ALMA MATER STUDIORUM

UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE

AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Natural Language Processing

BERT goes sustainable: an NLP approach to ESG financing

CANDIDATE	SUPERVISOR	CO-SUPERVISOR
Nicolaas Ruberg	Paolo Torroni	Vanessa Almeida (BNDES)

Academic year 2020-2021

Session 3rd

"The value of things is not the time they last, but the intensity with which they occur.
That is why there are unforgettable moments and unique people!"
Fernando Pessoa

To Gabriela, Sylvia, and Alice, for whom
every thought of mine has your names on it!

Summary

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my wife for the partnership and contributions to my studies at the master's and for being a safe haven where I found comfort and regained strength to go on. To my two lovely daughters, Alice and Sylvia, that moved to a different country, learned a foreign language and culture, and they were all joy and smiles for this adventure.

My sincere gratitude to the Brazilian Development Bank (BNDES) for the leave and for sponsoring my Master in Artificial Intelligence through the Data Science program. A particular thanks to Roney Lorentz Oliveira for the incentive on the proposal elaboration at BNDES. To my friend and co-supervisor, Vanessa da Rocha Santos Almeida, for the directions and contributions to the thesis work.

To my friends at BNDES Leonardo Cardoso, Rodrigo Leão e Felipe Curty for the motivation and support.

I would like to express my appreciation to my thesis supervisor Dr. Paolo Torroni for his guidance, motivating words, and continuous encouragement, from the early stage of this work to its very conclusion. I express my thanks: "Muito Obrigado!"

To Dr. Andrea Galassi and Dr. Federico Ruggeri for the valuable comments.

Special gratitude to the faculty of the Masters in Artificial Intelligence at UNIBO. For all the transmitted knowledge and also for the courses that pushed me to my best.

To my classmates and friends, Dr. Luca Lorello Salvatore, Dr. Giorgio Tsiotas, and Dr. Muhammad Salman Razzaq, I won't forget our group studies and time shared together at the university classrooms, Villa Spada Library, and on the meetings on Teams during the lockdown. We did an excellent team together!

A deep sense of gratitude to my father, a great inspiration which steps I try to follow, a Master in AI is one of those. To my mother, who raised me with commitment and dedication, values that were essential to accomplish this work.

Last, my deepest thanks to God and its mysterious ways to guide me in life, giving me more than I deserve and turning the bitter moments into lessons that make me a better person.

Abstract

Environmental, Social, and Governance (ESG) factors are a strategic topic for investors and financing institutions like the Brazilian Development Bank (BNDES). Currently, the Brazilian bank's experts are developing a framework based on those factors to assess companies' sustainable financing. In this work, we identify an opportunity to use Natural Language Processing (NLP) in this development. This opportunity arises from the observation that a critical document to the ESG analysis is the company annual activity report. This document undergoes a manual screening, and later it is decomposed, and its parts are redirected to specialists' analysis. Therefore, the screening process would largely benefit from NLP to automate the classification of text excerpts from the annual report.

The proposed automation solution is based on different Bidirectional Encoder Representations from Transformers (BERT) architectures, which rely on the attention mechanism to achieve optimal results on sentence-level analysis tasks. We devised a text classification task to enable the analysis of excerpts from the annual activity report of companies considering three categories, according to the ESG reference standard, the Global Reporting Initiative (GRI).

We produced the training and validation sets from Brazilian companies' annual reports from the GRI database to validate our experiments. To establish a benchmark, we implemented a baseline solution using a classic NLP approach, Naïve Bayes, which got a 51% accuracy and 50,33% F1-score. RoBERTa and BERT-large achieved 88% accuracy and almost 85% F1-score, the best results obtained from our experiments with different BERT architectures. Also, Albert showed to be a possible alternative for limited memory devices, with 85% accuracy and 78.5924% F1-score.

Finally, we experimented with a multilingual setup that would be interesting for a scenario where the BNDES wants a more generic model that can analyze English or Portuguese annual reports. Bert multilingual model reached almost 86% accuracy and 81.18% F1-score.

The proposed methodology to the GRI text classification and the BERT model selection for the ESG analysis of annual activity reports are significant contributions presented in this dissertation, aiming to improve the BNDES ESG framework substantially.

1 Introduction

1.1 Motivation

In recent years, there has been a shift in society's attitudes toward sustainability. This shift is impelling political pressure, a regulatory push, and technological advancements to build a more sustainable world economy. Consequently, a change in the investors' behavior is taking place, thus guiding a gradual capital reallocation (Hildebrand, Polk, Deese, & Boivin, 2020). In this new investment scenario, it is becoming crucial for investors to assess companies' sustainability (Napoletano & Curry, 2021).

Evaluating sustainability encompasses analyzing a company's initiatives, decisions, and actions concerning the Environmental, Social, and Governance (ESG) factors that improve long-term outcomes for the planet and future generations (Inderst & Stewart, 2018). This analysis involves a myriad of aspects with both quantitative and qualitative information, and experts usually carry it out. Supporting decision-making in sustainable investing is a difficult task that currently lacks proper tools to automatically assess ESG factors, particularly regarding analyzing qualitative information available on companies' reports.

Also, it is worth noting that this ESG analysis depends on the investor's perspective. For example, an individual can be more interested in the company's environmental impact on his community or region. At the same time, a larger investor – such as a hedge or pension fund, might address a company's sustainability by focusing on its governance and long-term sustainability. This same analysis is also needed for banks and government investment agencies. However, since these entities are strongly controlled and regulated, their focus usually relies on well-established sustainable standards to guide their investment decisions.

In this context, the Brazilian Development Bank (BNDES), as the major government investment agency in Brazil, is broadening its sustainability investment agenda by carrying out several projects focusing on ESG (BNDES, Sustainable Development, 2021). In 2017, BNDES was the first Brazilian Bank to issue a USD 1 billion green bond in the international market. Later, in 2020, BNDES became the first Brazilian financial institution to issue a Green Financial Note ("Letra Financeira Verde") in the domestic market, which proceeds finance expenditures related to wind and solar power plants. In April 2021, it released a Sustainability Bond Framework (SBF), in accordance with the Green, Social, and Sustainability Bond Principles (BNDES, BNDES Sustainability Bond Framework, 2021), under which it has set

the guidelines to issue Green, Social, and Sustainability Bonds in Brazil and abroad, as well as the use of proceeds regarding finance and refinance expenditures.

In order to grant a loan or finance a project under this sustainable financing umbrella, the BNDES has to perform a sustainability assessment of the companies, which requires an expert understanding of the loaner companies' ESG initiatives, actions, and strategies. A key input to this assessment process is the annual activity report, a document in which companies state the activities developed in the preceding year. Nevertheless, the ESG analysis of annual activity reports is usually time-consuming and subject to bias depending on the analyst's background and knowledge.

1.2 Thesis objectives

This thesis proposes the application of natural language processing (NLP) techniques to improve the ESG analysis of available qualitative information in the sustainability assessment process. Considering sustainable financing at the most significant Brazilian development bank, we target at the ESG analysis carried out by expert screeners on companies' reports, which is a task vulnerable to bias and performance bottlenecks.

We propose an automatic ESG classifier that uses an NLP technique based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2018), a method that relies on attention to achieve optimal results on sentence-level tasks. To enable the ESG understanding of companies' annual activity reports, we explore a well-established ESG standard for impact reporting, the Global Reporting Initiative (GRI) (Wikipedia, Global Reporting Initiative, 2021).

In this context, we investigate the main English BERT-like architectures, namely: BERT (Devlin, Chang, Lee, & Toutanova, 2018), RoBERTa, (Liu, et al., 2019), ELECTRA (Clark, Luong, Le, & Manning, 2020), DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2019) and ALBERT (Lan, et al., 2019). We examine these architectures with different model sizes based on their accuracy and F1-score and compare them to the results obtained with a baseline Naïve-Bayes classifier. Moreover, we consider two multilingual models to enable screening reports written in Portuguese, the Brazilian official language.

To evaluate the proposed solution, we set up several experiments using a collection of annual activity reports for 31 companies from the GRI public database[1], available in PDF

[1] https://www.globalreporting.org/how-to-use-the-gri-standards/register-your-report/

format. We developed a scraper to produce the training and validation sets by extracting document excerpts and their corresponding ESG categories from PDF files.

1.3 Text organization

This thesis is structured as follows. In Section 2, we motivate the relevance of the sustainable agenda in credit concession and financial project analysis. We outline the ESG factors used to assess companies' sustainability and the importance of the GRI standard for understanding these factors in companies' reports. We contextualized the role of the BNDES in the Brazilian financial system and its current agenda on sustainable finance.

Section 3 presents BNDES' workflow process to analyze ESG financing and propose an approach to incorporate NLP tasks into this workflow. We describe the role of the document screener and detail the main characteristics of the GRI compliant annual activity reports, their usual structure in the PDF format, and the GRI classification codes used in these reports.

In Section 4, we address the problem of text classification and explain the NLP techniques considered in this thesis. We describe the traditional Naïve-Bayes classifier and the BERT transfer learning technique's main implementation architectures and models.

In Section 5, we detail our experiments with a setup of 31 companies' annual activity reports from the GRI database. We present the scraping method to produce the training and validation sets of GRI annotated paragraphs and describe the baseline experiment using a Naïve-Bayes classifier as the minimum expected benchmark for the BERT experiments. We analyze our experiments and their results in terms of accuracy and F1-score and devise some guidelines for the BNDES in choosing the best solution to improve the ESG classification workflow.

Finally, Section 6 concludes by presenting this thesis's main results and contributions, the related work, and some improvement perspectives and research directions.

2 Sustainable Financing

Next, we describe the importance of Environment, Social, and Governance (ESG) investment and detail the Global Reporting Initiative (GRI) standard, which is an independent international standard that helps companies to be more transparent about their economic, social, and environmental impacts (Wikipedia, Global Reporting Initiative, 2021). Then, we introduce the Brazilian Development Bank (BNDES) and its role as a financing agent in the economy of Brazil. Finally, we contextualize the BNDES agenda to foster ESG financing using the GRI standard.

2.1 ESG Criteria

According to FORBES magazine (Napoletano & Curry, 2021), environmental, social, and governance (ESG) investing is a strategy where socially conscious investors want to put their money to work with companies that strive to make the world a better place.

ESG investing relies on independent ratings that help investors assess a company's behavior and policies regarding environmental performance, social impact, and governance issues. In other words, ESG investing aims to influence positive changes in society with the invested money. Thus, ESG investing fosters companies and corporations with good scores on scales of factors on environmental, societal, and governance responsibility. Third-party entities determine those scores.

Those entities approach the three criteria used to evaluate companies for ESG investing (Napoletano & Curry, 2021):

- **Environment** – what kind of impact does a company have on the environment? This can include a company's carbon footprint, toxic chemicals, water usage involved in its manufacturing processes, and sustainability efforts that make up its supply chain;
- **Social** – How does the company improve its social impact, both within the company and in the broader community? Social factors include everything from LGBTQ+ equality, racial diversity in both the executive suite and staff overall, and inclusion programs and hiring practices. It even looks at how a company advocates for social good in the broader world beyond its limited sphere of business; and

- **Governance** – How do the company's board and management drive positive change? Governance includes everything from issues surrounding executive pay to diversity in leadership and how well that leadership responds to and interacts with shareholders.

It is worth noting that these three criteria encompass a broad range of behaviors and policies.

ESG criteria can also help to avoid investing in companies with high financial risks due to environmental or other practices related to social responsibility.

Nowadays, there are several independent entities that provide objectives sustainable reports standards. (Meager, 2021) describes the most relevant ones. Among them, the Global Reporting Initiative (known as GRI) is the oldest one with the most significant number of participants.

2.2 The Global Report Initiative (GRI)

The GRI is an international independent standards organization that helps businesses, governments, and other organizations understand and communicate their impacts on issues such as climate change, human rights, and corruption (Wikipedia, Global Reporting Initiative, 2021). In other words, GRI assesses companies and organizations on scales of factors on ESG criteria and produces a score.

Nowadays, different groups of stakeholders like governments, clients, investors, and organizations like the Organization for Economic Co-operation and Development (OECD) and the World Economic Forum (WEF) require that companies be more transparent about the environmental, economic, and social impacts of their business.

Therefore, in September 2020, the WEF developed a core set of common metrics and disclosures on non-financial factors for their investors and other stakeholders; those metrics became the basis of several standards.

To address that, many companies regularly publish a sustainability report, also known as a corporate social responsibility (CSR) or environmental, social, and governance (ESG) report. Companies usually publish this CSR report on their annual activity report. To guide the production of the report, the GRI's framework for sustainability reporting helps companies identify, gather, and report this information in a transparent and comparable manner. Launched in year two thousand, GRI's sustainability reporting framework is now the most widely used

by multinational organizations, governments, small and medium enterprises (SMEs), Non-Governmental Organizations (NGOs), and industry groups in more than 90 countries.

2.3 GRI dimensions on an ESG analysis

The Global Sustainability Standards Board (GSSB) developed the GRI standards for sustainability reporting and launched them in October 2016. These reporting standards have been designed with a modular structure to keep the company dynamic up to date.

The Standards are designed as an easy-to-use modular set, starting with the Universal Standards. Topic Standards are then selected based on the organization's material topics – economic, environmental, or social. This process ensures that the sustainability report provides an inclusive picture of content topics, their related impacts, and how they are managed (GRI, 2021).

As mentioned before, GRI reporting is about companies' disclosure information in a standard way so that stakeholders can compare information concerning economic, environmental, and social performance. Moreover, the fact of clearly reporting it, by taking a systematic approach, strategies to improve the company shows up, like:

- Improved sustainability performance.
- Improved risk management and investor communications.
- Engagement with stakeholders and improved stakeholder relations.
- Motivated and engaged employees.
- More substantial credibility as a committed and effective corporate citizen.
- More robust internal data management and reporting systems.
- Improved sustainability strategy and selection of performance indicators and targets.
- Develop a benchmark sustainability performance.

A set of interrelated modules structure the GRI Standards that can be referenced and used together. In Figure 1, we observe that the standard is divided into "universal standards" and "topic-specific standards." The universal standards are known as GRI's 100 series, as the topic-specific standards are the series 200, 300, and 400.

