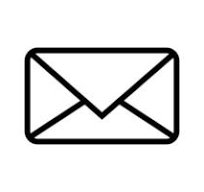
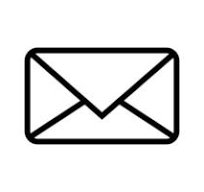
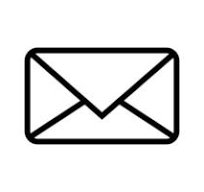
Data Descriptor

Assessment of TCFD voluntary disclosure compliance in the Spanish energy sector: A text mining approach to climate change financial disclosures.

Matías Domínguez-Quiñones 1\*[](mailto:matiasdo@ucm.es)[](https://orcid.org/0000-0001-5746-2212), Iñaki Aliende 2 [](mailto:ialiende@ucm.es)[](https://orcid.org/0000-0002-5187-0044)and Lorenzo Escot 3 [](mailto:escot@ucm.es)[](https://orcid.org/0000-0002-6734-6327)

|  |
| --- |
| Received: date  Revised: date  Accepted: date  Published: date  **Citation:** To be added by editorial staff during production.  **Copyright:** © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). |

1 Faculty of Statistical Studies, Complutense University of Madrid, Madrid, Spain

2 Universidad Complutense de Madrid (UCM), Faculty of Economics and Business, Campus de Somosaguas, 28224 Madrid, Spain.

3 Department of Applied, Public and Political Economics, Faculty of Statistical Studies, Complutense University of Madrid, Madrid, Spain

**\*** Correspondence: [matiasdo@ucm.es](mailto:matiasdo@ucm.es)

**Abstract:** This study investigates the voluntary implementation of the Task Force on Climate-related Financial Disclosures (TCFD) framework in 64 annual sustainability reports (2020-2023) of six Spanish energy companies listed on IBEX-35. The methodology includes advanced text mining (TM) techniques, such as named entity recognition (NER) and full-text searches (FTS), to ensure a comprehensive analysis. We evaluated the 11 recommended disclosures to assess their quality, extent, and relevance, as well as 70 specific concepts based on TM analyses and previous report evaluations, to construct an index of TCFD compliance. The results show year-on-year improvements in compliance with the TCFD. The TM technique reveals that Iberdrola and Repsol lead governance and risk disclosure, whereas Enagás and REE show inconsistencies in resilience and emissions, posing reputational risks. Significant progress has been made in 11 aspects of reporting quality, scope, and relevance to stakeholders. The index shows disclosure inequality for 70 specific concepts. The conclusions are that the energy sector drives political-social change against climate change, progress in opportunities and challenges remains, and reinforces the need for mandatory climate financial reporting standards. Future research will analyze the TCFD framework to assess intangible business assets and the impact of regulatory implementation on sustainability reports using TM. The originality, implications, and empirical evidence provide a multidisciplinary perspective using text mining, revealing key patterns, and promoting transparency for stakeholders.

Dataset: <https://github.com/MATDOMI/TCFD-Energy-Sector-TM.git>

**Dataset License:** license under which the dataset is made available (CC0, CC-BY, CC-BY-SA, CC-BY-NC, etc.)

**Keywords: c**limate change, CSRD, energy sector, financial reporting, NER, TCFD, text mining.

1. Summary

This study investigated whether the analyzed companies comply with the Task Force on Climate-Related Financial Disclosures (TCFD) framework using text mining (TM) in the financial reporting of annual and sustainability reports. We collected and preprocessed a dataset of 64 reports from six publicly traded companies in the Spanish energy sector of IBEX-35 between 2020 and 2023. The analyzed PDF reports were obtained from six corporate website officials of the companies studied.

This study employs data science with TM and has future potential because of the need to analyze numerical data and unstructured textual documents, such as financial and sustainability reports [1].

The methodology employed was Named Entity Recognition (NER). Our model adopts a rule-based approach to identify relevant entities or "word bags" in sustainability reporting and TCFD contexts, similar to the methodology used by Moreno and Caminero (2020) [2] in their analysis of climate-related disclosures in the Spanish banking sector. This approach incorporates domain-specific knowledge, while leveraging data-driven models aligned with broader trends in text mining and NLP for sustainability reporting analysis [3].

Prior to 2000, a few academics, entrepreneurs, and science communicators analyzed business opportunities to combat climate change. The Global Reporting Initiative (GRI) in 2000 was the first framework for non-financial and sustainability information disclosure in public-private companies [4]. ESG disclosure has evolved through various international protocols, particularly in the US and Europe. Between 2017 and 2023, the G20, major companies, and economic sectors promoted the TCFD model [5] proposed by the Financial Stability Board (FSB), highlighting climate change risks to financial stability and the global economy. Climate change concerns have prompted government actions to engage businesses in mitigation through opportunities such as a circular economy, renewable energy, sustainable infrastructure, education, and awareness, representing a systemic risk that requires coordinated action for sustainable economic transition [6].

This study is justified by applying data science techniques to analyze textual datasets in financial reports, advancing the discipline and enhancing financial data interpretation. These methods can help non-financial professionals and stakeholders to understand companies' actions and investments in addressing climate change. This approach identifies patterns and trends that are not apparent through conventional analyses, and contributes to companies' data and perceptions by quantifying transparency and intentionality. The TCFD significantly influences climate-related financial information quality [5]. As of November 2022, over 4000 companies from more than 100 countries, with a market capitalization of 28 trillion dollars, have adopted it in their reports [7,8]. However, only 4% of the companies complied with all 11 TCFD recommendations, according to the "2023 Status Report" [5]. Although disclosure has increased across sectors, studies have highlighted the challenges in communicating processes, progress to stakeholders and investors, and addressing multiple information requests. This literature review found a gap requiring analysis, namely, voluntary compliance in Spain’s energy sector [8,1,7,9].

Understanding TCFD compliance among Spanish energy companies is crucial for policymakers, analysts, investors, stakeholders, and academics to address gaps in scientific literature and their importance in climate-related discourse [8,1,7,9].

Studies on the TCFD frameworks in academic databases are scarce. Other sectors have been studied in relation to climate change. In the Indian energy sector, a study of 22 listed companies from to 2018-2020 found moderate climate disclosure per the TCFD recommendation, which is positively related to financial performance [10]. In the aviation sector, climate reports increased from 2015 to 2018; however, TCFD compliance was deficient, particularly in this strategy [11]. The construction sector is responsible for over 40% of global emissions and struggles to integrate climate-related risks owing to complexities and data collection issues [12]. More research is needed on logistics, tourism, fintech such as blockchain (Bitcoin and Ethereum), and technology companies for sustainable data center management for Artificial Intelligence (AI) and/or metaverse use [13].

