UCLouvain

LOUVAIN
School of Management

Louvain School of Management

The thematic content of ESG reports

A topic modelling approach

CONFIDENTIAL

Author: Thomas Grava
Supervisor: James Thewissen
Academic year 2021-2022
Dissertation for the Master in Management
Majoring in Corporate Finance
Daytime schedule

Abstract

ESG (Environmental, Social and Governance) reports are becoming mainstream, but their thematic content remains unknown to investors. Based on a sentence-based topic modelling, we open the black box and identify 30 topics contained in ESG reports.

We find that the thematic content of ESG reports contains information value to better estimate future ESG and financial performance. After categorizing topics between the economical, social and governance pillars, we find that investors place the most value on social topics, with an emphasis on the easiest to measure subjects. An imbalance in the reporting of E, S or G pillar is associated with poorer future financial performance and a lower ESG score. Finally, we look at the ESG reports of energy companies and observe that investors react primarily to environmental themes.

Finally, we conclude our paper by drawing key lessons and contextualizing our analysis in light of current trends in the financial sphere.

Acknowledgment

This intense and intellectually rewarding project would not have been possible without the help of my promoter, Professor Thewissen, who continuously supported me by making himself available to answer my countless questions.

One of the biggest challenges has been to remain consistent in my effort throughout these past months. For this reason, I would also like to thank my relatives and friends for the endless exchanges that inspired me when I encountered difficulties.

"Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things" (Sir Isaac Newton)

Keywords: ESG; Topic Modelling; Information value; Sustainability ; Energy firms

Contents

1. Introduction

Environmental, Social and Governance (ESG) reporting is voluntary corporate non-financial financial information that provides in-depth information about the firm's social and environmental performance, such as employee well-being, diversity, community outreach or the environment.[1] ESG reports have become a "one-stop source of CSR performance information for stakeholders" Du and Yu (2020) and are increasingly disclosed by companies both within the United States and globally King and Wim (2015). Despite an increasing number of firms reporting on their ESG activities, the lack of regulatory enforcement from public institutions and the opportunity to manipulate the content of such reports remain ongoing hurdles for the trustworthiness and credibility of such reports. Often compared to financial information, there are distinct informational challenges regarding ESG reports, as investors base their investment decisions and assess the riskiness of their investment on information that remains unaudited, voluntary and unregulated Allen and Ramanna (2013). However, despite the importance that investors and other stakeholders attach to the ESG report, there has been no systematic effort to examine its actual content. By relying on unsupervised machine learning, this study investigates ESG reports' content and its informativeness in predicting firms' financial and ESG performance.

Given the prominence of ESG reports in investors' decision-making, it is essential to identify its key factors that influence stakeholders' decisions. Recently, a growing body of research examines whether narratives and the associated language in ESG reports may help influence investors' behavior by conveying a sustainable identity to the firm, with mixed results. For instance, Du and Yu (2020) show that the tone and readability of ESG reports contain information value to predict both ESG and financial performance. Nazari et al. (2017) and Wang et al. (2019) further find a positive correlation between ESG performance and readability, while Clarkson et al. (2020) show that ESG report length (proxied by the number of words and sentences), as well as tone and readability, can predict ESG and financial performance. In contrast, there is also evidence of reports being used in order to greenwash investors' expectations regarding the firm's ESG commitment. For instance, Patten (2002) and Cho and Patten (2007) find that poorer environmental performance is associated with higher levels of disclosures, indicating that managers mostly use environmental disclosures as a legitimizing tool.

[1] As indicated by Du and Yu (2020), various terms are used to refer to company reports that provide information on social and environmental performance, such as corporate social responsibility (CSR) reports, corporate responsibility/citizenship reports, sustainability reports, and so on. For simplicity, we use the term "ESG reports" to broadly refer to these reports.

In addition, Muslu et al. (2019) relies on computer-based textual analysis to estimate ESG reports' content in terms of length, readability, tone, numeracy, and horizon. While they observe that ESG reporters with high disclosure scores are associated with more accurate forecasts, they find no difference in capital market consequences between ESG reporters with low disclosure scores and non-ESG reporting firms. Against this backdrop, whether the narrative content of ESG reports contain information value regarding future financial and ESG performance remains an open question. This study aims at unpacking the black box that constitute ESG reports and investigate the information value of ESG reports by examining *what* is contained in report instead of *how* information is communicated.

We propose that the thematic composition of ESG reports contains information value to explain future ESG and financial performance. The information content of ESG reports, such as tone, the number of words or readability, have frequently been highlighted as a key factors to explain future firm financial and ESG performance. However, if word count or length may capture the amount of information contained in a report, it is a weak proxy, if at all, for the quality and type of information conveyed by a document. In fact, such linguistic attributes ignore the underlying meaning or context of the disclosure and consider words as independent and informative units, thereby limiting the inferences that can be drawn (Loughran and Mcdonald, 2016). In addition, the approach of counting words to measure informativeness needs to be questioned in the context of ESG reports given their level of heterogeneity and the lack of a formal control mechanism or standardization. As argued by Hoberg and Lewis (2017), the flexible nature of disclosure content requires a more extensive set of dimensions along which we could analyze white paper narratives. Therefore, there is an urgent need to examine the content of ESG reports and demonstrate methods that accommodate contextual subtleties and informational differences across documents. In light of this background, in this study, we explore three research questions (i) what are the topics discussed in ESG reports, (ii) which topics matter to investors (iii) is the thematic content informative to predict future financial and ESG performance.

The thematic content of a white paper is defined as the distribution of underlying topics in a document, and is measured as the number of sentences relating to a specific theme. To identify the topics included in white papers, we use a novel process-based approach based on the Latent Dirichlet Allocation (LDA) at the sentence-level (sentLDA) introduced by Bao and Datta (2014). This method reasonably assumes that a sentence is the smallest integral unit of text that conveys a complete and meaningful idea (Ivers, 1993), and thus incorporates the information in sentence boundaries while

identifying the topic clusters (see Bao and Datta, 2014). Consequently, it allows us to identify a topic for each sentence, as opposed to solely estimating the topic distribution of the entire document. Because the purpose of our paper is to identify and quantify the white papers' content to understand its information value, rather than simply classifying the documents, sentLDA is particularly suited for the task. The topical composition is then used to discover the cues that influence investors' reaction at the release of the ESG report. Correspondingly, an investigation of the relationship between the discovered topics and financial performance and ESG ratings is included as a part of the study.

2. Motivation and literature review

2.1. ESG reports – A voluntary disclosure

ESG reporting has become increasingly prevalent in the past 30 years both in academic research and business strategy. This success is clearly visible in the exponential growth in the number of firms that measure and report ESG information (Cho et al., 2015). According to the Government Accountability Institute (2019), whereas less than 20 companies disclosed ESG data in the early 1990s, the number of firms issuing sustainability or integrated reports had increased to nearly 9,000 by 2016. Simultaneous to the supply of such reports, ESG information is increasingly used by investment managers to establish portfolios. Amel-Zadeh and Serafeim (2018) find that the clear majority of portfolio managers (82%) use ESG information because it is financially material to investment performance, which supports the fact that ESG reporting has become a mainstream source of information.[2,3] In their report on the evolution of ESG reporting, Tschopp and Huefner (2015) state that ESG reporting is increasingly destined to "become a key part of the overall accounting reporting framework, joining external financial reporting, income tax reporting, regulatory reporting, and internal reporting."

This focus on ESG reporting is largely motivated by the pressure from stakeholder groups, such as non-governmental organizations, heightened government regulation and increased investor interest in ESG-related information, but mostly because ESG information complements financial reporting (Vitolla et al., 2019, 2020). Relative to other corporate disclosures studied in prior management literature, such as the 10-K, CEO letter to shareholders, 8-K or the MD&A (Boudt and Thewissen,

[2]Sustainable investment assets globally rose by 34% in just two years to $30.7 trillion in 2018 (Lang and Electris, 2018), with half of ESG assets concentrated in the European market.

[3]According to another survey conducted by the CFA Institute (2015), the majority of survey participants (63%) cited management of investment risk as the reason why they incorporate ESG into investment and analysis.

2019; Arslan-Ayaydin et al., 2016, 2020; Arslan-Ayaydin et al., 2021; Shrestha et al., 2021), ESG reports are unique because they provide a deeper and larger vision on corporate social performance in critical domains in the firm, such as employee welfare, diversity, community outreach, product safety, the use of energy and the impact on the environment. Compared to financial reporting that provides information to the investor community and specializes on providing financial quantitative data, the target of the ESG report consists of a large array of stakeholders, such as customers, suppliers, labor unions and investors. This information is primarily provided in the form of textual information regarding the firm's policies, practices and performance in social, environmental and governance domains (Dhaliwal et al., 2010).

Another important contrast with mandatory financial information is that ESG reports are voluntary, largely unregulated and do not fit in a widely enforced reporting framework (Perrini and Tencati, 2006; Tschopp and Huefner, 2015). In spite of significant attempts to enforce a standardized ESG reporting topology (Allen and Ramanna, 2013), managers keep a substantial discretion in whether, what type and how they report ESG-related information. In fact, there is no standardized disclosure format dictating what a ESG report should or should not include. This lack of regulation regarding ESG reporting explains the diverse reporting practices with respect to length, performance indicators and topics contained in the report. In addition, the audit of such reports is neither comprehensive, nor stringent compared with their verification of corporate annual reports. Even if the report is audited, the scope of the audit relates to the process and less on the information itself (Muslu et al., 2019). A recent study by EY (2020) shows that most global and U.S. investors are dissatisfied with the persisting heterogeneity in ESG reporting practices.