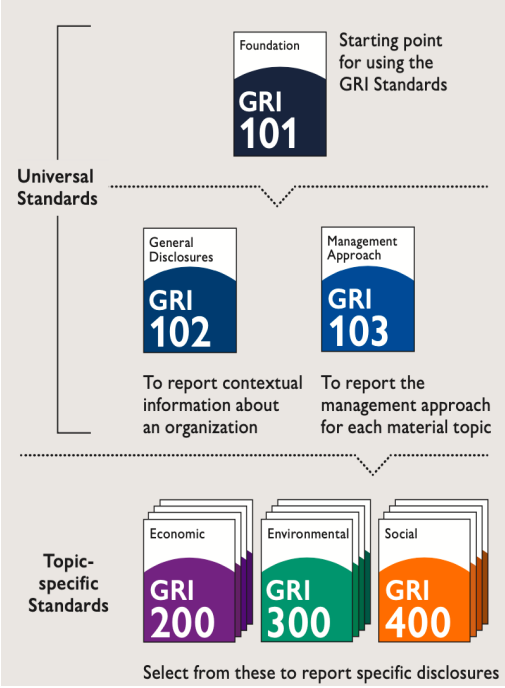

Figure 1– GRI series

This GRI 100 series are universal standards that apply to every organization preparing a sustainability report, addresses the company foundation, public disclosures, and management approach.

On the topic-specific standard GRI's 200 series, the company should report its economic dimension, i.e., the economic performance, market presence, indirect impact of its activities, procurement practices, anti-corruption efforts and behavior, and tax disclosure. In Table 1, we present all the topics for the 200 series.

When reporting the GRI 300 series topic-specific standard, the company reports their environmental situation and strategy, presenting the relevant aspects to the standard on used and produced materials, energy and water use and management, impacts and initiatives on biodiversity, gas emissions and strategy, waste management, overall environment compliance, and suppliers' environmental assessment. Table 2 details the topics list addressed by GRI 300 series.

Last, the GRI 400 series topic-specific standard addresses the company social aspects, in the following perspectives: employees (labor relations, health, and safety, education, opportunity, labor union, etc.); human rights for native peoples and local communities; suppliers, customer and government relations; and socioeconomic compliance. In Table 3, we present the category applied to each topic of the GRI 400 series.

Table 1 – GRI 200 detailed topics.

GRI 200: Economic	
GRI 201	Economic Performance
GRI 202	Market Presence
GRI 203	Indirect Economic Impacts
GRI 204	Procurement Practices
GRI 205	Anti-corruption
GRI 206	Anti-competitive Behavior
GRI 207	Tax

Table 2 – GRI 300 detailed topics.

GRI 300: Environmental	
GRI 301	Materials
GRI 302	Energy
GRI 303	Water and Effluents
GRI 304	Biodiversity
GRI 305	Emissions
GRI 306	Waste
GRI 307	Environmental Compliance
GRI 308	Supplier Environmental Assessment

Table 3 – GRI 400 detailed topics.

GRI 400: Social	
GRI 401	Employment
GRI 402	Labor/Management Relations
GRI 403	Occupational Health and Safety
GRI 404	Training and Education
GRI 405	Diversity and Equal Opportunity
GRI 406	Non-discrimination

GRI 407	Freedom of Association and Collective Bargaining
GRI 408	Child Labor
GRI 409	Forced or Compulsory Labor
GRI 410	Security Practices
GRI 411	Rights of Indigenous Peoples
GRI 412	Human Rights Assessment
GRI 413	Local Communities
GRI 414	Supplier Social Assessment
GRI 415	Public Policy
GRI 416	Customer Health and Safety
GRI 417	Marketing and Labeling
GRI 418	Customer Privacy
GRI 419	Socioeconomic Compliance

After that, the GRI standard details subtopics to provide a complete evaluation of the disclosures on each topic. And for the disclosures, the GRI standard gives guidance on the requirements to be present on the report and recommendations to what should be considered on the assessment. For example, in Figure 2, we extract the GRI guidance to report GRI 302 topic on its disclosure GRI 302-4.

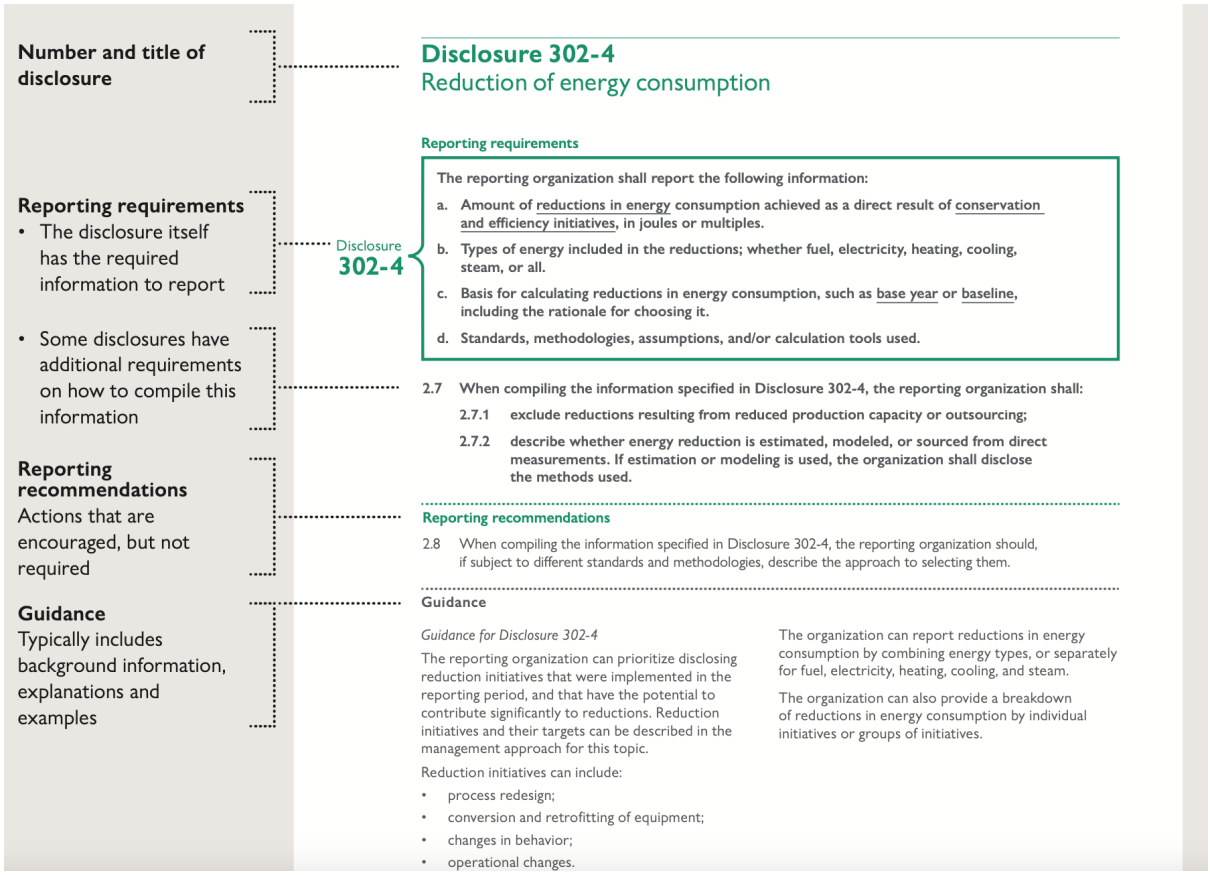

Figure 2 – Disclosure for a GRI topic-specific.

2.4 Sustainable financing at the Brazilian Development Bank

The Brazilian Development Bank (BNDES) is the leading financing agent for development in Brazil. Since its foundation in 1952, the BNDES has played a fundamental role in stimulating the expansion of industry and infrastructure in the country. More recently, the Bank expanded its activities in other areas, like exports of Brazilian goods, products and services, technological innovation, sustainable socio-environmental development, and public administration modernization.

The Bank offers several financial support mechanisms to Brazilian companies of all sizes and public administration entities, enabling investments in all economic sectors. In any supported financing, from the initial phase, where the project is assessed and discussed with the client, up to its implementation, the BNDES emphasizes three factors that are considered strategic: innovation, local development, and socio-environmental development.

In the 21st century, the BNDES aligns its operations with the reality of a globalized world, with economies deeply connected, and intensifies its efforts to take on roles and duties that surpass Brazilian borders, aligned with the increase of Brazil market insertion on the global

market. The BNDES also finances the expansion of national companies far beyond the country's borders and seeks to diversify the sources of its resources on the international market. In addition, the BNDES has strengthened its core activities, such as financing exports of Brazilian goods and services in projects carried out overseas and institutional fundraising through multilateral organizations, cooperating with international institutions to share experience and promote new opportunities to Brazilian companies.

Therefore, with its extensive knowledge, stemming from its vast experience allied with its technical and highly skilled workforce, the BNDES also plays the role of a fundamental partner for investors who want to understand and access opportunities offered by the Brazilian economy. And, with the expertise acquired after supporting Brazilian economic growth since it was created, through financing relevant investments, the BNDES reaches a new operational level, on a worldwide scale, where it consolidates its role as financing agent to the economic and social development of our country (BNDES, The Brazilian Development Bank, 2021).

BNDES is developing several projects focusing on ESG, such as: fundraising with the New Development Bank (NDB) (BNDES, Sustainable financing solutions, 2021) to create a climate change fund; developing an ESG framework for issuing ESG bonds; and contracting a third-party rating agency to assess BNDES ESG compliance.

As part of BNDES's effort to get involved in GRI reporting, the BNDES 2020 Annual Activity Report highlights the effort in broadening the agenda on sustainable financing. The report presents BNDES’s strategy and performance relating to the United Nations Sustainable Development Goals and reports them consistently to the GRI standard. This change emphasizes the impact of the reported information on social and environmental actions, aligned with its effort in fostering a sustainable agenda.

BNDES's effort to develop an ESG framework will enable a better and unified assessment of companies and investment projects, emphasizing sustainability. The reference framework adopted was the GRI standard.

To develop this ESG framework, BNDES must address the issue of competencies and expertise in analyzing sustainable aspects according to the GRI standard. Consequently, capturing the sustainable characteristics of loaners, beneficiaries, and the client’s annual activity report is critical, which means that the BNDES right specialist has to be involved in the company’s analysis for the framework's success.

Next, we detail the ESG process and framework that is being constructed by BNDES to better assess companies, clients, and products on their sustainable aspects based on the GRI standard.

3 ESG analysis process and corpus

The BNDES effort to develop an ESG framework is crucial for a better and unified assessment of companies and investment projects, emphasizing the sustainability perspective on investments decisions. On the design of this framework, BNDES adopted the use of the GRI standard as a guideline, where a team of BNDES specialists should assess loaners, beneficiaries, and clients' annual activity report to observe sustainable aspects on them.

Consequently, a vital part of the ESG process and analysis is to involve the right BNDES specialist in the company's assessment. Next, we present the BNDES current workflow designed for assessing ESG aspects.

3.1 The ESG analysis workflow

The current BNDES analysis workflow of an annual activity report from its beneficiaries, clients, or partners is divided into two main phases: the document screening phase; and the specialist evaluation phase. The outcome of this two-phase is an ESG analysis produced for further considerations in the concession of a loan, investment decision, or credit rating.

In Figure 3, we depict this two-phase process having as a final product an ESG analysis. From the Figure, we can observe how vital the screener's role is since it is responsible for pre-classifying and redirecting parts of the report to specialists for further analysis. Also, in the first phase, we notice that the screener is a bottleneck to the workflow process. Hence, the performance of the whole process depends on his ability to redirect the specific sections or items on the activity report to the right specialist. Moreover, the skill of the screener is acquired over time and is dependent on the person conducting the process. Also, new BNDES screeners or changes on the screener during an investment (company) evaluation can significantly increase an assessment time.

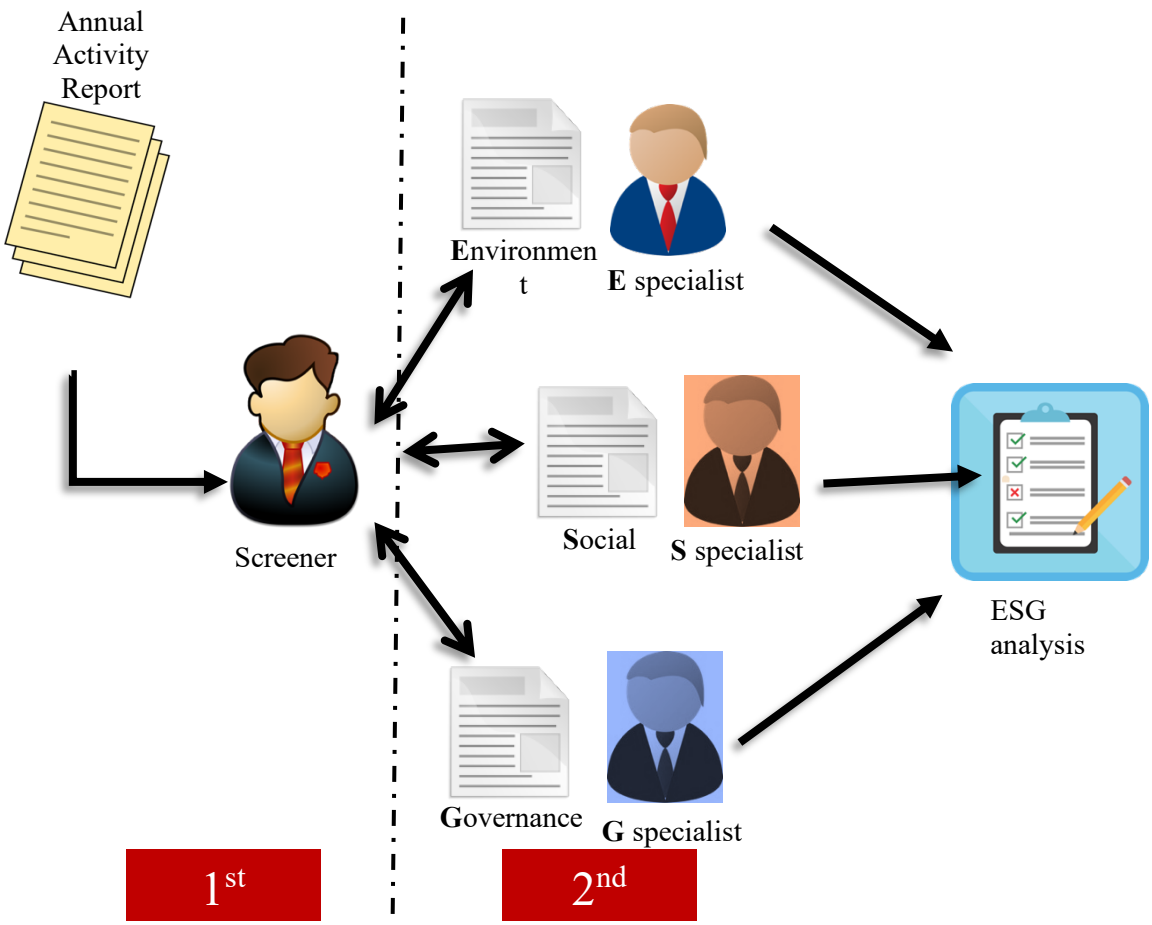

Figure 3 – BNDES specialists and the ESG workflow analysis.

