MSc Artificial Intelligence
Master Thesis

# Data Integration and Predictive Modeling for Impact Investing

by
Qingzhi Hu
13167200

June 30, 2022

Number of Credits: 48[1]
Period in which the research was carried out:
1st November 2021 - 1st July 2022

*Internal Supervisor:*
MSc. D. Daza
Dr. R. de Haan

*External Supervisor:*
Dr. L.A.P. Swinkels
MSc. K. Ūsaitė
MSc. R.J. 't Hoen

*Assessor:*
Prof. Dr. P.T. Groth

Universiteit van Amsterdam

ROBECO
The Investment Engineers

[1] Without figures and tables, the thesis's body is around 33-35 pages long. Following a conversation with my supervisors, we agreed to retain these schemas and figures since they make it easy for readers to comprehend the algorithms and ideas.

**Abstract**

Ensuring SDG alignment of investments could help build a more resilient, sustainable, and inclusive economy that produces returns and enhances economic development. The current SDG Alignment tools are designed to assign a score to the contribution of companies to each of the 17 SDGs. However, several issues are obstructing SDG alignment, such as lack of transparency, accountability, and coherence. The purpose of this project is to assess the present SDG framework produced by prominent financial institutions, as well as to aggregate diverse resources and automate the framework with the aid of artificial intelligence (AI), which could potentially help alignment with the SDGs and allow for informed investment decisions.

Firstly, we gathered data from disparate information sources, including company sustainability reports mined from their official websites, product information extracted from their Wikipedia pages, company-related news, and the Wikidata knowledge graph that connects companies through various types of relationships. Then, we employed transformers, which power down-streaming tasks like information retrieval and natural language inference to extract information relevant to each SDG from the acquired data. After preprocessing all the acquired data, we attempted to reverse-engineer the current SDG framework using classification to see how well our data could predict their scores. Incorporating all of the information we acquired significantly improves the F1 micro and macro scores compared to the baseline models, which simply use MSCI sector data. The information extracted from Wikipedia about a company's business is very helpful for predicting SDG product scores, while news headlines and sentiments could assist certain SDGs in achieving improved classification performance for SDG operational scores. In addition, we found that the use of graphs in classification improves the performance of minority classes. Lastly, we propose a new customizable SDG framework that can generate and forecast companies' SDG scores based on the existing frameworks and provide individual-specific explanations for the SDG scores.

*Keywords: Sustainable Development Goals (SDGs), Impact Investing, AI for Climate Change, Natural Language Processing, Information Retrieval, Knowledge Graph, Graph Neural Networks, Data Mining, Imbalanced Classification, Clustering, Explainable Artificial Intelligence (XAI), Dataset Construction*

**Acknowledgements**

Throughout the course of writing my master thesis, I have received a great deal of support and assistance from the following people who have helped me undertake this research:

I would like to thank my supervisors, Dr. Ronald de Haan and Daniel Daza, for providing guidance and feedback throughout this project. Daniel's expertise was invaluable, and his insights and knowledge into the subject matter steered me through this research. I am especially grateful for thoughtful discussions with Daniel. He often organized my jumbled ideas and recommended several important adjustments.

I would like to thank my internship supervisors Laurens Swinkels, Kristina Ūsaite and Robbert-Jan 't Hoen from Robeco. I was impressed by their expertise in the field of sustainable investing and their efforts in creating a valuable project that could potentially make an impact. Their support, guidance, and overall insights in this field shaped the outcome of this research and have made this an inspiring experience for me.

I would also like to thank my colleagues at Robeco who provided valuable feedback during research seminars and gave me the opportunities to talk to different experts in this field. I also want to express my gratitude to Robeco for providing me with the opportunity to work on a genuine business challenge. Finally, I would want to express my gratitude to my family for always being there for me, as well as to my friends for providing intriguing talks and enjoyable distractions to ease my mind outside of my research.

Qingzhi Hu
1st June, 2022

# Contents

# List of Tables

# List of Figures

3figure.caption.4

# Chapter 1

# Introduction

In recent years, machine learning (ML) and artificial intelligence (AI) have been applied to social and global concerns, but more research is needed to understand how these tools may best be deployed to combat climate change. The rise of artificial intelligence (AI) and its increasing impact on several industries necessitates an assessment of its impact on the Sustainable Development Goals (SDGs). Sustainability is a long-term driver for change, especially for big investment firms that wish to create a real impact on society by integrating SDGs into their investment strategies.

The United Nations launched the 17 Sustainable Development Goals (SDGs) in 2015 as part of the 2030 Sustainable Development Agenda. The 17 SDGs (as shown in Figure 1.1) address global issues such as poverty, food scarcity, health, education, climate change, responsible production, clean energy, and inequality[1]. The United Nations Climate Change Conference 2021 was held in Glasgow in November, bringing together parties to accelerate action toward the Paris Agreement's and UN Framework Convention on Climate Change's goals[2]. A new global agreement – the Glasgow Climate Pact – was formed during the COP26 session with the purpose of minimizing the worst consequences of climate change[3]. Recent developments have increased the importance of sustainable investing: firms have the duty and resources to create a more sustainable and resilient future by "putting their money where their mouth is".

In 2019, the G7 commended the joint proposal of the OECD and UNDP to develop a consistent framework for aligning finance with the Sustainable Development Goals (SDGs)[5]. The COVID-19 dilemma presents challenges and opportunities for tying our economy to the Sustainable Development Goals and the Paris Agreement. Ignorance of systemic hazards, ineffective risk management, and low investment in catastrophe risk mitigation are all associated with ineffective sustainability and equity management across industries. This demonstrates the vital importance of SDG alignment. The SDGs are increasingly considered a common global language, with the potential to guide and shape investment strategies.