This lack of regulation and audit leads the way to opportunistic behavior and selective reporting to manage stakeholders' impression regarding the firm's ESG performance. An increasing number of studies shows that firms tend to opportunistically emphasize the positive side of their ESG initiatives. Supporting this view of selective reporting, Li et al. (2006) find that firms with stronger ESG performance have higher levels of discretionary ESG disclosures. Similarly, Nazari et al. (2017) find that ESG performance is positively related to the length of CSR reports, suggesting that firms with superior ESG performance disclose more whereas those with inferior ESG performance disclose less. Similarly, Mishra and Modi (2013) find that managers highlight the positive ESG and ignore the negative sides of their ESG performance. Piecing this opportunistic behavior together with the lack of regulatory oversight raises the question of whether ESG reports contain, in fact, credible information.

2.2. The information value of ESG reports

This question resonates with a longstanding debate in financial accounting between voluntary versus mandatory disclosures (see, Healy and Palepu, 2001). Theoretical research argues that the disclosure of information leads to liquid and efficient financial markets, resulting in a lower cost of capital for firms (Grossman and Hart, 1980; Milgrom, 1981). A central prediction of information economics is that these incentives would lead managers to provide information relevant to investors voluntarily. However, as often observed in traditional markets, without oversight, in non-ideal conditions, firms are disinclined to make adequate disclosures (see, Beyer et al., 2010). Many studies evidence that increased volume of ESG disclosure is an unreliable and misleading indicator because self-serving firms can manipulate disclosure volume to emphasize positive aspects of their ESG performance, while minimizing the negative ESG aspects of their management. For instance, Muslu et al. (2019) relies on computer-based textual analysis in order to measure ESG report content in terms of length, readability, tone, numeracy, and horizon, and use these features in order to form an ESG disclosure quality score. While they observe that ESG reporters with high disclosure scores are associated with more accurate forecasts, they find no difference in capital market consequences between ESG reporters with low disclosure scores and non-ESG reporting firms. Patten (2002) and Cho and Patten (2007) further find that poorer environmental performance is associated with higher levels of disclosures, indicating that managers mostly use environmental disclosures as a legitimizing tool. Overall, these results indicate that corporate ESG disclosure is used as a mechanism to opportunistically manage stakeholder impressions, which could indicate greenwashing and a lack of informational content of ESG reports.

In contrast, there is also evidence suggesting that the content of ESG disclosures provide information to investors (Werther and Chandler, 2005). Recent research documents that ESG disclosures are associated with significant capital market benefits such as reduced cost of equity capital (Plumlee et al., 2015), higher share prices (Matsumura et al., 2014) and more accurate financial analysts' forecasts (Grewal et al., 2020). Furthermore, Du and Yu (2020) show that the tone and readability of ESG reports contain information value to predict both future ESG and firm performance. In particular, they find that one-year-ahead ESG performance is positively associated with the changes in both ESG report readability and tone, suggesting that more readable text and more optimistic tone in a ESG report indicate better future ESG performance. Furthermore, consistent with the view that ESG reports communicate important value relevant information to the market, they find a significant

market reaction to report readability and tone around the release of ESG reports. In line with this evidence, Nazari et al. (2017) and Wang et al. (2019) both show a positive relationship between ESG performance and readability, suggesting that better ESG performers are likely to have CSR reports with better readability. Given the contrast in findings between the two groups of studies, the extent to which ESG reports are informative to investors is unclear and remains investigated.

2.3. Opening the black box of ESG reports

Despite the ongoing debate on the information value of ESG disclosures, there is an important gap in the literature. Information content of ESG reports, such as tone, the number of words or readability, is repeatedly highlighted as a key determinant of the informativeness of the reports. However, measures used to represent information content have been weak, often based only on the count of words in the text. If word count or length may capture the amount of information contained in a ESG report, it is a weak proxy, if at all, for the quality and type of information conveyed by a document. In fact, such linguistic attributes make abstraction of the underlying meaning and consider words as independent and informative units (Loughran and Mcdonald, 2016). In addition, whilst the analysis of tone or readability could arguably be applied to analyze standardized audited financial corporate disclosures (e.g., 10-K), this approach to measure informativeness is particularly problematic for ESG reports given the level of heterogeneity and the lack of a formal control mechanism or standardization. The flexible nature of disclosure content requires a more extensive set of dimensions along which we could analyze ESG narratives (Hoberg and Lewis, 2017). There is therefore an urgent need to open the black box that constitute ESG reports by analyzing their thematic content.

We propose that the information content of ESG reports depends not only on the quantity of information, but also on the type of information provided in the report. While considering investing in a firm, an investor seeks information on a large array of aspects of the company, including information on the environmental impact, the governance issues, the employee welfare and the customer service. As such, we suggest that ESG information needs to be represented on the basis of the granular concepts that it represents, and not just on the count of words it contains. The study of Hummel et al. (2017) is arguably the closest paper to ours in attempting to understand the content of ESG reports. While they rely on content analysis to classify the content of ESG reports across four pre-defined topics (education, philanthropy, parental and climate), we distinguish ourselves from this paper by conducting a holistic, replicable and fine-graded analysis of the topics covered in ESG reports by

relying on recent advances in topic modeling. Moreover, topic modeling allows us to highlight an array of topics and to identify a large heterogeneity of themes covered in ESG reports. In addition, we set out to further leverage the information contained in the thematic content to predict future ESG and finance performance, as well as to enhance our ability to identify the topics of value-relevance to investors.

In this study, we first consider topic modeling to summarize ESG reports into individual topics (representing concepts). Topic modeling allows richer analysis than the standard text mining approach of representing information through keywords (Zhang et al., 2008). We then test whether the thematic content contains information value to predict future ESG and financial performance, beyond what can be achieved using quantitative financial metrics and aggregate measures of textual style features. Furthermore, which components of the Environment-Social-Governance construct do investors value the most? We, therefore, contribute to the literature by addressing the following research questions:

Research Question 1: What are the topics discussed in ESG reports?

Research Question 2: Is the thematic content of ESG reports informative to predict future ESG and financial performance, relative to other firm-specific characteristics and ESG reports' style features?

Research Question 3: Which topics contained in ESG reports are value-relevant to investors around the release of the ESG report? Which facet(s) of the ESG construct is valued the most/least by investors?

2.4. Additional analyses

Sustainability is a paradigm for thinking about the future in which environmental, societal and economic considerations are balanced in the pursuit of an improved quality of life (UNESCO, 2015). This paradigm is a complex construct and has led to more than 300 definitions in prior literature (Santillo et al., 2007). Yet, despite the heterogeneity in the definitions, most share a common element in that the ESG construct is a *balanced* integration of intra and inter-generational economic, social and environmental performance. This definition assumes that a company is sustainable only if it finds a proper equilibrium between the people, profit and planet by establishing causality and positive

feedback between the three spheres (McKelvey, 2002). If a company conducts a balanced ESG policy, this balance should reflect in the ESG report, where the pillar E, S and G would be given similar importance. Such company would therefore benefit from a superior ESG rating. On the contrary, if a firm draws more focus on the 'E' component of its sustainability policy at the expense of the 'S´ or 'G´ pillars, it is likely to be less sustainable. We therefore expect firms with an unbalanced distribution of topics regarding their pillars E, S and G to suffer from a lower ESG score.

Research Question 4: Do firms with a balanced distribution of topics across their E, S and G pillars benefit from a higher ESG score?

Another observation that can be made is about firms dealing in the Energy industry, comparatively to firms dealing in other sectors. Previous studies revealed that investors willing to invest in energetic firms should pay more attention to pollution improvement to decrease their risk Wu and Hu (2019). Another one described how Governance thematics are drivers to firms' ESG performance Shahbaz et al. (2020). The question is therefore to analyze if firms from the Energy sector are discussing different topics than firms from other industries.

Research Question 5: What are the topics discussed in ESG reports from the Energy sector?

3. Research Methodology

3.1. Overview of the data and the method

As described previously in "ESG reports – A voluntary disclosure", firms disclose their ESG reports on an annual basis without any mandatory auditing confirming its content. The first challenge is therefore to find a consistent way to build a representative sample within a specific time frame, regardless of their industry or their country of origin. By parsing publicly-listed companies websites, we built a large corpus of 5200 ESG reports from multiples countries by referencing their name, their International Securities Identification Number, and tickers, which we will subsequently detail in 4.1.

3.2. Sent-LDA

Among unsupervised machine-learning algorithm , we use the Latent Dirichlet Allocation. LDA is a non-linear generative probabilistic model for collections of discrete data (Blei et al., 2003). It has the advantage of reducing the dimensional complexity of features' classification by representing a document to a fixed set of words even when involves with multidimensional structures, as it is often the cases in unregulated reports. Yet, the LDA framework assumes that the order of words in a document is irrelevant, arguing that the informational value lies in the word itself and not the sentence. Sentences have a structure, and we want to identify the information value of a topic, it is best to approach the question from a sentence-based perspective instead (Thewissen et al., 2021). In this paper, we will be using its derivative, namely a Sentence Latent Dirichlet Allocation. Its advantage lies within the hypothesis that it sets, by taking into account the order of words in a sentence and reducing a sentence to a single topic, rather than its counter-part, who describes a sentence by a plurality of topics. In a more formal manner, following Wang et al. (2019):

1 For each topics k with $k \in \{1,...,\mathrm{K}\}, with \beta \sim Dirichlet(\eta)$ drawing a distribution over vocabulary words.