Moreover, in the second phase, we can observe from the workflow process, BNDES needs several competencies to achieve an ESG beneficiary final assessment. And each competence (Environment, Governance, and Social) is a broad domain segment, which means that a specialist in one domain is not necessarily a good specialist in another one. So, a good quality screening will minimize the specialist effort and reduce the time of his analysis, consequently reducing the overall analysis time.

3.2 Proposed NLP workflow analysis

As introduced in Section 3.1, it is crucial for a good result of the beneficiary ESG analysis the screener role, since his work's quality and efficiency depend on the final quality of the analysis in the whole workflow. Besides, the screener analysis's efficiency directly impacts the overall process time since it is a bottleneck to the process.

Therefore, we propose automatizing the screener role with a Natural Language Processing for the beneficiary annual activity report. In Figure 4, we present the revisited BNDES workflow analysis by introducing an automated NLP model for the GRI screening problem.

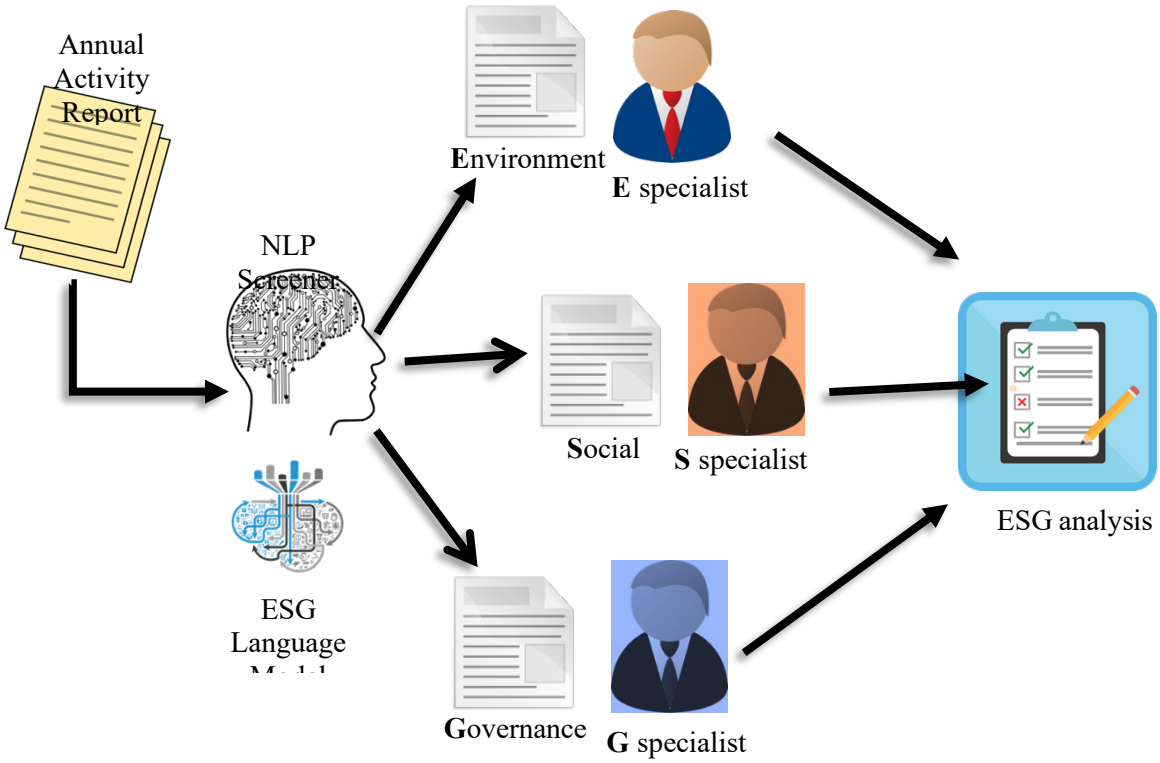

Figure 4 – NLP screener role in BNDES beneficiary ESG analysis.

In this new workflow process, the NLP screener is supported by a GRI Language Model, trained on a set of annual activity reports in compliance with the GRI standard database.

Taking a closer look at this workflow process, one can observe that the proposed workflow misses the feedback from the expert to the screener. So, to incorporate this feedback into our automated process, a periodic enrichment of our GRI Language Model can address this limitation. In Figure 5, we present the final workflow process.

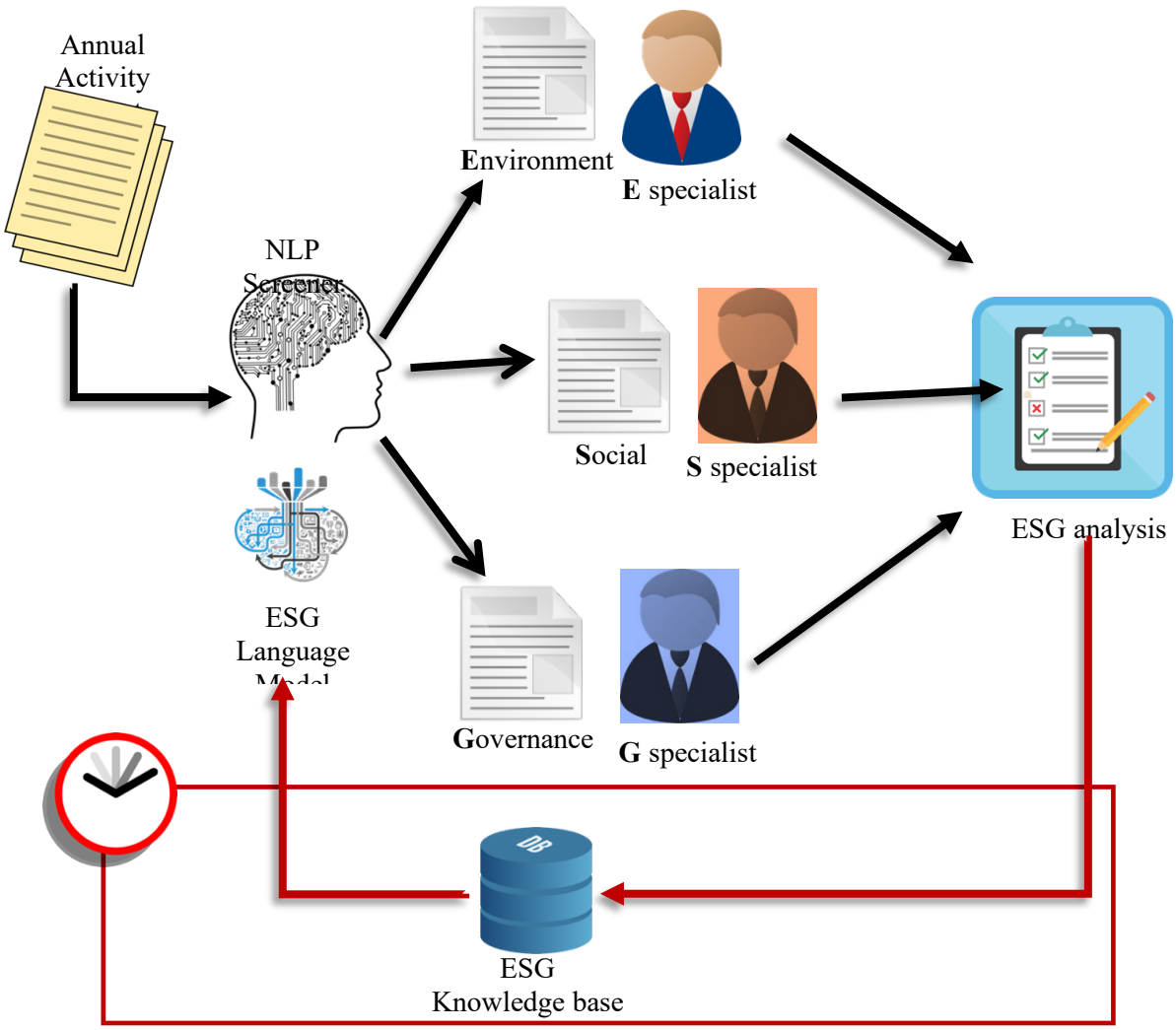

Figure 5 – Feedback on the BNDES beneficiary ESG analysis workflow.

3.3 The Annual Activity Report

The annual activity report, or annual report, is a survey of the company's performance in the preceding year. Those reports are meant to provide information to shareholders and stakeholders about their activities and financial performance. Although it is not a consensus that every company should produce an annual report, more and more compliance agencies, market regulators, and financing institutions require the document. Moreover, if a company is listed on a stock exchange, they are required to produce an annual report (Wikipedia, Annual Report, 14).

An essential requirement for the annual activity report is that the document presentation must be pleasant and easy to read. The desired audience are stakeholders, shareholders, government agencies and units, partners, and society. This audience is broad and not always

composed of specialists, a different audience from those expected to read the companies' financial statements, generally accountants and market experts. Consequently, to produce a document that is easy to read and portable, companies choose PDF files to present their annual activity reports.

3.4 Data description

As mentioned before, annual activity reports are PDF files. They are considered an ideal digital alternative for paper-based documents since they have excellent compatibility across devices and operating systems. So, they are heavily assumed for exchanging digital business documents. The key advantage of PDFs is that they are portable, platform-independent, and human-readable. However, this format is unstructured, making it difficult to access further analysis (i.e., extract data).

To extract data and information from PDFs for other computing processes is called scraping. PDF scraping has been researched for many years; for example, in 2001 (Bergmark, Phempoonpanich, & Zhao, 2001) tried to analyze ACM Digital Library to extract reference links from the library documents. And more recently, (Lu & Unpingco, 2021; HuggingFace, The AI community building the future, 2021) scraped PDF files to extract patients' data to prepopulate the COVID-19 vaccination forms. These examples show that PDF scraping depends on the target document, so it is necessary to understand some of the document structure to extract its data adequately.

For that, next, we present more details on the GRI annual reports that companies provide.

3.4.1 GRI annual reports in PDF files

Annually, companies and corporations present their annual activity reports to shareholders, stakeholders, clients, and society. As this document target a broad audience, it is provided as a PDF file. To get an idea of such a document, in Figure 6, we present an extract of 2017 Vale company, a mineral and coal Brazilian mining company. It is four pages out of 179 of its annual report. It calls attention, the visual presentation and how the text is formatted to look pleasant to the reader, with columns, graphs, and photos.

Though PDF files allow a nice visualization, regarding the content, when observing the adherence of a document to the GRI standard, a topic related to a disclosure (GRI topic reference) should be identified on the document. In Figure 7, we observe a page from the 2017 Vale annual report citing GRI concepts 102-10, 102-11, and 201-2. In this example, the GRI

disclosure code introduces the text that refers to the concept. Also, on the reports, the concept can appear within the text or paragraph.

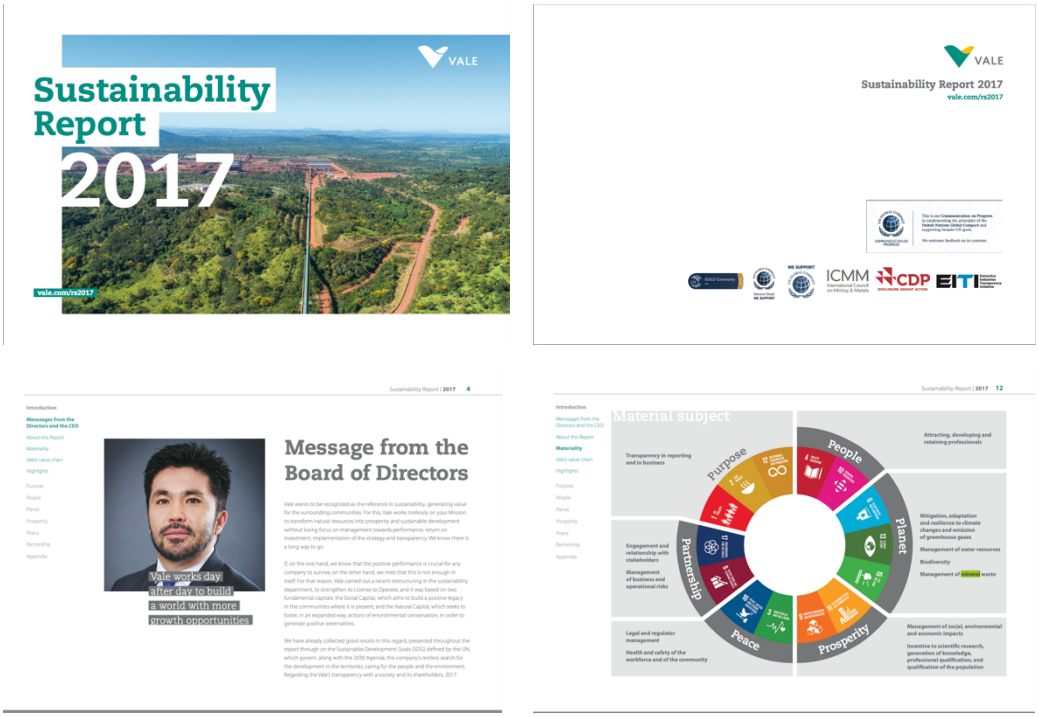

Figure 6 – Vale company PDF annual activity report extract.

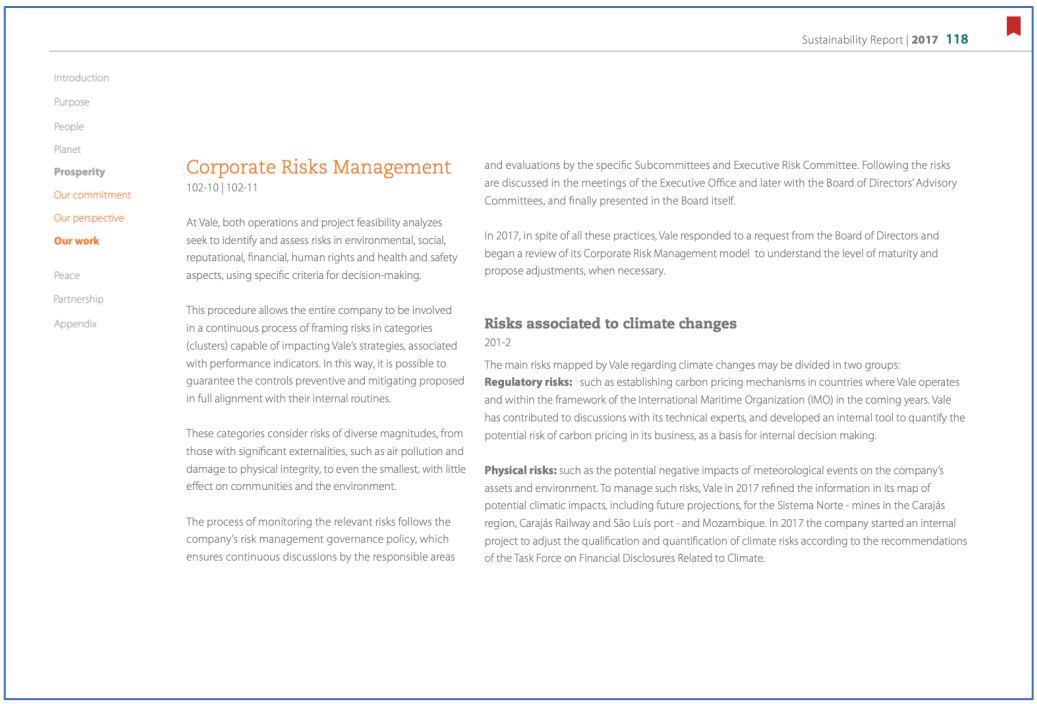

Figure 7 – Page describing Vale's GRI 201-2 topic.

