*0,00044*

		CSR Grade								
Grade		**MLR**	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
MAE	**Average**	3,77	3,79	1,71	1,45	1,26	1,42	0,81	2,03	7,14
	Variance	*0,08909*	*0,08396*	*0,16909*	*0,02163*	*0,25207*	*0,00049*	*0,31777*	*0,13344*	*0,01127*
COR	**Average**	0,35	0,62	0,47	0,14	0,32	0,25	0,61	0,39	0,26
	Variance	*0,01102*	*0,06120*	*0,02750*	*0,06047*	*0,11052*	*0,01828*	*0,08891*	*0,05399*	*0,00474*
BIAS	**Average**	-3,77	-3,79	-1,05	0,21	0,37	0,41	-0,04	-1,09	-7,14
	Variance	*0,08854*	*0,08312*	*0,04715*	*0,10978*	*0,00649*	*0,14660*	*0,01816*	*0,07141*	*0,01121*
PRECISION	**Average**	0,13	0,14	0,37	0,34	0,28	0,12	0,71	0,30	0,09
	Variance	*0,00059*	*0,01403*	*0,03182*	*0,04845*	*0,07576*	*0,00222*	*0,05455*	*0,03249*	*0,00018*
RECALL	**Average**	0,13	0,17	0,42	0,16	0,26	0,13	0,53	0,26	0,05
	Variance	*0,00060*	*0,01160*	*0,04718*	*0,00442*	*0,05510*	*0,00117*	*0,10131*	*0,03163*	*0,00012*

		Annual & CSR Grade								
Grade		**MLR**	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
MAE	**Average**	3,91	3,79	1,71	1,45	1,26	1,42	0,81	2,05	7,14
	Variance	*0,07300*	*0,08396*	*0,16909*	*0,02163*	*0,25207*	*0,00049*	*0,31777*	*0,13114*	*0,01127*
COR	**Average**	0,46	0,76	0,58	0,11	0,41	0,32	0,77	0,49	0,36
	Variance	*0,00225*	*0,02148*	*0,01358*	*0,05404*	*0,07625*	*0,00043*	*0,03058*	*0,02838*	*0,00050*
BIAS	**Average**	-3,91	-4,00	-1,01	0,58	0,37	0,97	0,00	-1,00	-6,89
	Variance	*0,07293*	*0,11776*	*0,03221*	*0,01064*	*0,01847*	*0,00448*	*0,00397*	*0,03721*	*0,00692*
PRECISION	**Average**	0,13	0,18	0,37	0,22	0,29	0,11	0,77	0,30	0,11
	Variance	*0,00006*	*0,01167*	*0,01871*	*0,07121*	*0,07537*	*0,00023*	*0,03631*	*0,03051*	*0,00064*
RECALL	**Average**	0,12	0,21	0,44	0,13	0,26	0,14	0,60	0,27	0,09
	Variance	*0,00020*	*0,00814*	*0,02545*	*0,00401*	*0,06139*	*0,00052*	*0,08184*	*0,02594*	*0,00050*

2. Detailed Performance metrics

2.1. MAE

Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	3,94	3,98	2,19	1,60	1,58	1,49	1,28	2,30	7,01
2018	3,88	3,92	2,04	1,54	1,70	1,49	1,05	2,23	7,17
2019	3,82	3,86	1,90	1,56	1,62	1,52	0,79	2,15	7,25
2020	**3,44**	**3,41**	**1,44**	**1,34**	**0,58**	**1,49**	**0,01**	**1,67**	**7,22**
Average	**3,77**	**3,79**	**1,89**	**1,51**	**1,37**	**1,50**	**0,78**	**2,09**	**7,16**
Variance	*0,05132*	*0,06742*	*0,10670*	*0,01347*	*0,28228*	*0,00031*	*0,30872*	*0,08042*	*0,01179*

Annual Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	3,72	4,01	2,52	1,71	1,83	1,64	1,25	2,38285714	7,04
2018	3,7	3,87	2,3	1,65	1,97	1,66	1	2,31	7,18
2019	3,63	3,77	2,14	1,65	1,85	1,67	0,71	2,20	7,27
2020	**3,47**	**3,51**	**2,09**	**1,57**	**0,71**	**1,62**	**0**	**1,85**	**7,34**
Average	**3,63**	**3,79**	**2,26**	**1,65**	**1,59**	**1,65**	**0,74**	**2,19**	**7,21**
Variance	*0,01287*	*0,04453*	*0,03749*	*0,00330*	*0,34800*	*0,00049*	*0,29207*	*0,05490*	*0,01676*

CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	3,98	3,97	2,03	1,55	1,46	1,41	1,3	2,24285714	6,99
2018	3,91	3,94	1,91	1,49	1,57	1,4	1,08	2,19	7,17
2019	3,87	3,9	1,78	1,52	1,5	1,45	0,83	2,12	7,24
2020	**3,33**	**3,36**	**1,11**	**1,23**	**0,51**	**1,43**	**0,01**	**1,57**	**7,16**
Average	**3,77**	**3,79**	**1,71**	**1,45**	**1,26**	**1,42**	**0,81**	**2,03**	**7,14**
Variance	*0,08909*	*0,08396*	*0,16909*	*0,02163*	*0,25207*	*0,00049*	*0,31777*	*0,09694*	*0,01127*