The SDGs can be adopted by investors through their investment decisions and reporting. When we consider problems of sustainability and the capacity to handle environmental, social, and governance concerns, we can already see how the landscape of investment has shifted significantly in a very short amount of time. The goal of impact investing[6] is to make a financial

[1] https://sdgs.un.org/goals

[2] https://www.un.org/en/climatechange/cop26

[3] https://www.bbc.com/news/science-environment-56901261

[5] https://www.oecd.org/development/financing-sustainable-development/Framework-for-SDG-Aligned-Finance-OECD-UNDP.pdf

[6] https://thegiin.org/impact-investing/need-to-know/#what-is-impact-investing


Figure 1.1: 17 Sustainable Development Goals (The figure is taken from United Nations website[4])

return while also having a beneficial effect on the environment and society. In particular, Eccles, Ioannou, and Serafeim (2014) found that sustainability companies outperformed their peers in terms of stock market performance as well as accounting success over the long run.

Investment firms have grown more interested in investments that both help to achieve the SDGs and provide financial benefits over the last few years. Multiple SDG frameworks such as RobecoSAM SDG Score[7] and MSCI SDG Alignment Tool[8] have been developed to assess the junction between the SDGs and business activities. SDG ratings may be utilized to construct impact investing products with industry-specific criteria and company-specific data. With the help of artificial intelligence (AI), this project aims to evaluate the current SDG framework developed by prominent financial institutions and to aggregate various resources and automate the framework, which may help SDG alignment in terms of transparency, accountability, and coherence for making informed investment decisions.

[7] https://www.robeco.com/en/key-strengths/sustainable-investing/sustainable-investing-research/robecosam-sdg-score.html

[8] https://www.msci.com/documents/1296102/20848268/MSCI-SDG-Net-Alignment.pdf/3dd59d08-3de3-e7e0-7f94-f47b5b93a9ed

## 1.1 Problem Statement - Three Steps

It is difficult to approach and quantify a company's sustainability impact. Issues like greenwashing pose a threat to impact investing. To measure a company's sustainability impact, experts must collect data from a variety of sources. Additionally, the relationships between various companies are intricate. For example, if a downstream supplier of a large organization receives a low rating for its sustainability performance, this could have a detrimental effect on the enterprises affiliated with this supplier. In recent years, companies have responded to climate change by reporting their activities towards it, and such information is documented in more comprehensive reports that encompass environmental, social, and governance risks and exposures (ESG) reports. However, with so much data, sustainability analysts must go through hundreds of pages of reports to locate meaningful data. As a result, substantial human effort would be necessary to manually zoom into reports, news, and other resources in order to unearth useful information for evaluating a company's sustainability performance. This manual approach cannot readily be scaled to a broader universe of companies in multiple markets, since sustainability reporting is not adequately regulated, and there is no central repository for all data. For instance, platforms such as the UN Global Compact collect just a small percentage of corporations' sustainability reports. In other words, relevant information may surface in a variety of locations throughout the web.

Another issue arising is that we are unable to gather all the necessary information from a single site due to fragmented information sources. For instance, not every business has their sustainability report published on official websites such as the global compact[9]. The sustainability reports are most likely to be accessible through each company's official website. However, some firms may not have sustainability reports in PDF format on their websites, but instead outline their sustainability operations in plain text on some of their official webpages. Especially for companies in emerging markets, obtaining these reports is challenging since websites like UN Global Compact do not gather their data and linguistic barriers must be overcome. It's also difficult to find and extract concise and valuable information from extensive reports. In this situation, data mining and Natural Language Processing (NLP) techniques may be able to assist in overcoming the cost and time constraints associated with organizing qualitative data.

In order to score firms' sustainability related activities, several financial corporations provide SDG or ESG ratings. However, Robeco points out that the use of ESG ratings has some shortfalls compared to a more comprehensive SDG assessment[10]. Although the methodology is opaque, and the ratings may or may not correlate with each other. For example, Morgan Stanley Capital International (MSCI) assesses company alignment with UN SDGs through the MSCI SDG Alignment framework[11] providing SDG Net Alignment assessments (including Strongly Aligned, Neutral, Misaligned and Strongly Misaligned) for each of the 17 global goals. They consider a diverse pool of firms in various markets. However, only a proportion of organizations' SDG targets have an analyst-generated classification score with short explanations provided for the score. Therefore, we require an automated technique to generate missing classification scores for firms overseen by current analyst scores. Additionally, it is critical to offer an entity-specific explanation for how the algorithm generates each missing classification score, and the explanations must be extremely accessible to humans, as investment firms frequently receive inquiries from clients such as "why did you include this firm (stock) in this portfolio?".

[9] https://www.unglobalcompact.org/

[10] See footnote 8

[11] https://www.msci.com/documents/1296102/20848268/MSCI-SDG-Net-Alignment.pdf/3dd59d08-3de3-e7e0-7f94-f47b5b93a9ed

One key impediment is the fact that several variables obstruct SDG alignment. Lack of alignment with the SDGs begins with a lack of a shared vocabulary and interpretation of the SDGs' underlying objectives across the public and corporate sectors. OECD[12] has identified three impediments to SDG alignment:lack of transparency, lack of accountability and lack of coherence. In terms of transparency, SDG greenwashing is possible if standards and practices are opaque. Standardization of the framework starts with transparency and comparability. In the meantime, insufficient accountability and oversight of non-financial returns by financial intermediaries and enterprises may hinder the push for sustainable investment. The execution of the framework and overall SDG alignment should be monitored. Lack of coherence refers to inconsistency where assets are only partially aligned as a result of a lack of incentives and fragmented regulation as a result of information asymmetries.