2 For each document d, $\Theta \sim Dirichlet(\alpha)$ drawing a vector of topic proportions:

3 For each sentence s in document d, $z_{d,s}$ ~Multinomial(Θ_d), drawing a topic assignment.

4 For each word $w_{d,s,n}$ in sentence s, $w_{d,s,n}$ ~Multinomial($\beta_{z,d,s}$), drawing a word.

The inferential issue, emerging from any Bayesian distributions is reflected in the posterior distribution of the hidden variables' computation Θ, representing the topic proportions, and ζ, representing the topic assignments, given its prior parameters w, α and β. As a result, the distribution is intractable to compute (Blei et al., 2003), meaning we need to approximate the poster distribution through learning algorithms. In this study, we will use the Variational Expectation Maximization learning algorithm, since it outperforms other state-of-the-art algorithms, in terms of clustering and estimation accuracy (El Assaad et al., 2016). In the first stage, the EM algorithm is applied on the dataset, and during the second stage, it's applied iteratively on the parameters of the distribution.

3.3. Regression analyses on the predictive value of SentLDA ESG reports topics

Following our first hypothesis, as stated in "Opening the black box of ESG reports", we extract the information value from the thematic composition of ESG reports by using a multiple linear regression

with our K topics, previously selected from the sentLDA algorithm, through the following model, as per Li et al. (2006) and Thewissen et al. (2021) :

$$Y_j = \alpha + \gamma \cdot ESGControls_j + \beta \cdot YearControls_j + \beta \cdot IndustryControls_j + \upsilon_j \quad (1)$$

Where Y represents the different ESG outcome variables, and υ is the Newey-West standard errors.

The company financial performance variables are expressed through the first-days cumulative abnormal return, through the Tobin and the ESG. More precisely, to measure the informational value of our corpus of document. Formally, following Li et al. (2006) and Thewissen et al. (2021):

$$Y_j = \alpha + \beta \cdot \sum Topics_{k,j} + \gamma \cdot ESGControls_j + \beta \cdot YearControls_j + \beta \cdot IndustryControls_j + \upsilon_j \quad (2)$$

Where the variable *Topics* is the number of sentences referring to topic k in the ESG reports, υ is the Newey-West standard errors, and *YearsControls* and *IndustryControls* being the dummy variables. The independend variables, represented by the 30 topics, are furthermore aggregated into the three main pillars of the ESG paradigm by computing their proportion such as Topic E = Number of Sentences in E divided by the total number of sentences.
We are following the same methodology for the two other pillars, namely S and G.

Following the second hypothesis, we want to find if thematic content of ESG report is informative to predict futur financial performance. Using Tobin's as dependent variable reflects underpricing or overpricing from a company market value and asset, which is *ipso facto* a measure of a company performance. Tobin is forward looking and has been proven to be an adequate variable to assess futur financial performance, as stated in Gaio and Henriques (2020) or Margaretha and Rachmawati (2016). On the other hand, *Tobin* "may be sensitive to variation that are independent of the operations and social activities of firms, like macroeconomics shock or industries-specific factors such as a shift in industry demand or restriction on supply" (Cavaco and Crifo, 2014). To compensate that downside, we use Industry and Years as dummies variable.

Following the third hypothesis, we aim at identifying which topics are of value-relevance for investors.In that regard, following Du and Yu (2020), we define $CAR[-5;+5]$ as our dependent variable Y. The cumulative abnormal return tests the initial reaction from investors by measuring the market response to the newly issued report (Du and Yu, 2020). It takes a window of 10 days around the report's date of publication and measures any price diversions. A stock price divergence transcends a surprise from investors on information that other mandatory disclosure reports failed to disclose. From that perspective, and assuming the Efficient Market Hypothesis, Cumulative Abnormal Return could reflect the long term success from the company's objectives.

3.4. Control Variables

Our dataset is international, with multiple industries, therefore the control variables need to be adapted to that framework. Following prior literature, we will use the same control variables than Du and Yu (2020) for the linear regression with cumulative abnormal return as dependent, the control variables from Li et al. (2006) for the *ESG* and *Tobin* as dependent variable. We start by controlling for *Fin* which is defined as the amount of debts raised by the firm at time t+0. A company, especially if their business is international, may leverage debts to reduce the amount of net taxes paid (Fatica et al., 2013), and therefore needs to be adjusted across our sample. We then control for *Volat* which is computed as the standard deviation on *ROA* for a period of 5 years. There is *Size*, expressed as the natural logarithm of the company's total assets. Most of our companies are publicly listed companies, hence that factor itself can influence drastically our dependent variable. We also control for *Proportion*, which is the logarithm of the total number of sentences. Other 5 control variables, out of the 11 ones, are described in *Table 1*. Associated with those control variables, we also include dummy variables for the years, and the industries (Clarkson et al., 2020; Li et al., 2006) .

The $\Delta Read$ is the change of readability in ESG reports. In this study, we use the FOG index as developed by Robert Gunning (Strong, 1986), computed as the average number of words per sentence added with the percentage of words per sentence. Following prior literature (Du and Yu, 2020), a high FOG index score reflects a more difficult text comprehension, and vice versa, a lower score reflects an easier text readability. The main objective from the FOG index is "to reflect the number of years that a reader of average intelligence would need to read and understand the text" (Du and Yu, 2020). Following prior literature, such as (Du and Yu, 2020) we know that Fog index is limited since it doesn't take into account the reader's background. For instance, company disclosures are

using seemingly complicated financial terms, yet, they are fairly easy to comprehend for a financial analyst. In this study, following (Lang and Stice-Lawrence, 2015; Nazari et al., 2017; Muslu et al., 2019; Dhaliwal et al., 2010), we will nonetheless choose the FOG Index method since disclosure length is "positively correlated with transparency and informativeness, especially for ESG reports" (Nazari et al., 2017).

The $\Delta Tone$ measures the change in ESG reports tone, it's a measure of sentiment of a report. According to previous literature (Du and Yu, 2020), Lougrhan and McDonald's list of positive and negative words is more adequate for business disclosure, especially when words are context-dependent (Loughran and Mcdonald, 2016). For instance, *nuclear* is relative to the industry and yet, using a different vocabulary list would have been interpreted as negative. Following prior literature, we define POS (NEG) as the number of positive (negative) words scaled by the number of total words.

4. Data and summary statistics

4.1. Data

Contrary to 10-Ks, ESG reports aren't centralized which makes its collection particularly challenging. Although there are few websites trying to compensate for that need, such as https://database.globalreporting.org/, reports generally tend to be shared on the company's own website, and yet most of them aren't machine-readable. One of the ways to make it machine-readable used by IFRS (2019) or European is to digitally tag each report by associating a date and an author. We hereby extract multiple reports all over the world, yearly, in English, and digitally tag each of those reports by extracting PDF's meta-data. Among other information, PDF meta-data offers a date of publication, author name, and date of the latest modification. As a robustness test, we first check if the date of the latest modification differs from its date of publication, then we manually check their publication's date by looking it over the newspaper.

Having gathered enough reports giving us confidence in our sample to infer a causal relationship, we then collect the machine-readable ESG reports that are still available from the initial collection. We thereafter proceed to transform our PDF format into a text format (Thewissen et al., 2021), and start our filtering process to optimize the topic modelling procedure. The common way to clean the data, as described by previous literature, is to remove non-English words by using a

standard stop-words list, which also takes care of the highly frequent words such as "the", "are", "and" carrying no informational value 'per se'. Having extract our corpus of words, we then filter words that are specific to a document - since our study identifies common themes across ESG reports - thanks to a term frequency-inverse document frequency (tf-idf) statistical measure which transcends how important a word is to a document (Kim and Gil, 2019).

4.2. Summary statistics

Table 3 offers an insight into our control and dependent variables. We also provide (*Table 2*) a geographical analysis by ordering reports by continents, and the result shows that the trend of ESG reports keeps growing over the years with a substantial presence from the North American continent, as shown in the latest KPMG report (Threlfall et al., 2020).

5. The topic modeling of ESG reports

We hereby provide our main results for each of the three main hypotheses. First, we validate our topic modeling implementation, then we detail which topics are contained in the ESG reports. Finally, we evaluate whether thematic content can predict future ESG and financial performance, and more precisely, which topics are valued the most around the release of the ESG report.

5.1. Implementation of topic modeling

Our sentLDA algorithm takes cleaned documents as input and generates 30 topics as output. Each topic is expressed as a list of weighted words, which can be further analyzed through their cosine (two-word phrases excluding stopwords) similarities. Formally, following Thewissen et al. (2021) :

$$Topic_k = TopicWeigh_k.Word_z$$

$$TopicAssignments_S = K|max \sum_{w=1} TopicWeight_k.Word_w$$

With K being the K topics, Z being a word in the vocabulary, s being a sentence in the ESG report, W being the total number of words in sentences S.