Previous studies [8,1,7,9] have focused on various sectors adopting TCFD, with limited attention paid to the energy sector, particularly Spain. This study addresses this gap using text mining (TM) techniques to analyze how Enagás, Endesa, Iberdrola, Naturgy, REE, and Repsol have adopted TCFD recommendations in their annual and sustainability reports from 2020 to 2023. These six companies represent the Spanish energy sector in IBEX-35, and this research aims to analyze the implementation of information related to the voluntary TCFD framework. The textual components of financial reports provide valuable indicators of a company's current and future projections [14.15].

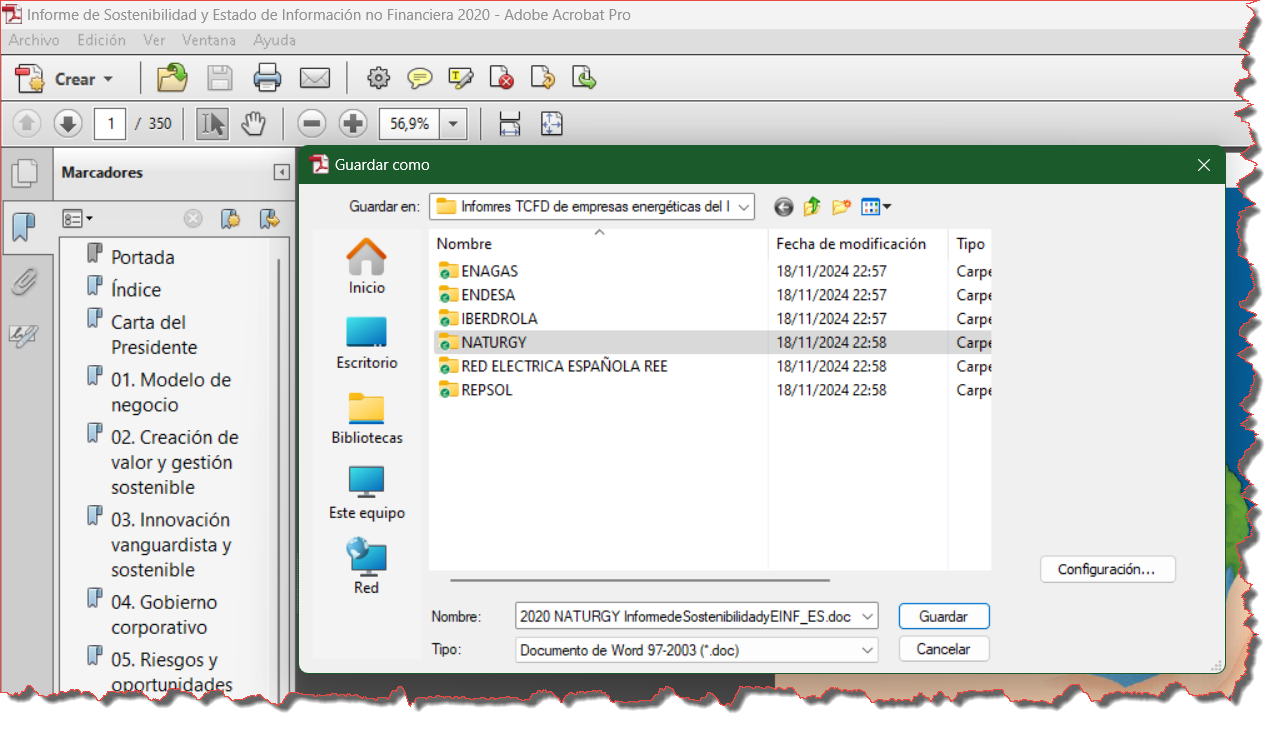
Financial reporting regulations began at the beginning of the century, especially in Europe and the United States, with the GRI [4] between 1997-2000 and progressed to the voluntary TCFD framework in 2017 [9]. In Europe, the Non-Financial Reporting Directive of 2014 [16] includes tangible and intangible assets such as Communication, Brand, or Corporate Reputation (CR) [17]. The Corporate Sustainability Reporting Directive [18] and European Sustainability Reporting Standards [19] have been mandatory for listed companies since January 1, 2024 [20].

The analysis of text and sentiment in the digital era originated with text mining and natural language processing (NLP) [21], which initiated new research and historiographical analyses. In 1995, Dr. Picard, R. *Affective Computing* [22] published on artificial emotional intelligence, stating, "Emotions play an essential role in rational decision-making, perception, interaction, and human intelligence. These facts, combined with computers' abilities in written expression and affect recognition, open new research areas." This marks a new era in text and sentiment analysis in the corporate world. McKee [23] defines textual analysis as a process that "involves making informed conjectures about the most likely interpretations of a text." As a subset of qualitative analysis, methods such as computational linguistics, natural language processing (NLP), and content analysis extract information from written sources. TM is promising for financial reporting because it focuses on textual and numerical data [24,25]. There has been an increasing proliferation of unstructured online documents, including corporate financial reporting, agency reports, press articles, audits, and analyst reports [9].

2. Data Description

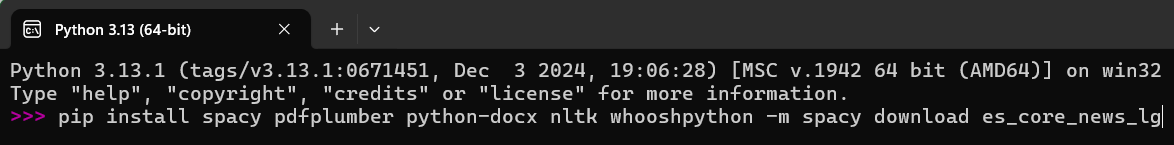
2.1. Definition and Objectives: This study examines the voluntary adoption of the Task Force on Climate-related Financial Disclosures (TCFD) framework in 64 annual sustainability reports (2020-2023) of six Spanish energy firms listed on the IBEX-35 index. The analyzed PDF reports were obtained from six corporate website officials of the companies studied. Web site Enagás: <https://www.enagas.es/es/>; Web site Endesa: <https://www.endesa.com/>;Web site Iberdrola: <https://www.iberdrola.es/> ;Web site Naturgy: <https://www.naturgy.com/> ; Web site REE: <https://www.ree.es/es> ; Web site Repsol: <https://www.repsol.com/es/index.cshtml>

**2.2. Tools and Resources used:** Reports were manually downloaded in PDF format and hosted in a shared Microsoft OneDrive space. They were sorted and formatted using Adobe Pro X and converted into Microsoft Word (figure 1).