The SDGs present a challenge as well as an opportunity for researchers and decision-makers across fields and sectors. AI could be used to address several SDG challenges given current technical advancements and data explosion. In a study published in Nature Communications (Vinuesa et al., 2020), researchers found that artificial intelligence (AI) may assist meet 79 percent of the Sustainable Development Goals (SDGs). With the assistance of AI, we may overcome some difficulties associated with SDG alignment by employing artificial intelligence to facilitate the process of leveraging multiple data sources as well as improve the transparency, accountability, and coherence of the current SDG frameworks. Therefore, this project has three sub-objectives:

- Constructing a sustainability benchmark dataset to monitor or measure companies' SDG performance.
- Classification of companies with different methods for each SDG in order to obtain insight into current SDG frameworks. Investigating whether graphs could help forecast SDGs.
- Improving the existing SDG framework by integrating more data, exploiting relations between entities and applying new approaches to preprocessing text data. Additionally, this framework is able to intuitively explain the individual classification score.

### 1.1.1 Step I - Dataset Construction

In this project, we aim to address the previous issues by approaching it systematically in three steps. To address information asymmetry and a lack of accountability, we need to provide a pipeline for collecting important information that could be used to measure the SDG performance of firms of interest in a methodical manner. The absence of high-quality public sustainability datasets stymies progress on this effort. Thus, we must begin by establishing a sustainability dataset, which is the first step toward establishing a common foundation for quantitative sustainability research. Hence, the first step is to construct the benchmark sustainability dataset. The constructed dataset should be able to cover almost all the firms of interest while allowing for a more comprehensive assessment of the company's SDG performance. Specifically, the dataset contains data describing the complicated structural relationships between companies, as well as descriptions of the firms' sustainability-related actions. Multiple data sources are used to evaluate a firm's sustainability performance, including company descriptions, sustainability reports, news, and Wikidata knowledge graph.

[12]See footnote 5

### 1.1.2 Step II - Investigating existing SDG frameworks through classification

The second step is to classify firms based on their SDG ratings, produced by common SDG frameworks. To conduct the predictive analysis, we must first identify the variables that are effective in tracking and quantifying businesses' sustainability impacts as reflected by SDG scores. Considering that we have to extract a small amount of desired information from a lot of text data before feeding it to the classification models, modern information retrieval methods are applied to mimic the human procedures of reading long documents. As a result, preprocessing data to eliminate irrelevant information from the data we collected is a critical step before classification. In other words, we need to automate the human procedure of reading lengthy documents into a data-driven procedure that is more scalable and faster than humans. Furthermore, an ablation research is necessary to determine which feature contributes the most to the prediction of SDG scores, whether Wikidata knowledge graph network topology is utilized by existing SDG frameworks, as well as the fundamental distinctions between the various SDG frameworks.

### 1.1.3 Step III - A customizable SDG framework: generating SDG scores with individual-specific explanations

Finally, the objective is to build a new SDG framework that is able to solve the limitations of existing frameworks. When it comes to improving transparency and accountability, the entire process should be thoroughly documented, from data collection to automatic SDG grading, and the framework should be able to synthesize data from rich resources. The final scores for each company should follow the general logic of the existing SDG framework while also incorporating additional relevant data. Additionally, interpretability has risen to prominence as more machine learning models are deployed and frequently utilized to make critical choices. A highly explainable AI model is desired, notably in the finance sector, where investment decisions are made. In our particular scenario, we produce a detailed explanation of the score generated for every company in the graph. In other words, the explanations must be entity-specific and intuitive. To evaluate the framework's capacity to explain the SDG scores, we perform two case studies to quantitatively and qualitatively demonstrate how the explanations are generated.

## 1.2 Research Scope & Questions

In order to define the most important aspects within which this research project is carried out, we seek to answer the following questions (RQs) in order to evaluate firms' sustainability impacts based on text evidence in their sustainability disclosures, news, and other sources:

- RQ1: How can a company's contribution to sustainability be measured? How does the present SDG framework classify companies in terms of the SDGs? What is the most important factor that present SDG frameworks consider when classifying companies?
- RQ2: How to evaluate the company's sustainability performance when it has not been evaluated by human experts? What causes firms' SDG ratings to differ from one another?
- RQ3: How can the SDG scores be automated? How can a new framework be created that preserves the most significant aspects of the present framework while still being adaptable enough to accommodate new data? How can a framework be designed to improve transparency, accountability, and coherence?

## 1.3 Challenges & Contributions

Firstly, we provide an overview of the academic innovations of the thesis by introducing how our work fills gaps in the literature. In this dissertation, we mix numerous AI algorithms from several subfields in a unique manner to address the problem of sustainable investment in the financial arena. The challenges are solved by combining many AI approaches in an original way, which has not traditionally been done in earlier studies. The main areas of artificial intelligence (AI) that are covered by this thesis are (1) traditional machine learning algorithms for classification; (2) graph machine learning for classification, generative models and Explainable Artificial Intelligence (XAI); (3) Information Retrieval (IR) through transformers. The contributions are summarized in bullet points at the end of this section.