Table 1.: Description of Variables used in this Study

Dependent Variables	**Definition and source**
ESG	Industry adjusted, scored from 0 to 10, retrieved from MSCI.
CAR	Cumulative Abnormal Return on t+5 and t-5
Tobin	A relationship between market valuation and intrinsic value, measuring the market value of a company divided by its assets, on t+1
Independent Variables	**Definition and source**
E	Proportion of sentence from environmental topics divided by the total number of topics' sentences.
S	Proportion of sentence from social topics divided by the total number of topics' sentences.
G	Proportion of sentence from corporate governance topics divided by the total number of topics' sentences.
Control Variables	**Definition and source**
Fin	Amount of debts raised by the firm at time t+0
Volat	We apply a standard deviation on ROA for a period of 5 years.
ROA	Classified by company beforehand, we apply a one year lag for ROA.
Size*	Natural logarithm of assets
New	The asset newness measured as the ratio of net properties, plant, equipment divided by the gross properties, plant and equipment
Capin	Capital intensity, measured as a ratio of capital spending divided by total sales revenues
Lev	The leverage ratio, measured as the total debt divided by the total assets
RD	The research and development intensity measures through their expenses, deflated by sales for year t+0
Tone	Change in tone, which is a proportion of positive minus negative words in ESG report based on the dictionaries provided by Loughran and Mcdonald (2011).
Read	Change in readability, which is the proportion of positive minus negative words in ESG report based on the dictionaries provided by Loughran and Mcdonald (2011).
Proportion*	Logarithm of total number of sentences.

Note : * = Natural logarithmic values used in the linear regression models
Variables' description come from Li et al. (2006); Loughran and Mcdonald (2011); Clarkson et al. (2013)
MSCI = Morgan Stanley Capital International, see https://www.msci.com/

Table 2.: ESG reports by Continents

Years	Total	North America	South Africa	Africa	Oceania	Asia	Europa
2007	30	30	0	0	0	0	0
2008	50	49	0	0	0	1	0
2009	65	63	0	0	0	2	0
2010	95	77	2	0	0	9	7
2011	115	93	2	0	1	10	9
2012	167	137	4	0	1	14	11
2013	195	166	5	0	1	13	10
2014	244	209	5	0	1	14	15
2015	281	237	7	1	2	16	18
2016	350	301	7	1	2	19	20
2017	427	365	9	1	3	23	26
2018	546	477	10	1	4	27	27
2019	188	162	1	0	1	10	14
Total	2,753	2366	52	4	16	158	157

Table 3.: Summary Statistics of Dependent and Control Variables

Statistic	N	Mean	Median	St. Dev.	Min	Max
Dependent Variables						
Tobin	2,547	1.932	1.517	1.321	0.440	13.586
CAR	1,437	0.001	0.001	0.033	−0.093	0.122
ESG	2,720	5.426	5.467	2.214	0.000	10.000
Control Variables						
Fin	2,331	−0.009	−0.011	0.085	−0.849	0.797
Volat	2,751	0.044	0.032	0.043	0.0003	0.342
ROA	2,753	0.050	0.046	0.073	−0.590	0.486
Size	2,753	9.810	9.696	1.687	4.779	14.780
New	2,535	0.503	0.479	0.158	0.054	1.000
Capin	2,574	0.049	0.039	0.042	0.000	0.286
Lev	2,732	0.271	0.256	0.172	0.000	2.439
RD	2,753	0.032	0.000	0.232	0.000	11.538
$\Delta Tone$	2,013	−0.0001	−0.00002	0.006	−0.098	0.092
$\Delta Read$	2,015	0.035	0.083	9.311	−238.503	235.182
Proportion	2,753	6.215	6.332	0.997	0.693	8.510

Note : This table presents the summary statistics (mean, standard deviation, minimum and maximum).

To compensate for the unsupervised condition from our Machine Learning algorithm, we need to quantitatively and qualitatively validate our algorithm output. Measuring the algorithm efficiency quantitatively means we need to extract the perplexity, which reflects the classification precision from our output. The lower perplexity is, the more precise our classification is. Formally, following Nokel and Loukachevitch (2015) :

$$Perplexity(D) = exp(\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} lnp(w|d)) \tag{3}$$

With D the set of documents in our cleaned dataset, n_{dw} the number of occurence of the words w in the document d and $p(w—d)$ the conditional probability of an occurence of w in d.

Measuring the efficiency qualitatively involves a subjective approach through our judgment, which is *ipso facto* one of its limitations. Nonetheless, previous literature heavily supports semantic interpretability of the results over quantitative measures (Thewissen et al., 2021).

Following Lau et al. (2014) and Thewissen et al. (2021), we extract 1000 sentences per topic based on each topic's weighted words. We sort our list by weight and length, then extract the middle tercile as representative sentences of typical length. We then generate 20 most frequent bi-grams from those 334 mid-length sentences. Furthermore, we sort our sentences based on the cosine similarity between sentences. That way, we can now label each topic based upon the bi-grams (see *Appendix 2*), as represented in the word clouds (see *Appendix 1*).

As a robustness test, following Thewissen et al. (2021), we can graphically validate our results by extracting a network graph wherein each weighted line represents the correlation between adjoining topics based on the weights assigned to all the words. An overlap highlights the words they emphasize, which allows us afterward to detect clusters among documents. *Figure 1a* shows that Collaboration and Strategy correlate to each other, same goes by for *Financial* and *Supplychain* or *Community* and *Employee Welfare*. Those topics fall into the same category, as described in *Table 4*, namely *Environment*, *Governance* or *Social*. *Figure 1b* displays the co-occurrence of topics across documents with the size of the nodes representing proportionally the number of sentences assigned to the given topics within the whole dataset. In this instance, we can see how popular topics about *Transparency*, *Climate Change*, *Community* or *Work Safety* are relatively to other.

5.2. Hypothesis 1 – The content of ESG reports

In order to identify the thematic content of ESG reports, we first use the summary statistic table describing various topic variables. We find that the standard deviation around the mean is substantial which indicates a large variability within the informational content from a topic. No topics have been used in each document, however, the high observed maximum transcends a willingness from document authors to highlight a specific topic relative to others, such as *Compliance* which has been used more than 1.8 times for a single document.

The most frequent thematics are *Climate Change* with 62.1 sentences (8.5%), *WorkSafety* with 40 sentences (5.5%), and *Community* with 35.1 sentences (4.8%). Those three topics have also the highest mean in their respective paradigm, *i.e.* E or S or G. The less frequent thematics are *Emission* with 12.2 sentences (1.67%), *Training* with 13.4 sentences (1.8%), and *Compliance* with 14.5 topics (2%). The less frequent thematics are represented in two out of the three pillars since we have two of them in E and one in G. This indicates, overall, a prevalence for the *Social* pillar over the other two pillars.

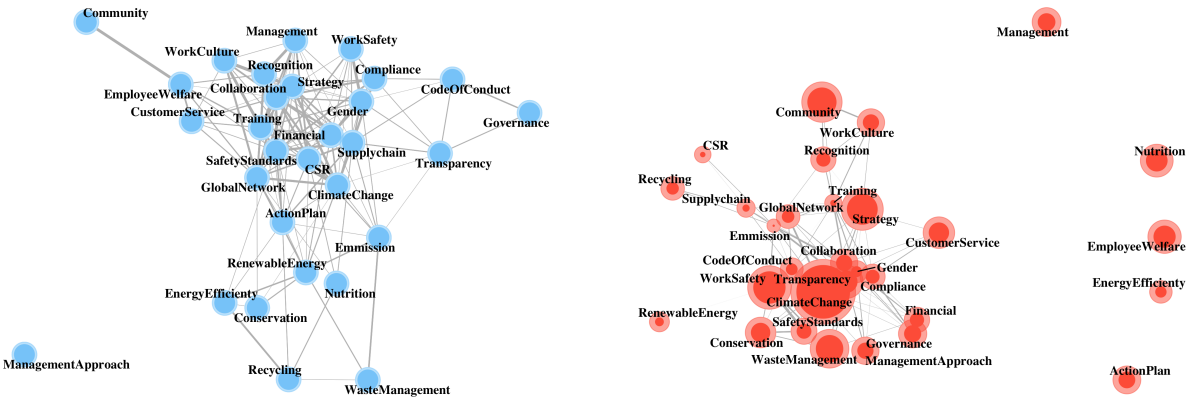
An interesting observation is the non-linearity within the distribution of thematics frequencies across the pillars. We can observe outliers in our sample, such as *Climate change* being 3.5% larger than the second one being 0.7% larger than the third one, although those outliers are not related to each other since they are part of different categories. We observe substantial differences in the distribution of topics across ESG reports over time, as reported in Figure 2. There is a clear increase in the number of sentences used in the ESG reports around the *Social* thematic since 2010, to the detriment *Environment* while *Governance* remains limited. However, we find that this trend has substantially shifted in 2008 and 2009 have seen important, which hasn't been explained by other studies yet, but might be correlated to the market crisis' consequence during 2008.

In summary, most of the important topics echo the current themes we see in the news every day, both positive and negative. We can see a substantial evolution from the *Social* comparatively to the *Environment* and the *Governance* pillars over the years, with a large emphasis on *Work Safety* or *Employee Welfare*. It might transcend a willingness from the management to impact its image to its employees, themselves shareholders through their employee stock options.