**Figure 1:** Adobe Pro X converted to Microsoft Word format. 

Source: Author's own elaboration

The software used was, Python v. 3.11 [26], Python-docx v.1.1.2, and pdfplumber v.0.11.4 libraries were used to read the documents, whereas the spaCy v.3.8.4 library [27] facilitated tokenization. Whoosh v.2.7.4 [28] was employed for full-text searches. Microsoft Excel was used to compile and analyze the results, as shown in Figure 2. The phrase consists of a series of commands used in a programming environment, specifically Python, to install libraries and download a natural language processing model.

**Figure 2:** Installation of the Python library 

Source: Author's own elaboration

The meaning of each command is as follows:

* spaCy: Natural language processing library.
* pdfplumber: Library for extracting text, images, and table data from PDF files.
* Python-docx: A library for creating and modifying Word (. docx) documents.
* NLTK [29]: The Natural Language Toolkit, a library for working with natural language processing in Python. Download stopwords of NLTK (nltk.download('stopwords') stop\_words = set(stopwords.words('spanish'))
* whoosh: A library for indexing and searching text within Python applications.
* python -m spacy download es\_core\_news\_lg: This command downloads a large Spanish language-processing model (es\_core\_news\_lg) using spacy.

These commands ensure that the Python environment is set up to work with natural language processing, PDF and Word document handling, and text indexing and searching in the Spanish language.

**2.3 Text Preparation:** PDF files were successfully converted to text using Adobe X, pdfplumber, and Python-docx in 95.31% of cases, except those with graphics, tables, and text boxes. Conversion errors were manually resolved to adjust the success rate and minimize human bias, as shown in Table 1.

Table 1. No. of reports and file conversion

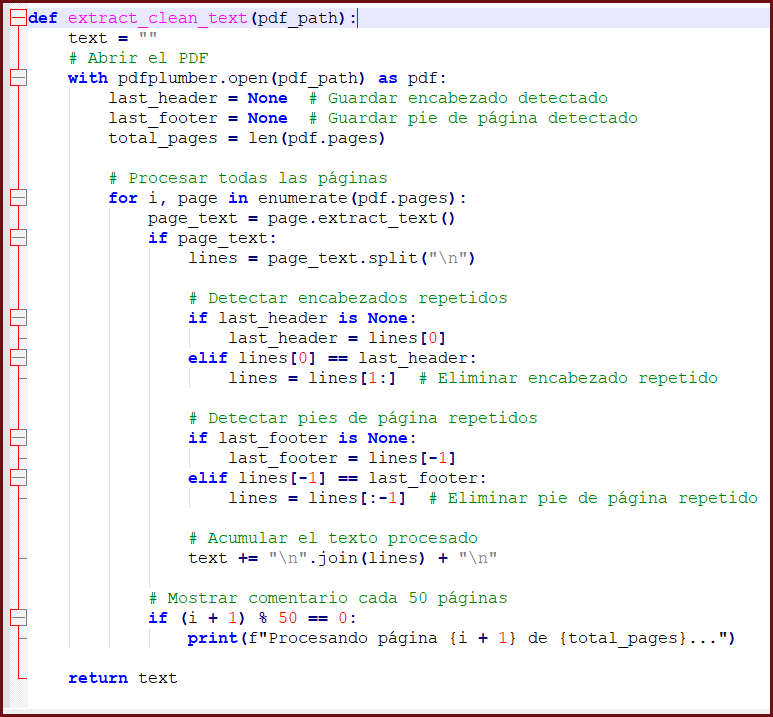
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Enagás** | **Endesa** | **Iberdrola** | **Naturgy** | **REE** | **Repsol** | **TOTAL** | |
| No. of reports | 9 | 11 | 12 | 11 | 9 | 12 | 64 | |
| Success Conversion | 94,80% | 96,00% | 95.23% | 95.90% | 95.50% | 95,45% | 95,31% | |
| Cause of Failure | Graphs and information tables not converted properly. Tables not converted properly; Bullets require manual correction. | | | | | |  | |
| Resolution | Manual - Minimizations of human bias / Improved post-error resolution. | | | | | | |  |

Source: Author's own elaboration

The following is an example script, figure 3, using pdfplumber on one of the reports in

PDF reports.

Figure 2: Script to parse a pdf with the pdfplumber library



Source: Author's own elaboration with NotePad++ v 8.6.4

**2.4 Structured text Representation and Extraction: This** script extracts text from a PDF file, removes repeated headers and footers, cleans text by removing unwanted elements, and saves both clean and dirty texts in separate files. Each part of the script is detailed below and an example of the result is shown in the script. This function clean text cleans up text by removing multiple spaces, empty lines, page numbers, URLs and normalises inverted commas.

def clean\_text(text):

# Reemplazar múltiples espacios por uno solo

text = re.sub(r'\s+', ' ', text)

# Eliminar líneas vacías

lines = [line.strip() for line in text.split("\n") if line.strip()]

# Eliminar líneas que sean solo números (posibles números de página o índices)

lines = [line for line in lines if not re.fullmatch(r'\d+', line)]

# Eliminar solo las URLs, pero no toda la línea

lines = [re.sub(r'http[s]?://\S+', '', line) for line in lines]

# Normalizar comillas y caracteres raros

text = "\n".join(lines).replace("“", "\"").replace("”", "\"").replace("’", "'")

return text ef extract\_clean\_text(pdf\_path):

text = ""

# Abrir el PDF

with pdfplumber.open(pdf\_path) as pdf:

last\_header = None # Guardar encabezado detectado

last\_footer = None # Guardar pie de página detectado

total\_pages = len(pdf.pages) # Total de páginas en el PDF

# Procesar todas las páginas

for i, page in enumerate(pdf.pages):

page\_text = page.extract\_text()

if page\_text:

lines = page\_text.split("\n")

# Detectar encabezados repetidos

if last\_header is None:

last\_header = lines[0]

elif lines[0] == last\_header:

lines = lines[1:] # Eliminar encabezado repetido

# Detectar pies de página repetidos

if last\_footer is None:

last\_footer = lines[-1]

elif lines[-1] == last\_footer:

lines = lines[:-1] # Eliminar pie de página repetido

# Acumular el texto procesado

text += "\n".join(lines) + "\n"

# Mostrar comentario cada 50 páginas

if (i + 1) % 50 == 0:

print(f"Procesando página {i + 1} de {total\_pages}...")