Annual & CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	4,13	3,97	2,03	1,55	1,46	1,41	1,3	2,26428571	6,99
2018	4,04	3,94	1,91	1,49	1,57	1,4	1,08	2,20	7,17
2019	3,95	3,9	1,78	1,52	1,5	1,45	0,83	2,13	7,24
2020	**3,52**	**3,36**	**1,11**	**1,23**	**0,51**	**1,43**	**0,01**	**1,60**	**7,16**
Average	**3,91**	**3,79**	**1,71**	**1,45**	**1,26**	**1,42**	**0,81**	**2,05**	**7,14**
Variance	*0,07300*	*0,08396*	*0,16909*	*0,02163*	*0,25207*	*0,00049*	*0,31777*	*0,09432*	*0,01127*

2.2. *Correlation*

Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,38	0,55	0,33	0,00	0,20	0,22	0,44	0,30	0,27
2018	0,39	0,57	0,29	0,09	0,11	0,23	0,60	0,32	0,27
2019	0,44	0,68	0,42	0,09	0,24	0,24	0,73	0,40	0,31
2020	**0,57**	**0,97**	**0,54**	**0,42**	**0,78**	**0,34**	**1,00**	**0,66**	**0,32**
Average	**0,45**	**0,69**	**0,39**	**0,15**	**0,33**	**0,26**	**0,69**	**0,42**	**0,29**
Variance	*0,00771*	*0,03839*	*0,01208*	*0,03353*	*0,09250*	*0,00312*	*0,05581*	*0,02689*	*0,00086*

Annual Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,4571	0,5409	0,0329	0,0900	0,1826	0,2271	0,4352	0,28	0,2244
2018	0,4365	0,584	0,107	0,190	0,051	0,206	0,619	0,31	0,2521
2019	0,5189	0,675	0,186	0,209	0,142	0,148	0,751	0,38	0,2986
2020	**0,6868**	**0,972**	**0,189**	**0,291**	**0,730**	**0,244**	**1,000**	**0,59**	**0,2666**
Average	**0,52**	**0,69**	**0,13**	**0,20**	**0,28**	**0,21**	**0,70**	**0,39**	**0,26**
Variance	*0,01289*	*0,03776*	*0,00553*	*0,00682*	*0,09438*	*0,00175*	*0,05641*	*0,01899*	*0,00095*

CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,2704	0,4478	0,4287	-0,0555	0,1419	0,1240	0,3076	0,2378	0,1848
2018	0,2910	0,4520	0,3067	0,0590	0,0662	0,1825	0,4601	0,2596	0,2173
2019	0,3411	0,6028	0,4396	0,0507	0,2530	0,2485	0,6631	0,3713	0,2784
2020	**0,5022**	**0,9741**	**0,7007**	**0,4987**	**0,8006**	**0,4355**	**0,9974**	**0,7013**	**0,3406**
Average	**0,35**	**0,62**	**0,47**	**0,14**	**0,32**	**0,25**	**0,61**	**0,39**	**0,26**
Variance	*0,01102*	*0,06120*	*0,02750*	*0,06047*	*0,11052*	*0,01828*	*0,08891*	*0,04580*	*0,00474*

Annual & CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,4115	0,6502	0,5138	-0,0273	0,2708	0,3154	0,5824	0,3881	0,3862
2018	0,4426	0,6755	0,4538	0,0105	0,2247	0,2952	0,7089	0,4016	0,3316
2019	0,4685	0,7574	0,6195	0,0020	0,3204	0,3290	0,7690	0,4665	0,3595
2020	**0,5236**	**0,9728**	**0,7177**	**0,4589**	**0,8187**	**0,3443**	**1,0000**	**0,6909**	**0,3558**
Average	**0,46**	**0,76**	**0,58**	**0,11**	**0,41**	**0,32**	**0,77**	**0,49**	**0,36**
Variance	*0,00225*	*0,02148*	*0,01358*	*0,05404*	*0,07625*	*0,00043*	*0,03058*	*0,01968*	*0,00050*

2.3. Accuracy

Table 34: Accuracy per year, on average, and variance for every model regardless of disclosure type (2017-2020)

Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,16	0,18	0,24	0,21	0,24	0,25	0,44	0,25	0,05
2018	0,18	0,20	0,29	0,24	0,21	0,26	0,53	0,27	0,06
2019	0,18	0,22	0,30	0,24	0,24	0,26	0,64	0,30	0,06
2020	**0,26**	**0,54**	**0,45**	**0,35**	**0,76**	**0,30**	**1,00**	**0,52**	**0,04**
Average	**19,44%**	**28,39%**	**32,06%**	**26,17%**	**36,40%**	**26,63%**	**65,26%**	**33,48%**	**5,15%**
Variance	*0,184%*	*3,020%*	*0,752%*	*0,376%*	*7,154%*	*0,040%*	*6,006%*	*1,608%*	*0,007%*

Table 35: Accuracy per year, on average, and variance for every model when fitted on Annual Reports (2017-2020)

