Acknowledgements

I would like to start by thanking Professor Grando, not only for the support and rapid feedback provided during the completion of this work, but also for the inspiring company visits and guests speeches he provided us with throughout the course.

I want to thank my parents and my sister for the unconditional love, support and trust they have given me during these three years, even at a distance. I also want to thank my aunt for hosting me regularly in Milan and for being an inspiration in the field of corporate sustainability.

These past three years have been an incredible journey, and I want to thank all my friends and colleagues who walked alongside me on this path. Bunker study sessions, free breakfasts at gud, bocconi giusti lunches, hackathons, and spritzes in politecnico and statale will not be forgotten.

Are sustainability claims enough? A text-mining based analysis on the relationship between the extent of sustainability reports and ESG performance

Contents

1 Introduction

Over the last decade, the concept of sustainability has grown in relevance in the context of business management. Firms are now under pressure from regulators and by consumers to improve their production standards, both from an environmental and social standpoint.

Since the introduction of Freeman's stakeholder theory (Freeman & Mcvea 2001) an increasing number of companies has started to consider the sustainability dimension in the decision making process, starting to account for the stakeholders welfare as well as the shareholders'.

Around 2008, after the financial crisis, a number of firms started to publish sustainability reports. Initially they were almost completely dedicated to the governance standards implemented in the banking sector, as the scope of the tool was to prove the firm's compliance to investors and stakeholders. Shortly after many players started reporting on other management areas, environmental and social impact, to inform the stakeholders of the decisions taken.

As of 2023, sustainability reporting has become a standard in most industries. Most developed and developing countries started requiring large firms to report on the ESG practices in place (KPMG 2022). For instance the EU, in 2022 passed the Non Financial Reporting Directive that now requires 11.700 companies to publish a sustainability report (EU 2023).

One of the drawbacks of sustainability reporting is the tendency to overstate the effects of the actions taken by the firm. Overstatement and exaggeration in sustainability reports can cause discrepancies between the disclosure and actual performance. Even though sustainability is becoming increasingly more important, in most cases, it is not management's top priority. Sustainability reporting and CSR communication can function as a signaling tool (Uyar et al. 2020) that allows to achieve a lower cost of equity (Dhaliwal et al. 2011). This works as an incentive for firms to overstate the sustainability results.

However, discrepancies between CSR communication and performance, even more the inability to deliver on the claims made can be extremely harmful for businesses and result in backlash, damaging brand's reputation and stakeholder's trust (Rahman et al. 2015).

Identifying discrepancies starting from the content of the sustainability reports could be a useful solution for investors, regulators and managers. Therefore, the scope of this study is to further investigate the relationship between sustainability reports extent and the actual performance of the firm.

2 Literature Review

In the past decade, corporate social responsibility has become an increasingly pressing area of concern for corporations worldwide (KPMG 2022) and effective communication with stakeholders is a crucial component of corporate social responsibility.
The purpose of this chapter is to investigate the current research on the topic of sustainability reporting with a focus on what sustainability reporting is and its importance, the current EU and Italian legal framework related to sustainability reporting, and how text-analysis models can be leveraged to extract useful information from sustainability reports.

2.1 Sustainability reporting

The practice of disclosing ESG information finds its roots in Freeman's stakeholder's theory, which states that a firm shaping its policies and decisions should not only take the shareholders into account, but also consider every group or individual with an interest (Freeman & Mcvea 2001).
The first corporations to disclose on sustainability were banks after the 2008 crisis, using them as a way to build back the trust that was lost to the crisis, explaining and publishing the sustainability in governance practices. Although the publication of sustainability reports was initially voluntary, it has now shifted to a more strict and regulated setting. In both the EU and the US, large companies are required to disclose their sustainability practices (Baldissera 2023).

Ioannou & Serafim (2011) and Wang et al. (2018) are attempts at analyzing the effects of mandatory CSR reporting. The studies found a positive link between firm transparency and disclosure quantity and quality. These studies also found a positive relation between disclosure quantity and quality and a firm's financial results.

2.2 EU legal framework and reporting standards

In the European Union the non-financial reporting directive is the main piece of legislation on the regulating sustainability disclosure. It regulates the sustainability reporting of 11.700 large companies (EU 2023). The directive requires large companies to disclose information on: environmental matters, social matters and treatment of employees, respect for human rights, anti-corruption and bribery, diversity on company boards.

In January 2023 a new piece of legislation was added to expand the scope of the rules on sustainability reporting to large companies and listed SMEs. The new directive now requires approximately 50.000 companies to report on sustainability. The first companies will have to apply the new rules for the first time in the 2024 financial year, for reports published in 2025 (EU 2023).

The Italian government adopted the directive with the decree n. 254 on the 30th of December 2016. The decree states that companies with more than 500 employees will be required to produce and publish sustainability reports. The thematic areas that require disclosure are: environmental impact, social problems, human rights, corruption, and money laundering. At the current state the decree does not force to use a specific reporting standard nor a sector specific reporting standard (Borsa-Italiana 2016).

There is a growing amount of literature on the effect of compulsory disclosure for SMEs and smaller companies is yet unknown. Recent studies on the effect of NFD found that the directive had a positive impact on sustainability reporting quality (Mion & Loza Adaui 2019). Other studies found NFD to positively increase the amount of CSR activities firms with low CSR level, however no evidence was found to support the hypothesis that mandatory disclosure corresponds to a decrease in irresponsible actions (Jackson et al. 2020).

2.3 GRI - Sustainability reporting standards

The Global Reporting Initiative (GRI) provides guidelines and a framework for sustainability reporting. starting from 2000, the organization has been developing a modular set of Standards for sustainability reporting, such as the universal standards, the sector standards and topic standards.

The GRI framework had a positive impact on sustainability disclosure quality (SDQ), which according to recent literature has a positive effect on the quality of the earnings as well as the market value (Pereira et al. 2023) (Nwaigwe et al. 2022).

As found by both Clarkson et al. (2008) and Cormier & Magnan (2007) following guidelines likely improves a firm's sustainability performance, as they found a positive relationship between sustainability disclosure and actual performance.

Moreover, recent literature shows that following guidelines allows stakeholders to obtain a more thorough view of the company's impact (Papoutsi & Sodhi 2020). This practice also plays an important role in reducing asymmetry of information among shareholders and managers on sustainability performance (Isaksson & Steimle 2009) (Manetti 2011).

2.4 ESG Ratings

ESG ratings are a measure of a company's environmental, social, and governance performance. There is a growing interest in ESG ratings, and it is important to understand how they are calculated and what factors are taken into account.