# Limpiar el texto extraído

return text

pdf\_path = "2020naturgy.pdf"

output\_clean\_txt = "2020naturgy\_clean.txt"

output\_dirty\_txt = "2020naturgy\_dirty.txt"

# Extraer texto de todas las páginas

dirty\_text = extract\_clean\_text(pdf\_path)

# Limpiar el texto extraído

cleaned\_text = clean\_text(dirty\_text)

# Guardar el texto "sucio" sin limpiar

with open(output\_dirty\_txt, "w", encoding="utf-8") as f:

f.write(dirty\_text)

# Guardar el texto limpio

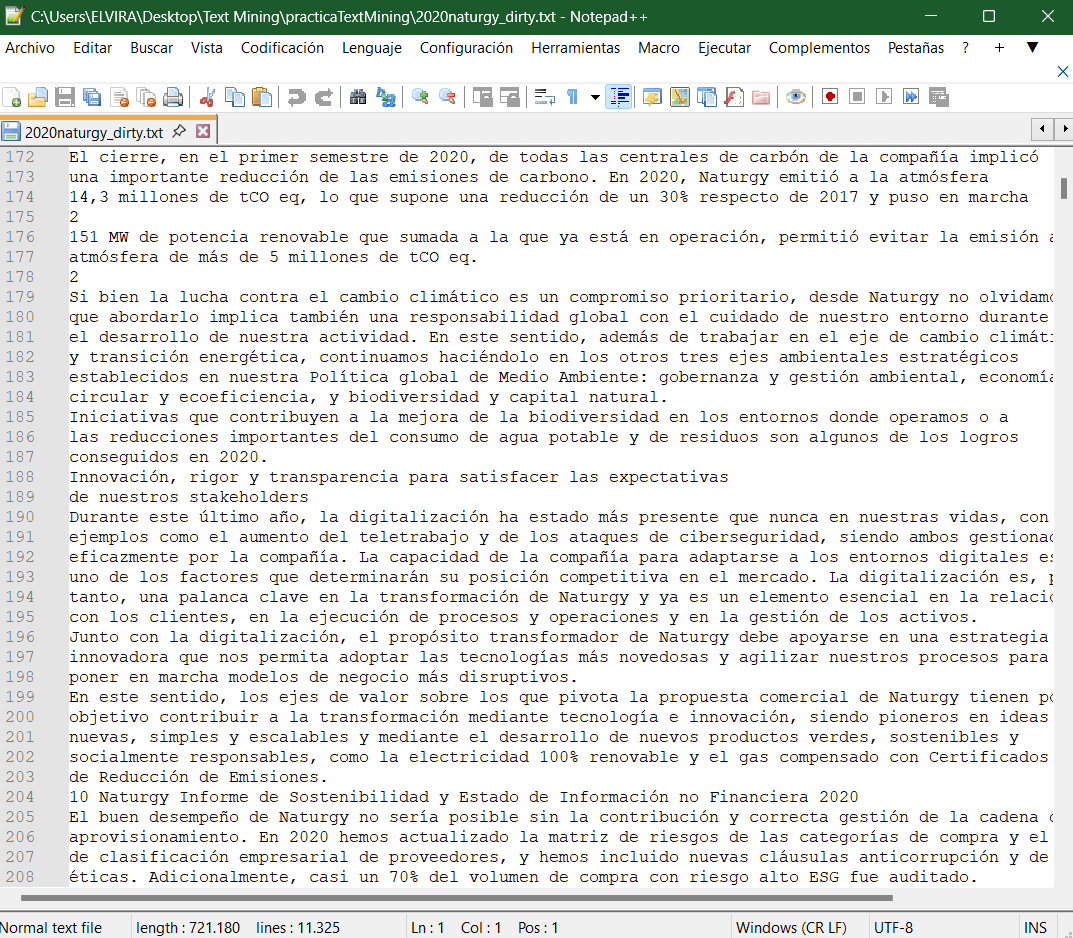
with open(output\_clean\_txt, "w", encoding="utf-8") as f:

f.write(cleaned\_text)

print(f"Texto sucio del PDF guardado en {output\_dirty\_txt}")

print(f"Texto limpio del PDF guardado en {output\_clean\_txt}")

Figure 3: Example of data extraction



Source: Author's own elaboration with NotePad++

2.5 Calculation of compliance with the TCFD framework using rules: This script calculates and stores a TCFD [30] compliance index based on key terms found in a preprocessed text and assigns scores to different categories according to the rules defined.

def calculate\_tcfd\_compliance(results):

scores = {}

for category, matches in results.items():

if len(matches) < 3:

scores[category] = 1

elif len(matches) < 6:

scores[category] = 2

else:

scores[category] = 3

return scores

# Cargar el texto preprocesado

with open("2020naturgy\_preprocessed.txt", "r", encoding="utf-8") as f:

processed\_text = f.read()

# Buscar menciones de términos clave TCFD

matches = search\_text\_for\_tcfd(processed\_text, TAXONOMY)

# Calcular el índice de cumplimiento basado en las reglas definidas

tcfd\_compliance = calculate\_tcfd\_compliance(matches)

# Guardar los resultados en un archivo

with open("tcfd\_compliance\_results.txt", "w", encoding="utf-8") as f:

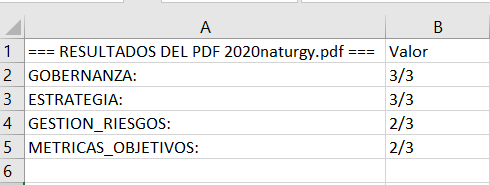
f.write("=== RESULTADOS DEL ANÁLISIS DE CUMPLIMIENTO TCFD ===\n")

for category, score in tcfd\_compliance.items():

f.write(f"{category.upper()}: {score}/3\n")

print("\nAnálisis completado. Resultados guardados en 'tcfd\_compliance\_results.txt'.")

The obtained results were converted from \*. txt format into an Excel spreadsheet for further analysis, as shown in figure 4.

Figure 4: TXT to Microsoft Excel data conversion 

Source: Author's own elaboration with Microsoft Excel

2.6 Revision and Validation in Taxonomy Creation: This script sets up an advanced natural language processing model in Spanish using spaCy, defines a taxonomy of terms related to TCFD [30], searches for matches of these terms in a given text, and returns a dictionary with the matches found in each category.

import spacy

import re

from collections import defaultdict

# Cargar modelo avanzado de spaCy para español

nlp = spacy.load("es\_core\_news\_lg")

# Definir la taxonomía con categorías de TCFD

TAXONOMY = {

"gobernanza": [

"consejo administración", "comité de sostenibilidad", "directivos", "responsabilidad",

"supervisión", "gobernanza", "estrategia corporativa", "decisiones estratégicas",

"transparencia", "información financiera", "comités de riesgos", "cultura corporativa",

"junta", "gerencia", "directorio", "remuneración", "periodicidad", "seguimiento"

],

"estrategia": [

"riesgos climáticos", "oportunidades ambientales", "impacto financiero", "escenarios climáticos",

"planificación a largo plazo", "transición energética", "resiliencia", "modelos de negocio sostenibles",

"innovación tecnológica", "adaptación al cambio climático", "estrategias de mitigación",

"sostenibilidad", "carbono neto cero", "descarbonización", "riesgos de cambio climático",