Annual Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	19,70%	18,18%	20,96%	20,71%	24,49%	26,01%	45,96%	25,14%	4,29%
2018	20,63%	21,12%	23,54%	23,06%	19,66%	25,97%	53,88%	26,84%	6,80%
2019	22,20%	25,30%	24,58%	22,91%	21,00%	25,78%	66,35%	29,73%	6,44%
2020	**28,99%**	**53,38%**	**27,05%**	**27,54%**	**73,67%**	**28,26%**	**100,00%**	**48,41%**	**4,83%**
Average	**22,88%**	**29,49%**	**24,03%**	**23,55%**	**34,71%**	**26,50%**	**66,55%**	**32,53%**	**5,59%**
Variance	*0,176%*	*2,621%*	*0,064%*	*0,082%*	*6,789%*	*0,014%*	*5,678%*	*1,157%*	*0,015%*

Table 36: Accuracy per year, on average, and variance for every model when fitted on CSR Reports (2017-2020)

CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	13,91%	11,65%	21,43%	23,31%	25,94%	25,94%	39,85%	23,15%	3,38%
2018	15,74%	16,07%	30,49%	28,85%	23,93%	26,89%	51,48%	27,63%	4,59%
2019	14,75%	18,88%	31,56%	28,02%	25,66%	25,37%	59,88%	29,16%	4,42%
2020	**26,39%**	**56,94%**	**56,94%**	**43,33%**	**79,44%**	**30,83%**	**99,72%**	**56,23%**	**3,61%**
Average	**17,70%**	**25,89%**	**35,11%**	**30,88%**	**38,75%**	**27,26%**	**62,73%**	**34,04%**	**4,00%**
Variance	*0,341%*	*4,376%*	*2,326%*	*0,749%*	*7,370%*	*0,061%*	*6,756%*	*2,253%*	*0,004%*

Table 37: Accuracy per year, on average, and variance for every model when fitted on Annual & CSR Reports (2017-2020)

Annual & CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	14,65%	22,98%	31,06%	19,70%	20,20%	24,49%	45,96%	25,58%	6,82%
2018	16,26%	22,09%	32,28%	20,87%	20,87%	24,76%	54,37%	27,36%	6,55%
2019	18,38%	21,48%	35,32%	21,24%	25,54%	25,54%	65,63%	30,45%	6,21%
2020	**21,74%**	**52,66%**	**49,52%**	**34,54%**	**76,33%**	**29,71%**	**100,00%**	**52,07%**	**3,86%**
Average	**17,76%**	**29,80%**	**37,05%**	**24,09%**	**35,74%**	**26,12%**	**66,49%**	**33,86%**	**5,86%**
Variance	*0,094%*	*2,326%*	*0,723%*	*0,490%*	*7,380%*	*0,059%*	*5,640%*	*1,514%*	*0,018%*

2.4. Macro F1-score

Table 38: F1-score per year, on average, and variance for every model regardless of disclosure type (2017-2020)

Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,08	0,10	0,20	0,05	0,12	0,09	0,33	0,14	0,05
2018	0,09	0,11	0,24	0,08	0,09	0,09	0,49	0,17	0,06
2019	0,10	0,11	0,27	0,08	0,12	0,09	0,61	0,20	0,06
2020	**0,15**	**0,33**	**0,48**	**0,20**	**0,67**	**0,15**	**1,00**	**0,42**	**0,04**
Average	**0,1064**	**0,1611**	**0,2970**	**0,1029**	**0,2512**	**0,1028**	**0,6073**	**0,2327**	**0,0519**
Variance	*0,0008*	*0,0122*	*0,0148*	*0,0047*	*0,0767*	*0,0009*	*0,0807*	*0,0168*	*0,0001*

Table 39: F1-score per year, on average, and variance for every model when fitted on Annual Reports (2017-2020)

Annual Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1180	0,1008	0,1221	0,0510	0,1342	0,0909	0,3208	0,1340	0,0484
2018	0,1353	0,1140	0,1671	0,0661	0,0788	0,0924	0,5532	0,1724	0,0729
2019	0,1467	0,1332	0,1880	0,0679	0,0933	0,0906	0,6814	0,2002	0,0667
2020	**0,2041**	**0,3339**	**0,1977**	**0,1002**	**0,7079**	**0,1482**	**1,0000**	**0,3846**	**0,0554**
Average	**0,1510**	**0,1705**	**0,1687**	**0,0713**	**0,2536**	**0,1055**	**0,6389**	**0,2228**	**0,0608**
Variance	*0,0014*	*0,0120*	*0,0011*	*0,0004*	*0,0923*	*0,0008*	*0,0802*	*0,0124*	*0,0001*

Table 40: F1-score per year, on average, and variance for every model when fitted on CSR Reports (2017-2020)

CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,0655	0,0527	0,2124	0,0610	0,1123	0,0696	0,3143	0,1268	0,0296
2018	0,0717	0,0764	0,2902	0,1146	0,1031	0,0723	0,4111	0,1628	0,0406
2019	0,0594	0,0868	0,2683	0,1148	0,1163	0,0665	0,4859	0,1712	0,0421
2020	**0,1281**	**0,3133**	**0,6423**	**0,2759**	**0,6373**	**0,1528**	**0,9979**	**0,4497**	**0,0359**
Average	**0,0812**	**0,1323**	**0,3533**	**0,1416**	**0,2423**	**0,0903**	**0,5523**	**0,2276**	**0,0370**
Variance	*0,0010*	*0,0148*	*0,0382*	*0,0087*	*0,0694*	*0,0017*	*0,0932*	*0,0223*	*0,0000*

Table 41: F1-score per year, on average, and variance for every model when fitted on Annual & CSR Reports (2017-2020)

Annual & CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,0698	0,1413	0,2752	0,0370	0,1219	0,0991	0,3673	0,1588	0,0658
2018	0,0766	0,1301	0,2696	0,0508	0,0972	0,0976	0,5034	0,1750	0,0641
2019	0,0896	0,1187	0,3452	0,0607	0,1595	0,1097	0,6526	0,2194	0,0630
2020	**0,1126**	**0,3319**	**0,5856**	**0,2347**	**0,6528**	**0,1434**	**1,0000**	**0,4373**	**0,0385**
Average	**0,0872**	**0,1805**	**0,3689**	**0,0958**	**0,2578**	**0,1125**	**0,6308**	**0,2476**	**0,0579**
Variance	*0,0004*	*0,0103*	*0,0220*	*0,0087*	*0,0700*	*0,0005*	*0,0741*	*0,0166*	*0,0002*

2.5. Bias

Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	-3,94	-4,10	-1,17	0,32	0,40	0,62	0,01	-1,12	-6,94
2018	-3,88	-3,99	-0,88	0,40	0,48	0,70	0,07	-1,01	-7,07
2019	-3,81	-3,89	-0,79	0,46	0,58	0,76	0,13	-0,94	-7,17
2020	**-3,44**	**-3,46**	**-0,69**	**0,69**	**0,36**	**0,98**	**0,00**	**-0,79**	**-7,14**
Average	**-3,7699**	**-3,8612**	**-0,8813**	**0,4686**	**0,4543**	**0,7645**	**0,0534**	**-0,9674**	**-7,0808**
Variance	*0,0510*	*0,0767*	*0,0421*	*0,0260*	*0,0095*	*0,0247*	*0,0035*	*0,0191*	*0,0101*

Annual Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	-3,7172	-4,0101	-0,8788	0,5631	0,6591	0,8712	0,3232	-0,8842	-7,0429
2018	-3,6990	-3,8689	-0,6602	0,6044	0,6189	0,9223	0,2549	-0,8325	-7,1845
2019	-3,6325	-3,7709	-0,4726	0,6253	0,7255	0,9165	0,2076	-0,7716	-7,2745
2020	**-3,4710**	**-3,5121**	**-0,3333**	**0,6884**	**0,4517**	**0,9348**	**0,0000**	**-0,7488**	**-7,3357**
Average	**-3,6299**	**-3,7905**	**-0,5862**	**0,6203**	**0,6138**	**0,9112**	**0,1964**	**-0,8093**	**-7,2094**
Variance	*0,0125*	*0,0441*	*0,0560*	*0,0027*	*0,0136*	*0,0008*	*0,0194*	*0,0037*	*0,0162*

CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	-3,9774	-3,9699	-1,3571	-0,1090	0,2895	0,0639	-0,2068	-1,3238	-6,9925
2018	-3,9115	-3,9410	-1,0295	0,0721	0,3672	0,2426	-0,0590	-1,1799	-7,1705
2019	-3,8673	-3,9027	-0,9027	0,2065	0,4838	0,3894	0,1150	-1,0683	-7,2419
2020	**-3,3306**	**-3,3639**	**-0,8917**	**0,6667**	**0,3583**	**0,9500**	**0,0083**	**-0,8004**	**-7,1639**
Average	**-3,7717**	**-3,7944**	**-1,0452**	**0,2091**	**0,3747**	**0,4115**	**-0,0356**	**-1,0931**	**-7,1422**
Variance	*0,0885*	*0,0831*	*0,0471*	*0,1098*	*0,0065*	*0,1466*	*0,0182*	*0,0490*	*0,0112*

Annual & CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	-4,1313	-4,3056	-1,2626	0,5076	0,2551	0,9116	-0,0859	-1,1587	-6,7929
2018	-4,0364	-4,1626	-0,9466	0,5243	0,4393	0,9393	0,0194	-1,0319	-6,8592
2019	-3,9451	-4,0095	-1,0000	0,5442	0,5346	0,9666	0,0644	-0,9778	-6,9857
2020	**-3,5193**	**-3,5169**	**-0,8406**	**0,7295**	**0,2681**	**1,0652**	**0,0000**	**-0,8306**	**-6,9251**
Average	**-3,9080**	**-3,9987**	**-1,0125**	**0,5764**	**0,3743**	**0,9707**	**-0,0005**	**-0,9998**	**-6,8907**
Variance	*0,0729*	*0,1178*	*0,0322*	*0,0106*	*0,0185*	*0,0045*	*0,0040*	*0,0185*	*0,0069*