Understanding the text structure is crucial to designing a good scraper for the annual report PDF files and correctly annotating the text and paragraphs for future text classification processing. Next, we introduce the concepts of Natural Language Processing for text classification, which are the theoretical basis to address the classification task to solve our GRI screening problem.

4 Natural language processing and text classification

Natural language processing (NLP) exploits knowledge of language and computation by building useful technologies (Bird, Klein, & Loper, 2009). To devise this language knowledge, NLP relies on Artificial Intelligence (AI) to understand and interpret human language, enabling new tools and technologies.

NLP is present in several day-to-day applications, like text autocompletion, in most search engines to predict the sequel of query and provide multiple options that the user can select. Or in chatbots, a technology that uses AI to facilitate written online interactions and conversations between computers and humans, such use is standard on e-commerce websites to provide customer assistance (like Amazon, e-Bay, Alibaba, etc.). Another example of NLP on our day-to-day use is digital voice assistance like Apple Siri, Amazon Alexa, and Microsoft Cortana, which are voice-activated chatbots that use speech rather than text.

Moreover, the field of NLP covers a broad range of tasks and methods like information extraction, concept discovery and representation (ontologies), speech processing, text translation, document summarization, etc.

An NLP particular field of our interest is text classification or text categorization, which is the processing that assigns the text documents into a set of predefined text categories; hence, it is an automated process for classifying text into predefined categories. In other words, for a given text, the text classification goal is to assign a discrete label from a set of possible tags. To bring up a more concrete example, recently, on social media and e-commerce websites is vital to understand the customers' feelings about a product, a post, or a comment. Therefore, classifying customer feedback into good, bad, or neutral is essential for the seller or service provider to make future decisions. So, this text classification is called sentiment analysis, an NLP task that, given a text, classifies it into the categories: good, bad, or neutral (the costumers' sentiment or feeling).

4.1 Text representation and Naïve Bayes for text classification

Before going deeper into the classification techniques, a basic understanding of how NLP technologies can represent a document is necessary. The simple representation for a text is the bag of words. The document is described as a vector of words counts, i.e., every vector cell represents a word, and its content expresses the number of occurrences of that word in the document or corpora. A remark on this technique is that this representation loses the order in which the words appear in the text. Nevertheless, this approach can be pretty efficient in text

classification since particular words can be strong predictors; for example, in a movie context, the word “X-men” can be a strong predictor that we are referring to a Marvel movie.

(Dhar, Mukherjee, Dash, & Kaushik, 2021) presents a survey of a wide variety of algorithms used for categorizing text documents. On the conventional methods, among the various classification algorithms, it is worth mentioning: Naïve Bayes (NB), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). On the novel techniques, we can list Recurrent Neural Networks (RNN), long short-term memory (LSTM), gated recurrent networks, and Transformers (Vaswani, et al., 2017). In particular, Transformers are the building block of a series of powerful transfer learning techniques like Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformers 2 (GPT-2), and more recently Generative Pre-trained Transformers 3 (GPT-3). The first one, BERT, was designed by Google, while the other two were designed by OpenAI, where the last version, GPT-3 was released in 2020 (Brown, et al., 2020).

Next, we present details of the techniques used throughout our work on the GRI classification problem. And our first approach to text classification is the Naive Bayes model, from where we will devise the baseline reference to the experiments. Later, we describe Transformers, BERT, and some of its architecture variations, also used in our experiments.

4.1.1 Naïve Bayes

The Bayes theorem states that a conditional probability is a probability that something will happen, given that something else has already occurred. The conditional probability can give us the probability of an event using its prior knowledge. So, the conditional probability states that:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)},$$

Where *P(A)* is the probability of the hypothesis being true, the prior probability. *P(B)* is the probability of the evidence. *P(A|B)* is the probability of the evidence given that the hypothesis is true. And P*(B|A)* is the probability of the hypothesis given that the evidence is true.

Then, using the Bayes theorem, we can develop a naïve Bayes classifier that segregates different objects based on certain features of variables, where the goal is to predict a tag or class from a text extract. So, the probability is computed for a given sample, and then the tag with the highest chance is selected.

The Naïve Bayes (NB) approach makes a simple assumption on how the features interact. Since the document text is represented as an unordered set of words, and for each word, its frequency is annotated. This strategy losses the order that words appear in the document; consequently, it is considered a bag of words.

Next, we introduce some concepts and more sophisticated strategies for NLP and text classification.

4.2 Transformers, Bert, and HuggingFace

An essential aspect of natural language processing (NLP) is that communication and language are temporal phenomena; hence, a sequence of input streams needs to be sequentially processed to understand the written text or verbal communication. Traditional machine learning techniques for NLP like Naïve Bayes, Support Vector Machines, and K-Nearest Neighbors don't have this temporal nature. As mentioned in Section 4.1, the bag of word model and/or similar models are used by those techniques; with those models, the input is accessed simultaneously.

Even feedforward neural networks, which had great success in computer vision and image processing, do not capture this temporal aspect. These fully connected networks use fixed-size inputs and associated weights to capture all the relevant parts of an example at once. This simultaneity makes it challenging to deal with sequences of varying lengths and fails to capture relevant temporal aspects of natural language (Martin & Jurafsky, 2008).

A neural network architecture that can capture a preceding sequence is a Recurrent Neural Network (RNN). This property is achieved due to a hidden layer with a recurrent link; this link provides some memory from previous states. This memory turned RNN into a trend in NLP, particularly in tasks like named-entity recognition (NER) and part of speech (POS) tagging. In Figure 8, we present a standard RNN structure; we can observe that the hidden layers are updated according to the information received from the input layer and the activation from the previous forward propagation. In Figure 8, *Wx* is the weight matrix connecting the input layer and the hidden layer; *Wh* is the weight matrix connecting two consecutive hidden states*; Wy* weight matrix connecting the hidden state and the output layer*; xt* is the input at *t* time; and *yt* is the predicted output at time *t* (Fang, Xu, & Xu, 2019).

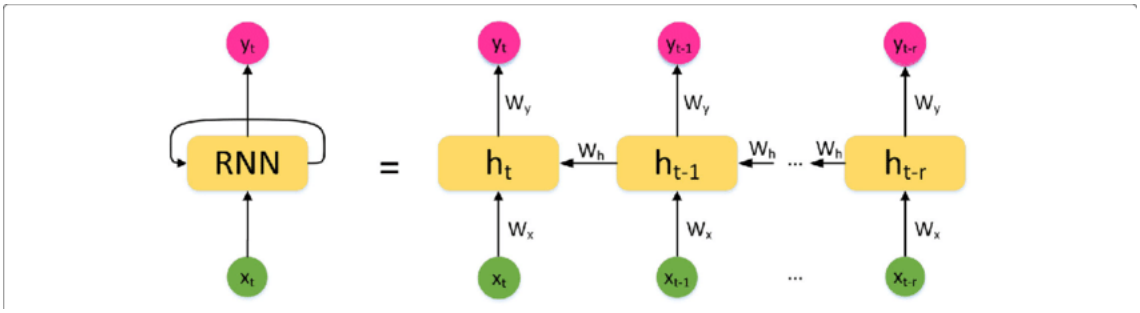

Figure 8 – Standard RNN structure[2].

Nevertheless, RNNs only capture recent information, not lingering the communication context. We need a different neural network architecture to improve the retention problem: the Long-Short Term Memory Cell (LSTM). LSTM networks have memory cells instead of a hidden layer update. This kind of architecture is better at finding and exposing long-range dependencies in data which is imperative for sentence structures (Bajpai, 2019).

The drawback of RNNs and LSTMs is the computational factor and memory use derived from the input retention to generate the output sequence. Several hidden states are needed to keep track of previous states sequentially due to the alignment of the hidden states as a function of the prior state.

To reduce the number of sequential computations, a more recent approach is the mechanism of self-attention that relates different positions of a single sequence to compute a representation of the sequence. Transformer models (Vaswani, et al., 2017) mainly explore self-attention to extract features of each sequence to determine the word importance concerning the previous word or sentence. The computation can be efficiently parallelized since the architecture does not use recurrent units, just weighted sums, and activations.

The Transformer architecture scales with training data and model size due to its efficient parallel training and ability to capture long-range sequence features (Wolf, et al., 2020). Therefore, this architecture is convenient to pre-train on huge corpora, increasing the accuracy on most NLP tasks like text classification, language understanding, translation, and summarization.

Moreover, the use of a pre-trained model speeds up and simplifies the production deployment of the system. So, as mentioned before, two big AI players released their transformer models: Open AI, with GPT 2 and 3 (Generative Pre-Training), and Google with

[2] Drawing from https://www.researchgate.net/figure/A-standard-unfolded-RNN-structure-at-time-t_fig1_333294428.

BERT (Bidirectional Encoder Representations from Transformers). The latter, BERT, is an open-source and deeply bidirectional of an unsupervised language representation, which is pre-trained solely using a plain text corpus (Ajay, 2020).

As an open-source project, BERT found an excellent acceptance in HuggingFace[3], an NLP community. They are constantly experimenting and training new NLP models. And developed libraries and models and a framework that simplifies the usage of state-of-the-art transformers and pre-trained for NLP tasks.

Next, we introduce BERT model architecture and some model variations inspired or adapted from BERT used in our experiments.

4.2.1 BERT base model (uncased)

The BERT base model is pre-trained on the English language using a masked language modeling (MLM) strategy (Devlin, Chang, Lee, & Toutanova, 2018). The model is available at HuggingFace, bert-uncased[4]. Also, we chose the uncased model for our experiments since case sensitiveness is unnecessary to our problem; hence, the model makes no distinctions from CaPS, CAPS, or caps.

4.2.1.1 Model description

The BERT is a transformer model pre-trained on a large corpus of English data, Wikipedia and BooksCorpus (Zhu, et al., 2015), with more than 980 million words in 11.038 books. The pre-training is done on raw text, with no manual labeling, where an automatic process generates inputs and labels from those texts. More precisely, it was pre-trained with two objectives:

- Masked language modeling (MLM) to train the model to understand a bidirectional representation of the sentence. 15% of the words are randomly masked in the input and then run the entire masked sentence through the model, which predicts the masked words.
- Next Sentence Prediction (NSP), where the model is trained to predict if the two input sentences are next to each other. Two masked input sentences are concatenated and provided as inputs, randomly they were together or not.

These two objectives allow the model to learn the English language representation to extract features useful for other NLP tasks.

---

[3] https://huggingface.co/

[4] https://huggingface.co/bert-base-uncased.

Specifically for text classification, after a fine-tunning procedure, the final hidden state of the **[CLS]** token will have a fixed-dimensional representation of the sequence (Figure 9), which is fed to the classification layer. The classification layer is the only new parameter added and has a dimension of K x H, where K is the number of classifier labels and H is the size of the hidden state (Seth, 2019). In our experiments, the label probabilities are computed with a sigmoid.

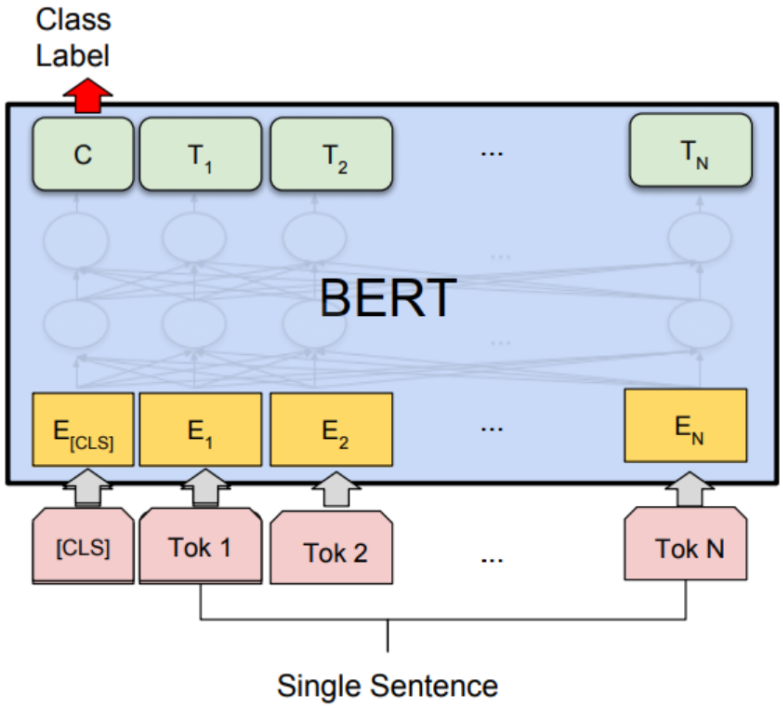

Figure 9 – BERT model for text classification.

4.2.2 RoBERTa base model (cased)

The RoBERTa base model is pre-trained on the English language, using masked language modeling (MLM) (Liu, et al., 2019). In HuggingFace, RoBERTa is available as a case-sensitive model; despite it, we conducted our fine-tuning experiments with uncased text.

4.2.2.1 Model description

Like BERT, RoBERTa is a transformer model pre-trained on a large corpus of English data. The pre-trained is done on raw text, with no manual labeling; this allows vast amounts of publicly available data for the training. An automatic process generates inputs and labels from those texts. Unlike BERT, RoBERTa uses just a Masked language modeling (MLM) objective, with dynamic masking that randomly generates the mask every time a sample is fed into the

model; 15% of the input sentences are masked. Also, the RoBERTa has a much simpler structure than BERT since it removed the next sentence prediction (NSP) task (Gatto, 2021).

4.2.2.2 Analysis

RoBERTa authors observed that BERT was undertrained, and they did a careful evaluation and adjustment of the hyperparameters and increased the training dataset size. Therefore, RoBERTa's pre-trained strategy included:

- Longer model training, with bigger batches, and with more data;
- Removal of the Next Sentence Prediction (NSP) objective;
- Longer sequences on training; and
- Dynamically changing the masking pattern on the training data.

The final RoBERTa model superseded BERT's result with both GLUE and SQuAD datasets.

4.2.3 ELECTRA base model (uncased)

The ELECTRA is a BERT-like model pre-trained in a configuration resembling a generative adversarial network (GAN) (Clark, Luong, Le, & Manning, 2020), the available model in HuggingFace is case-insensitive.