In contrast to the existing sustainability datasets created by prior research (Friederich et al., 2021; Corringham et al., 2021; Mishra and Mittal, 2021), learning to rank is utilized to increase the speed and accuracy of web-scraping sustainability reports, allowing us to produce the largest sustainability datasets with company specific information. In addition, our datasets took into account a variety of factors to assess organizations' sustainability performance, in contrast to other studies that mainly focused on examining news (CzvetkÃ³ et al., 2021) or sustainability reports (Chen et al., 2021). In the research published by Kheradmand et al. (2021), they summarize how AI approaches can be used to extract climate risk information from company text data, and we further extend their work by providing concrete examples of how we combine work in information retrieval to analyze unstructured data and extract relevant sentences from reports and wikipedia through transformers (Vaswani et al., 2017) powered downstreaming applications such as semantic search (Reimers and Gurevych, 2019) and Natural Language Inference (NLI) (Bowman et al., 2015).

To investigate in the existing SDG frameworks, the classification of companies' SDG scores is performed. In contrast to prior studies (Chen et al., 2021) that only used binary classification to classify companies' sustainability performance from reports, we also take the multi-class scenario and the problem of imbalanced data into account,. Apart from adding new information to improve the performance of classification, we seek to explain the text classifier results to summarize rules that influence the SDG scores of the company. This is crucially important for AI to be used in finance, as finding new insights from data may be more important than a few percentage points of statistical improvement. Furthermore, we also investigate when it is effective to apply graph algorithms and whether applying graph classification algorithms contributes to classifying the SDG performance of the company. We also learn that building training/validation/test sets for machine learning algorithms on graphs and traditional machine learning methods require distinct approaches. Thus, model performance comparisons must be done carefully, and applying graph algorithms may reveal new insights into the problem.

Inspired by network economics (Knieps, 2015) and the previous work about the potential application of knowledge graphs in achieving the Sustainable Development Goals (SDGs), we try to come up with a new framework which could improve the existing SDG frameworks through aggregating more information and providing better explainability. The previous research Stamou (2021) and Palacios and Piedra (2019) only suggested the potential use of KG or undertake the research at country scope while our work suggests a (knowledge) graph-based solution or implementation that facilitates SDG alignment for companies by uniquely interconnecting the work of GCN (Kipf and Welling, 2017), Graph clustering with GNN (Shchur and Günnemann, 2019) and GNNExplainer (Ying et al., 2019).

The main challenges of this project are:

1. Fragmented information resources: it is difficult to capture sustainability-related information for every company as the sustainability report, company news, and fundamental product and service information are spread over the corners of the web.

2. Analyzing sustainability reports or company news requires finding and identifying relevant information from long documents is like looking for a needle in a haystack, which demands and requires applying suitable modern NLP techniques.

3. Understanding the present framework is required to build a new SDG framework on top of it, and integration of various forms of data resources is required for the construction of a new framework.

This work's major contributions are summarized as follows:

1. We proposed a pipeline for constructing sustainability datasets. The collection includes around 2000 companies (and can be easily extended to quickly collect data for more companies), their recent sustainability reports, websites connected to sustainability, Wikipedia, news, and relevant subgraphs from Wikipedia knowledge graphs. Our research demonstrates one possible application of this dataset. This dataset may potentially serve researchers in conducting other sustainability related research beyond this paper.

2. We proposed a process for extracting SDG-related information from different parts of the data we collected.

   (a) Semantic search and natural language inference to model sustainability reports.

   (b) Modeling the Wikipedia through extracting relevant product information.

   (c) Modeling the news through proposed aggregated sentiment scores for finding impactful news.

3. We further conducted empirical research to study existing SDG frameworks through forecasting the MSCI SDG index and RobecoSAM SDG index with the extracted sustainability related information from text in order to determine the most relevant features that may be used to predict SDG scores.

4. We proposed a novel graph algorithm for generating SDG ratings on knowledge graphs, allowing users to tailor input characteristics and different types of linkages. This new SDG framework is proposed by including more data components and allowing for some degree of flexibility. As a result, the present SDG framework is expanded to provide a more transparent and configurable SDG framework.

5. We demonstrated the potential of how varied NLP and machine learning approaches may be used to tackle real-world business challenges in the sphere of impact investment. The applied AI algorithms in this project vary from conventional random forest classification to popular transformers-based NLP techniques to machine learning algorithms on graphs. We further discussed the industrial concerns in applying AI to finance domain, and address the issues such as copyrights of web data, abstract labels and the need for continuous update and improvement of the AI system in the financial sector.

## 1.4 Outline

In Chapter 2, we firstly present SDGs and two industry-leading SDG frameworks developed by MSCI and RobecoSAM. Then, we look at different public and private sustainability datasets to prove our datasets' value. The remainder of this chapter covers a wide range of methods, including imbalanced classification, Natural Language Processing (NLP), Information Retrieval (IR) and machine learning on graphs. In Chapter 3, we present the data engineering pipeline and data preprocessing pipeline for retrieval of relevant information from acquired data. In Chapter 4, we examine the current frameworks through classification and propose a new framework for generating SDG scores. Additionally, we also conduct case studies to show that our proposed method is able to generate sensible scores. Finally, in Chapter 5, we summarize our findings from the preceding sections in order to answer the research questions in Section 1.2. We also explore the difference between industrial and academic concerns when AI is applied to the finance domain, and further point out the future research directions for this subject. A flow chart of the engineering pipeline is presented in Figure 1.2, which shows how all of the components are linked together.

## 1.5 Implementation

You can download the full version of the code here[13]. The original code repository is 20+ GB, hence a light version of code repository is provided on GitHub[14] where all the large data files (including pretrained models) are removed. Readers can use the light version to quickly go through the README.md and check Jupyter Notebook results without the need to open these scripts locally.