2.6. *Macro Precision*

Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequer**
2017	0,14	0,10	0,23	0,12	0,18	0,09	0,49	0,19	0,09
2018	0,14	0,14	0,30	0,16	0,11	0,12	0,73	0,24	0,12
2019	0,14	0,14	0,30	0,18	0,17	0,11	0,78	0,26	0,11
2020	**0,18**	**0,33**	**0,50**	**0,50**	**0,75**	**0,15**	**1,00**	**0,49**	**0,09**
Average	**0,1493**	**0,1798**	**0,3349**	**0,2405**	**0,3006**	**0,1138**	**0,7510**	**0,2957**	**0,1035**
Variance	*0,0004*	*0,0104*	*0,0135*	*0,0312*	*0,0919*	*0,0006*	*0,0432*	*0,0172*	*0,0002*

Annual Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1460	0,1330	0,1716	0,1572	0,2516	0,0931	0,5141	0,2095	0,0672
2018	0,1688	0,2095	0,3161	0,1301	0,0812	0,0893	0,7349	0,2471	0,1357
2019	0,1932	0,2182	0,2682	0,1181	0,1101	0,0865	0,8244	0,2598	0,1285
2020	**0,2538**	**0,3419**	**0,3040**	**0,2395**	**0,8669**	**0,1420**	**1,0000**	**0,4497**	**0,1075**
Average	**0,1905**	**0,2256**	**0,2650**	**0,1612**	**0,3275**	**0,1027**	**0,7683**	**0,2915**	**0,1097**
Variance	*0,0022*	*0,0075*	*0,0043*	*0,0030*	*0,1349*	*0,0007*	*0,0409*	*0,0116*	*0,0009*

CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1369	0,0461	0,2348	0,1326	0,1239	0,0599	0,4268	0,1659	0,0732
2018	0,1222	0,0933	0,3116	0,2646	0,1379	0,1594	0,7125	0,2573	0,1054
2019	0,0968	0,0951	0,2932	0,3228	0,1687	0,1174	0,7013	0,2565	0,0944
2020	**0,1542**	**0,3106**	**0,6305**	**0,6505**	**0,6927**	**0,1595**	**0,9988**	**0,5138**	**0,0908**
Average	**0,1276**	**0,1363**	**0,3675**	**0,3426**	**0,2808**	**0,1241**	**0,7098**	**0,2984**	**0,0909**
Variance	*0,0006*	*0,0140*	*0,0318*	*0,0484*	*0,0758*	*0,0022*	*0,0546*	*0,0225*	*0,0002*

Annual & CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1383	0,1336	0,2959	0,0615	0,1509	0,1072	0,5386	0,2037	0,1306
2018	0,1197	0,1311	0,2751	0,0990	0,1065	0,1012	0,7460	0,2255	0,1193
2019	0,1300	0,1072	0,3455	0,0927	0,2176	0,1139	0,8143	0,2602	0,1166
2020	**0,1314**	**0,3387**	**0,5726**	**0,6171**	**0,6998**	**0,1356**	**1,0000**	**0,4993**	**0,0729**
Average	**0,1299**	**0,1776**	**0,3723**	**0,2176**	**0,2937**	**0,1145**	**0,7747**	**0,2972**	**0,1099**
Variance	*0,0001*	*0,0117*	*0,0187*	*0,0712*	*0,0754*	*0,0002*	*0,0363*	*0,0187*	*0,0006*

2.7. Macro Recall

Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,12	0,13	0,28	0,10	0,15	0,13	0,32	0,17	0,07
2018	0,13	0,14	0,30	0,12	0,12	0,12	0,45	0,20	0,08
2019	0,13	0,15	0,34	0,12	0,14	0,12	0,55	0,22	0,08
2020	**0,17**	**0,34**	**0,58**	**0,21**	**0,63**	**0,18**	**1,00**	**0,44**	**0,07**
Average	**0,1358**	**0,1891**	**0,3741**	**0,1354**	**0,2564**	**0,1387**	**0,5788**	**0,2583**	**0,0774**
Variance	*0,0005*	*0,0098*	*0,0187*	*0,0023*	*0,0617*	*0,0008*	*0,0875*	*0,0154*	*0,0000*

Annual Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1354	0,1243	0,1988	0,0996	0,1420	0,1278	0,3085	0,1623	0,0723
2018	0,1464	0,1384	0,2458	0,1097	0,1010	0,1265	0,5088	0,1967	0,0792
2019	0,1530	0,1608	0,2826	0,1107	0,1120	0,1260	0,6288	0,2248	0,1144
2020	**0,2017**	**0,3428**	**0,3051**	**0,1355**	**0,6450**	**0,1832**	**1,0000**	**0,4019**	**0,0687**
Average	**0,1592**	**0,1916**	**0,2581**	**0,1139**	**0,2500**	**0,1409**	**0,6115**	**0,2464**	**0,0836**
Variance	*0,0009*	*0,0104*	*0,0022*	*0,0002*	*0,0697*	*0,0008*	*0,0845*	*0,0114*	*0,0004*

CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1126	0,0950	0,2898	0,1124	0,1630	0,1209	0,2990	0,1704	0,0683
2018	0,1315	0,1211	0,3256	0,1400	0,1325	0,1156	0,3793	0,1922	0,0423
2019	0,1005	0,1287	0,3181	0,1331	0,1290	0,1064	0,4318	0,1925	0,0544
2020	**0,1567**	**0,3284**	**0,7445**	**0,2594**	**0,6100**	**0,1815**	**0,9972**	**0,4683**	**0,0516**
Average	**0,1253**	**0,1683**	**0,4195**	**0,1612**	**0,2586**	**0,1311**	**0,5268**	**0,2559**	**0,0541**
Variance	*0,0006*	*0,0116*	*0,0472*	*0,0044*	*0,0551*	*0,0012*	*0,1013*	*0,0202*	*0,0001*

Annual & CSR Grade									
	MLR	**RF Reg**	**NB**	**SVM Poly**	**SVM Gau**	**LDA**	**RF Class**	**Average**	**Words Frequency**
2017	0,1093	0,1769	0,3380	0,0946	0,1306	0,1336	0,3469	0,1900	0,0823
2018	0,1169	0,1628	0,3380	0,1002	0,1142	0,1273	0,4538	0,2019	0,1280
2019	0,1221	0,1482	0,4268	0,1037	0,1670	0,1383	0,5918	0,2426	0,0810
2020	**0,1427**	**0,3415**	**0,6755**	**0,2260**	**0,6308**	**0,1776**	**1,0000**	**0,4563**	**0,0866**
Average	**0,1228**	**0,2073**	**0,4446**	**0,1311**	**0,2607**	**0,1442**	**0,5981**	**0,2727**	**0,0945**
Variance	*0,0002*	*0,0081*	*0,0254*	*0,0040*	*0,0614*	*0,0005*	*0,0818*	*0,0155*	*0,0005*

3. Hypothesis Tests

3.1. Kruskal-Wallis and Tukey's Tests

3.1.1. Average Accuracy

Table 42: Results from the Kurskal Wallis Test on Mean Accuracy of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	df	p-value
Mean Accuracy: Words Frequency vs. MLR vs. RF Reg vs. … vs. RF Class	59.728	7	1.711e-10

Pair Comparison using Tukey's Test

	MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class
MLR	-	-	-	-	-	-	-
RF Reg	0,7348	-	-	-	-	-	-
NB	0,3471	0,9000	-	-	-	-	-
SVM Poly	0,9000	0,9000	0,9000	-	-	-	-
SVM Gau	0,0677	0,8325	0,9000	0,6037	-	-	-
LDA	0,9000	0,9000	0,9000	0,9000	0,6506	-	-
RF Class	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	-
Words Frequency	0,2003	***0,0023***	***0,0010***	***0,0084***	***0,0010***	***0,0065***	***0,0010***

3.1.2. Average F1-score

Table 43: Results from the Kurskal Wallis Test on Mean F1-score of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	df	p-value
Mean F1: Words Frequency vs. MLR vs. RF Reg vs. … vs. RF Class	61.75	7	6.747e-11

Pair Comparison using Tukey's Test

	MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class
MLR	-	-	-	-	-	-	-
RF Reg	0,9000	-	-	-	-	-	-
NB	***0,0464***	0,3422	-	-	-	-	-
SVM Poly	0,9000	0,9000	***0,0395***	-	-	-	-
SVM Gau	0,2637	0,7913	0,9000	0,2358	-	-	-
LDA	0,9000	0,9000	***0,0393***	0,9000	0,2350	-	-
RF Class	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	***0,0010***	-
Words Frequ	0,9000	0,6082	***0,0029***	0,9000	***0,0310***	0,9000	***0,0010***

3.2. Friedmann and Nemenyi's Tests

Table 44: Results from the Friedmann Test comparing the Mean F1-score, Accuracy, MAE, Bias, Correlation, Precision, and Recall of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Nemenyi's test (below).

Friedmann rank sum Test

	Kruskall-Wallis Chi Squared	df	p-value
Words Frequency vs. MLR vs. RF Reg vs. … vs. RF Class	28.762	6	0.0001598

Pair Comparison using Nemenyi's Test

	MLR	RF Reg	NB	SVM Poly	SVM Gau	LDA	RF Class
MLR	-	-	-	-	-	-	-
RF Reg	0,9000	-	-	-	-	-	-
NB	0,6357	0,9000	-	-	-	-	-
SVM Poly	0,9000	0,9000	0,7662	-	-	-	-
SVM Gau	0,5053	0,9000	0,9000	0,6357	-	-	-
LDA	0,9000	0,9000	0,8967	0,9000	0,7662	-	-
RF Class	***0,0466***	0,2984	0,8967	0,0864	0,9000	0,1490	-
Words Freque	0,7010	0,2407	***0,0165***	0,5705	***0,0077***	0,4352	***0,0010***