"riesgos de transición", "riesgos físicos", "préstamos verdes", "clima extremo", "reducción de costos",

"impacto", "riesgos reputacionales", "estándares", "estrategia", "uso de tecnología", "energía renovable"

],

"gestion\_riesgos": [

"evaluación de riesgos", "planes de mitigación", "factores de riesgo", "exposición financiera",

"riesgos físicos", "riesgos transicionales", "gestión de la cadena de suministro", "estrategias de adaptación",

"gestión de impacto climático", "impacto económico de los riesgos", "seguridad climática", "gestión de desastres",

"riesgo legal", "riesgo reputacional", "riesgo financiero", "regulación", "acuerdos internacionales",

"respuesta al riesgo", "materialidad", "precio del carbono", "litigios", "energías renovables",

"costos de transición", "sistemas integrados de gestión", "control de gestión de riesgos"

],

"metricas\_objetivos": [

"huella de carbono", "emisiones de CO2", "objetivos de reducción", "indicadores de sostenibilidad",

"emisiones netas", "información climática", "energías renovables", "medición de huella ecológica",

"rendimiento sostenible", "reporte de sostenibilidad", "certificación ambiental", "contribución climática",

"objetivos de sostenibilidad", "reducción de CO2", "emisiones de GEI", "consumo de energía",

"consumo de agua", "consumo de combustible", "intensidad"

]

}

def search\_text\_for\_tcfd(text, taxonomy):

results = {category: [] for category in taxonomy.keys()}

for category, keywords in taxonomy, and items ()

for keyword in keywords:

pattern = rf"\b{keyword}\b|\b{keyword.replace(' ', '[\_-]?')}\b"

matches = re.findall(pattern, text, re.IGNORECASE)

if matches:

results[category].append((keyword, len(matches)))

return results

2.7 Evaluation of results (Model Accuracy): To obtain an evaluation of the results of a machine learning model in terms of accuracy (precision), recall (completeness) and F1-Score, we use the scikit-learn library in Python. We import the metrics, the functions precision\_score, recall\_score, and f1\_score of the sklearn.metrics [31]. The categories were "gobernanza,” "estrategia,” "gestión de riesgos,” "métricas y objetivos" while the labels were 1 (presente), 0 (ausente), so we would run the following script:

pip install scikit-learn

from sklearn.metrics import precision\_score, recall\_score, f1\_score

# Convertir listas a un solo vector de etiquetas verdaderas y predichas

true\_labels\_flat = [label for sublist in true\_labels for label in sublist]

predicted\_labels\_flat = [label for sublist in predicted\_labels for label in sublist]

# Calcular precisión, exhaustividad y F1-Score

precision = precision\_score(true\_labels\_flat, predicted\_labels\_flat, average='weighted')

recall = recall\_score(true\_labels\_flat, predicted\_labels\_flat, average='weighted')

f1 = f1\_score(true\_labels\_flat, predicted\_labels\_flat, average='weighted')

print(f"Precision: {precision:.2f}")

print(f"Recall: {recall:.2f}")

print(f"F1-Score: {f1:.2f}")print(f"Recall: {recall:.2f}")

print(f"F1-Score: {f1:.2f}")

3. Methods

We developed three subsections: Describe the study design, analytical procedures, taxonomy construction, applications, and techniques. The aspects of the 64 reports were evaluated using domain-specific rigor. Generate an index from the analysis of 70 SC applied to the text mining (TM) results

3.1 Description of the TM Process

The methodology used was Named Entity Recognition (NER). Our model employs a rule-based approach to identify relevant entities in sustainability reports and the TCFD context, similar to the method used by Moreno and Caminero (2020) for climate-related disclosures in the Spanish banking sector. This approach combines domain-specific knowledge with data-driven models, aligned with broader trends in text mining and NLP for sustainability reporting analysis.

3.1.1. Document collection and data preparation: The analyzed PDF reports were obtained from six corporate website officials of the companies studied. The final selection included 64 studies published between 2020-2023. The inclusion criteria were as follows: (i) >250 employees; (ii) external auditor review; (iii) public corporate website publication; and (iv) PDF format. Analyzed reports included "Annual, consolidated, or integrated reports," "non-financial information and/or sustainability reports" or “Environmental, Social, and Governance (ESG) reports”, and "Corporate Governance Reports." Other reports with potential TCFD-related data were also included. The exclusion criteria were as follows: (i) non-annual reports; (ii) non-PDF reports; (iii) reports not available on official websites; (iv) infographic reports; and (v) reports that did not contribute to the study after manual review. The files were organized according to the entity, year, and report type. These reports helped to create an initial taxonomy through a manual review. No single report on official websites covers all the TCFD sections. Text mining extracts key themes, words, and data from TCFD-related reports and involves data preparation when analyzing PDF texts from corporate websites. The extracts were identified by leveraging key concepts from a specially designed sustainability taxonomy tailored to the energy sector, using the Global Reporting Initiative and TCFD [4,30] as a reference.

3.1.2. Structured text representation and extraction: We used Python script with SpaCy [26,27], an efficient NLP library, to break paragraphs and sentences into manageable extracts, identifying and analyzing paragraphs, sentences, bullets, objects, and tables without manipulating the text. We used a total of 14, 503 extracts, as shown in Table 2.

Table 2. Structured text representation and extraction

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Enagás** | | **Endesa** | **Iberdrola** | **Naturgy** | **REE** | **Repsol** | **TOTAL** |
| No. of sentences  (Key and thematic sentences) | | 1.264 | 1.897 | 3.152 | 2.837 | 1.645 | 2.465 | 13.260 |
| Number of bullets  (Elements identified) | 57 | | 62 | 71 | 68 | 54 | 86 | 398 |
| No. of objects  (Prioritised list) | 89 | | 97 | 112 | 75 | 106 | 134 | 613 |
| No. of Tables  (structured data) | 33 | | 42 | 40 | 37 | 35 | 45 | 232 |
| TOTAL (%) | ≈13,7% | | ≈13,1% | ≈21,7% | ≈19,6% | ≈11,3% | ≈20,6% | 14.503 (100%) |

Source: Author's own elaboration based on data compiled

3.1.3. Analysis and Information Extraction: Python with spaCy was used to define relevant NER for reports, indexing them as environmental metrics, financial indicators, risk categories, phrases, words in the 'bag of words,' titles, and independent tables for tokenization. SpaCy, a powerful NLP library, allows for efficient identification and classification of named entities We started with 11 TCFD-recommended disclosures and subdivided them into 70 detailed disclosures based on domain knowledge and TCFD guidelines, as listed in Table 4 [30]. For each detailed disclosure, we established a rule to determine its presence in corporate reports, assigning values from 1 to 3 based on the frequency of follow-up related to the 'entity,’ as in Table 3. These values reflect the relevance, frequency of follow-up, and industry standards, ensuring a prioritized evaluation of key disclosures.