[13] https://drive.google.com/drive/folders/15Cju0hlG1vCmh9ZUeVQpK-flenVyJND2?usp=sharing
[14] https://github.com/QingzhiHu/Thesis_Qingzhi

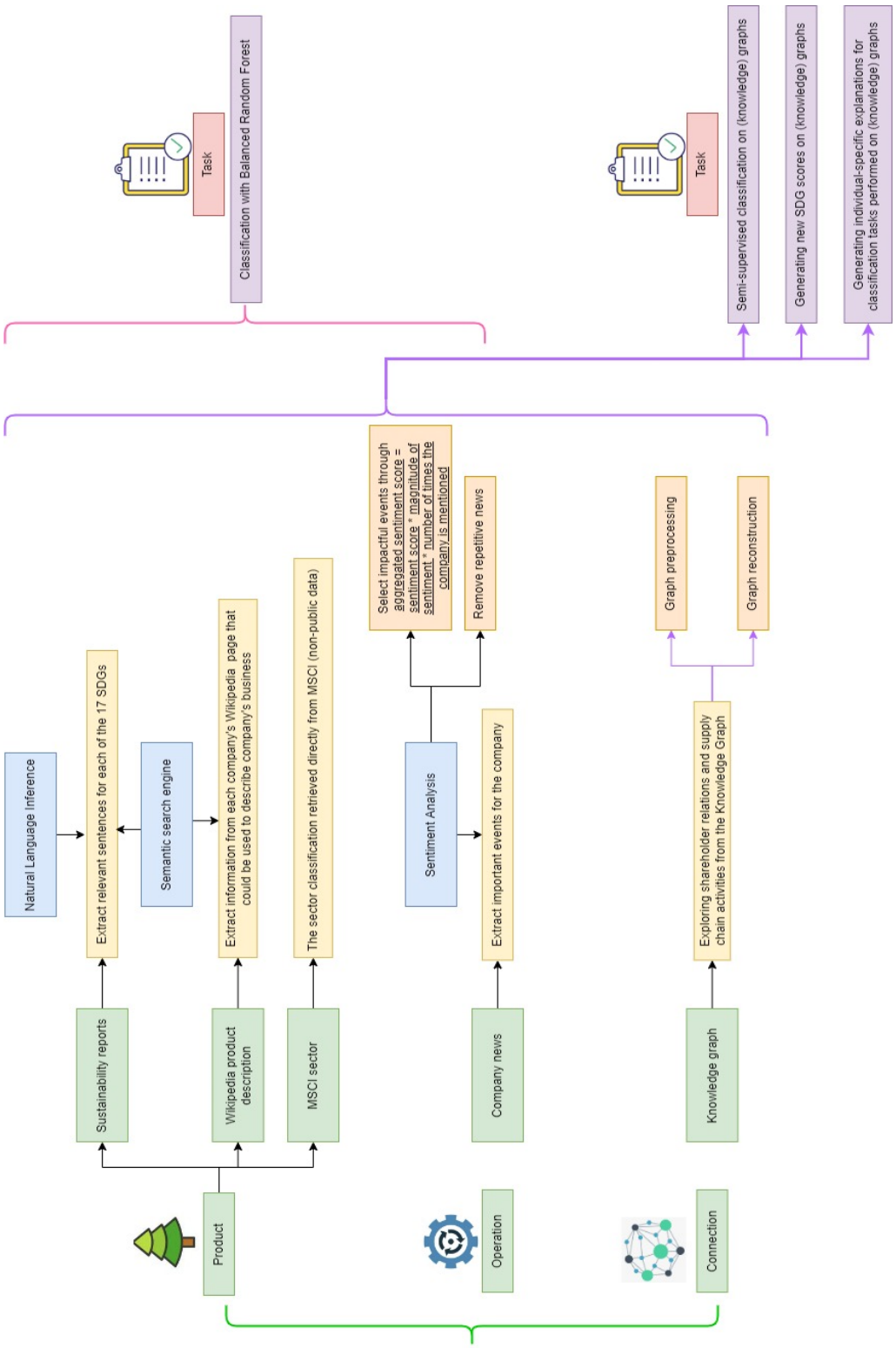



Figure 1.2: The flow chart of the procedures involved in the thesis. Throughout the data collection and analysis, the product information, operations, and shareholder or relevant supply chain activities of the companies are considered.

# Chapter 2

# Preliminaries & Related Work

This chapter establishes the groundwork of our investigation by examining the previous studies. We begin by introducing well-known SDG frameworks and datasets for assessing the sustainability performance of companies. After that, we discuss text preprocessing using NLP approaches and classification algorithms for predicting SDG scores for firms. As we use graph algorithms for both classification and generating SDG scores, in order to better comprehend the graph algorithms we employ in this study, we dedicate an entire section to reviewing the relevant literature in this section. Before reading this chapter, we advise readers to first go through Section 1.3 to get an overview of the academic innovations of our work as well as how the most relevant literature is interlinked to solve our business problem.

## 2.1 Sustainable Development Goals Frameworks

In 2015, the United Nations issued a global call to action to eliminate poverty, protect the environment, and promote peace and prosperity by 2030 with the adoption of the Sustainable Development Goals (SDGs). The 17 SDGs are interconnected, recognizing that progress must balance social, economic, and environmental sustainability. Investment firms have committed to prioritizing progress through establishing SDG frameworks to evaluate companies' 17 SDG performance. SDG Framework is able to give a comprehensive assessment of how each of the 17 UN Sustainable Development Goals is being addressed by companies (SDGs). Assessing SDG Alignment is a complex process that takes into account all aspects of a company's operations and procedures, including its products, services, policies, and practices. There are two SDG frameworks we investigate:

1. **MSCI SDG Alignment Tool**[1] **:** According to MSCI's SDG Alignment methodology, each of the 17 global objectives may be classified as Strongly Aligned, Aligned, Neutral, Misaligned and Strongly Misaligned. Additional evaluations and ratings are available for each firm and each of the 17 objectives based on product and operation alignment. According to them, their framework is based solely on data that has been made public.