Part 3: Comparison of Disclosure format

1. Category Ratings

1.1. Summary of Performance Metrics

Table 45: Performance metrics by type of report

		Comparison Category						
		MAE	**COR**	**ACC**	**F1**	**BIAS**	**PRECISION**	**RECALL**
Annual	Average	**0,6111**	**0,4276**	**0,5692**	**0,3824**	**-0,2746**	**0,4912**	**0,4011**
	Variance	*0,0046*	*0,0131*	*0,0053*	*0,0094*	*0,0016*	*0,0124*	*0,0068*
CSR	Average	**0,6018**	**0,3434**	**0,5495**	**0,3082**	**-0,3451**	**0,4143**	**0,3515**
	Variance	*0,0100*	*0,0330*	*0,0116*	*0,0164*	*0,0121*	*0,0135*	*0,0109*
Annual & CSR	Average	**0,6268**	**0,4231**	**0,5442**	**0,3336**	**-0,3114**	**0,4334**	**0,3770**
	Variance	*0,0058*	*0,0076*	*0,0053*	*0,0072*	*0,0052*	*0,0046*	*0,0047*

1.2. Kruskall-Wallis and Tukey's Test – Annual vs. CSR vs. Annual & CSR (global)

1.2.1. Accuracy

Table 46: Results from the Kruskal Wallis Test on Mean Accuracy of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	**df**	**p-value**
Mean Accuracy: Annual vs. CSR vs. Annual & CSR	6.7181	3	0.03477

Pair Comparison using Tukey's Test

	Annual	**CSR**
Annual	-	-
CSR	0,8702	-
Annual & CSR	0,8958	0,9000

1.2.2. F1-score

Table 47: Results from the Kruskal Wallis Test on Mean F1-score of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	**df**	**p-value**
Mean Accuracy: Annual vs. CSR vs. Annual & CSR	1.8647	3	0.39362

Pair Comparison using Tukey's Test

	Annual	**CSR**
Annual	-	-
CSR	0,2832	-
Annual & CS	0,6342	0,7817

1.3. Friedmann and Nemenyi's tests – Annual vs. CSR vs. Annual & CSR (global)

Table 48: Results from the Friedmann Test comparing the Mean F1-score, Accuracy, MAE, Bias, Correlation, Precision, and Recall of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Nemenyi's test (below).

Friedmann rank sum Test

	Kruskall-Wallis Chi Squared	df	p-value
Mean Accuracy: Annual vs. CSR vs. Annual & CSR	6.0	2	0.0498

Pair Comparison using Nemenyi's Test

	Annual	CSR
Annual	-	-
CSR	***0,0427***	-
Annual & CSR	0,6849	0,2444

1.4. Kruskal-Wallis and Tukey's tests – Model by Model – Accuracy

Table 49: Results from the Kurskal Wallis Test on Mean Accuracy of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test for Models for which difference was significant (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	p-value	df
MLR	5,653846154	**0,059194711**	3
RF Reg	3,038461538	0,218880192	3
NB	2,192307692	0,334153822	3
LDA	2,346153846	0,309413434	3
SVM Poly	3,115384615	0,210621561	3
SVM Gau	5,760526316	**0,056119992**	3
RF Class	1,076923077	0,583645478	3
Words Frequency	6,961538462	**0,030783722**	3

MLR

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1	0,13666344	***0,052538087***
CSR	0,13666344	1	0,800768188
Annual & CSR	***0,0525381***	0,80076819	1

SVM Gau

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1,000	0,181	0,900
CSR	0,181	1,000	0,190
Annual & CSR	0,900	0,190	1,000

Words Frequency

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1,000	***0,071***	0,235
CSR	***0,071***	1,000	***0,005***
Annual & CSR	0,235	***0,005***	1,000

1.5. Kruskal-Wallis and Tukey's tests – Model by Model – F1-score

Table 50: Results from the Kurskal Wallis Test on Mean F1-score of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test for Models for which difference was significant (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	p-value	df
MLR	7,730769231	**0,020954861**	3
RF Reg	3,846153846	0,146156557	3
NB	2,192307692	0,334153822	3
LDA	3,576923077	0,167217229	3
SVM Poly	4,769230769	**0,092124405**	3
SVM Gau	4,192307692	0,122928321	3
RF Class	2,807692308	0,245650336	3
Words Frequency	7,730769231	**0,020954861**	3

MLR

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1	***0,0022457***	***0,001264368***
CSR	***0,0022457***	1	0,9
Annual & CSR	***0,0012644***	0,9	1

SVM Poly

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1,000	0,181	***0,084***
CSR	0,181	1,000	0,864
Annual & CSR	***0,084***	0,864	1,000

Words Frequency

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1,000	***0,017***	0,519
CSR	***0,017***	1,000	***0,003***
Annual & CSR	0,519	***0,003***	1,000

2. Grade Ratings

2.1. Summary of Performance Metrics

Table 51: Performance metrics by type of report (Grade Ratings)

		Comparison Grade						
		MAE	**COR**	**ACC**	**F1**	**BIAS**	**PRECISION**	**RECALL**
Annual	Average	**2,1864**	**0,3893**	**0,3253**	**0,2228**	**-0,8093**	**0,2915**	**0,2464**
	Variance	*0,0549*	*0,0190*	*0,0116*	*0,0124*	*0,0037*	*0,0116*	*0,0114*
CSR	Average	**2,0296**	**0,3925**	**0,3404**	**0,2276**	**-1,0931**	**0,2984**	**0,2559**
	Variance	*0,0969*	*0,0458*	*0,0225*	*0,0223*	*0,0490*	*0,0225*	*0,0202*
Annual & CSR	Average	**2,0493**	**0,4868**	**0,3386**	**0,2476**	**-0,9998**	**0,2972**	**0,2727**
	Variance	*0,0943*	*0,0197*	*0,0151*	*0,0166*	*0,0185*	*0,0187*	*0,0155*