4.2.3.1 Model description

The ELECTRA model has a different pre-training approach, where the training involves two transformers: the generator and the discriminator. The generators replace tokens in a sequence and are trained as a masked language model (MLM). The discriminator is the model to be used as the pre-trained model, which tries to identify the tokens replaced by the generator.

4.2.3.2 Analysis

ELECTRA takes a different approach for the pre-training; instead of masking the input tokens, some input tokens are replaced with plausible alternatives for the token (by a generator network), Figure 10. Both BERT and RoBERTa use the masked language modeling (MLM) pre-training methods, where 15% of the input tokens are replaced with [MASK] and then train a model to reconstruct the original tokens. This strategy produces good results when used in transfer learning for NLP tasks. The drawback of this approach is that a large amount of data is necessary to have an efficient model. Therefore, a discriminative model is trained to predict if the generator replaced the token. As a result, the contextual representations learned by the model are better than the ones discovered by BERT and RoBERTa when we observe the

dimensions: model size, input data, and computational training time (HuggingFace, Electra, 2020).

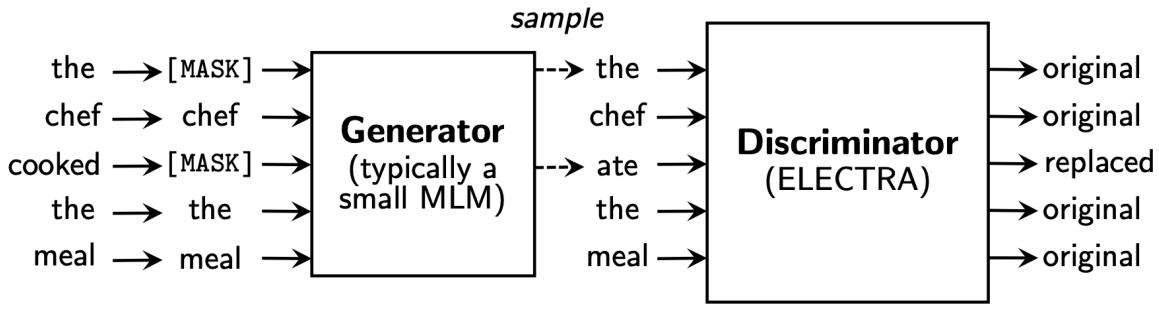

Figure 10 – Electra replaced token strategy.

4.2.4 DistilBERT base model (uncased)

DistilBERT model is a distilled version of the BERT base model. The available model in HuggingFace is uncased (Sanh, Debut, Chaumond, & Wolf, 2019).

4.2.4.1 Model description

When compared to BERT, DistilBERT is a smaller and faster transformers model. It was pre-trained on the same corpus from BERT in a self-supervised fashion. BERT base model was used as a teacher. It was pre-trained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts using the BERT base model. More precisely, it was pre-trained with three objectives:

- Distillation loss: the goal is to achieve a model with the same probabilities as the BERT base model.
- It uses the same strategy as BERT of a Masked language modeling (MLM)
- With cosine embedding loss, the hidden states are trained to be similar to the BERT base model.

Hence, DistilBERT achieves the same inner language representation as its teacher model, with the advantage of being faster on traditional NLP tasks, fine-tuning, and producing a smaller output model.

4.2.5 ALBERT V2 base model (uncased)

ALBERT model is pre-trained on the English language using a masked language modeling (MLM) objective (Lan, et al., 2019). The model available in HuggingFace is case-insensitive.

4.2.5.1 Model description

As BERT, ALBERT is a transformers model pre-trained on a large corpus of English data in a self-supervised fashion. The model was pre-trained with two objectives:

- As in BERT, the Masked language modeling (MLM).
- Unlike BERT, ALBERT looks for the Sentence Ordering Prediction (SOP), i.e., the model tries to predict the order of two consecutive segments of text.

Also, those objectives are meant to capture the internal representation of the language. Hence, ALBERT is suited for standard classifiers using the features produced by the model as inputs. ALBERT has a particular design, sharing layers across its Transformers. With the same weight layers, that results in a smaller memory footprint. Nevertheless, the computational cost remains similar to a BERT-like architecture with the same number of hidden layers. It has to iterate through the same number of (repeating) layers (HuggingFace, Albert - albert-base-v2, 2020; HuggingFace, The AI community building the future, 2021).

4.2.6 BERT multilingual base model (uncased)

The BERT multilingual base model is pre-trained on a large corpus of multilingual data using a masked language modeling (MLM) strategy (Devlin, Chang, Lee, & Toutanova, 2018). The model is available at HuggingFace, bert-base-multilingual-uncased[5]. As in the others model selection, we chose the uncased model for our experiments since case sensitiveness is unnecessary to our problem; hence, the model makes no distinctions from CaPS, CAPS, or caps.

4.2.6.1 Model description

BERT multilingual is a transformers model pre-trained on the 102 languages with the largest Wikipedias. As in BERT, BERT multilingual model was pre-trained with the same two objectives:

- Masked language modeling (MLM) to train the model to understand a bidirectional representation of the sentence. 15% of the words are randomly

[5] https://huggingface.co/bert-base-multilingual-uncased

masked in the input and then run the entire masked sentence through the model, which predicts the masked words.

- Next Sentence Prediction (NSP), where the model is trained to predict if the two input sentences are next to each other. Two masked input sentences are concatenated and provided as inputs, randomly they were together or not.

With such objectives, the model learns to capture the internal representation of the languages in the training set.

4.2.7 DistilBERT base multilingual model (cased)

This model is a distilled version of the BERT base multilingual model. This model is cased: it does make a difference between Caps and caps. Although our experiments do not need case sensitiveness, that was the only DistilBERT multilingual model available at HuggingFaces, by then (distilbert-base-multilingual-cased[6]).

4.2.7.1 Model description

DistilBERT multilingual is trained on the concatenation of Wikipedia in 104 different languages listed here. The model has six layers, 768 dimensions, and 12 heads, totalizing 134M parameters (compared to 177M parameters for BERT-base-multilingual). On average, DistilBERT is two times faster than BERT-base-multilingual.

BERT base multilingual model was used as a teacher to generate DistilBERT multilingual. As in the English DistilBERT, it was pre-trained with three objectives:

- Distillation loss: the goal is to achieve a model with the same probabilities as the BERT base model.
- It uses the same strategy as BERT of a Masked language modeling (MLM)
- With cosine embedding loss, the hidden states are trained to be similar to the BERT base model.

Hence, DistilBERT multilingual achieves the same inner language representation as its teacher model, with the advantage of being faster on traditional NLP tasks, fine-tuning, and producing a smaller output model.

[6] https://huggingface.co/distilbert-base-multilingual-cased

5 Experiments and analysis

In this section, we present the implemented tools to generate and evaluate several language models to be used on the ESG screener. To conduct our experiments, we needed to extract the ESG data from the annual activity report PDF files; we implemented a scraper. We describe the computational set up to run our experiments, a baseline experiment, and the experiments conducted in two groups of BERT-like architectures. Finally, we analyze the results of our experiments with different perspectives. We point out the best model for the development of the screener and present alternatives for the screener evolution.

5.1 Scraping the PDF's Annual Activity Report

As presented in Section 3.4, the annual activity report PDF files are produced so that they are pleasant documents for human reading, not necessarily for data extraction and analysis. Therefore, to conduct our experiments and produce the supervised data for our training and validation sets, we developed a scraping tool to extract the GRI concepts and paragraphs from PDF annual report.

In Figure 11, we present the data preparation process. This two-phase process generates annotated text from the annual activity report that our NLP model's training will later consume on the GRI classification task.

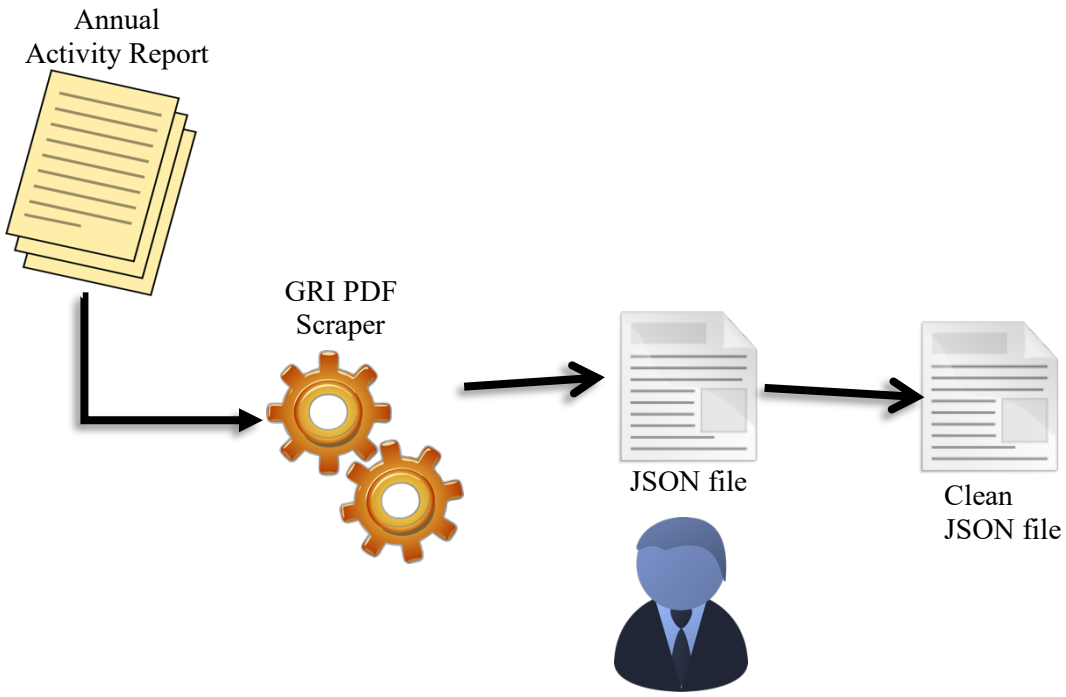

Figure 11 – PDF scraping and data pre-processing.

PDF files are a file format to present documents, including formatted text (text with different font types and sizes), tables, and images. Consequently, to extract the GRI content from the PDF file, we need to identify correlated text and paragraphs and recognize parts of the text according to the GRI reference on the document. So, our scraper has two distinct tasks: i) text identification; ii) and GRI text segmentation.

The font size is used to identify the paragraphs; this is how our scraper distinguishes the content text from titles, descriptions, table and figure captions, etc. We can say that the most frequent font size on the document is, in fact, the most relevant content on the document. Also, on our processing, it is irrelevant minor variations on text font within paragraphs like boldfaces or word emphasis, so this format information is discarded. In the end, the output of this task is a list of font sizes and text paragraphs.

Next, we do the GRI text segmentation task (ii), we iterate over the list generated on the first step, looking for the GRI disclosure (identification codes like 401-2, 305-1, 305-2, etc.), when found we create a tuple with the GRI code and the text. Also, if the following text paragraph font size does not change, we consider this paragraph part of the identified GRI concept.

The second phase of our data preparation is a manual intervention. We observed that our scraper extracts index, summaries, and cross-reference tables that do not contain meaningful information to our process. Therefore, we manually removed this content from the data. We can improve this phase of manually removing such kind of text; BNDES can later make further improvements in this phase.

To exemplify the operation of the scraper, Figure 12 presents one of Vale's company report pages with GRI's concepts 304-2 and 304-4. Also, in Table 4, the text extract with our scraping tool. We adopted a conservative approach, getting the concept for the nearest text paragraph from the text. If the concept is described in several paragraphs, we might not capture all that information. It is also important to point out that our output is an expansion of concepts and paragraphs, i.e., when a paragraph reference two GRI codes (e.g., concepts 304-2 and 304-4, from Figure 12), the data processing creates two tuples as the output of the process.

Figure 12 – Vale's annual activity reporting concept 304-2 and 304-4.

Table 4 – Scraped text for concepts 304-4 on Vale's report.

304-4	Biodiversity			MM2Vale considers biodiversity and ecosystem services 1 as essential topics intrinsic to its business. Based on its Sustainability Policy, Vale develops a strategy for managing biodiversity based on commitments: ‚Ä¢ To know and to monitor the regions in which it operates. ‚Ä¢ To manage risks and impacts to biodiversity, by adopting prevention, mitigation, compensation and monitoring measures that are prioritized when planning socio-environmental expenditures. ‚Ä¢ To act with transparency regarding the company‚Äôs practices and performance in protecting biodiversity.

Last, there is a lot of room for improvements on the scraping phase, either by reducing human intervention, a better precision on the concept detection, and improvement to detect tables on the text. For the last, a promising python tool is (Tabula, 2021).

5.2 Training and Validation Data

In our experiments, we manually downloaded verified annual activity reports from the GRI database[7] to produce the data set for our supervised training. From the website, we downloaded thirty-one GRI Brazilian companies' reports from 2016 to 2019, which represent nearly all English language published documents of the database. The purpose of using those reports is because the BNDES clients and loaners are from Brazil. Using them as a data set, we can produce a language model that can recognize GRI concepts fitted for Brazilian companies. Another aspect to consider was the English language choice since we wanted to explore a large variety of BERT strategies. For the English language, there are several BERT language representation models available. We could experiment with many different pre-trained strategies and look for the best alternative suited to our problem.

For the sake of our training and validation set, the obtained classification can be considered a *Gold Standard* since we extract the classified text from the official GRI database, so those documents had their classification well scrutinized and validated by GRI experts. Nevertheless, since our scraping strategy can miss some text or tags due to PDF format, we might not get all concepts related text on the document. Therefore, our approach is not complete in producing all annotations from the documents. For our problem, missing some annotation is not a big issue since we are more interested in good examples to train our model.

The output of our scraper is a set of paragraphs associated with one of three categories. This set will be used to train our classifier. In our experiments and for the BNDES initial use of the screener, there is no need for a fourth class, representing the set of out of the categories. In this first setup, the BNDES will work with companies that provide only the ESG relevant part of their annual activity reports to the Bank. Therefore, the classification task required for the screener will always be to choose from one of three GRI categories.

After the data preparation of the thirty-one Annual Activity Reports, we classified 3,247 paragraphs into the GRI categories (200, 300, and 400). In Figure 13, we observe the distribution text occurrences of those categories.