2. **RobecoSAM SDG Scores**[2]**:** The RobecoSAM SDG Scores are based on industry-specific criteria and company-specific data. Products, corporate policies and practices, and controversies are all examined via a three-step process by the researchers. Companies are evaluated in accordance with a comprehensive set of standards and Key Performance

[1] https://www.msci.com/documents/1296102/20848268/MSCI-SDG-Net-Alignment.pdf/3dd59d08-3de3-e7e0-7f94-f47b5b93a9ed

[2] https://www.robeco.com/en/key-strengths/sustainable-investing/sustainable-investing-research/robecosam-sdg-score.html

Indicators (KPI). The final SDG score is derived using Robeco's SDG framework (as shwon in Figure 2.1[3]), which considers the performance of companies from three distinct angles, namely products, operations, and controversies.

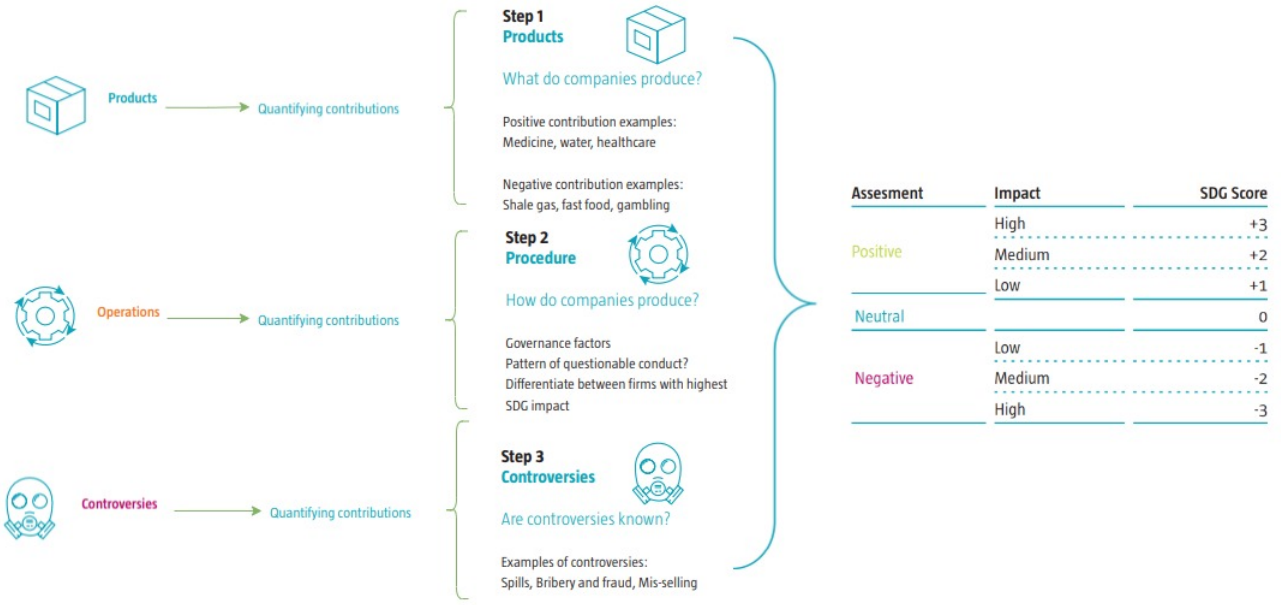



Figure 2.1: RobecoSAM SDG Framework

## 2.2 Sustainability Dataset

First and foremost, we examine the existing sustainability datasets that have been released by AI researchers. In order to identify types of climate risk disclosures in corporate reports, Friederich et al. (2021) have gathered a collection of over 120 manually annotated annual reports from European corporations to detect and categorize climate-related risks by training several classification algorithms to identify and categorize climate-related threats. Corringham et al. (2021) use BERT to classify sentences on a dataset of Paris Agreement climate action plans web-scrapped from Climate Watch[4]. They use "weak" labels, which are generated for each sentence by exploiting the nested headers, sub-headers, and table structures within the HTML documents. Mishra and Mittal (2021) create the climate change knowledge graph (KG) directly without human supervision from news articles. However, all the datasets do not contain company-specific information. The datasets mentioned above are also of small size.

Following that, we examine the sustainability datasets and ratings that have been released by government agencies and financial institutions. Several large corporations' communications on progress reports are collected and updated on the participation webpage by the United Nations Global Compact[5]. Bloomberg also provides sustainability reports for Russell 1000 businesses. Global Reporting Initiative (GRI)[6] Sustainability Disclosure Database covers companies' financial and sustainability disclosures such as Corporate Social Responsibility (CSR) reports. Besides MSCI and RobecoSAM, Morningstar Sustainalytics, ISS SDG Impact Rating,

[3]This information is obtained through https://www.robeco.com/media/4/6/b/46b2a4b697dd4ab51f93cd9e9a7e5d07_202109-capturing-sdg-impact-robeco-sdg-framework_tcm17-31594.pdf

[4]https://www.climatewatchdata.org/

[5]https://www.unglobalcompact.org/

[6]https://www.globalreporting.org/

and ESG & SDG frameworks by FTSE Russell, RepRisk, Refinitiv ASSET4 and Vigeo Eiris (Moody's) also provide ESG or SDG ratings. However, the ratings and their frameworks are commercial and not accessible for free to the public.