The environmental ratings evaluate the performance of a company in terms of energy efficiency, greenhouse gas emissions, waste of water and other resource. There is already a large amount of literature investigating and trying to define the relationship between environmental and financial performance. One of the main studies in the field is Manrique & Martí-Ballester (2017) found that more eco-friendly firms were benefiting from higher stock returns than similar non-eco-friendly companies.

The social ratings attempt to measure all the policies and actions that directly and indirectly affect employees' satisfaction, as well as individuals, groups, and society. It is an important pillar as, according to Edmans (2011), there exists a clear positive relationship between employee's satisfaction and long-run stock return. Other studies, Cho et al. (2019), found a positive relationship between the individuals, groups, and society satisfaction and firm value. The governance ratings are related to aspects such as the independence of the board of administration, shareholders' rights, managers' remuneration, control procedures and anti-competitive practices, as well as the respect of the law. The idea that stronger governance practices have a positive impact on companies' profitability is supported by a large number of studies, the most notable ones are Gompers et al. (2003) and Velte (2017). Where similar results were found, with Velte focusing on the German market and Gompers on the US one.

Although ESG ratings seem to be an efficient way to measure the sustainability performance, in practice they present some shortcomings. One of the most evident ones is the variety of ESG rating systems. Recent studies have shown that lack of common metrics among rating agencies and heterogeneity in judgment can lead agencies to assign even opposite ratings to a given company (Billio et al. 2021) (Dorfleitner et al. 2015).

2.5 Sustainability reports impact on ESG ratings

There is a growing amount of literature analyzing the impact of sustainability reporting quality and extent of their coverage over various financial and sustainability. However, given the complexity of the data, results often cannot be generalized and there is still room for improvement.

Cheng et al. (2014) found that higher sustainability disclosure quality is related to higher ESG performance scores, using Asset4 as a proxy. Similarly, Gao et al. (2016) finds that firms with better CSR performance tend to provide higher quality CSR disclosures, given the greater amount of resources and stronger corporate governance. Another piece of literature on the matter is Papoutsi & Sodhi (2020) that focuses on finding the impact of the level of disclosure on actual sustainability performance. A significant positive link was found between the quality of the information disclosed and third-party ratings. This paper's main limitations were the use of a scale based on qualitative data and the size of the sample dataset.

Other related studies in terms of methodology, are Nwaigwe et al. (2022) (2022) and Kiriu & Nozaki (2020). Nwaigwe focuses on Sustainability Disclosure Quality impact on market value. The study analyses data coming from 39 Nigerian companies, from 2010 to 2019. The study found that financial results worsen as firms disclose more, establishing a negative link between sustainability disclosure and firm value.

While Kiriu & Nozaki (2020) investigate the relationship between qualitative data from sustainability reports and firm value in the Japanese market. This approach chosen for the analysis is innovative as it leverages text mining. Which allows for a more uniform and less subjective evaluation of the data extracted. The study found a link between the two variables, however further studies are needed to establish the nature of the relationship.

Starting from the literature investigated above, we can see that a number of studies already investigate the content of sustainability reporting (Papoutsi & Sodhi 2020) (Nwaigwe et al. 2022). Nevertheless, there is still room for improvement and a possible way to advance our knowledge in this field could be by applying Natural Language Processing (NLP) techniques to the research on sustainability reporting. This addition could help in removing subjectivity in the evaluation from the studies mentioned above and introduce a standardized way to evaluate reports.

Kiriu & Nozaki (2020) is a first attempt at applying text mining to sustainability reporting; however, the study implemented a word matching scoring system, which is a less accurate technique when compared with the results of other language models such as GPT or BERT.

Gathering from the literature that was analysed, we can hypothesize the following:

H1 – CSR disclosure is associated with ESG performance

H2 - The extent of disclosure, meaning the quantity and variety of disclosure, is positively related to ESG performance.

3 Methodology

3.1 Research Question

As shown in the previous chapter, there is a number of studies focusing on the effects of sustainability disclosure quality and quantity. However, given the as the complexity and difficulties in interpreting the large amount of data in sustainability reports, results are not certain and in some cases are exposed to a high degree of subjectivity.

One of the main limitations reported in these studies, apart from the reduced sample size, is the methodology used. Many of the studies rely on a manual evaluation of the data or, in other cases, on an exact word match. Which respectively expose the analysis to a degree of subjectivity and possible errors.

The proposal of this study is to follow a similar structure as the one used by Kiriu & Nozaki (2020), by using text-mining to investigate the impact of Sustainability Disclosure Quantity on a firm's actual sustainability performance. To be more precise the scope of the study will be to investigate the relationship between the number and type of topics mentioned in the reports and actual ESG performance.

The focus of the study will be the 20 Italian largest listed companies, which have been publishing reports since 2017. For each company, five sustainability reports from 2017 to 2021 and corresponding ESG ratings will be analysed. The analysis will be carried out on the English version of the sustainability reports for simplicity.

3.2 Variables

The dependent variable of this study will be the actual ESG performance of a firm. To give a quantitative measure 3 proxies will be used, one for each sustainability area. The data will be gathered from the Bloomberg terminal.

The independent variables will be the number of mentions of each of the topic found by the BERTopic model. The control variables chosen are the revenues of the company and two dummy variables for the different industry to account for the lower ratings of the banking and utilities sector. Another control that is going to be considered is amount of sentences labeled in each document.

3.3 Data Sources

In the context of this study, it is important that the dependent variable comes from a third party. Since Bloomberg's ESG ratings have been designed to be objective and independent, we can assume that they are third-party data and therefore use them as a proxy of actual ESG performance. Sustainability reports will be downloaded from each company's website sustainability page and transformed into plain text using python library "PDFPlumber".

3.4 Data Processing Steps

1. Gathering sustainability reports from the major listed Italian companies, which have been publishing sustainability reports on a voluntary basis from 2017 to 2021.

2. Fetch ESG scores from Bloomberg terminal

3. Converting PDFs to plain text using 'pdfplumber' python library

4. Parsing the text and cleaning from special characters

5. Using sentence-transformer tokenizer to divide text into sentences, creating a Pandas DataFrame to store each sentence and the name of the document from which it was

extracted

6. Using the same process to tokenize text, we create embeddings with sentence-transformers
7. Using countvectorizer to transform each sentence into a vector containing 300 values from 0 to 1.
8. Using UMAP library to reduce the number of dimensions from 300 to 15 for the actual analysis and to 3 dimensions for data visualization, to visually check the relation between the data points.
9. Using HDBSCAN for fast data clustering
10. Finally, using BERTopic for labeling data and extract topic labels, removing the ones with low certainty.
11. As some of the topics recognized do not provide useful information we can remove them, for instance one of the clusters is related to page number.
12. We keep only the most frequent topics generated by BERTopic and remove those that contain noise.
13. We store these results in a DataFrame containing both the by the total number of sentences per topic and the total number of labeled sentences in each document.
14. As expected, we obtain a DataFrame object with on each row the sustainability report for a company in a given year, with as many columns as the topics found plus the control variables.
15. We implement OLS and Ridge regression to investigate the relationship.