[7] https://database.globalreporting.org

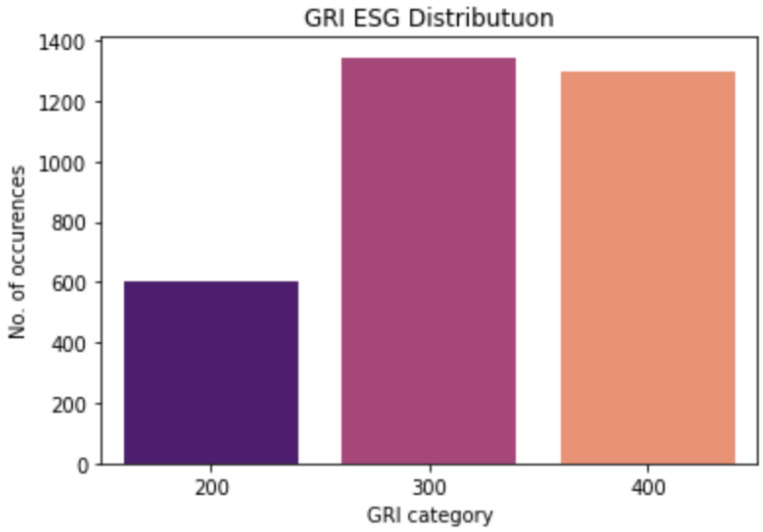

Figure 13 – Dataset GRI categories distributions.

To better understand our data, in Figure 14.a, we present the number of published annual activity reports on the GRI dataset since 2016. We can observe an increasing number of documents indicating the adoption of GRI standards among Brazilian companies. Another dimension of this aspect is captured in Figure 14.b, where we plot the number of GRI paragraphs per year obtained by our scraper. The two figures' data distribution is almost the same, meaning that our scraper preserved the year distribution over the paragraphs.

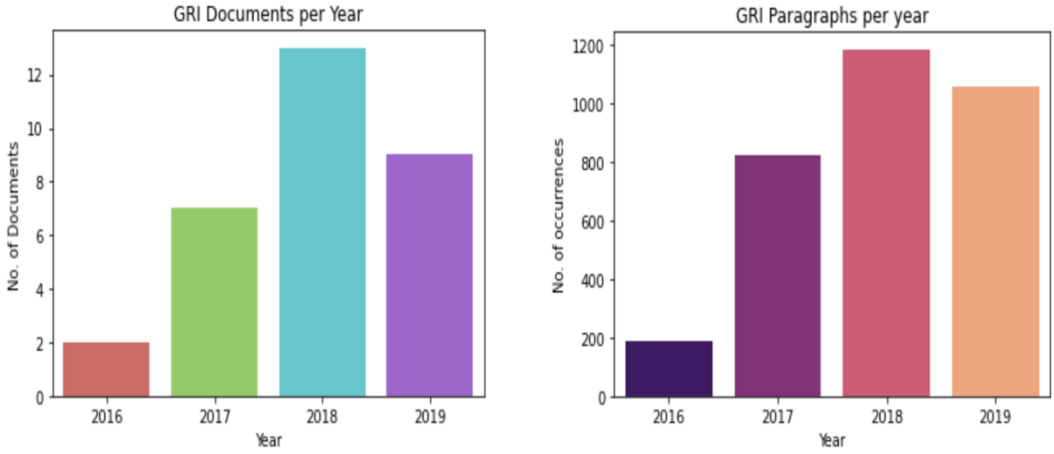

Figure 14 – a) nº of reports per year; and b) GRI paragraphs per year.

Finally, from Figure 13, we observed that the categories are imbalanced, which will be reflected in the accuracy of our model. Nevertheless, since the total number of documents is reduced, we opted to use all the available documents.

5.2.1 Running environment

We run our experiment in Google Colab PRO using TPU with the following hardware:

- Forty processors Intel(R) Xeon(R) CPU @ 2.30GHz, family 6 (model 63); and
- A memory of 36Gb with 2Gb cache.

Since we experimented with larger models like **bert-large**, **electra-large-generator**, and **albert-large**, we needed more memory RAM to run our experiments. Therefore, we used Google Colab PRO. All the other models were trained with the standard Google Colab setup, with the following configuration:

- Two processors Intel(R) Xeon(R) CPU @ 2.30GHz, family 6 (model 63); and
- A memory of 12Gb with 46Mb cache.

5.2.2 Conducted experiments

To better understand which model is more suited to capture our GRI concepts on English annual reports from Brazilian companies, we conducted a total of 23 experiments:

- One experiment, setting our baseline: Naïve Bayes,
- Six small models' setups:
  - Two electra-small,
  - Two albert-base-v2,
  - Two distilbert-uncased
- Six standard models' setups:
  - Two bert-uncased,
  - Two roberta-base, and
  - Two electra-base.
- Six large models' setups:
  - Two bert-large,
  - Two electra-large-generator,
  - Two albert-large-v2.
- Four multilingual setups:
  - Two distilbert-multilingual,
  - Two bert-base-multilingual.

We did each experiment twice, one time without any text pre-processing and another with text pre-processing. Next, we detail each experiment.

5.3 Baseline experiment

A traditional approach for text classification is using a Naïve Bayes classifier (Martin & Jurafsky, 2008). Depending on the problem, Naïve Bayes can achieve impressive results; for example, (Rahman & Akter, 2019) reaches a 91.8% accuracy using Naïve Bayes on a subset of Amazon's product reviews. Therefore, to better understand our classification problem and produce a baseline for our experiments, we implemented a Naïve Bayes to classify our paragraphs with its corresponding GRI ESG concept class.

In our experiments, we did some text pre-processing commonly used to improve the NB approach (Phuc & Phung, 2007) (Rahman & Akter, 2019), like removing stopwords, numbers, leading spaces, punctuation, and special characters. To get a broader perspective of our dataset, in Figure 15, we present the word cloud after pre-processing all documents' text.

Figure 15 – Words from GRI Annual reports.

Highlighted words like "company" and "employee" reveal the main topics of our annual reports. But also, some words show the ESG dimension of the reports like "management," "compliance," "health," "safety," among others.

We used *sklearn*[8] *CountVectorizer()* to prepare our bag-of-words for the *GaussianNB()* classifier for the experiment. We randomly split our prepared data, 90% for training and 10% for validation. The result of our experiment achieved 51% accuracy and 50,33% macro F1-score, and in Table 5, we present the classification report.

[8] https://scikit-learn.org/stable/index.html

Table 5 – Classification report.

Class	Precision	Recall	F1-score
200	17%	45%	25%
300	71%	67%	69%
400	72%	47%	57%

The results obtained are not impressive, and the Naïve Bayes approach would need much more effort to reach higher levels of accuracy. Next, we present the set of experiments with Bert.

5.4 Bert experiments

Due to the complex structure of our documents in expressing GRI concepts from Governance, Environment, and Social, we experimented with a family of newer and sophisticated classifiers: BERT (Devlin, Chang, Lee, & Toutanova, 2018), RoBERTa, (Liu, et al., 2019), ELECTRA (Clark, Luong, Le, & Manning, 2020), DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2019) and ALBERT (Lan, et al., 2019). Specifically, we used some Hugging Face (HuggingFace, The AI community building the future, 2021) transformers models and frameworks to train our GRI ESG classifier.

In this experiment, we compared five different transformer-based architectures (BERT, RoBERTa, ELECTRA, DistilBERT, and ALBERT) for the text classification of our GRI dataset. Our models were fine-tuned from pre-trained models and tested on a 90–10 train-validation randomly split. We conducted two groups of experiments, the first one experimenting with the five architectures with different model sizes and with no text pre-processing. We repeated the set of experiments with the same text pre-processing used in the baseline experiment in the second group. This basic text pre-processing consisted of removing multiple white spaces and punctuation, removing articles, special characters, and stopwords, lowering the text case, and removing numbers.

We implemented a generic architecture that we could plug different Bert architectures with the same experimental setup. We devised a modular neural architecture that feeds a tokenized input (constituted by the concatenation of sentence and context, i.e., **<cls>**text**<sep>**) to a transformer connected to a dense layer computing the predicted classification, with a

sigmoid activation trained on a *binary*_crossentropy[9] loss. Since transformers have a fixed maximum size length for input tokens, we trained our model with 512 tokens; if our sentence is longer than that, we discard it. On our modular architecture, we made the following adjustments according to the particular BERT architecture model:

- The original BERT requires three tensors as the input, *input_ids* (the token embeddings), *token_type_ids* (in our case, the text/paragraph tokenized with the model-specific tokenizer), and *attention_mask* (used in our case simply as a padding mask). The *token_type_ids* is an input requirement of BERT; the other models do not require the *token_type_ids* tensor.
- The next-sentence prediction classifier of BERT is connected to the final embedding of the **<cls>** token and processed with a *tanh* dense layer (called *pooler_output*). The *pooler_output* embeds the **<cls>** special token, representing the complete sentence or the predicted class. The *pooler_output* is not present on all models, just on BERT and RoBERTa. Hence, for the other models, we flatten the *last_hidden_state* to produce the predicted class.

5.4.1 Experiment setup

Each model was trained for ten epochs with the same optimizer hyperparameters to minimize the biases in the evaluation phase. Also, we split the documents into paragraphs, and this set of paragraphs are separated into a 90 − 10% train-validation set, which is randomly chosen. We assessed our model performance by analyzing the macro F1-score and accuracy. During the text, we refer to the macro F1-score just as F1-score.

As mentioned in the previous section, we conducted two groups of experiments with the models BERT, RoBERTa, ELECTRA, DistilBERT, and ALBERT. We also segmented our models according to their sizes: standard, large, small, and multilingual.

It is also important to remark that the learning rate proved to be extremely important for convergence (all models diverged using Adam's default hyperparameters, a value of lr = 5 · 10−5, and the default ones for β1 and β2, proved to be good to make all the transformers converge).

[9] For future work, another loss function can be assessed like SigmoidFocalCrossEntropy.

5.4.2 1st group of experiments – without text pre-processing

We conducted the first group of experiments without any text pre-processing; we took inspiration from some NLP forum discussions that have obtained good results in scraping PDF documents without any pre-processing, like the orientation from the PRODIGY[10] support team (Support, 2020).

We fine-tuned all the transformer experiments using the maximum token lengths of 512 on Google Colab's and Colab's PRO TPU. We divided our experiments according to the model sizes: standard, large and small models. We also experimented with multilingual DistilBERT and BERT.

We evaluated (using a 90–10 train-validation randomly split) each model by accuracy and F1-score. Next, we present our results for each category split by model size:

5.4.2.1 Standard models:

In Table 6, we present the results of our experiments with the standard models. The **roberta-base** achieved the best accuracy (83.33%) and the best F1-Score (81.43%) among the tested models. The model **bert-uncased** score was not far from **roberta-base**, with an 81,08% accuracy and 78.94% F1-score. It is relevant to point out that **roberta-base** is a cased model, but our trained data is uncased. Nevertheless, the model was able to present a good result.

Table 6 – Standard model results w/o text pre-processing.

Bert Model	Accuracy	Precision	Recall	F1-score	Size
bert-uncased	81.0811%	77.8638%	84.9919%	78.9436%	417.9MB
roberta-base	**83.3333%**	**79.5455%**	**84.1132%**	**81.4275%**	**475.8MB**
electra-base	75.0000%	72.7569%	76.0193%	71.4635%	420MB

5.4.2.2 Small models:

In Table 7, we present the results of our experiments with the small models. Small models are convenient in a scenario that a client has limited memory and storage. For the moment, we expect to carry out our screener ESG evaluations on a desktop computer.

[10] https://prodi.gy/ - Efficient machine teaching. An annotation tool powered by active learning.

Nevertheless, in the future, we foresee that BNDES teams will do an assessment while visiting the company headquarter or facilities; and using a mobile device to help with the analysis.

Distilbert-uncased achieved an accuracy of 87% and 79.85% F1-score, with a generated model size of 257.8MB. **Albert-base-v2** achieved an accuracy of 85.25% and 78.59% F1-score, with a much smaller model (49.1Mb). Therefore, ALBERT may be an interesting model for mobile devices.

It is also important to point out that **distilbert-uncased**'s accuracy is higher than the **roberta-base** base case in this experiment without text pre-processing. Hence, **distilbert-uncased** can be compared to our standard models, being very appropriate for a desktop system setup.

Table 7 – Small model results w/o text pre-processing.

Bert Model	Accuracy	Precision	Recall	F1-score	Size
electra-small	78.6667%	76.4845%	65.4516%	67.1300%	53.2MB
albert-base-v2	85.2459%	78.0388%	79.2186%	78.5924%	49.1MB
distilbert-uncased	**87.1429%**	**86.4444%**	**75.9695%**	**79.8549%**	**257.8MB**

5.4.2.3 Large models:

In Table 8, we present the results of our experiments using large models. We wanted to verify if larger models (i.e., pre-trained with larger corpora) could better understand our GRI concepts. Indeed, we could confirm that the **bert-large** model achieved the best accuracy in our 1st group of experiments with an accuracy of 88% and an F1-score of 84.89%.

It is interesting to notice that despite the good results in the small model, **albert-large-v2** performed worse than albert-base.

Finally, in our experiment with large models, we tried our experiment setup with **roberta-large**. Despite the extra amount of memory (32GB) provided by Google Colab PRO, we could not run our experiment due to a lack of memory.

Table 8 – Large model results w/o text pre-processing.

Bert Model	Accuracy	Precision	Recall	F1-score	Size
bert-large	**88.1579%**	**84.0912%**	**86.1574%**	**84.8941%**	**1.25GB**
electra-large-generator	85.1351%	86.5919%	74.7811%	78.5729%	196MB
albert-large-v2	62.8205%	20.9402%	33.3333%	25.7218%	73,5MB

5.4.2.4 Multilingual model:

In our last experiment, we target the performance of a multilingual model. This analysis perspective is relevant to the BNDES since soon ESG evaluations are planned to be done over Portuguese annual reports.

We assessed **distilbert-multilingual** and **bert-base-multilingual**; the latter reached an accuracy of 84.93% and an F1-score of 81.18%, with lower accuracy and F1-score than our previous experiment with a large model, where **bert-large** got 88.18% accuracy and 84.89% F1-score. Nevertheless, in case the BNDES needs the development of an English/Portuguese ESG classifier, **bert-base-multilingual** points out an up-and-coming solution.

Table 9 – Multilingual model results w/o text pre-processing.

Bert Model	Accuracy	Precision	Recall	F1-score	Size
distilbert-multilingual	82.8571%	79.2970%	79.2970%	79.2970%	518.6MB
bert-base-multilingual	**84.9315%**	**81.3083%**	**81.5657%**	**81.1794%**	**638.7MB**

5.4.3 2nd group of experiments – with text pre-processing

We conducted the second group of experiments with the exact text pre-processing applied in our baseline experiment. So, the pre-processing done was to remove multiple white spaces, punctuation, articles, special characters, numbers, and stopwords. Finally, the text was lower-cased.

As in the first experiment, all the transformer models were fine-tuned using the maximum token lengths of 512 on Google Colab's and Colab's PRO TPU. As in the first experiment, we separated our experiments according to their model sizes: standard, large and small models. Again, we experimented with **distilbert-multilingual** and **bert-base-multilingual**.