Table 4.: Topic description

Statistic	Description	Mean	Median	St. Dev.	Min	Max
Environment						
ClimateChange	Initiatives taken to reduce climate change	62.141	48	55.778	0	467
Supplychain	With a focus on international network	18.118	12	20.067	0	183
EnergyEfficienty	Reduction in energy from factory	15.947	8	22.498	0	239
Compliance	Regulation of companies policy	14.523	0	80.569	0	1,830
RenewableEnergy	Alternative energy sources	18.107	9	28.639	0	415
WasteManagement	Reduction of wastes for raw material	27.956	17	32.074	0	244
Recycling	Reducing consumption of raw material	20.619	11	29.557	0	525
Emmission	Initiatives taken to reduce emission of CO2 within the atmosphere	12.212	1	25.466	0	413
CSR	Companies taking their responsibilities toward society	18.286	1	62.026	0	813
Conservation	Energy conservation thanks to an increase in life cycle	22.301	10	36.531	0	591
Social						
EmployeeWelfare	Intellectually or socially improving employees comfort	29.225	15	40.603	0	334
GlobalNetwork	Impact from a decision on the whole society	21.428	1	69.942	0	1,106
Nutrition	Increasing consumption of local food rather than industrial	26.042	17	29.038	0	278
Collaboration	Increasing collaboration among employees	22.484	2	51.703	0	514
CodeOfConduct	Global standard within the company	18.555	12	21.697	0	206
WorkCulture	Corporate culture, explicitly described	23.694	13	33.236	0	639
SafetyStandards	On health, whether physiological or psychological	21.573	14	26.300	0	270
Gender	Reducing discrimination through fair representation of people with different gender	17.125	8	23.868	0	234
CustomerService	From customer satisfaction to support	27.196	17	30.958	0	345
WorkSafety	Reducing risk of injure at work through higher standard	40.041	26	45.854	0	440
Governance						
Recognition	Whether external or internal, on talent and performance	21.551	6	44.767	0	656
Community	From employees to employee's relatives	35.058	20	42.167	0	515
Strategy	General expectation on short to mid-term	32.462	16	45.485	0	417
Financial	Main financial indicators	22.585	3	54.453	0	568
Management	Hierarchy displayed through objectives	26.947	4	83.395	0	858
Transparency	Measured as the willingness to disclose	28.615	7	61.562	0	688
ManagementApproach	Details relating to management	20.528	1	66.143	0	1,263
ActionPlan	Details relating to strategy	20.199	10	41.601	0	942
Governance	Rights and responsibilities of the company	30.643	20	37.874	0	382
Training	Whether soft skills or hard skills	13.361	8	16.097	0	181

Note : This table presents the summary statistics of topics (description, mean, standard deviation, minimum and maximum).



(a) Word correlation among topics

(b) Co-occurence of topics across documents

Figure 1.: Network graphs on topic interlinkages

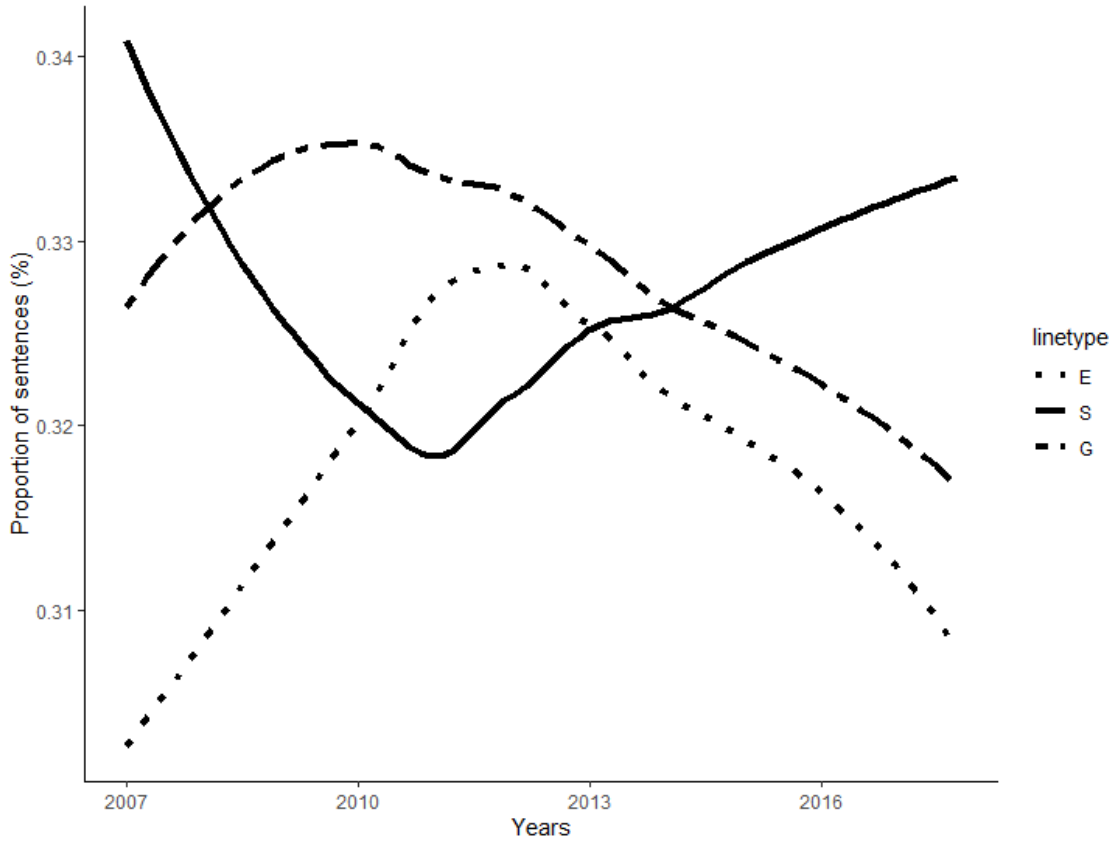


Figure 2.: Evolution of ESG thematic content over time

5.3. Hypothesis 2 – The informational value of ESG reports

5.3.1. The information value to predict future ESG ratings

As stated in "Opening the black box of ESG reports", we ought to test whether the thematic content of ESG reports is informative to predict ESG future and financial performance. Table 5 displays the statistical significance from the multiple linear regressions with *ESG t+1* as the dependent variable. Since the first part of our hypothesis focuses on the ESG future predictability, by taking the ESG score a year ahead, we need to see whether the topics and the aggregated topics improve the R square.

The results in columns one and two display that *Recognition* with a positive coefficient of 0.005, *Community* with a coefficient of 0.006, and *Compliance* with a coefficient of 0.003 are thematics positively impacting and predicting the ESG future performance. On the other side, *Conservation* from the *Environment* pillar is negatively impacting ESG's future with a coefficient of 0.008. Through the Anova, we can see a net improvement in its p-value with a result closes to 0, confirming the first part from our Hypothesis, and the prior literature Li et al. (2006).

We now emphasize our previous findings by regressing on the aggregated topics, namely *Environment, Social, Governance*, as displayed in Table 6. We can see a net improvement in its p-value between the two linear regressions with a result closes to 0. Each of the ESG pillars shares an equally substantial coefficient, respectively 6.5 for *E*, 5.1 for *S*, and 6.9 for *G*.

5.3.2. The information value to predict future firm performance

We can now analyze whether those same topics are equivalent to predict the company's stability in the longer term through Tobin's dependent variable. If we take a look at Table 5 with *Tobin t+1* as the dependent variable, we see that *Renewable Energy* and *Energy Efficiency* are negatively associated while Energy Conservation is positively associated with the prediction of the company's long term success.

The null p-value from the Anova reflects that our model fits our dataset by rejecting the null hypothesis, and therefore, confirms the premise from our second hypothesis that ESG can predict the future financial performance of a company, as previously confirmed by De Lucia et al. (2020).

Let's now confirm our previous findings through Table 6, wherein the three pillars' coefficient from the ESG paradigm are significantly positive, with 1.349 for *E*, 2.521 for *S*, and 1.919 for G. Its 0.002 p-value confirms that there is - comparatively to the one without - a substantial difference with the model including independent variables.

To summarize, we have seen that our independent variables, whether aggregated, can predict the long-term company's stability, which is also consistent with the first sub-question from our second hypothesis. We can see a predominance from the *Social* pillar, confirming our previous finding that companies emphasized substantially that thematic over the years (see *Figure 2*). Furthermore, we can infer that entities investing in each pillar rather than specializing in a single one would perform relatively better than others.

5.4. Hypothesis 3 – The impact of ESG reports publication on investors

As stated in "Opening the black box of ESG reports", our third hypothesis ought to predict the initial investor reaction from the disclosure ESG report by taking the Cumulative Abnormal Return with a window of 10 days.

From Table 5, when regressing on 30 topics, we learn that *Work Safety*, *Safety Standards*, *Collaboration*, and *Work Culture* are the most significant thematics impacting investors' early reactions. Interestingly, those topics are all part of the *Social* pillar.
The 0.010 p-value highlights a statistical difference between our two models, transcending its difference, and *ipso facto*, its usefulness.

From Table 6, we see two topics emerging substantially more than their counter-part, namely the ones on *Environment* and *Social*. Investors favoring those topics reflect a reduction in asymmetry relating to the environment and social trends. This is explained by the lack of information extracted from other mandatory disclosure, emphasizing financial variables rather than ESG variables.