3.5 Libraries

3.5.1 PDFplumber

PDF plumber is a python library for object extraction from PDF files. It allows to extract detailed information about each char, rectangle, line and easily extract text and tables.
It enables us to easily extract text from sustainability reports and remove non relevant objects such as images and lines.
Sustainability reports often contain images, characters and hyperlinks that are difficult to parse using a library such as the one selected for the study, but, as sustainability reports are long and heavy documents, the solution identified offers an optimal level of both performance and low computation time.

3.5.2 Sentence-Transformer

A Transformer is a model architecture that eschews recurrence and instead relies entirely on an attention mechanism to draw global dependencies between input and output (Vaswani et al. 2017). Attention Mechanisms are components used in neural networks to model long-range interaction, for example across a text in NLP. The key idea is to build shortcuts between a context vector and the input, to allow a model to attend to different parts (Vaswani et al. 2017). Sentence-transformers is a Python library that enables to create sentence, text, and image embeddings. Embeddings are real-valued vectors that can be referred to as the representation of a word. Embeddings can be easily used for common tasks like semantic text similarity, semantic search, and paraphrase mining.

3.5.3 UMAP: Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection is a dimension reduction technique that can be used for visualisation, but also for general non-linear dimension reduction (*How UMAP Works* n.d.).
UMAP can be used as an effective preprocessing step to boost the performance of density

based clustering.

3.5.4 HDBSCAN

HDBSCAN is a clustering algorithm developed by Campello et al. (2013). It extends DBSCAN by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering based in the stability of clusters.

3.5.5 BERTopic

BERTopic is a topic modeling technique that leverages embedding models and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions (Grootendorst 2022*b*).

BERTopic generates document embedding with pre-trained transformer-based language models, clusters these embeddings, and finally, generates topic representations with the class-based TF-IDF procedure. BERTopic generates coherent topics and remains competitive across a variety of benchmarks involving classical models and those that follow the more recent clustering approach of topic modeling (Grootendorst 2022*a*).

4 Empirical Study

4.1 Data Preparation: Bloomberg ESG ratings

As mentioned in the Variables section of the methodology, Bloomberg ESG data will be used as a proxy for actual ESG performance. The data gathered from Bloomberg Terminal have been organized in a CSV file:

	name	year	e	s	g
0	prysmian	2017	4.40	1.45	5.68
1	prysmian	2018	5.95	1.45	6.34
2	prysmian	2019	5.62	1.45	6.81
3	prysmian	2020	5.79	1.45	7.63
4	prysmian	2021	6.06	1.45	6.95
5	unipol	2017	0.28	1.43	5.73
6	unipol	2018	0.64	1.78	5.63

Figure 1: First 7 rows of ESGscores.csv

4.2 Data Preparation: Sustainability reports metrics

The sustainability reports, in PDF format, have been downloaded manually from the sustainability page of each company's website. Only sustainability reports from 2017 to 2021 will be analysed.

4.2.1 Word processing and cleaning

To extract the text from each PDF file, a python library named "PDFplumber" has been used. As mentioned in section 4.5.1 this library allows to extract all characters from a PDF page.

The process used to obtain all the text within a PDF file consists in extracting the text using the aforementioned library, after this step number and special characters can be removed, we then use spaCy to divide the collected text into sentences, leveraging the accuracy of the "en-core-web-sm" model. After having divided the text into sentences we can start cleaning the text. We start by changing all text to lowercase form and removing new lines and indentations. As the next step we remove the remaining punctuation and stopwords.

For each of the PDFs we append all the cleaned sentences that are longer than 75 characters to a CSV file, along with the name of the company and the year of the report from which the sentence was extracted.

After repeating this process for all the sustainability reports selected we are left with a CSV file with approximately 92.000 sentences.

text	name	year
order guarantee maximum transparency market re...	unicredit	2021
social respect human rights fight active corru...	unicredit	2021
well principal generated perceived risks relat...	unicredit	2021
signed principles associated indicators refle...	unicredit	2021
full report disclosing social impacts group co...	unicredit	2021
materiality analysis updated directors togethe...	unicredit	2021

Figure 2: First rows of dataset.csv

4.2.2 Topic Modeling

BERTopic, the model selected to assign a specific topic label to each sentence, works on pre-processed data.

To prepare the data we will use a vectorizer to convert text documents in a matrix of token counts, an Embedding model to transform words into real-valued vectors, the UMAP library to reduce the dimensionality of the vectors, and HDBSCAN for efficient clustering.

Vectorizer The vectorizer chosen is CountVectorizer from sklearn feature extraction library. It allows to choose how many combination of words should be explored and it groups them accordingly. The parameters chosen for the vectorizer are the following:

```
vectorizer_model = CountVectorizer(ngram_range=(1, 3), stop_words="english")
```

Figure 3: CountVectorizer function

Embedding Model For creating sentence embeddings, the Sentence-Transformer model will be used. As mentioned in the methodology, embeddings are real-valued vectors that can be referred to as the representation of a word.

The sentence-transformers model chosen is ”all-MiniLM-L6-v2”. It maps sentences and paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search. For reduced computation time the model runs on the GPU, using PyTorch.

```
embedding_model = SentenceTransformer('all-MiniLM-L6-v2', device=torch.device('mps'))
```

Figure 4: Sentence-Transformer Model

UMAP As explained in section 4.5.3, UMAP is used for dimensionality reduction. In this case we will be reducing each of the embeddings from 384 to 20 dimensions. The model is configured as follows:

```
umap_model = umap.UMAP(n_neighbors=100, n_components=20, min_dist=0.075)
```

Figure 5: UMAP configuration

HDBSCAN For the clustering model HDBSCAN will be used. The "euclidean" metric has been chosen in favor of other methods as it resulted to be faster and performed better. The clustering model was implemented with the following configuration:

```
hdbscan_model = hdbscan.HDBSCAN(min_cluster_size=80, min_samples=40, metric='euclidean',
                                prediction_data=True, gen_min_span_tree=True)
```

Figure 6: HDBSCAN configuration

Now that all the preprocessing models are set up we can integrate them all in the BERTopic model.