We evaluated (using a 90–10 train-validation randomly split) each model by accuracy and F1-score. Next, we present our results for each group.

5.4.3.1 Standard models:

In Table 10, we present the results of our experiments with the standard models. The **roberta-base** achieved the best accuracy (88%) with the best F1-score (84.85%) among the tested models, an improvement of almost 3% in the F1-score when compared to the same model without text pre-processing. Also, the **bert-uncased** improved its F1-score by 2% when compared to its model without pre-processing.

Table 10 – Standard model results - 2nd experiment.

Bert Model	Accuracy	Precision	Recall	F1-score	Size
bert-uncased	85.7143%	82.0295%	80.7029%	81.2157%	417.9MB
roberta-base	**88.0000%**	**84.7617%**	**85.4853%**	**84.8465%**	**475.8MB**
electra-base	77.3333%	75.5310%	71.3255%	71.5120%	420MB

5.4.3.1 Small models:

In Table 11, we present the results with our experiments with the small models with text pre-processing.

Text pre-processed **distilbert-uncase** achieved an F1-score of 82.72%, almost 2% higher than the standard **distilbert-uncase** without text pre-processing. Also, **electra-small** and **albert-base** did not benefit or even lost accuracy and F1-score with text pre-processing.

Indeed, as observed in the experiments with no text pre-processing, **distilbert-uncase** is the best within the group of small models. Therefore, a **distilbert-uncase** without text pre-processing should be the right choice on a setup with some memory constraints. For a configuration with stringent memory restrictions, then the **albert-base** without text-processing is the model to choose.

Table 11 – Small model results - 2nd experiment.

Bert Model	Accuracy	Precision	Recall	F1-score	Size
electra-small	70.1299%	69.7807%	71.1121%	67.7874%	53.2MB
albert-base-v2	76.9231%	71.1490%	74.8402%	72.3019%	49.1MB
distilbert-uncased	**85.3333%**	**82.3082%**	**84.0360%**	**82.7158%**	**257.8MB**

5.4.3.1 Large models:

In Table 12, we present the results with our experiments with the large models with text pre-processing. We achieved the best accuracy (86.49%) and F1-score with the **electra-large** model. **Albert-large** had a fantastic gain from its model trained without pre-processing, both in accuracy and F1-score.

Although, the **bert-large** lost accuracy and F1-score when trained with text pre-processing. It scored 84.41% of accuracy, while the model trained without it scored 88.16%. The same happened with the F1-score, decreasing from 84.89% to 79.88%.

As in our **roberta-large** experiment without text pre-processing, the **roberta-large** with pre-processing could not be run on Google Colab PRO.

Table 12 – Large models results - 2nd experiment.

Bert Model	Accuracy	Precision	Recall	F1-score	Size
bert-large	84.4156%	81.1288%	78.8510%	79.8756%	1.25GB
electra-large-generator	**86.4865%**	**82.2948%**	**82.6731%**	**82.4086%**	**196MB**
albert-large-v2	85.2941%	57.3718%	61.0507%	59.0986%	73,5MB

5.4.3.1 Multilingual model:

As in the 1st experiment group, we evaluated a multilingual model. We assessed **distilbert-multilingual** and **bert-base-multilingual** models with text pre-processing. We obtained an accuracy of 85.71% of accuracy (an increase of almost 1%), but with a much lower

F1-score (decrease of nearly 8%). Therefore, for multilingual models, using a model trained without text pre-processing is the best option.

Table 13 – Multilingual model result - 2nd experiment.

Bert Model	Accuracy	Precision	Recall	F1-score	Size
distilbert-multilingual	83.5616%	79.9825%	70.0483%	72.2728%	518.6MB
bert-base-multilingual	**85.7143%**	**82.9832%**	**70.4308%**	**73.3333%**	**638.7MB**

5.4.1 Experiment analysis

In Figure 16 and Figure 17, we present comparisons on accuracy and F1-score among the most relevant models experimented with or without text pre-processing. Those graphs show the standard and large models, and we also included **distilbert-uncase** in our analysis.

5.4.1.1 Accuracy analysis

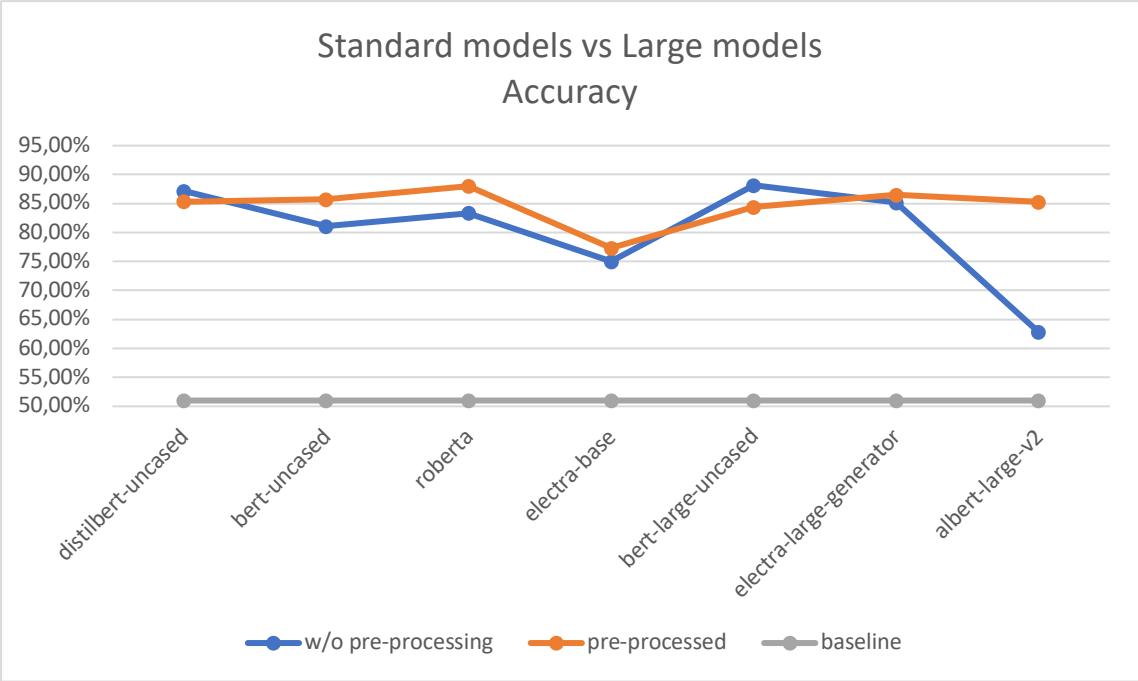

Figure 16 – Comparison between Standard and Large models (Accuracy).

In the graph above we can conclude that text pre-processed experiments performed better than those without text pre-processing. In a closer look at our text data, we observe that the text pre-processing extracts irrelevant information, keeping the essence representation of the text to the model training. An exception was **bert-large**, which got the highest accuracy (88.16%); a possible explanation is that **bert-large** pre-training included similar situations where the model is polluted with numbers and spurious information.

In a scenario where memory and storage are restricted, like smartphones and portable devices, there is **distilbert-uncase**, one of the smallest models with the size of 257.8MB, that achieved an accuracy of 85.33% for the pre-processed model and 87.14% accuracy without text pre-processing, just 1% below our best model. Also, **albert-base-v2**, without text pre-processing, one of the small models, with 49.1MB five-times smaller than **distilbert-uncase**, was able to reach 85.25% accuracy. Hence, those two are exciting choices for a running setup with small devices.

5.4.1.2 F1-score analysis

Another relevant aspect to be considered when evaluating our models is the F1-score, a metric that captures the model performance on the class distribution.

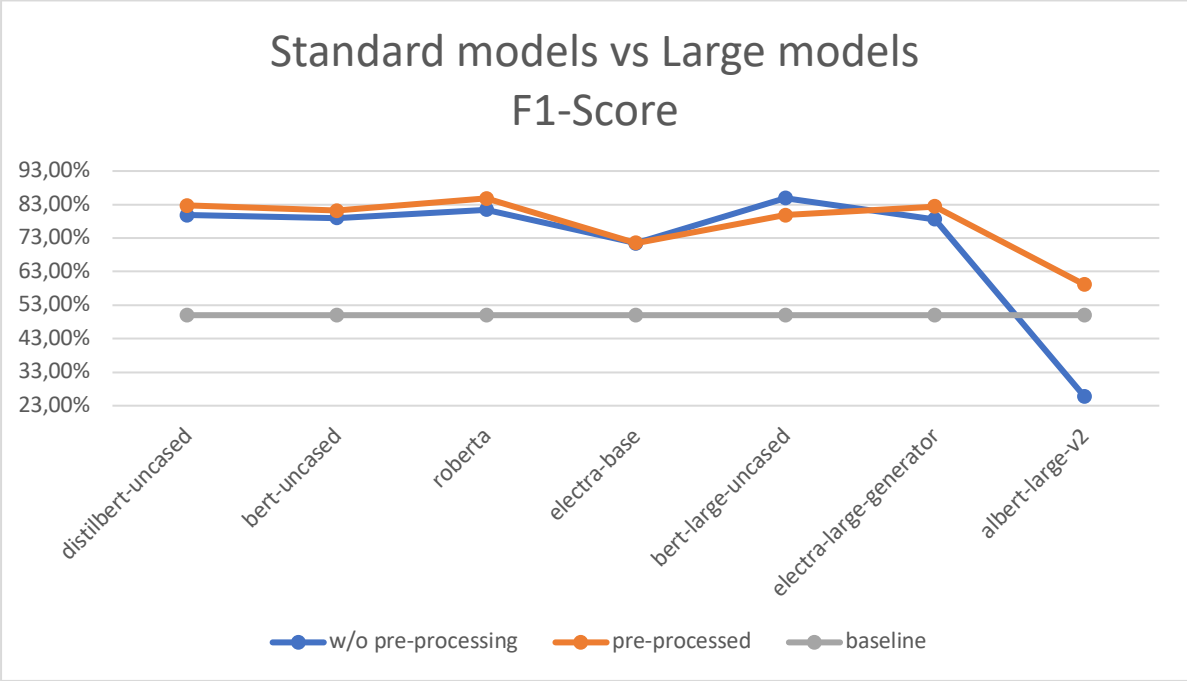

Figure 17 – Comparison between Standard and Large models (F1-Score).

From the graph above, as in our accuracy analysis, we can also observe that text pre-processed experiments performed better than those without text pre-processing. Again, the exception is **bert-large**, which had a better performance without text pre-processing. When

considering the two best models, **roberta-base** and **bert-large** are tied; roberta-base had an F1-score of 84.85%, while **bert-large** achieved 84.89%.

Again, taking a closer look to the **distilbert-uncased**, since it is a small model (257.8MB), it achieved an F1-score of 82.72% for the pre-processed model and 79.85% F1-score without text pre-processing, just 2% below than **bert-large** and **roberta-base**, our two top F1-score models.

5.4.1.3 Final analysis

Comparing text pre-processed **roberta-base** and **bert-large** without text pre-processing scores are tied (in accuracy and F1-score). But, when considering the model size, **bert-large** has the size of 1.25GB, while **roberta-base** has 475.8MB, almost three times smaller. Consequently, choosing a **roberta-base** is the best alternative to a desktop or web server setup for our ESG evaluation solution.

On a limited space setup, like smartphones and portable devices, **albert-base** and **distilbert-uncased** are clear choices. Considering, the F1-score the pre-processed **distilbert-uncased** is the best alternative since its F1-score is 82.72% with an 85.33% accuracy, while **albert-base-v2** reached 78.59% F1-score with an 85.25% accuracy. Now, choosing the suitable model is not so simple; if we select the **distilbert-uncased**, we prefer a more accurate and consistent model but use more memory space on the system. Since the BNDES ESG screener application requirements for portable devices are not defined yet, we should revisit this criterion at the proper time.

5.5 Experiments summary

To evaluate our work, we implemented several text classification methods and set up our experiments using a collection of thirty-one annual activity reports from the GRI public database. We developed a tool to scrap the ESG paragraphs and their corresponding GRI classification for each annual activity report in our testbed in order to produce training and validation sets. To establish a baseline for our experiments, we implemented a solution using the Naïve Bayes method, which achieved a 51% accuracy and a 50,33% F1-score. We fine-tuned five BERT-like architectures to the ESG classification task using the models available on the HuggingFace open web portal, namely: BERT; RoBERTa; ELECTRA; ALBERT; and DistillBERT. We run several experiments on them with different model sizes, considering both with and without text pre-processing in our training and validation sets.

The best models were **roberta-base**, a text pre-processed RoBERTa model, and **bert-large**, a BERT-large model without text pre-processing. Both achieved about 88% accuracy and almost an 85% F1-score. However, regarding the model size, while the BERT-large model has 1.25GB, RoBERTa has 475.8MB, which is nearly three times smaller. Hence, the RoBERTa model was the most interesting alternative to a desktop or web server set up in our ESG evaluation solution.

Moreover, we experimented with multilingual models that can provide a single solution to analyze both English and Portuguese annual activity reports. In this case, the **bert-base-multilingual** model without text pre-processing reached 85% accuracy with an 81% F1-score.

An interesting feature for the sustainability analysis on the BNDES ESG framework would be to enable the annual activity report screening process to be performed on the client's facilities (e.g., by allowing the system to be run on a smartphone or a portable device). To evaluate a resource-constrained device setup with limited memory space, we experimented with several small-size models and came up with two possible best alternatives: the **albert-base-v2** model, generated without text pre-processing; and the pre-processed **distilbert-uncase** model. The first one reached 85% accuracy with a 78.59% F1-score, while the latter got 85% accuracy with an 82.71% F1-score. Again, the best choice depends on the memory availability: **albert-base-v2** has a model size of 49.1MB, while the **distilbert-uncase** has 257.8MB (that is, five times bigger).

6 Related work

Although there is relevant attention to the opportunities of applying NLP in finance (Mayor, 2020), automatic analysis of ESG factors with qualitative information to support sustainable investing strongly depends on domain-specific NLP techniques and is still an open research topic. In this context, compared to state-of-the-art academic proposals and industrial solutions related to text classification, our work introduces a novel BERT-based approach to support the sustainability assessment process by exploring the GRI standard in an ESG classification methodology. In that sense, the following related work provides directions for further improvements.

6.1.1 Claudette

A noticeable project using NLP on the task of understanding legal text is Claudette[11]. The project provides two online tools for service consumers: the first one focuses on online privacy policies and their compliance assessment regarding the European General Data Protection Regulation (GDPR) (Liepin, et al., 2019); and the second tool detects unfair clauses in online consumer contracts (namely, ToS – Terms of Service) (Lippi, et al., 2017).