The inconsistency of sustainability ratings may hinder the further development of a more transparent, accountable, and coherent SDG framework. According to Berg, Kölbel, and Rigobon (2019), it is hard to assess companies' sustainability performance when their ratings substantially disagree with each other. The research reveals that ESG rating disparity is caused not just by differing perspectives, but also by differing data. As an example, this discrepancy is seen in certain ESG performance sub-categories such as Human Rights and Energy. Thus, their findings advocate for "aggregate confusion" and call for greater transparency for rating organizations. The future SDG frameworks may be improved by making the framework, the data used to build the framework, and the ratings publicly accessible.

## 2.3 Imbalanced Classification

Classifying SDG scores requires tackling the class imbalance problem in the presence of multiple class settings. When classifying data with an uneven distribution of classes, most standard classifier learning algorithms perform poorly. Lots of classification problems may exhibit a substantial imbalance in the distribution of classes, such as fraud detection (Bolton and Hand, 2002), anomaly detection (Chandola, Banerjee, and Kumar, 2009) in the field of medical imaging (Larrazabal et al., 2020), sensor networks (Patel et al., 2020), biology (Yang et al., 2011). Readers can find a comprehensive review on this problem in Haixiang et al. (2017) and Sun, Wong, and Kamel (2009).

### 2.3.1 Sampling methods for imbalanced dataset

One solution to class imbalance is to randomly resample the training sets. Several random oversampling methods are proposed (Chawla et al., 2002; Han, Wang, and Mao, 2005; Nguyen, Cooper, and Kamei, 2011; He et al., 2008; Menardi and Torelli, 2014). However, oversampling the minority class may lead to overfitting. In the meantime, several random undersampling methods are proposed (Tomek, 1976; Mani and Zhang, 2003; Kubat and Matwin, 1997; Wilson, 1972). Due to the fact that this technique eliminates samples from the majority class, it is likely that the selected samples from the majority class are skewed, resulting in a biased and erroneous conclusion that does not accurately represent the distribution of the original data. Re-sampling the training sets may help with binary classification problems as well as multi-class classification problems that have one or more majority or minority classes involved.

### 2.3.2 Balanced Random Forest

To uncover hidden patterns in data, traditional bagging and boosting machine learning methods and neural networks might be employed. However, these techniques are most effective in a variety of situations. The use of Random Forests (RF) (Ho, 1995) is more appropriate for structured datasets, while deep neural networks, on the other hand, may do better with data that is mostly unstructured or homogenous, such as time series or pixels in images. Additionally, training Random Forest is less computationally intensive than training neural networks, and explaining Random Forest predictions is often simpler than explaining neural network predictions.

Choosing an appropriate classifier is contingent upon the data and the nature of the task at hand. In our scenarios, we intend to discover patterns from text representations that are sufficiently interpretable to humans to comprehend how the keywords or phrases in text contribute to forecasting existing SDG scores. This goal could be achieved by using Random Forests in combination with tools such as LIME (Ribeiro, Singh, and Guestrin, 2016) and SHAP (Lundberg and Lee, 2017) for explaining the text classifiers.

In order to tackle the imbalanced data problem encountered in this project, we use the Balanced Random Forest (BRF) algorithm proposed by Chao, Liaw, and Breiman (2004). BRF integrates downsampling from the majority class and ensemble learning seamlessly. It downsamples the majority class to the same size as the minority class for each tree in RF. All the training data is utilized if there are enough trees. There is no loss of information, and it is computationally efficient since each tree only sees a tiny portion of the training data. We use the implementation of this algorithm from Lemaître, Nogueira, and Aridas (2017).

### 2.3.3 Evaluation metrics for Imbalanced Classification

Forman and Scholz (2010) produced an illuminating work on comparing evaluation metrics for class imbalances problems. They discovered through simulations that when k-fold cross-validation was utilized, computing F1 yielded the "most unbiased" estimate. The purpose of the F1-score metric (Sasaki, 2007) is to strike a compromise between precision and recall, which is particularly beneficial in the majority of cases when dealing with unbalanced datasets. There are several types of F1 scores, we use macro-$F_1$ and micro-$F_1$ as evaluation metrics in this paper. Macro-F1 weights each class label equally, whereas Micro-F1 offers equal weight to all occurrences in the averaging process. Let $TP_t, FP_t, FN_t$ denotes the true-positives, false-positives and false-negaives for class-label $t \in T$. Let $P$ denote precision and $R$ denote recall. $P_t$ and $R_t$ denote the precision and recall for class t. The macro-average $F_1$ is calculated through Huang and Ling (2005):

$$P_t = \frac{\mathrm{TP}_t}{\mathrm{TP}_t + \mathrm{FP}_t}, R_t = \frac{\mathrm{TP}_t}{\mathrm{TP}_t + \mathrm{FN}_t}, \text{Macro-F}_1 = \frac{1}{|T|} \sum_{t \in T} \frac{2P_t R_t}{P_t + R_t}$$

The micro-average $F_1$ is calculated through

$$P = \frac{\sum_{t \in T} TP_t}{\sum_{t \in T} TP_t + FP_t}, R = \frac{\sum_{t \in T} TP_t}{\sum_{t \in T} TP_t + FN_t}, \text{Micro-}F_1 = \frac{2PR}{P + R}$$

The performance of unbalanced multi-class classification may be evaluated using curves. The curves commonly incorporate information from the confusion matrix, which can be used to calculate recall, precision, accuracy, and AUC-ROC curve. For each recall threshold, one may determine the average precision scores by utilizing Area Under the Precision-Recall Curve (PR AUC) (Keilwagen, Grosse, and Grau, 2014). When the classes are extremely unbalanced, precision-recall is an effective metric for assessing prediction performance. A large area under the curve indicates good recall as well as precision. In a multi-class setting, the micro-average curve reflects the mean of PR-Curves from all classes when data imbalance is taken into account, whereas the macro-average curve represents the mean PR-Curve when data imbalance is not taken into account. Each class can also have its own PR curve. Furthermore, ISO-F1 curves might be utilized to depict precision recall space lines with the same F1 values.