Contrary to prior literature, we do not find a significant association between the change in tone, readability, and the CAR at the ESG release. This could be due to the differences between our sample in Du and Yu (2020) who analyzed 500 reports. Our sample is substantially larger with 2753

ESG reports and covers an international set of firms.[4]

To summarize, our results reveal that the *Social* and *Environment* thematics are positively impacting the investors' initial reactions, with a predominance from the *Social* pillar. Those results are also consistent with the second hypothesis wherein the *Social* pillar was statistically more significant than others to predict the financial and ESG future performances.

6. Additional Analysis

6.1. Hypothesis 4 – Sustainability paradigm

We now analyze whether a company with a balanced distribution of topics across their E, S and G pillars benefit from a higher ESG score.

By using a similar model to the one referenced in the previous section "The impact of ESG reports publication on company's ESG performance", Table 7 displays the results for *ESG* taken a year ahead as the dependent variable, and *Standard Deviation ESG* as the independent variable. As a reminder, the independent variable is computed as the standard deviation from *E, S and G*.

We observe a positive impact from *Standard Deviation ESG* with a negative coefficient of 0.295. We can hereby confirm our additional hypothesis that a more balanced distribution of topics around the sustainable pillars E, S and G favorably impacts the company's ESG performance. Additionally, we have *CAR* and *Tobin* for the robustness test, with a statistically significant regression for both dependent variables, with a coefficient of -0.003 and -0.067 respectively.

6.2. Hypothesis 5 – Energy Sector

The summary table (Table 8) compares main statistics from entities within the Energy Sector and within the Non-Energy Sector. By reducing our sample to the energy sector, following the same classification methodology as Siccode (1998), we regress our *ESG* score variable with the same control and independent variables as our main hypothesis.

[4]We conducted the analysis in Section X on US firms only and do not find evidence of a relationship between Tone and CAR either.

Table 5.: The informational value of ESG reports and its impact on investors' initial reaction

	ESG	ESG	CAR	CAR	Tobin	Tobin
(Intercept)	**5.496*****	**6.511*****	0.028	0.040	**1.928*****	**1.482***
Fin	0.725	0.784	0.018	0.017	**−1.776*****	**−1.741*****
Volat	−3.888	−3.251	0.001	0.029	1.390	1.545
ROA	**2.272****	**2.781*****	0.063	0.075	**4.754*****	**4.726*****
Lev	**−1.316****	**−1.092****	0.001	−0.008	**1.194*****	**1.214*****
Size	**−0.196***	**−0.259****	0.000	0.001	**−0.117****	−0.098
New	−0.311	0.337	−0.010	−0.016	0.451	0.247
Capin	**−6.730****	**−6.307****	0.054	0.032	**−2.728****	−2.026
RD	0.090	0.087	**0.005***	0.004	0.111	0.113
$\Delta Tone$	7.383	7.108	−0.264	−0.150	−2.210	−1.693
$\Delta Read$	0.004	0.000	−0.000	−0.000	−0.002	−0.001
Proportion	**0.395*****	0.117	−0.001	−0.007	−0.052	−0.031
Recognition		**0.005****		0.000		−0.002
Community		**0.006****		−0.000		0.001
EmployeeWelfare		−0.000		0.000		−0.002
Strategy		−0.001		0.000		−0.001
ClimateChange		0.001		0.000		−0.001
Supplychain		0.000		0.000		−0.001
Financial		0.001		0.000		−0.000
EnergyEfficienty		−0.004		−0.000		**−0.006****
Management		0.001		−0.000		0.000
Training		0.006		−0.000		−0.001
Transparency		−0.003		0.000		−0.001
ManagementApproach		0.001		0.000		−0.000
GlobalNetwork		0.002		0.000		−0.000
Compliance		**0.003***		−0.000		−0.000
Nutrition		0.004		−0.000		0.002
RenewableEnergy		0.002		0.000		**−0.003***
WasteManagement		0.001		−0.000		0.001
Collaboration		0.002		0.000		0.001
CodeOfConduct		−0.002		−0.000		−0.000
WorkCulture		0.002		**0.000***		−0.001
Recycling		−0.002		0.000		0.002
Emmission		0.001		−0.000		0.001
CSR		0.001		−0.000		0.001
ActionPlan		0.003		0.000		−0.000
Governance		−0.004		−0.000		0.001
SafetyStandards		−0.005		**0.000****		0.003
Gender		−0.007		−0.000		0.000
CustomerService		0.005		0.000		0.002
Conservation		**−0.008*****		0.000		**0.003***
WorkSafety		0.002		**−0.000****		0.001
Industry and Year	Contr.	Contr.	Contr.	Contr.	Contr.	Contr.
Model Comparison						
R^2	0.512	0.551	0.132	0.184	0.645	0.661
Adj. R^2	0.440	0.474	−0.049	−0.022	0.592	0.602
Num. obs.	1599	1599	1015	1015	1563	1563
ANOVA						
F(p-value)		4.012***(0.000)		1.720**(0.010)		41.546***(0.000)

Note : $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.
Standard errors have been removed for spacing but are available upon request.

Table 6.: The informational value of ESG reports and its impact on investors' initial reaction

	ESG	ESG	CAR	CAR	Tobin	Tobin
Fin	0.725	0.741	0.018	0.019	$\mathbf{-1.776}^{***}$	$\mathbf{-1.791}^{***}$
	(0.676)	(0.676)	(0.021)	(0.021)	(0.361)	(0.361)
Volat	−3.888	$\mathbf{-3.965}^{***}$	0.001	0.003	1.390	$\mathbf{1.388}^{**}$
	(2.581)	(2.521)	(0.053)	(0.051)	(1.459)	(1.449)
ROA	$\mathbf{2.272}^{**}$	$\mathbf{2.318}^{***}$	0.063	$\mathbf{0.057}^{**}$	$\mathbf{4.754}^{***}$	$\mathbf{4.778}^{***}$
	(1.102)	(1.104)	(0.049)	(0.048)	(1.083)	(1.078)
Size	$\mathbf{-0.196}^{*}$	$\mathbf{-0.202}^{***}$	0.000	0.000	$\mathbf{-0.117}^{**}$	$\mathbf{-0.111}^{***}$
	(0.107)	(0.109)	(0.002)	(0.002)	(0.058)	(0.057)
New	−0.311	−0.193	−0.010	−0.012	0.451	0.378
	(0.856)	(0.850)	(0.020)	(0.019)	(0.497)	(0.503)
Capin	$\mathbf{-6.730}^{**}$	$\mathbf{-6.769}^{***}$	0.054	0.065	$\mathbf{-2.728}^{**}$	$\mathbf{-2.903}^{***}$
	(2.816)	(2.783)	(0.085)	(0.083)	(1.385)	(1.415)
Lev	$\mathbf{-1.316}^{**}$	$\mathbf{-1.282}^{***}$	0.001	−0.002	$\mathbf{1.194}^{***}$	$\mathbf{1.187}^{***}$
	(0.607)	(0.611)	(0.011)	(0.011)	(0.271)	(0.267)
RD	0.090	0.082	$\mathbf{0.005}^{*}$	0.005	0.111	0.117
	(0.091)	(0.093)	(0.003)	(0.003)	(0.131)	(0.129)
$\Delta Tone$	7.383	4.983	−0.264	−0.174	−2.210	−0.231
	(4.962)	(5.365)	(0.336)	(0.343)	(3.195)	(3.319)
$\Delta Read$	0.004	0.003	−0.000	−0.000	−0.002	−0.001
	(0.002)	(0.003)	(0.000)	(0.000)	(0.002)	(0.002)
Proportion	$\mathbf{0.395}^{***}$	$\mathbf{0.408}^{***}$	−0.001	−0.000	−0.052	$\mathbf{-0.078}^{***}$
	(0.087)	(0.092)	(0.003)	(0.002)	(0.052)	(0.052)
E		$\mathbf{6.519}^{***}$		$\mathbf{0.067}^{**}$		$\mathbf{1.349}^{***}$
		(1.347)		(0.035)		(0.523)
S		$\mathbf{5.086}^{***}$		$\mathbf{0.097}^{***}$		$\mathbf{2.521}^{***}$
		(1.289)		(0.032)		(0.751)
G		$\mathbf{6.884}^{***}$		0.014		$\mathbf{1.919}^{***}$
		(1.264)		(0.035)		(0.604)
Industry and Year	Contr.	Contr.	Contr.	Contr.	Contr.	Contr.
Model Comparison						
R^2	0.933	0.933	0.132	0.143	0.897	0.898
Adj. R^2	0.923	0.924	−0.050	−0.040	0.881	0.882
Num. obs.	1599	1599	1015	1015	1563	1563
ANOVA						
F(p-value)		$3.442^{**}(0.032)$		$4.933^{***}(0.007)$		$6.299^{***}(0.002)$

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 7.: Sustainability paradigm