As in our work, Claudette relies on standards, regulations, and laws for the understanding of domain-specific documents such as contracts and legal terms (or, in our case, the annual activity reports). However, Claudette is a mature project that offers a broad platform to process and understand documents in a layered fashion, first approaching three dimensions of unlawfulness, which are then subdivided into categories according to specific criteria. Compared to our work, Claudette's dimensions understanding corresponds to the classification of the GRI categories, and Claudette's categories understanding is equivalent to subclassification tasks that specialists manually carry out at the BNDES ESG framework. Following Claudette's methodology, we are considering in our system roadmap to further automatize the specialist's activities.

6.1.2 ESG-BERT

ESG-BERT (Mukherjee, 2020) is a pre-trained domain-specific BERT model. Namely, the author trained this language model with an ESG domain corpora to be fine-tuned to perform downstream NLP tasks on sustainable investing text data. Since ESG factors have a unique

[11] http://claudette.eui.eu.

vocabulary, she approached the problem of understanding the text domain by retraining BERT on texts related to sustainability.

The ESG-BERT model seems promising since it embeds sustainable concepts and can be further fine-tuned to other NLP tasks such as sentiment analysis, text classification, and even question and answer. However, the author did not specify which corpora were used to train the language model, thus making it challenging to assess its limitations, biases, and suitability. For example, we could not tell from the ESG-BERT documentation if the "investment" perspective considered when generating the ESG-BERT model would carry some bias that could not adequately consider the "financing" perspective that is necessary to our work.

Nonetheless, a specific language model such as ESG-BERT has the potential to improve the understanding of the GRI categories in our work. Still, it would require further experiments to investigate the involved trade-offs and validate the performance of this approach.

6.1.3 MemBERT

MemBERT (Ruggeri, Lippi, & Torroni, 2021) extended BERT with a Memory Augmented Neural Network architecture (MANN) and achieved good results in the context of knowledge injection in the domain of Legal Analytics. The architecture has two intermediate layers on top of the BERT component: a memory-lookup layer that produces a set of similarity scores; and a memory extraction layer that summarizes the attention scores in a weighted vector. From these components, a fully-connected layer yields the MemBERT classification output.

Moreover, MemBERT has an additional BERT layer to pre-process the slot input on each memory slot of the MANN architecture. The authors claim that, although MemBERT experiments are based on pre-trained DistilBERT, the technique applies to any other BERT-like architecture.

Based on the results from MemBERT, a possible improvement to our work is to explore the injection of ESG domain knowledge to provide an interpretable model with classification explanations to support the sustainable financing specialist. This way, the extra information could improve the quality of her decisions on the ESG subclassification tasks.

7 Conclusion

7.1 Main results

As sustainability is gaining momentum in society and causing changes in people's attitude towards Environmental, Social, and Governance (ESG) concerns, a reshaped investors' behavior is guiding a gradual capital reallocation to more sustainable companies. Nevertheless, the question of how to assess companies' sustainability efforts remains an issue since it requires investors to analyze a broad range of aspects regarding initiatives, decisions, and actions on ESG-related factors that could improve long-term outcomes, using both quantitative and qualitative information. Hence, the success of ESG investing – also called socially responsible investing, strongly relies on the adoption of more structured and preferably automatic ways to evaluate these factors.

In this context, our work addressed the Brazilian Development Bank (BNDES) current initiative to establish an ESG framework for assessing companies' sustainability strategies. We proposed an automation approach based on a natural language processing (NLP) method to improve the analysis of ESG factors by exploring the Global Reporting Initiative (GRI), a well-established standard that has been used to structure companies' annual activity reports regarding ESG aspects. The proposed solution targets at the analysis carried out by domain experts on these reports and introduces an automatic screener in order to overcome an important bottleneck in the BNDES' current ESG assessment process.

The proposed automation solution is based on Bidirectional Encoder Representations from Transformers (BERT), which relies on the attention mechanism to achieve optimal results on sentence-level analysis tasks. We devised a text classification task to analyze excerpts from the annual activity report of companies considering three categories, according to the GRI standard: Environment; Social; and Governance/Economic.

We experimented with several BERT-like architectures (BERT, RoBERTa, DistilBERT, ALBERT, and ELECTRA) with different model sizes for English and Multilingual, generating the classifier to the GRI categories. Our results highlight and quantify the trade-off between model size and performance, assessed by the accuracy and F1-score.

We observed in our experiments that **roberta-base** trained with pre-processed text is the best model with an F1-score (84.85%) and size 475.8MB. **Bert-large** achieved the same F1-score but with a much larger size of 1.25GB.

When memory is a concern (i.e., smartphones), we came up with two possible alternative models, with a clear trade-off in size and performance. These two models were **albert-base-v2** and **distilbert-uncased**, the first with a size of 49.1MB and 78.59% F1-score, while the second has a size of 257.8MB and 82.72% F1-score, both trained without text pre-processing.

Last, we experimented with multilingual models. We observed that **bert-base-multilingual** is the best multilingual model with an 81.18% F1-score and size 638.7MB, with good performance compared with our best English model, **roberta-base**. These results enable a single model to screen documents both in English and Portuguese, the official Brazilian language.

Summing up, we developed an NLP system that tackles the sustainability assessment problem on screening companies' annual activity reports at the Brazilian Development Bank by providing an automatic ESG classification strategy based on the GRI standard and on BERT techniques. We implemented several BERT-like language models with different architectures to provide a comprehensive analysis of the proposed solution, as well as to address various system requirements (e.g., memory setups). Also, we investigated multilingual models that can be a good alternative as an English and Portuguese classifier. Such a multilingual model would enable the BNDES ESG framework to have a single screener, therefore a less complex configuration.

7.2 Final remarks and perspectives

Sustainability is an ultimate issue for economic and social development in Brazil. The BNDES is the leading Brazilian investment bank and therefore plays a critical role in fostering sustainable investing in the country. The proposed NLP-based solution contributes to the development of a modern and efficient framework to evaluate ESG factors for companies, thus strengthening the BNDES sustainable financing agenda.

We intend to extensively explore further automation based on NLP techniques in the BNDES ESG framework, such as in the subclassification tasks. In that sense, we believe that besides the typical performance evaluation metrics (e.g., accuracy and F1-score), the adoption of proposed solutions should also consider their costs in terms of energy and computational resources usage, especially those related to the training steps. In the context of sustainability assessment, we defend that proposed solutions must also be sustainable, and we intend to include these metrics in our future studies.

A necessary improvement for the proposed system is regarding the PDF scraping tool. To generate our training and validation sets, we have implemented a straightforward scraping tool tailored for annual activity reports based on the GRI standard. However, in the future, the BNDES ESG framework will need clever tools to extract text excerpts from possibly other formats and to improve dealing with the challenges of paragraphs identification (Sporleder & Lapata, 2004). Particularly, since PDF files have properties to enhance visual aspects of the documents, when extracting information, eventually coherent text excerpts are split by elements such as an image or a table, thus losing their boundaries. The PDF scraping tool should cover these situations. It is worth noting that NLP techniques can also provide good results for this problem. For example, once the scraper identifies incomplete paragraphs, an additional BERT Next-Sentence-Prediction classifier can be used to link them to their matching pairs.

Another improvement to the BNDES ESG framework is on the evolution lifecycle of the language model, which happens over time based on the experts' feedback. An annotation tool is an interesting way to support this cycle, and a product that can help it is Prodigy (Prodigy, 2021). The periodic enrichment of the language model of our system can benefit from more structure annotation support. The language model evolution would be particularly interesting to the mid and long-term evolution of the ESG framework.

Finally, it is worth mentioning that our work is the first initiative at the BNDES to use NLP techniques for companies' assessment in general, and it opens several new opportunities to automatize and improve business intelligence at this important investment bank.

8 Bibliography

Ajay, D. (2020, 12 3). *Watson Blog.* Retrieved from How BERT and GPT models change the game for NLP: https://www.ibm.com/blogs/watson/2020/12/how-bert-and-gpt-models-change-the-game-for-nlp/

Bajpai, A. (2019, 02 23). *Recurrent Neural Networks: Deep Learning for NLP*. Retrieved from Towards Data Science: https://towardsdatascience.com/recurrent-neural-networks-deep-learning-for-nlp-37baa188aef5

Bergmark, D., Phempoonpanich, P., & Zhao, S. (2001, 09). Scraping the ACM Digital Library. *SIGIR Forum*, pp. 1-7.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly.

BNDES. (2021, 04 01). *BNDES Sustainability Bond Framework.* Retrieved 11 16, 2021, from https://www.bndes.gov.br/wps/wcm/connect/site/82a951ca-d413-4abf-8a62-f7337624d4f3/BNDES_011_GEDIT_framework_INGLES_final.29.03.pdf?MOD=AJPERES&CVID=nyed8sy

BNDES. (2021, 11 16). *Sustainable Development*. Retrieved from BNDES Homepage: https://www.bndes.gov.br/wps/portal/site/home/desenvolvimento-sustentavel

BNDES. (2021). *Sustainable financing solutions.* Retrieved 11 16, 2021, from https://www.bndes.gov.br/wps/portal/site/home/desenvolvimento-sustentavel/solucoes-financas-sustentaveis/solucoes-de-financas-sustentaveis

BNDES. (2021, 09 08). *The Brazilian Development Bank*. Retrieved from BNDES Portal: https://www.bndes.gov.br/SiteBNDES/bndes/bndes_en/Institucional/The_BNDES/

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Krueger, G. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 1877-1901.

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ICLR 2020 Conference Blind Submission.*

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*.

Dhar, A., Mukherjee, H., Dash, N., & Kaushik, R. (2021). Text categorization: past and present. *Artificial Intelligence Review volume 54* (pp. 3007–3054). Springer.

Fang, X., Xu, M., & Xu, S. (2019). A deep learning framework for predicting cyber attacks rates. *EURASIP Journal on Information Security*.

Gatto, J. (2021, 05 14). *An Overview of the Various BERT Pre-Training Methods.* Retrieved from Analytics Vidhya: https://medium.com/analytics-vidhya/an-overview-of-the-various-bert-pre-training-methods-c365512342d8

GRI. (2021, 09 28). *The global standards for sustainability reporting*. Retrieved from Global Reporting: https://www.globalreporting.org/standards

Hildebrand, P., Polk, C., Deese, B., & Boivin, J. (2020, 02). *Sustainability: The tectonic shift transforming investing*. Retrieved from BlackRock: https://www.blackrock.com/institutions/en-zz/insights/blackrock-investment-institute/sustainability-in-portfolio-construction

HuggingFace. (2020). *Albert - albert-base-v2*. Retrieved from Hugging Face: https://huggingface.co/albert-base-v2

HuggingFace. (2020, 09 21). *Electra*. Retrieved from HuggingFace: https://huggingface.co/transformers/model_doc/electra.html

HuggingFace. (2021, 09 12). *The AI community building the future*. Retrieved from Hugging Face: https://huggingface.co

Inderst, G., & Stewart, F. (2018). *Incorporating Environmental, Social and Governance Factors into Fixed Income Investment*. Retrieved from World Bank: https://documents1.worldbank.org/curated/en/913961524150628959/pdf/Incorporating-environmental-social-and-governance-factors-into-fixed-income-investment.pdf

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. Retrieved from arXiv: http://arxiv.org/abs/1909.11942

Liepin, R., Contissa, G., Drazewski, K., Lagioia, F., Lippi, M., Micklitz, H. -W., . . . Torroni, P. (2019). GDPR privacy policies in CLAUDETTE: Challenges of omission, context and multilingualism. *3rd Workshop on Automated Semantic Analysis of Information in Legal Texts, ASAIL 2019* (p. 2385). Montreal: Universite de Montreal.

Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Panagis, Y., . . . Torroni, P. (2017). Automated Detection of Unfair Clauses in Online Consumer Contracts. *Legal Knowledge and Information Systems*, 145-154.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019, 07 26). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* Retrieved from arXiv.org : https://arxiv.org/abs/1907.11692v1

Lu, H., & Unpingco, J. (2021). How PDFrw and fillable forms improves throughput at a Covid-19 Vaccine Clinic. *Proc. of the 20th Python in Science Conference (SCIPY 2021)*.

Martin, J. H., & Jurafsky, D. (2008). *Speech and Language Processing.* Pearson International Edition.

Mayor, T. (2020, 11 3). *Why finance is deploying natural language processing.* Retrieved from MIT Management Sloan School: https://mitsloan.mit.edu/ideas-made-to-matter/why-finance-deploying-natural-language-processing

Meager, E. (2021, 08 31). *Capital Monitor*. Retrieved from Capital Monitor AI: https://capitalmonitor.ai/institution/a-guide-to-sustainable-reporting-standards/

Mukherjee, M. (2020, 09 10). *ESG-BERT: NLP meets Sustainable Investing.* Retrieved from Towards Data Science: https://towardsdatascience.com/nlp-meets-sustainable-investing-d0542b3c264b

Napoletano, E., & Curry, B. (2021, 03 01). *FORBES Advisor*. Retrieved from FORBES: https://www.forbes.com/advisor/investing/esg-investing/

Phuc, D., & Phung, N. T. (2007). Using Naïve Bayes Model and Natural Language Processing for Classifying Messages on Online Forum. *2007 IEEE International Conference on Research, Innovation and Vision for the future*, pp. 247-252.

Prodigy. (2021, 10 20). *Prodi.gy*. Retrieved from https://prodi.gy

Rahman, M. A., & Akter, Y. A. (2019). Topic Classification from Text Using Decision Tree, K-NN and Multinomial Naïve Bayes. *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1-4.

Ruggeri, F., Lippi, M., & Torroni, P. (2021, 09 02). *MEMBERT: Injecting Unstructured Knowledge into BERT.* Retrieved from arXiv.org: https://arxiv.org/abs/2110.00125v1

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.* Retrieved from http://arxiv.org/abs/1910.01108

Seth, Y. (2019, 07 19). *BERT Explained – A list of Frequently Asked Questions.* Retrieved from A blog on data science, machine learning and artificial intelligence: https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/

Sporleder, C., & Lapata, M. (2004, July). Automatic Paragraph Identification: A Study across Languages and Domains. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 72–79.

Support, P. (2020, 09 20). *Using prodigy with PDF documents.* Retrieved from Prodigy Support: https://support.prodi.gy/t/using-prodigy-with-pdf-documents/322/2

Tabula. (2021, 09 30). *Read tables in a PDF into DataFrame*. Retrieved from Tabula: https://tabula-py.readthedocs.io/en/latest/getting_started.html

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Wikipedia. (14, 09 2021). *Annual Report*. Retrieved from Wikipedia - EN: https://en.wikipedia.org/wiki/Annual_report

Wikipedia. (2021, 08 09). *Global Reporting Initiative*. Retrieved from Wikipedia - EN: https://en.wikipedia.org/wiki/Global_Reporting_Initiative

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Brew, J. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *Proceedings of the 2020 EMNLP (Systems Demonstrations)* (pp. 38–45). Association for Computational Linguistics.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015, 06 22). *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books.* Retrieved from arXiv.org: https://arxiv.org/abs/1506.06724