4.2.3 Topic Selection

The model created in the previous chapter will output the 20 most occurring topics and it assigns a topic label and a probability to each sentence. After training the model with the data gathered in the previous section we obtain a vector of labels and probabilities with the

```python
# Train the BERTopic model
topic_model = BERTopic(
    vectorizer_model=vectorizer_model,
    embedding_model=embedding_model,
    umap_model=umap_model,
    hdbscan_model=hdbscan_model,
    nr_topics=20,
    min_topic_size=10,
    calculate_probabilities=True,
    verbose = True
)
```

Figure 7: BERTopic model

same length of the training dataset.

	text	name	year	labels	prob
	assurance cover information required article a...	unicredit	2021	4	0.095273
	order guarantee maximum transparency market re...	unicredit	2021	-1	0.136838
	social respect human rights fight active corru...	unicredit	2021	-1	0.053489
	well principal generated perceived risks relat...	unicredit	2021	2	0.116600
	signed principles associated indicators refle...	unicredit	2021	-1	0.038178

Figure 8: Prediction results

Now, the non-labeled sentences, non relevant topics and the ones that contain noise can be manually removed, for instance topics such as ”7-tyre-tyres-cars” and ”-1-group-financial-management’.

For the scope of the study all sentences with probability lower than 0.3 were removed. The resulting DataFrame contains around 11.000 labeled sentences for 9 different topics.

To use these results as a metric to measure the extent of sustainability reporting, the entries

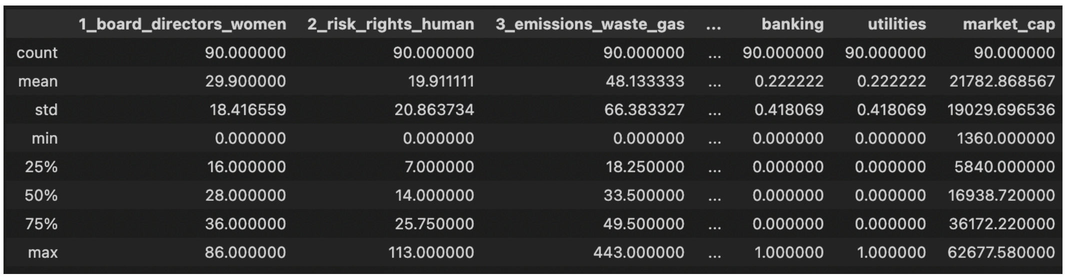
	1_board_directors_women	2_risk_rights_human	3_emissions_waste_gas	...	banking	utilities	market_cap
0	17	7	48	...	1	0	36172.22
1	29	2	35	...	1	0	36172.22
2	30	7	35	...	1	0	36172.22
3	31	13	50	...	1	0	36172.22
4	32	9	36	...	1	0	36172.22
...	...	...	...	...	...	...	...

Figure 9: data correlation

in the DataFrame have been grouped and counted based on company name and year of the document from which they were extracted.

After selecting and narrowing down the number of topics, we obtain a dataset containing the number of mentions of the 9 selected topics for each of the 100 files analysed.

4.2.4 Data quality and Data visualization



	1_board_directors_women	2_risk_rights_human	3_emissions_waste_gas	...	banking	utilities	market_cap
count	90.000000	90.000000	90.000000	...	90.000000	90.000000	90.000000
mean	29.900000	19.911111	48.133333	...	0.222222	0.222222	21782.868567
std	18.416559	20.863734	66.383327	...	0.418069	0.418069	19029.696536
min	0.000000	0.000000	0.000000	...	0.000000	0.000000	1360.000000
25%	16.000000	7.000000	18.250000	...	0.000000	0.000000	5840.000000
50%	28.000000	14.000000	33.500000	...	0.000000	0.000000	16938.720000
75%	36.000000	25.750000	49.500000	...	0.000000	0.000000	36172.220000
max	86.000000	113.000000	443.000000	...	1.000000	1.000000	62677.580000

Figure 10: data distribution

After consolidating the ESG ratings and the metrics obtained from the previous analysis on a single dataset, we can observe that topic labeling on such a small set of firms has clearly some flaws. One in particular is that some of the topics are industry or company specific and in some cases we have a low occurrence of topics. For instance, the 75th percentile for topic 5 and 9 is close to one occurrence per document.

Moreover, we can observe from figure 10 that the data present high level of correlation correlated and this is going to shape the decision of the model used later on.

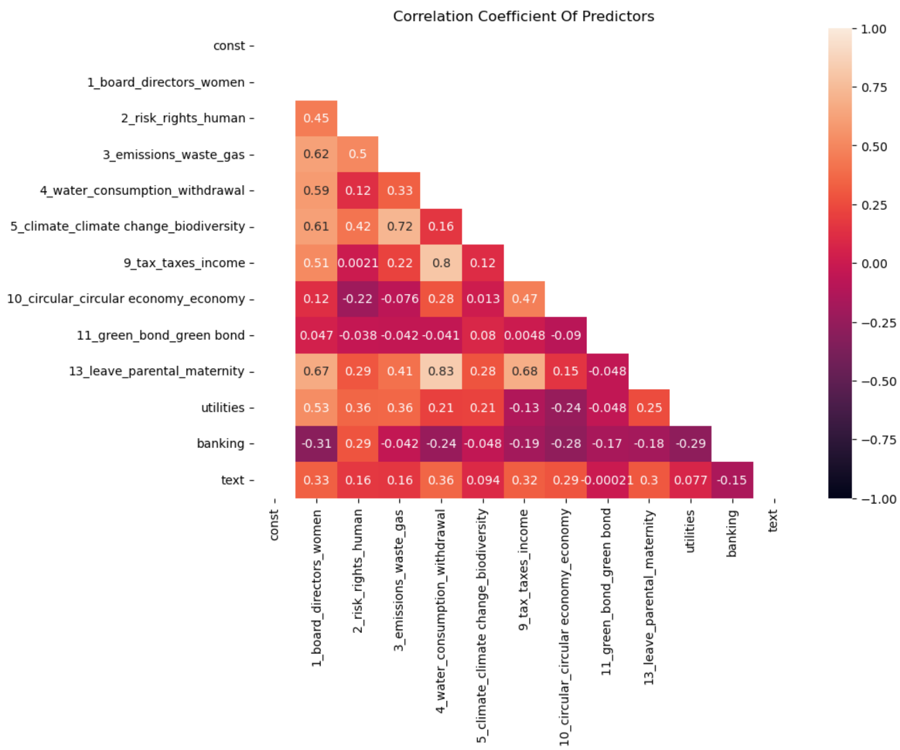


Figure 11: data correlation

To better understand the distribution of the datapoints presented, using UMAP we can reduce the dimensionality from 10 to 3, and display the results using plotly, this will help in visualizing the similarity of the data points. It can be observed that the reports coming from the same company are mostly distributed in the same area, as one could expect given the similarity in the topics and in the level of disclosure.