2.2. Kruskall-Wallis and Tukey's Test – Annual vs. CSR vs. Annual & CSR (global)

2.2.1. Accuracy

Table 52: Results from the Kruskal Wallis Test on Mean Accuracy of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	**df**	**p-value**
Mean Accuracy: Annual vs. CSR vs. Annual & CSR	0.66888	3	0.715739

Pair Comparison using Tukey's Test

	Annual	**CSR**
Annual	-	-
CSR	0,9000	-
Annual & CSR	0,9000	0,9000

2.2.2. F1-score

Table 53: Results from the Kruskal Wallis Test on Mean F1-score of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	**df**	**p-value**
Mean Accuracy: Annual vs. CSR vs. Annual & CSR	0.16101	3	0.92265

Pair Comparison using Tukey's Test

	Annual	**CSR**
Annual	-	-
CSR	0,9000	-
Annual & CS	0,9000	0,9000

2.3. Friedmann and Nemenyi's tests – Annual vs. CSR vs. Annual & CSR (global)

Table 54: Results from the Friedmann Test comparing the Mean F1-score, Accuracy, MAE, Bias, Correlation, Precision, and Recall of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Nemenyi's test (below).

Friedmann rank sum Test

	Kruskall-Wallis Chi Squared	df	p-value
Mean Accuracy: Annual vs. CSR vs. Annual & CSR	7.1428	2	0.02811

Pair Comparison using Nemenyi's Test

	Annual	CSR
Annual	-	-
CSR	0,3762	-
Annual & CS	***0,0206***	0,3762

2.4. Kruskal-Wallis and Tukey's tests – Model by Model – Accuracy

Table 55: Results from the Kurskal Wallis Test on Mean Accuracy of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test for Models for which difference was significant (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	p-value	df
MLR	3,576923077	0,167217229	3
RF Reg	1,5	0,472366553	3
NB	5,653846154	**0,059194711**	3
LDA	1,846153846	0,397294713	3
SVM Poly	4,192307692	0,122928321	3
SVM Gau	1,884615385	0,389727426	3
RF Class	0,464788732	0,792633474	3
Words Frequency	4,192307692	0,122928321	3

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1,000	0,320	0,222
CSR	0,320	1,000	0,900
Annual & CSR	0,222	0,900	1,000

2.5. Kruskal-Wallis and Tukey's tests – Model by Model – F1-score

Table 56: Results from the Kurskal Wallis Test on Mean F1-score of Text Mining based ESG scoring models (above). Table of p-values returned by the pair comparison by Tukey's test for Models for which difference was significant (below).

Kruskal-Wallis rank sum Test

	Kruskall-Wallis Chi Squared	p-value	df
MLR	6,730769231	**0,034548726**	3
RF Reg	3,038461538	0,218880192	3
NB	7,538461538	**0,023069802**	3
LDA	2,461538462	0,292067824	3
SVM Poly	3,230769231	0,198814189	3
SVM Gau	0,5	0,778800783	3
RF Class	1,051754386	0,591036684	3
Words Frequency	6	**0,049787068**	3

MLR

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1	***0,0239665***	***0,036979895***
CSR	***0,0239665***	1	0,9
Annual & CSR	***0,0369799***	0,9	1

NB

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1,000	0,216	0,173
CSR	0,216	1,000	0,900
Annual & CSR	0,173	0,900	1,000

Words Frequency

Pair Comparison using Tukey's Test

	Annual	CSR	Annual & CSR
Annual	1	***0,0241438***	0,9
CSR	***0,0241438***	1	***0,045570137***
Annual & CSR	0,9	***0,0455701***	1

Abstract :

In this Master thesis, we propose to measure the ESG performance of a company by relying on the text mining analysis of its sustainability disclosure. More precisely, we are deriving ESG ratings from the number of occurrences of ESG terms in Annual and CSR reports. Of course, one cannot expect such a technique to provide a fully accurate indicator of ESG performance. Yet, in light of the importance of ESG described above and given the large number of firms investors have to analyze, this is a first step towards a more efficient and qualitative assessment of ESG performance.

Concretely, we are evaluating the performance of ratings built by a set of 8 different models (the sum of words frequency, the Multiple Linear Regression, the Naive Bayes, the Support Vector machine with non-linear kernels, the Random forest regressor and classifier, and the Linear Discriminant Analysis) against the official ESG ratings from the firm Refinitiv. Eventually, we find that this approach of measuring ESG is at best when relying on a Random Forest classifier. Yet, even in this optimal situation the predictions made by the model did not perfectly approximate Refinitiv's rating. Despite these disappointing results, our study allowed us to highlight certain interesting avenues for research on the development of Text Mining based ESG ratings.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique
www.uclouvain.be/lsm