Explanation of Frequency of Follow-up: This refers to how often terms related to an entity are mentioned in corporate reports. A higher frequency indicates greater relevance. The values assigned to queries reflect the importance of the term to the company.

Table 3. The examples of these three rules demonstrate the flexibility of their search concepts.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Areas** | **Recommended**  **Disclosures (RD)** | **Specific**  **Disclosures** | **Query** | **Value** |
| A3  Risk  Assessment  and  Management | RD8 - Corporate resilience and reputation plan against climate change | Frequency of follow-up  in relation to  the 'reputation' | reputation AND plan AND climate change  reputation AND plan AND resilience  reputation AND plan AND (strategy OR sustainable) | 3    2  1 |

Source: Author's own elaboration

For multiple query matches, only the maximum value was considered. The rules ensured 'climate change' disclosures scored higher than 'sustainable' ones. The queries were defined using a simple query language with Boolean operators. After the iterations, 103 queries were established using 70 distinct concepts to identify 57 detailed disclosures, each linked to one of the 11 RD listed in Table 4.

Table 4. Quantification of specific disclosures and inquiries pertaining to each recommended disclosure

|  |  |  |  |
| --- | --- | --- | --- |
| **4 Areas** | **11 Recommended Disclosures (RD)** | **Detailed Disclosures** | **Total Queries** |
| A1 –  Government | RD1 - Structures and processes for monitoring climate risks  RD2 - Senior management involvement in climate management | 3  8 | 9  18 |
| A2 –  Strategies and Opportunities | RD3 - Impact of climate change on the organisation  RD4 - Development of green products for diversification, adaptation and resilience  RD5 - Identification of climate risks and opportunities. | 7  11  2 | 10  12  3 |
| A3 –  Risk Assessment and Management | RD6- Climate risk assessment and management  RD7 - Adaptation to changes in regulation  RD8 - Corporate resilience and reputation plan against climate change | 2  7  3 | 4  12  6 |
| A4 –  Metrics  and Targets | RD9 - Measuring and quantifying climate impact in the company  RD10 - GHG Emissions Protocol and Management, Scope 1, 2 and 3  RD11 - Objectives and actions for the use of renewable energies and business sustainability | 4  6  4 | 10  12  7 |

Source: Author's own elaboration based on TCFD [30]

3.1.4. Visualization of results with tokenization and indexing: Whoosh [28] implemented full-text searches (FTS) in the extract set, which were stored for searching via a graphical user interface (GUI). FTS is a technique used for comprehensive searches of a document’s text. Whoosh tokenizes and indexes the extracts, thereby enabling entity identification using specific words. The number of extracts for the 64 reports from the six analyzed entities was 14,503. Whoosh expedited and enhanced the precision of this study by processing a significant volume of data. The search criteria included keywords, phrases, expressions, and NER contexts, such as company names, environmental metrics, categorization by section (like TCFD areas), and relevance and frequency. The FTS results using Whoosh ensured the accurate identification of the relevant extracts from the 14,503 samples.

3.1.5. Revision and Validation in Taxonomy Creation: We manually reviewed entities in company reports using FTS and developed a concept taxonomy based on the TCFD recommendations. The concepts were organized by area and 11 TCFD disclosures, as listed in Table 5.

Table 5. Set of concepts for creating the taxonomy according to the TCFD framework, [30]

|  |  |  |
| --- | --- | --- |
| **4 Areas** | **11 Recommended Disclosures (RD)** | **70 Specific Concepts (SC)** |
| A1 - Government | RD1 - Structures and processes for monitoring climate risks  RD2 - Senior management involvement in climate management | SC1 - management, remuneration, periodicity, monitoring  SC2 - management, remuneration, periodicity,  monitoring, sustainability committee |
| A2 - Strategies and Opportunities | RD3 - Impact of climate change on the organisation  RD4 - Development of green products for diversification, adaptation and resilience  RD5 - Identification of climate risks and opportunities. | SC3 – criteria, strategy, effect, reputational risks, reporting standards, technological utilization, sustainable energy  SC4 - climate scenarios, temperature increase  SC5 - tangible risks, climate change risks, shift risk, cost reduction opportunities, , financing, mortgages, sustainable financing, short-term, medium, long-term, several weather events, |
| A3 Risk Assessment and Management | RD6- Climate risk assessment and management  RD7 - Adaptation to changes in regulation  RD8 - Corporate resilience and reputation plan against climate change al cambio climático | SC6 - environmental risks, prospects, shift risks, tangible risks, procedures, disclosure Guidelines, compliance risk, reputation risk, economic risk, legislative controls, international agreements  SC7 - Risk Management Strategies, materiality, Emission Valuation, litigation, severe weather events, renewable energy, transition costs  SC8 - Strategy, crisis response |
| A4 - Metrics and Targets | RD9 - Measuring and quantifying climate impact in the company  RD10 - GHG Emissions Protocol and Management, Scope 1, 2 and 3  RD11 - Objectives and actions for the use of renewable energies and business sustainability | SC9 - reduction, CO2emissions, refuse, energy use, water use, fuel use, renewable energies,  SC10 - Scope, CO2emissions, CO2units, CO2intensity, CO2eq.  SC11 - target, reduction, CO2emissions, refuse, energy use, water use, fuel use, renewable energies, renewable energies. |

Source: Author's own elaboration based on TCFD [30]

The objective was to streamline the process by labeling the excerpts. A manual review ensured taxonomic consistency with the TCFD guidelines and 70 disclosures were verified. Each word can be labeled once, and while the SC may recur, the rules prioritize longer texts. Rules with explicit text in the recommendations prevented the occurrence of false positives. The lexicon was lemmatized and converted to a lower case to account for term variations. Lemmatization reduces words to their base form, helping to consistently identify terms. Extracts were processed using Python and labeled according to the taxonomy.

3.1.6. Evaluation of results (Model Accuracy): Model Accuracy was evaluated with an F1-Score of 73%, indicating good precision, balancing accuracy, and comprehensiveness. The precision was 0.78 and the recall was 0.68. The model's performance was affected by ambiguous lexicon, as certain terms could apply to multiple concepts, such as "climate," "emissions," or "sustainable." The lack of a concise TCFD report complicates the unification of the 11 disclosures and confirms compliance levels. This methodology required comprehensive subject comprehension to establish rules, potentially introducing extraneous variables into the results [30].

3.1.7. Implementation and utilization: Its application, with domain-specific training, amplifies financial reporting analyses to measure metrics and disseminate reports to stakeholders, investors, and academics without extensive economic knowledge. Its implications relate to intangible business assets in marketing, communication, and departments such as reporting, risk management, and executive leadership.