Another key information that can be extracted is the distribution by industry, it can be seen that companies operating in the utilities and banking sectors, are respectively located in roughly the same areas of the plot.

4.3 Models

To explore the relationship between the metrics extracted from the sustainability reports and the ESG performance, two main models were chosen, Ordinary Least Squares, and Ridge Regression.

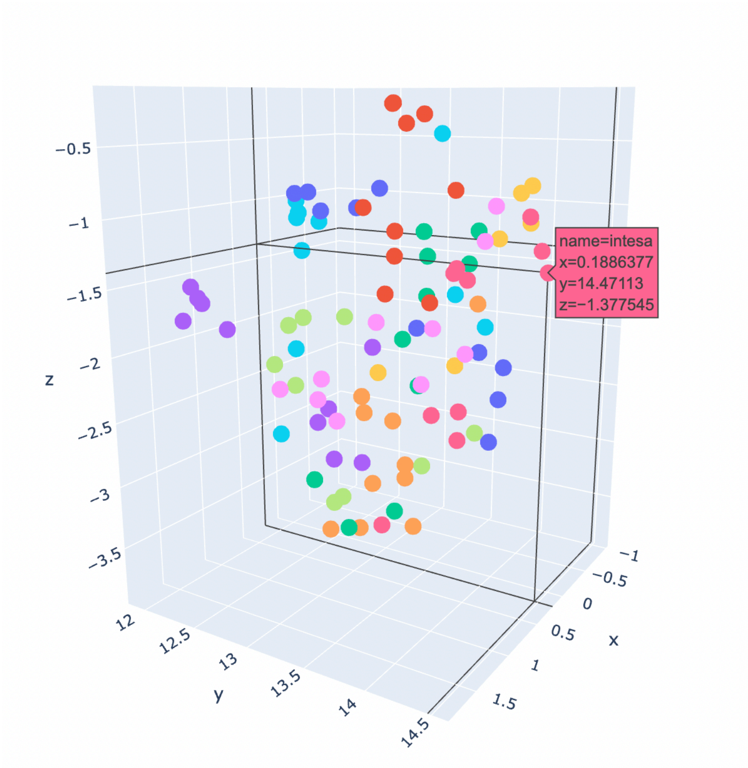


Figure 12: 3D plot of sustainability reports metrics

4.3.1 Ordinary Least Squares

OLS is a simple and easy to implement technique that can help in finding a linear relationship between independent and target variables. If the model is found to be significant it can be a useful way to gain insights on the the magnitude and individual effects of predictors.

When implementing OLS, however, we are making strong assumptions of linearity, independence and homoskedasticity that in more complex settings such as the subject of the study can be violated.

To implement the model we are going to use scikit-learn's OLS. After standardizing the data, we create three different models, each time changing the target variable, to estimate the impact on Environmental, Social and Governance ratings.

As observed in the previous section the data gathered are strongly correlated, which threatens the overall robustness of the model. For this reason we are going to implement a Ridge

```python
#get data & set const term
x = dataset[labels]
y = dataset['e']
x = sm.add_constant(x)

#fit linear regression model
model = sm.OLS(y, x).fit()

#view model summary
print(model.summary())
```

Figure 13: python code snippet for OLS with Environmental target variable

Regression as well.

4.3.2 Ridge Regression with Cross-Validation

The rationale behind the choice of this model is its ability to deal with multicollinear data while maintaining the ability of providing information on the magnitude of the impact of the predictors. Given the reduced size of the dataset at disposal we are going to implement cross validation which helps in avoiding overfitting. With ridge regression we don't need to evaluate the single coefficients and their significance as with do with OLS. For this reason we focus on the magnitude of the coefficients and the overall performance of the model, rather than implementing a traditional significance test. To evaluate the model we rely on R^2 Mean Squared Error (MSE), and Mean Absolute Error (MAE). The ridge regression with cross-valuation is implemented with python using scikit-learn's RidgeCV module.

```python
ridge = RidgeCV(alphas = [0.0001, 0.001,0.01, 0.1, 1, 10])
ridge.fit(x, y)

y_pred = ridge.predict(x)
print("R^2: "+str(r2_score(y,y_pred)))
print("MSE: "+str(mean_squared_error(y,y_pred)))
print("MAE: "+str(mean_absolute_error(y,y_pred)))
```

Figure 14: python code snippet for Ridge implementation

5 Results

Tables 1 and 2 contain the main results of the two models used. Starting from table 1, we can observe the OLS model results. The three models are significant as the F-statistic is smaller than 0.05. Moreover, the models yields good Adjusted R-squared scores and, as it can be observed in the appendix, some of the regressors are significant. Although the model has a relatively good fit to the data, the high condition number shows that there are serious multicollinearity issues, that make the result non robust.

Model	R^2	Adj. R^2	F-statistic	Prob	Cond. No.	Significant Regressors
Model (e)	0.574	0.510	8.924	2.92e-11	5.31e+05	5_climate change_biodiversity, 10_circular_economy, 11_green_bond , 13_leave_maternity,
Model (s)	0.570	0.505	8.771	4.30e-11	5.31e+05	4_water_consumption, 9_tax_income, 10_circular economy, 11_green_bond
Model (g)	0.469	0.389	5.847	1.34e-07	5.31e+05	10_circular_economy

Table 1: Summary of OLS Models

To solve these issues we use ridge regression, whose results are shown in table 2. We can see that the R-squared results are slightly lower, however the nature of the ridge regression allows us to interpret and make inferences using the sign and magnitude of the regressors' coefficients even when there is high multicollinearity.

Target Variable	R-squared (R^2)	MSE	MAE
Model (e)	0.571	2.011	1.033
Model (s)	0.567	1.514	1.013
Model (g)	0.468	0.215	0.360

Table 2: Summary of Ridge Regression Models

As it can be observed in the appendix, the ridge regression offers insights on the individual effects of the predictors.

Ridge results: E Ratings The model used is able to explain 57 percent of the variance in the data. The study finds a positive relationship between the number of mentions of topics such as water consumption, biodiversity, taxes and green bonds. Both banking and utilities companies have a negative relation, probably given by the nature of the sectors. Moreover, the total number of sentences in the document appears to be negatively related.

Ridge results: S Ratings With the social target variable the model explains 56 percent of the variance in the data. A positive link was found between the S rating and the number of mentions of topics such as women board directors, taxes and income taxes, circular economy while surprisingly a slightly negative coefficient was found for parental and maternity leave, but also for the human rights risk topic. Utilities companies have higher social scores while banking coefficient of 0.16 shows that the difference in the social ratings for banks is relatively low. The social target variable is positively affected by the quantity of reporting.