	ESG t+1	CAR t+5	Tobin t+1
(Intercept)	**5.626*****	0.024	**1.951*****
	(1.815)	(0.029)	(0.724)
Fin	0.779	0.016	**−1.765*****
	(0.677)	(0.015)	(0.360)
Volat	−4.068	0.010	1.349
	(2.513)	(0.041)	(1.456)
ROA	**2.317****	0.037	**4.760*****
	(1.087)	(0.027)	(1.075)
Lev	**−1.282****	0.003	**1.200*****
	(0.597)	(0.009)	(0.270)
Size	**−0.196***	−0.001	**−0.116****
	(0.105)	(0.002)	(0.058)
New	−0.232	−0.012	0.474
	(0.861)	(0.017)	(0.494)
Capin	**−6.703****	0.047	**−2.723***
	(2.795)	(0.065)	(1.390)
RD	0.070	**0.005****	0.107
	(0.092)	(0.002)	(0.131)
$\Delta Tone$	8.114	−0.273	−2.014
	(4.967)	(0.312)	(3.190)
$\Delta Read$	0.003	−0.000	−0.002
	(0.003)	(0.000)	(0.002)
Proportion	**0.374*****	−0.001	−0.057
	(0.087)	(0.002)	(0.052)
Standard Deviation ESG	**−0.295*****	**−0.003***	**−0.067***
	(0.078)	(0.002)	(0.034)
Industry and Year	Controlled	Controlled	Controlled
Model Comparison			
R^2	0.520	0.155	0.646
Adj. R^2	0.449	−0.022	0.593
Num. obs.	1598	1015	1563
ANOVA			
F(p-value)	24.922***(0.000)	2.756*(0.097)	5.162***(0.023)

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 9 displays our main findings, in which *Financial*, *Management*, *Renewable Energy* and *Safety Standards* are positively correlated with the dependent variable, while *Recognition*, *Nutrition*, *Waste Management* and *Work Culture* are negatively correlated.
We also provide the CAR as a robustness test, but most of its coefficients are closed to insignificant, while the p-values from its Anova are both significant with 0.036 and 0.048 respectively.

To summarize, several topics emerging from our results are echoing current topics trending in the world. On the other hand, investors react to a limited number of topics within the *Environment* dimension since only *Renewable Energy, and Supply Chain* are significant, transcending attention to matters more easily measurable for a stockholder.

Table 8.: Summary table from the Energy Sector, and Non-Energy Sector

Statistic	Median E	Median NE	St. Dev. E	St. Dev. NE	Min E	Min NE	Max E	Max NE
Tobin	1.197	1.543	0.609	1.340	0.556	0.440	5.849	13.586
CAR	0.007	−0.0004	0.039	0.049	−0.120	−0.887	0.235	0.332
ESG	5.442	5.467	1.999	2.224	1.175	0.000	9.300	10.000
Fin	0.011	−0.012	0.069	0.086	−0.175	−0.849	0.260	0.797
Volat	0.030	0.032	0.026	0.044	0.003	0.0003	0.126	0.342
ROA	0.027	0.048	0.044	0.074	−0.194	−0.590	0.168	0.486
Lev	0.310	0.253	0.132	0.173	0.000	0.000	0.703	2.439
Size	9.987	9.691	1.559	1.692	6.249	4.779	11.724	14.780
New	0.664	0.475	0.174	0.155	0.233	0.054	0.908	1.000
Capin	0.054	0.039	0.032	0.043	0.005	0.000	0.203	0.286
RD	0	0	0.036	0.237	0	0	0	12
$\Delta Tone$	0.0002	−0.00003	0.006	0.006	−0.023	−0.098	0.013	0.092
$\Delta Read$	0.177	0.073	3.246	9.494	−9.676	−238.503	18.857	235.182
Proportion	753	557	607.157	588.943	10	1	2,660	4,963

7. Conclusion

Through the performance of our Machine Learning algorithm, our large international dataset of ESG reports from listed entities, and our financial and accounting databases, we were able to tackle multiple challenging issues that have been previously left aside due to its dimensional complexity. To answer those challenges, we made 4 assumptions. One was dealing with topics discussed in the ESG reports, which revealed the predominance of the sustainability paradigm through topics about *Environment*, *Social* and *Governance*.

Our second hypothesis dealt with the ESG and financial future performance of a firm by predicting whether thematic content transcending from the ESG report was impacting a firm's performance. We noticed that both ESG and its financial performance were impacted by each of the three pillars. We also saw a predominance of topics emerging from the *Social* pillar.

Our third hypothesis investigated which thematics were impacting investor's initial reaction when a company releases its ESG report publicly. While topics about *Governance* weren't prevalent, our main findings were consistent with the previous hypotheses, namely that investors were reacting to subjects from the *Social* pillar which are not available in mandatory disclosures.

Our first additional hypothesis dealt with the sustainability paradigm *per se*, with the idea that a balanced distribution of sentences in each pillar would favorably impact a company in its ESG and financial performance. The main result confirmed our hypothesis, which is also consistent with our previous hypotheses.

We also provided another additional analysis, in which we tried to determine whether firms from the Energy sectors were fundamentally different from firms dealing in other industries. Our main results confirmed partly that hypothesis since other thematics were impacting the ESG future performance from an Energy firm.

In conclusion, the predominance of sentences dealing with the *Social* thematic transcends a willingness from entities to support their effort to invest on the human capital, by taking other people's opinion into account. It also allows a company to differentiate itself from its competitors by sharing information on other topics than the traditional 10-K in which they are usually preoccupied with *hard* metrics (accounting data) rather than *soft* metrics (*E, S, G*).

Table 9.: Energy Sector

	ESG t+1	ESG t+1	CAR t+5	CAR t+5
(Intercept)	−1.193	**18.571*****	0.058	−0.243
Fin	3.906	5.664	0.029	−0.056
Volat	5.688	−3.737	**0.419****	−0.036
ROA	11.069	**13.004***	**0.226***	0.147
Lev	**−7.326*****	−3.782	**0.086****	**0.182*****
Size	0.223	−0.123	**−0.009*****	0.006
New	−2.945	**−14.658*****	−0.052	0.117
Capin	**32.109****	25.140	−0.281	−0.089
RD	**36.223*****	**68.272*****	0.011	−0.283
$\Delta Tone$	−29.366	19.240	0.481	0.338
$\Delta Read$	−0.063	**−0.270****	0.000	−0.001
Proportion	0.031	**−2.957*****	−0.001	0.005
Recognition		**−0.035***		0.000
Community		0.003		**0.000***
EmployeeWelfare		−0.002		−0.000
Strategy		−0.019		−0.000
ClimateChange		0.013		−0.000
Supplychain		0.014		**−0.001****
Financial		**0.023***		0.000
EnergyEfficienty		0.054		−0.001
Management		**0.009****		−0.000
Training		0.017		0.000
Transparency		−0.017		**0.000****
ManagementApproach		0.042		−0.001
GlobalNetwork		−0.016		0.001
Compliance		−0.086		0.002
Nutrition		**−0.035***		0.000
RenewableEnergy		**0.054*****		**−0.000***
WasteManagement		**−0.032***		0.000
Collaboration		0.033		0.000
CodeOfConduct		0.055		−0.000
WorkCulture		**−0.088*****		0.000
Recycling		0.042		−0.000
Emmission		−0.026		0.000
CSR		0.026		0.003
ActionPlan		0.036		0.001
Governance		−0.014		**−0.001****
SafetyStandards		**0.073*****		−0.000
Gender		0.004		0.000
CustomerService		−0.020		**0.001*****
Conservation		0.071		−0.000
WorkSafety		0.011		−0.000
Industry and Year	Controlled	Controlled	Controlled	Controlled
Model Comparison				
R^2	0.631	0.923	0.545	0.900
Adj. R^2	0.438	0.687	0.309	0.595
Num. obs.	121	121	121	121
ANOVA				
F(p-value)		2.270**(0.036)		2.130**(0.048)

Note : $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.
Standard errors have been removed for spacing but are available upon request.

References

(2015). Building a sustainable investment management profession. Technical report, CFA.

Allen, A. and K. Ramanna (2013, February). Towards an understanding of the role of standard setters in standard setting. *Journal of Accounting and Economics 55*(1), 66–90.

Amel-Zadeh, A. and G. Serafeim (2018, July). Why and how investors use esg information: evidence from a global survey. *Financial Analysts Journal 74*(3), 87–103.

Arslan-Ayaydin, , N. Bishara, J. Thewissen, and W. Torsin (2020, November). Managerial career concerns and the content of corporate disclosures: An analysis of the tone of earnings press releases. *International Review of Financial Analysis 72*, 101598.

Arslan-Ayaydin, , K. Boudt, and J. Thewissen (2016, November). Managers set the tone: Equity incentives and the tone of earnings press releases. *Journal of Banking & Finance 72*, S132–S147.

Arslan-Ayaydin, , J. Thewissen, and W. Torsin (2021, January). Disclosure tone management and labor unions. *Journal of Business Finance & Accounting 48*(1-2), 102–147.

Bao, Y. and A. Datta (2014, June). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science 60*(6), 1371–1391.

Beyer, A., D. A. Cohen, T. Z. Lys, and B. R. Walther (2010, December). The financial reporting environment: Review of the recent literature. *Journal of Accounting and Economics 50*(2-3), 296–343.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003, March). Latent dirichlet allocation. *The Journal of Machine Learning Research 3*(null), 993–1022.

Boudt, K. and J. Thewissen (2019, March). Jockeying for position in ceo letters: impression management and sentiment analytics: impression management and sentiment analytics. *Financial Management 48*(1), 77–115.