3.2. Aspect-Based Evaluation

We evaluated environmental information disclosure in 64 reports from the IBEX-35 energy sector following the TCFD recommendations [30]. This assessment covers the quality, extent, and relevance of information using Beck et al. ’s [32] content analysis approach for sustainability reports. We used quantitative criteria such as "completeness of information", "details on climate policies", and "transparency on measurable results", alongside qualitative scales indicating high scores for "presence of specific examples" and "alignment with recognized standards" or low scores for "vague language" or "lack of detail in climate policies". The evaluation included all reports identifying the 11 RD, we weighted the coverage of TCFD issues based on existing literature and our expertise, considering climate change action details, reporting clarity, TCFD implementation degree, and contextualization in relation to each company's sustainability strategy.

3.3. Generation of an index based on the fulfilment of 70 specific concepts.

Based on the approach proposed by Halme and Huse [33] to assign a value to environmental disclosure in companies, we used an index from 0 to 1 to weigh a quantitative scale evaluating the level of TCFD disclosure in the six analyzed companies. To calculate this index in the four main areas, we assigned scoring indices to 70 SC derived from text mining analyses, the evaluation of analyzed aspects, and our subject knowledge. Each SC was weighted between 0 and 1, considering factors such as mention, relevance, explanation, quality, and detail of the text in each SC and comparing the 64 reports. The obtained values were summed across the four areas to derive a total score. The final index, which reflects the cumulative score of each company, provides a detailed analysis of the TCFD framework. Table 6 describes the SP in the 64 reports of the six companies and the final index reflects the total score of each company in the climate disclosure assessment.

**Table 6.** Index to assess mention, relevance, explanation, quality, and detail of text in the 70

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| YEARS: 2020 - 2021 -2022 -2023 | | | E1 | | E2 | E3 | E4 | E5 | | E6 |
| A1 | RD1 | SC1 - management, remuneration, periodicity, monitoring | | 1 | 1 | 1 | 1 | 1 | | 1 |
| RD2 | SC2 - management, remuneration, periodicity,  monitoring, sustainability committee | | 1 | 1 | 1 | 1 | 1 | | 1 |
| A2 | RD3 | SC3 – criteria, strategy, effect, reputational risks, reporting standards, technological utilization, sustainable energy | | 0 | 0 | 1 | 1 | 0 | | 1 |
| RD4 | SC4 - climate scenarios, temperature increase | | 0 | 1 | 1 | 1 | 1 | | 1 |
| RD5 | SC5 - tangible risks, climate change risks, shift risk, cost reduction opportunities, financing, mortgages, sustainable financing, short-term, medium, long-term, several weather events, | | 0 | 0 | 1 | 1 | 0 | | 1 |
| A3 | RD6 | SC6 - environmental risks, prospects, shift risks, tangible risks, procedures, disclosure guidelines, compliance risk, reputation risk, economic risk, legislative controls, international agreements | | 1 | 1 | 1 | 1 | 1 | | 1 |
| RD7 | SC7 - risk management Strategies, materiality, emission valuation, litigation, severe weather events, renewable energy, transition costs | | 1 | 1 | 1 | 1 | 1 | | 1 |
| RD8 | SC8 - strategy, crisis response | | 0 | 0 | 1 | 0 | 0 | | 1 |
| A4 | RD9 | SC9 - reduction, CO2 emissions, refuse, energy use, water use, fuel use, renewable energies, | | 0 | 0 | 1 | 1 | 0 | | 1 |
| RD10 | SC10 - Scope, CO2 emissions, CO2 units, CO2 intensity, CO2 eq. | | 0 | 0 | 0 | 0 | 0 | | 0 |
| RD11 | SC11 - target, reduction, CO2 emissions, refuse, energy use, water use, fuel use, renewable energies | | 1 | 1 | 1 | 1 | 1 | | 1 |
| Result / Index: | | | 5 | | 6 | 10 | 9 | | 5 | 10 |

**Source:** Author's own elaboration based on company analysis of the TCFD framework [30].

**Legend:** (E1) ENAGAS, (E2) ENDESA, (E3) IBERDROLA, (E4) NATURGY, (E5) REE, (E6) REPSOL

**Author Contributions:** Conceptualization, Iñaki Aliende and Lorenzo Escot; Data curation, Matías Domínguez-Quiñones, Iñaki Aliende and Lorenzo Escot; Formal analysis, Iñaki Aliende and Lorenzo Escot; Investigation, Matías Domínguez-Quiñones, Iñaki Aliende and Lorenzo Escot; Methodology, Matías Domínguez-Quiñones, Iñaki Aliende and Lorenzo Escot; Project administration, Iñaki Aliende and Lorenzo Escot; Supervision, Matías Domínguez-Quiñones, Iñaki Aliende and Lorenzo Escot; Val-idation, Matías Domínguez-Quiñones, Iñaki Aliende and Lorenzo Escot; Visualization, Iñaki Al-iende and Lorenzo Escot; Writing – review & editing, Matías Domínguez-Quiñones, Iñaki Aliende and Lorenzo Escot.

All authors have read and agreed to the published version of the manuscript.” Please turn to the [CRediT taxonomy](https://img.mdpi.org/data/contributor-role-instruction.pdf) for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** This research received no external funding

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** <https://github.com/MATDOMI/TCFD-Energy-Sector-TM.git>

**Conflicts of Interest:** The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

|  |  |
| --- | --- |
| AI | Artificial Intelligence |
| CSRD | Corporate Sustainability Reporting Directive |
| ESRS | European Sustainability Reporting Standards |
| CSR | Corporate Social Responsibility |
| ESG | Environmental, Social, and Governance |
| FSB | Financial Stability Board |
| FTS | full-text searches |
| G20 | Group of Twenty |
| GHG | Greenhouse Gas emission management |
| GRI | Global Reporting Initiative |
| NER | Named Entity Recognition |
| NLP | Natural Language Processing |
| NFRD | Non-Financial Reporting Directive |
| SDGs | Sustainable Development Goals |
| TCFD | Task Force on Climate-related Financial Disclosures |
| UNFCCC | United Nations Framework Convention on Climate Change |
| PDF | Portable Document Format |
| RD | Recommended Disclosures |
| SC | Specific Concepts |

References

[1] Cho, C. H., Swierszcz, A., Shah, J. M., Bohr, K., Partridge, K., & Choi, T. J. (2020). Advancing sustainability reporting in Canada: 2019 Report on Progress. Accounting Perspectives, 19(3), 181-204. <https://doi.org/10.1111/1911-3838.12232>