Ridge results: G Ratings Compared to the other models, the one with Governance rating as target yields a lower R-squared value, explaining 46 percent of the variance.
Most of the coefficients have a value smaller than 1, that indicates a small effect on the dependent variable. The main predictors that are positively linked are climate change and maternity leave topics, while those negatively related are waste and emissions and circular

economy. Banking and Utilities companies were found to have slightly lower ratings. The total number of sentences is positively linked with the Governance rating.

6 Conclusions and Discussion

The empirical study conducted focuses on the impact of the extent of disclosure of sustainability reporting on the actual ESG performance of the firm. The analysis was performed on 100 observation coming form 20 Italian publicly listed companies, on a time span of five years (from 2017 to 2021). The main finding of the study was a positive link between the extent of disclosure in certain thematic areas and the actual performance, as it was initially hypothesized. Moreover, with regards to hypothesis 1, the quantity of disclosure appears to have a positive effect on both the social and governance ratings, but it has a negative effect on the environmental one.

The results were partly in contrast with the findings of (Kiriu & Nozaki 2020), as the social and governance ratings were found to have a positive link, while the Environmental ratings were negatively affected by the total quantity of disclosure. However, going more into detail we can see that the some of the topics, such as water waste, climate change and biodiversity had a positive effect. Thus, we can say that the results of the study confirm Kiriu & Nozaki (2020) findings and improve current research by providing information on the effects of single topics and yielding higher R2 scores. Some of the coefficients may be highlighting possible discrepancies, especially those of human rights in relation to the social rating and the emission and waste topic in relation to the environmental one. However, in order to make any claims on this relationship, further analysis and data are needed. In conclusion, the results of the study are relevant for performance evaluation, investor decision making, stakeholder engagement, risk management and regulatory compliance. A possible future development of the study could be the detection of green-washing through the discrepancies between the reported performance and the actual performance.

6.1 Limitations

One of the main limitations of the study was the sample size, only 20 Italian companies were subject to the study, from a multitude of sectors. This sample was chosen for two main reasons, the reduced sample sized allowed fast and efficient computation as the process to extract text from PDFs requires a fair amount of computational power and the focus on the Italian public companies guaranteed the presence of at least 5 sustainability reports from 2017 to 2021.

Another limitation that was encountered is that the accuracy of BERTopic is lower compared to other language models, such as GPT-3.5 Turbo and GPT-4, therefore a study using these models would probably lead to better sentence extraction and classification. However, the application of these powerful models would come at a cost, as the APIs offered by OpenAI are offered as pay-as-you-go service. Considering the scale of the analysis would be .

An additional limitation can be found in the assumption that the performance can be measured by the use of a proxy such as the ESG ratings, as they are known for being an imperfect metric, with scores varying significantly between different rating systems.

Thus, the suggestion for future studies on the topic is to use a bigger and sector specific sample, but also to use a more powerful and accurate model for classification.

7 Appendix

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      e   R-squared:                       0.574
Model:                            OLS   Adj. R-squared:                  0.510
Method:                 Least Squares   F-statistic:                     8.924
Date:                Sat, 03 Jun 2023   Prob (F-statistic):           2.92e-11
Time:                        10:50:16   Log-Likelihood:                -176.44
No. Observations:                 100   AIC:                             380.9
Df Residuals:                      86   BIC:                             417.4
Df Model:                          13
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     2.4167      0.407      5.944      0.000       1.608       3.225
1_board_directors_women                  -0.7780      2.157     -0.361      0.719      -5.067       3.511
2_risk_rights_human                      -0.4908      0.949     -0.517      0.606      -2.376       1.395
3_emissions_waste_gas                    -3.3450      1.901     -1.759      0.082      -7.125       0.435
4_water_consumption_withdrawal            2.2003      1.712      1.285      0.202      -1.204       5.604
5_climate_climate change_biodiversity     7.2592      1.764      4.115      0.000       3.753      10.766
9_tax_taxes_income                        4.7554      1.972      2.411      0.018       0.834       8.676
10_circular_circular economy_economy     -2.6526      0.794     -3.339      0.001      -4.232      -1.073
11_green_bond_green bond                  1.9857      0.926      2.145      0.035       0.145       3.826
13_leave_parental_maternity              -4.1820      1.339     -3.123      0.002      -6.844      -1.520
market_cap                             4.794e-05   1.16e-05      4.147      0.000     2.5e-05    7.09e-05
utilities                                -0.1749      0.705     -0.248      0.805      -1.577       1.227
banking                                  -2.3070      0.476     -4.848      0.000      -3.253      -1.361
text                                     -0.0001      0.000     -0.490      0.625      -0.001       0.000
==============================================================================
Omnibus:                        6.277   Durbin-Watson:                   0.919
Prob(Omnibus):                  0.043   Jarque-Bera (JB):                6.583
Skew:                           0.388   Prob(JB):                       0.0372
Kurtosis:                       3.989   Cond. No.                     5.31e+05
==============================================================================
```

Figure 15: OLS results with target E

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      s   R-squared:                       0.570
Model:                            OLS   Adj. R-squared:                  0.505
Method:                 Least Squares   F-statistic:                     8.771
Date:                Sat, 03 Jun 2023   Prob (F-statistic):           4.30e-11
Time:                        10:53:16   Log-Likelihood:                -162.31
No. Observations:                 100   AIC:                             352.6
Df Residuals:                      86   BIC:                             389.1
Df Model:                          13
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     2.2736      0.353      6.441      0.000       1.572       2.975
1_board_directors_women                   1.2429      1.873      0.664      0.509      -2.481       4.966
2_risk_rights_human                      -1.1648      0.824     -1.414      0.161      -2.802       0.472
3_emissions_waste_gas                    -0.4099      1.651     -0.248      0.805      -3.692       2.872
4_water_consumption_withdrawal           -3.4592      1.487     -2.327      0.022      -6.415      -0.504
5_climate_climate change_biodiversity     1.8861      1.531      1.232      0.221      -1.158       4.931
9_tax_taxes_income                        6.9354      1.712      4.050      0.000       3.531      10.340
10_circular_circular economy_economy      0.9601      0.690      1.392      0.167      -0.411       2.331
11_green_bond_green bond                 -2.8435      0.804     -3.538      0.001      -4.441      -1.246
13_leave_parental_maternity              -0.6267      1.163     -0.539      0.591      -2.938       1.685
market_cap                             2.728e-06      1e-05      0.272      0.786   -1.72e-05    2.27e-05
utilities                                 2.6887      0.612      4.390      0.000       1.471       3.906
banking                                   0.1786      0.413      0.432      0.667      -0.643       1.000
text                                      0.0003      0.000      1.346      0.182      -0.000       0.001
==============================================================================
Omnibus:                        2.368   Durbin-Watson:                   0.935
Prob(Omnibus):                  0.306   Jarque-Bera (JB):                2.371
Skew:                           0.357   Prob(JB):                        0.306
Kurtosis:                       2.755   Cond. No.                     5.31e+05
==============================================================================
```