Cavaco, S. and P. Crifo (2014, September). CSR and financial performance: complementarity between environmental, social and business behaviours. *Applied Economics 46*(27), 3323–3338.

Cho, C. H., G. Michelon, D. M. Patten, and R. W. Roberts (2015, January). CSR disclosure: the more things change...? *Accounting, Auditing & Accountability Journal 28*(1), 14–35.

Cho, C. H. and D. M. Patten (2007, October). The role of environmental disclosures as tools of legitimacy: A research note. *Accounting, Organizations and Society 32*(7-8), 639–647.

Clarkson, P. M., X. Fang, Y. Li, and G. Richardson (2013, September). The relevance of environmental disclosures: Are such disclosures incrementally informative? *Journal of Accounting and Public Policy 32*(5), 410–431.

Clarkson, P. M., J. Ponn, G. D. Richardson, F. Rudzicz, A. Tsang, and J. Wang (2020, March). A textual analysis of us corporate social responsibility reports. *Abacus 56*(1), 3–34.

De Lucia, C., P. Pazienza, and M. Bartlett (2020, July). Does good esg lead to better financial performances by firms? Machine learning and logistic regression models of public enterprises in europe. *Sustainability 12*(13), 5317.

Dhaliwal, D. S., O. Z. Li, A. Tsang, and Y. G. Yang (2010). Voluntary nonfinancial disclosure and the cost of equity capital: the initiation of corporate social responsibility reporting. *SSRN Electronic Journal*.

Du, S. and K. Yu (2020, April). Do corporate social responsibility reports convey value relevant information? Evidence from report readability and tone. *Journal of Business Ethics*.

El Assaad, H., A. Samé, G. Govaert, and P. Aknin (2016, November). A variational Expectation–Maximization algorithm for temporal data clustering. *Computational Statistics & Data Analysis 103*, 206–228.

Fatica, S., T. Hemmelgarn, and G. Nicodeme (2013). The debt-equity tax bias: consequences and solutions. *Reflets et perspectives de la vie économique LII*(1), 5.

Gaio, C. and R. Henriques (2020, September). Social responsibility and financial performance: the case of stoxx europe index. In *Corporate Social Responsibility [Working Title]*. IntechOpen.

Grewal, J., C. Hauptmann, and G. Serafeim (2020, February). Material sustainability information and stock price informativeness. *Journal of Business Ethics*.

Grossman, S. J. and O. D. Hart (1980, May). Disclosure laws and takeover bids. *The Journal of Finance 35*(2), 323–334.

Healy, P. M. and K. G. Palepu (2001, September). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics 31*(1-3), 405–440.

Hoberg, G. and C. Lewis (2017, April). Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance 43*, 58–85.

Hummel, K., S. A. Mittelbach-Hoermanseder, C. H. Cho, and D. Matten (2017). Implicit versus explicit corporate social responsibility disclosure: a textual analysis. *SSRN Electronic Journal*.

IFRS (2019). IFRS - European Commission requires electronic tagging of financial statements using the IFRS Taxonomy.

Institute, G. A. (2019). 90% of s&p 500 index companies publish sustainability reports in 2019, g&a announces in its latest annual 2020 flash report.

Ivers, M. (1993, April). *Random House Guide to Good Writing* (Ballantine Books ed.).

Ki-Moon, B. (2015). Sustainable development.

Kim, S.-W. and J.-M. Gil (2019, December). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences 9*(1), 30.

King, A. and B. Wim (2015). Currents Of Change. Technical Report 9th edition, KPMG.

Lang, K. and C. Electris (2018). Global Sustainable Investment Review. Technical report, The Global Sustainable Investment Alliance.

Lang, M. H. and L. Stice-Lawrence (2015, February). Textual analysis and international financial reporting: large sample evidence. SSRN Scholarly Paper ID 2407572, Social Science Research Network, Rochester, NY.

Lau, J. H., D. Newman, and T. Baldwin (2014). Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, pp. 530–539. Association for Computational Linguistics.

Li, Y., P. Clarkson, G. D. Richardson, and F. P. Vasvari (2006). Revisiting the relation between environmental performance and environmental disclosure: an empirical analysis. *SSRN Electronic Journal*.

Loughran, T. and B. Mcdonald (2011, February). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance 66*(1), 35–65.

Loughran, T. and B. Mcdonald (2016, September). Textual analysis in accounting and finance: a survey: textual analysis in accounting and finance. *Journal of Accounting Research 54*(4), 1187–1230.

Margaretha, F. and B. Rachmawati (2016, April). The effect of corporate social responsibility to financial performance of company in indonesia. SSRN Scholarly Paper ID 2776257, Social Science Research Network, Rochester, NY.

Matsumura, E. M., R. Prakash, and S. C. Vera-Muñoz (2014, March). Firm-value effects of carbon emissions and carbon disclosures. *The Accounting Review 89*(2), 695–724.

McKelvey, B. (2002). Managing Coevolutionary Dynamics. Barcelona, Spain.

Milgrom, P. R. (1981). Good news and bad news: representation theorems and applications. *The Bell Journal of Economics 12*(2), 380.

Mishra, S. and S. B. Modi (2013, October). Positive and negative corporate social responsibility, financial leverage, and idiosyncratic risk. *Journal of Business Ethics 117*(2), 431–448.

Muslu, V., S. Mutlu, S. Radhakrishnan, and A. Tsang (2019, February). Corporate social responsibility report narratives and analyst forecast accuracy. *Journal of Business Ethics 154*(4), 1119–1142.

Nazari, J. A., K. Hrazdil, and F. Mahmoudian (2017, August). Assessing social and environmental performance through narrative complexity in CSR reports. *Journal of Contemporary Accounting & Economics 13*(2), 166–178.

Nelson, M. (2020, July). How will ESG performance shape your future? Technical Report Fifth global institutional investor survey, Ernst and Young.

Nokel, M. and N. Loukachevitch (2015). A method of accounting bigrams in topic models. In *Proceedings of the 11th Workshop on Multiword Expressions*, Denver, Colorado, pp. 1–9. Association for Computational Linguistics.

Patten, D. M. (2002, November). The relation between environmental performance and environmental disclosure: a research note. *Accounting, Organizations and Society 27*(8), 763–773.

Perrini, F. and A. Tencati (2006, September). Sustainability and stakeholder management: the need for new corporate performance evaluation and reporting systems. *Business Strategy and the Environment 15*(5), 296–308.

Plumlee, M., D. Brown, R. M. Hayes, and R. S. Marshall (2015, July). Voluntary environmental disclosure quality and firm value: Further evidence. *Journal of Accounting and Public Policy 34*(4), 336–361.

Santillo, D., P. Johnston, M. Everard, and R. Karl-Henrik (2007, January). Reclaiming the definition of sustainability(7 pp). *Environmental Science and Pollution Research - International 14*(1), 60–66.

Shahbaz, M., A. S. Karaman, M. Kilic, and A. Uyar (2020, August). Board attributes, CSR engagement, and corporate performance: What is the nexus in the energy sector? *Energy Policy 143*, 111582.

Shrestha, P., Arslan-Ayaydin, J. Thewissen, and W. Torsin (2021, March). Institutions, regulations and initial coin offerings: An international perspective. *International Review of Economics & Finance 72*, 102–120.

Siccode (1998). SIC Codes for energy.

Strong, R. A. (1986). Using gunning's fog index with term papers and outside reading lists. *Journal of Financial Education* (15), 63–67.

Thewissen, J., T. Wouter, P. Shrestha, and A. M. Pastwa (2021). Unpacking the black box of ICO white papers: A topic modeling approach (Working paper).

Threlfall, R., A. King, J. Shulman, and W. Bartels (2020, December). The time has come: The KPMG Survey of Sustainability Reporting 2020. Technical Report 11th edition, KPMG.

Tschopp, D. and R. J. Huefner (2015, March). Comparing the evolution of csr reporting to that of financial reporting. *Journal of Business Ethics 127*(3), 565–577.

Vitolla, F., N. Raimo, M. Rubino, and A. Garzoni (2019, September). How pressure from stakeholders affects integrated reporting quality. *Corporate Social Responsibility and Environmental Management*, csr.1850.

Vitolla, F., A. Salvi, N. Raimo, F. Petruzzella, and M. Rubino (2020, February). The impact on the cost of equity capital in the effects of integrated reporting quality. *Business Strategy and the Environment 29*(2), 519–529.

Wang, Q., J. Qu, B. Wang, P. Wang, and T. Yang (2019, December). Green technology innovation development in China in 1990–2015. *Science of The Total Environment 696*, 134008.

Wang, Y., B. Li, G. Li, X. Zhu, and J. Li (2019). Risk factors identification and evolution analysis from textual risk disclosures for insurance industry. *Procedia Computer Science 162*, 25–32.

Werther, W. B. and D. Chandler (2005, July). Strategic corporate social responsibility as global brand insurance. *Business Horizons 48*(4), 317–324.

Wu, C.-M. and J.-L. Hu (2019, May). Can CSR reduce stock price crash risk? Evidence from China's energy industry. *Energy Policy 128*, 505–518.

Zhang, C., H. Wang, Y. Liu, D. Wu, Y. Liao, and B. Wang (2008). Automatic keyword extraction from documents using conditional random fields. *4*(3), 1169–1180.

8. Appendix

See associated file.

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