[2] Moreno, A.-I., & Caminero, T. (2022). Application of text mining to the analysis of climate-related disclosures. International Review of Financial Analysis, 83, 102307. <https://doi.org/10.1016/j.irfa.2022.102307>

[3] Mohammadrezaei, M., Marques, J. C., & Huq, A. (2024). Use of text mining and natural language processing techniques in analyzing sustainability reports: A systematic literature review and assessment. <http://dx.doi.org/10.2139/ssrn.5071032>

[4] Global Reporting Initiative. (2024). GRI Standards. Retrieved from <https://www.globalreporting.org/standards/standards-development/sector-standard-for-oil-and-gas/>

[5] Financial Stability Board (FSB). (2015-2023). Working Group on Climate Finance Disclosures: Progress Report in 2019, 2020, 2021, 2022, and 2023. Retrieved from <https://www.fsb-tcfd.org/about/>

[6] Carney, M. (2015). Breaking the tragedy of the horizon: Climate change and financial stability. Speech presented at Lloyd’s of London, London, UK. Retrieved from: <https://www.bankofengland.co.uk/-/media/boe/files/speech/2015/breaking-the-tragedy-of-the-horizon-climate-change-and-financial-stability.pdf>

[7] Bose, S., & Hossain, A. (2021). An exploratory study on climate-related financial disclosures: International evidence. In M. Marzouk & K. Hussainey (Eds.) Corporate Narrative Reporting: Beyond the Numbers (pp. 1-31). Routledge Taylor and Francis Group. <https://doi.org/10.4324/9781003095385-16>

[8] Tyagi, S. (2018). What comes next: Responding to recommendations from the Task Force on Climate-related Financial Disclosure (TCFD). The APPEA Journal, 58, 633-636. <https://doi.org/10.1071/AJ17179>

[9] Borré, S. S., & Gelmini, L. (2024). Global corporate performance evaluation and sustainability reporting. Corporate Ownership & Control, 21(4), 129–136. <https://doi.org/10.22495/cocv21i4art11>

[10] Maji, S. G., & Kalita, N. (2022). Climate change financial disclosure and firm performance: Empirical evidence from Indian energy sector based on TCFD recommendations. Society and Business Review, 17(4), 594-612. <https://doi.org/10.1108/SBR-10-2021-0208>

[11] David, B., & Giordano-Spring, S. (2021). Climate reporting related to the TCFD framework: An exploration of the air transport sector. Social and Environmental Accountability Journal, 42(1-2), 18-37. <https://doi.org/10.1080/0969160x.2021.2007784>

[12] Siew, R. (2020). Briefing: Task force for climate financial disclosures (TCFD) for the property and construction industry. Sustainable Buildings, 5, 3. <https://doi.org/10.1051/sbuild/2020002>

[13] Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., & Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. Neurocomputing, 599, 128096. <https://doi.org/10.1016/j.neucom.2024.128096>

[14] Parveen, N. N., Kavitha, D., Maheswari, B. U., & Sujatha, R. (2023). Assessment of firm’s performance by employing text mining techniques. Proceedings of 3rd International Conference on Artificial Intelligence, Springer, Singapore. <https://doi.org/10.1007/978-981-19-7041-2_8>

[15] Sai, P., Gupta, P., Fernandes, S. (2019). Analysing performance of company through annual reports using text analytics. In 2019 International Conference on Digitization (ICD), Sharjah, United Arab Emirates, 2019, pp. 21-31. <https://doi.org/10.1109/ICD47981.2019.9105673>

[16] European Union. (2014). Directive 2014/95/EU of the European Parliament. Official Journal of the European Union. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014L0095>

[17] Godawska, J. (2023). The use of corporate external environmental reporting in the evaluation of the state environmental policy. Economics and Environment, 86(3), 8-30. <https://doi.org/10.34659/eis.2023.86.3.550>

[18] European Union. (2022). Corporate Sustainability Reporting Directive (CSRD – (EU) 2022/2464). Official Journal of the European Union 405/1. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2464&qid=1713287906082>

[19] European Union. (2023). Regulation (EU) 2023/2772 of the European Parliament as regards sustainability reporting standards. Official Journal of the European Union 330. Retrieved from <https://eur-lex.europa.eu/eli/reg_del/2023/2772/oj/eng>

[20] Čufar, M., Belak, J., & Primec, A. (2024). Beyond Compliance: Leveraging NFRD and CSRD for Transformative CSR Reporting and Stakeholder Empowerment in Corporate Governance. In M. Pucelj & R. Bohinc (Eds Corporate Governance and CSR Strategies for Sustainability. (pp. 233-261). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-5863-4.ch010>

[21] Stearns, C. Z., & Stearns, P. N. (1986). Emotionology: Anger. The struggle for emotional control in America’s history. Chicago, IL: University of Chicago Press. <https://doi.org/10.1126/science.234.4782.1447>

[22] Picard, R. W. (1997). Affective computing. MIT Press. https://doi.org/10.7551/mitpress/1140.001.0001

[23] McKee, A. (2003). Textual Analysis: A Beginner’s Guide. Sage

[24] Amani, F. and Fadlalla, A. (2017). Data mining applications in accounting: A literature review and organizational framework. International Journal of Accounting Information Systems, 24, 32-58. <https://doi.org/10.1016/j.accinf.2016.12.004>

[25] Mohammadrezaei, M., Marques, J. C., & Huq, A. (2024). Use of text mining and natural language processing techniques in analyzing sustainability reports: A systematic literature review and assessment. <http://dx.doi.org/10.2139/ssrn.5071032>

[26] Van Rossum, G.; Drake, F.L. Python 3 Reference Manual; CreateSpace: Scotts Valley, CA, USA, 2009; ISBN 978-1-4414-1269-0.

[27] Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing; To appear, 2017.

[28] Chapman, M. Whoosh: Fast, Featureful Full-Text Indexing and Searching Library Implemented in Pure Python; Available online: <https://whoosh.readthedocs.io/en/latest/>

[29] Bird, S.; Klein, E.; Loper, E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.

[30] Task Force on Climate-related Financial Disclosures (TCFD). (2017-2023). Recommendations of the Task Force on Climate-related Financial Disclosures. Retrieved from <https://assets.bbhub.io/company/sites/60/2021/10/FINAL-2017-TCFD-Report.pdf>

[31] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.

[32] Beck, A. C., Campbell, D., & Shrives, P. J. (2010). Content analysis in environmental reporting research: Enrichment and testing of the method in a British-German context. British Accounting Review, 42(2), 207-222. <https://doi.org/10.1016/j.bar.2010.05.002>

[33] Halme, M., & Huse, M. (1997). The influence of corporate governance, industry, and country factors on environmental reporting. Scandinavian Journal of Management, 13(2), 137-157.. <https://doi.org/10.1016/S0956-5221(97)00002-X>

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.