Figure 16: OLS results with target S

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      g   R-squared:                       0.469
Model:                            OLS   Adj. R-squared:                  0.389
Method:                 Least Squares   F-statistic:                     5.847
Date:                Sat, 03 Jun 2023   Prob (F-statistic):           1.34e-07
Time:                        10:54:03   Log-Likelihood:                -64.904
No. Observations:                 100   AIC:                             157.8
Df Residuals:                      86   BIC:                             194.3
Df Model:                          13
Covariance Type:            nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                     6.2968      0.133     47.248      0.000       6.032       6.562
1_board_directors_women                   0.7120      0.707      1.007      0.317      -0.694       2.118
2_risk_rights_human                      -0.5618      0.311     -1.807      0.074      -1.180       0.056
3_emissions_waste_gas                    -1.1241      0.623     -1.804      0.075      -2.363       0.115
4_water_consumption_withdrawal            0.0347      0.561      0.062      0.951      -1.081       1.151
5_climate_climate change_biodiversity     1.0557      0.578      1.826      0.071      -0.094       2.205
9_tax_taxes_income                        0.4775      0.647      0.739      0.462      -0.808       1.763
10_circular_circular economy_economy     -1.5770      0.260     -6.056      0.000      -2.095      -1.059
11_green_bond_green bond                 -0.1155      0.303     -0.381      0.704      -0.719       0.488
13_leave_parental_maternity               0.6514      0.439      1.484      0.141      -0.221       1.524
market_cap                             2.904e-06   3.79e-06      0.766      0.446   -4.63e-06    1.04e-05
utilities                                -0.0446      0.231     -0.193      0.848      -0.504       0.415
banking                                  -0.2808      0.156     -1.800      0.075      -0.591       0.029
text                                      0.0002   7.35e-05      2.125      0.036    1.01e-05       0.000
==============================================================================
Omnibus:                        5.129   Durbin-Watson:                   1.038
Prob(Omnibus):                  0.077   Jarque-Bera (JB):                4.681
Skew:                           0.390   Prob(JB):                       0.0963
Kurtosis:                       3.717   Cond. No.                     5.31e+05
==============================================================================
```

Figure 17: OLS results with target G

Ridge Regression results:
y = E

R^2: 0.5710413079217811
MSE: 2.010858932996214
MAE: 1.0333467469891913

	label	coef
0	1_board_directors_women	-0.097928
1	2_risk_rights_human	-0.613943
2	3_emissions_waste_gas	-2.415977
3	4_water_consumption_withdrawal	1.988956
4	5_climate_climate change_biodiversity	6.062955
5	9_tax_taxes_income	3.963959
6	10_circular_circular economy_economy	-2.387510
7	11_green_bond_green bond	2.036859
8	13_leave_parental_maternity	-3.763761
9	market_cap	0.000046
10	utilities	-0.338479
11	banking	-2.247302
12	text	-0.000119

Figure 18: Ridge results with target E

```
Ridge Regression results:
 y = S

R^2: 0.5673288397189433
MSE: 1.5138712665481118
MAE: 1.012906141333846
```

	label	coef
0	1_board_directors_women	1.697429
1	2_risk_rights_human	-1.127285
2	3_emissions_waste_gas	-0.305986
3	4_water_consumption_withdrawal	-2.685924
4	5_climate_climate change_biodiversity	1.519638
5	9_tax_taxes_income	5.768879
6	10_circular_circular economy_economy	1.055857
7	11_green_bond_green bond	-2.706263
8	13_leave_parental_maternity	-0.661327
9	market_cap	0.000004
10	utilities	2.421460
11	banking	0.166648
12	text	0.000240

Figure 19: Ridge results with target S

```
Ridge Regression results:
 y = G

R^2: 0.467817766783884
MSE: 0.21496683878854797
MAE: 0.3597878929107412
```

	label	coef
0	1_board_directors_women	0.678741
1	2_risk_rights_human	-0.540853
2	3_emissions_waste_gas	-0.886993
3	4_water_consumption_withdrawal	0.055194
4	5_climate_climate change_biodiversity	0.883146
5	9_tax_taxes_income	0.424828
6	10_circular_circular economy_economy	-1.506118
7	11_green_bond_green bond	-0.088301
8	13_leave_parental_maternity	0.631769
9	market_cap	0.000002
10	utilities	-0.038740
11	banking	-0.275917
12	text	0.000156

Figure 20: Ridge results with target G

References

Baldissera, A. (2023), 'Sustainability reporting in banks: History of studies and a conceptual framework for thinking about the future by learning from the past'.
URL: *https://onlinelibrary.wiley.com/doi/epdf/10.1002/csr.2491*

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; 29.
URL: *https://doi.org/10.1002/csr.2177*

Borsa-Italiana (2016), 'Non-financial disclosure obligations - borsa italiana'.
URL: *https://www.borsaitaliana.it/finanza-sostenibile/materiali/nfrd.en.htm*

Campello, R. J. G. B., Moulavi, D. & Sander, J. (2013), Density-based clustering based on hierarchical density estimates, *in* J. Pei, V. S. Tseng, L. Cao, H. Motoda & G. Xu, eds, 'Advances in Knowledge Discovery and Data', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 160–172. ID: 10.1007/978-3-642-37456-$2_1$4.

Cheng, B., Ioannou, I. & Serafeim, G. (2014), 'Corporate social responsibility and access to finance', *Strategic Management Journal* **35**(1), 1–23. https://doi.org/10.1002/smj.2131; 29.
URL: *https://doi.org/10.1002/smj.2131*

Cho, S. J., Chung, C. Y. & Young, J. (2019), 'Study on the relationship between csr and financial performance', *Sustainability* **11**(2).

Clarkson, P. M., Li, Y., Richardson, G. D. & Vasvari, F. P. (2008), 'Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis', *Accounting, Organizations and Society* **33**(4), 303–327. ID: 271665.
URL: *https://www.sciencedirect.com/science/article/pii/S0361368207000451*

Cormier, D. & Magnan, M. (2007), 'The revisited contribution of environmental reporting to investors' valuation of a firm's earnings: An international perspective', *Ecological Economics* **62**(3), 613–626. ID: 271867.

URL: *https://www.sciencedirect.com/science/article/pii/S0921800906003818*

Dhaliwal, D. S., Li, O. Z., Tsang, A. & Yang, Y. G. (2011), 'Voluntary nonfinancial disclosure and the cost of equity capital: The initiation of corporate social responsibility reporting', *The Accounting Review* **86**(1), 59–100. 2023.

URL: *https://doi.org/10.2308/accr.00000005*

Dorfleitner, G., Halbritter, G. & Nguyen, M. (2015), 'Measuring the level and risk of corporate responsibility – an empirical comparison of different esg rating approaches', *Journal of Asset Management* **16**(7), 450–466. ID: Dorfleitner2015.

URL: *https://doi.org/10.1057/jam.2015.31*

Edmans, A. (2011), 'Does the stock market fully value intangibles? employee satisfaction and equity prices', *Journal of Financial Economics* **101**(3), 621–640. ID: 271671.

URL: *https://www.sciencedirect.com/science/article/pii/S0304405X11000869*

EU (2023), 'Corporate sustainability reporting'.

URL: *https://finance.ec.europa.eu/capital-markets-union-and-financial-markets/company-reporting-and-auditing/company-reporting/corporate-sustainability-reporting$_e$n*

Freeman, R. & Mcvea, J. (2001), 'A stakeholder approach to strategic management', *SSRN Electronic Journal* .

Gao, F., Dong, Y., Ni, C. & Fu, R. (2016), 'Determinants and economic consequences of non-financial disclosure quality', *European Accounting Review* **25**(2), 287–317. doi: 10.1080/09638180.2015.1013049.

URL: *https://doi.org/10.1080/09638180.2015.1013049*

Gompers, P., Ishii, J. & Metrick, A. (2003), 'Corporate governance and equity prices*', *The*

Quarterly Journal of Economics **118**(1), 107–156. 2023.
URL: *https://doi.org/10.1162/00335530360535162*

Grootendorst, M. (2022*a*), 'Bertopic: Neural topic modeling with a class-based tf-idf procedure', *arXiv preprint arXiv:2203.05794* .

Grootendorst, M. P. (2022*b*), 'Bertopic algorithm'.

How UMAP Works (n.d.).
URL: *https://umap-learn.readthedocs.io/en/latest/how$_u$map$_w$orks.html*

Ioannou, I. & Serafim, G. (2011), 'The consequences of mandatory corporate sustainability reporting by ioannis ioannou, george serafeim'.
URL: *https://papers.ssrn.com/sol3/papers.cfm?abstract$_i$d = 1799589paper − citations − widget*

Isaksson, R. & Steimle, U. (2009), 'What does gri-reporting tell us about corporate sustainability?', *The TQM Journal* **21**(2), 168–181. 29.
URL: *https://doi.org/10.1108/17542730910938155*

Jackson, G., Bartosch, J., Avetisyan, E., Kinderman, D. & Knudsen, J. S. (2020), 'Mandatory non-financial disclosure and its influence on csr: An international comparison', *Journal of Business Ethics* **162**(2), 323–342. ID: Jackson2020.
URL: *https://doi.org/10.1007/s10551-019-04200-0*

Kiriu, T. & Nozaki, M. (2020), 'A text mining model to evaluate firms' esg activities: An application for japanese firms', *Asia-Pacific Financial Markets* **27**(4), 621–632. ID: Kiriu2020.
URL: *https://doi.org/10.1007/s10690-020-09309-1*

KPMG (2022), Kpmg global survey of sustainability reporting 2022, Technical report.
URL: *https://kpmg.com/it/it/home/insights/2022/12/kpmg-global-survey-of-sustainability-reporting-2022.html*

Manetti, G. (2011), 'The quality of stakeholder engagement in sustainability reporting: empirical evidence and critical points', *Corporate Social Responsibility and Environmental Management* **18**(2), 110–122. https://doi.org/10.1002/csr.255; 29.
URL: *https://doi.org/10.1002/csr.255*

Manrique, S. & Martí-Ballester, C.-P. (2017), 'Analyzing the effect of corporate environmental performance on corporate financial performance in developed and developing countries', *Sustainability* **9**(11).

Mion, G. & Loza Adaui, C. (2019), 'Mandatory nonfinancial disclosure and its consequences on the sustainability reporting quality of italian and german companies'.
URL: *https://www.mdpi.com/2071-1050/11/17/4612*

Nwaigwe, N. G., Ofoegbu, G. N., Dibia, N. O. & Nwaogwugwu, C. V. (2022), 'Sustainability disclosure: Impact of its extent and quality on value of listed firms in nigeria', *Cogent Business Management* **9**(1), 2079393. doi: 10.1080/23311975.2022.2079393.
URL: *https://doi.org/10.1080/23311975.2022.2079393*

Papoutsi, A. & Sodhi, M. S. (2020), 'Does disclosure in sustainability reports indicate actual sustainability performance?', *Journal of Cleaner Production* **260**, 121049. ID: 271750.
URL: *https://www.sciencedirect.com/science/article/pii/S0959652620310969*

Pereira, C., Monteiro, A., Silva, D. & Lima, A. (2023), 'Do the levels of environmental sustainability disclosure and indebtness affect the quality of earnings?', *Sustainability* **15**(4).

Rahman, I., Park, J. & Chi, C. G. (2015), 'Consequences of "greenwashing"', *International Journal of Contemporary Hospitality Management* **27**(6), 1054–1081. 05.
URL: *https://doi.org/10.1108/IJCHM-04-2014-0202*

Uyar, A., Karaman, A. S. & Kilic, M. (2020), 'Is corporate social responsibility reporting a tool of signaling or greenwashing? evidence from the worldwide logistics sector', *Journal of*

Cleaner Production **253**, 119997. ID: 271750.

URL: *https://www.sciencedirect.com/science/article/pii/S0959652620300445*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), 'Attention is all you need'. ID: DBLP:journals/corr/VaswaniSPUJGKP17.

URL: *http://arxiv.org/abs/1706.03762*

Velte, P. (2017), 'Does esg performance have an impact on financial performance? evidence from germany', *Journal of Global Responsibility* **8**(2), 169–178. 29.

URL: *https://doi.org/10.1108/JGR-11-2016-0029*

Wang, X., Cao, F. & Ye, K. (2018), 'Mandatory corporate social responsibility (csr) reporting and financial reporting quality: Evidence from a quasi-natural experiment', *Journal of Business Ethics* **152**(1), 253–274.

URL: *https://ideas.repec.org//a/kap/jbuset/v152y2018i1d10.1007*$_s10551-016-3296-2.html$