## 2.4 Natural Language Processing

### 2.4.1 Natural Language Processing for analyzing sustainability reports

In order to measure a public company's contribution to the SDGs, techniques in Natural Language Processing (NLP) are often used to analyze the sustainability reports. Using AI and machine learning to analyze climate risk can assist in analyzing large volumes of corporate disclosures. The application of NLP algorithms to ESG domains enables the extraction of SDG-relevant information (cutting reliance on commercial data sources) and therefore increases overall transparency. Kheradmand et al. (2021) mention that AI can assist in assessing climate risk from corporate disclosures, and there are three types of methodologies used to extract climate risk information from company text data such as disclosures, news, and earning calls: (1) traditional keyword-based approach, (2) AI/statistical-based approaches, and (3) contextual AI-based approach. These methods vary from creating simple embeddings to using more advanced transformer-based models.

The traditional approach involves classifying text based on predefined related keywords (heuristic search/fuzzy string matching). It can't capture the context and interpretation of the issues because it merely counts the keywords in a document. Topic patterns can be implicit and occasionally confusing in the context of climate change, where heuristic keyword-based models are inadequate. On the other hand, a statistical-based approach like Bag of Words (BoW) has been used to create bigrams to classify text (Kheradmand et al., 2021). Chen et al. (2021) uses word embeddings of sustainability reports generated from Word2Vec and Doc2Vec models with an SVM binary classifier to classify companies' alignments with the SDGs. However, these methods can not extract and measure the contextualized information.

In this project, we adopt a contextualized AI-based approach for retrieving relevant information from our data through pretrained models. As for the contextual AI-based approach, it often needs a large amount of labeled data, and the labels are often very abstract for the topics related to sustainability or SDGs. Creating such subjective labels would require hiring a group of financial analysts, and it takes much longer time compared to creating objective labels for classifying objects. Therefore, unsupervised learning methods or transferring learning through fine-tuning the pre-trained models for different down-streaming tasks are desired. For example, Contextualized Topic Models (CTM) (Bianchi, Terragni, and Hovy, 2021), a family of topic models that use pre-trained BERT to improve the coherence of topic models, could be considered as an alternative way to model the text. In addition, Luccioni, Baylor, and Duchene (2020) proposes a ClimateQA model which uses pre-trained models like ClimateBERT (Webersinke et al., 2021) trained on a corpus of climate-related text to identify climate-relevant sections in the sustainability reports based on a question-answering approach. They used a team of sustainability analysts to annotate financial reports and then fine-tuned ClimateBERT on the tagged data.

### 2.4.2 Text Representation

Research in natural language processing and other machine learning problems has long focused on obtaining effective and efficient word representations. A major challenge involves representing the textual information gathered from the reports, news, and Wikipedia descriptions. In the previous research of graph algorithms, the textual information is embedded into a feature

vector through a vector space model (Melucci, 2009) such as BoW (Harris, 1954) Word2Vec (Mikolov et al., 2013) and Doc2Vec (Le and Mikolov, 2014), which can be used for various tasks performed on the graph in this project, such as node classification and node clustering. For simple NLP tasks like text classification, BoW may be sufficient to convey the needed information for the model to make intelligent judgments. For other tasks, such as semantic search, natural language inference, sentiment analysis, translation, and question answering, where a better understanding of the context is required to achieve good results, BoW is ineffective, and contextualized word embeddings are required to achieve effective results.

Firstly, we introduce the most frequently used discrete text representation, namely One-Hot Encoding, Bag-of-Words representation (BOW). One-Hot encoding can lead to an expansion in the feature space, resulting in a sparse matrix representation of high dimension, but it is incapable of modeling the interaction between different words. However, the BOW CountVectorizer does not take into consideration the position of a word, and high-frequency stop words require further cleaning in order to reduce the size of the representation and capture precise information. TF-IDF is a newer iteration of the BOW algorithm and solves some limitations of CountVectorizer. According to TF-IDF, a weight is assigned to each word based not just on its frequency but also on its frequency in the corpus as a whole. However, the spatial information of the word is still missing. Because TF-IDF suppresses high frequency words and ignores low frequency ones, there is a need to normalize the weights of words proportionately.

Word Embeddings (Wang, Zhou, and Jiang, 2019) are a more sophisticated approach used to better capture the semantic meanings of words. For instance, Word2Vec, which includes Continuous Bag of Words (CBOW) and Skip-gram, is one of the most popular used word embedding. However, Word Embeddings approaches such as Word2Vec, Doc2Vec or GloVe are incapable of handling polysemy, where identical words may communicate various meanings depending on context. In order to address this issue, contextualized embeddings such as LMo, GPT-2, BERT, BART and XLNet are invented for more complicated language tasks that requires capturing positional information of the words or involves ambiguous words and unseen words in the training set (Arora et al., 2020). Contextualized embeddings can also be used to encode information across different languages (Liu, Kusner, and Blunsom, 2020). In this research, BoW is used to model the text for classification, whereas contexuliazed embeddings are used for various preprocessing tasks involving text data.

## 2.5 Information Retrieval: Transformers

Using the concept of self-attention, a transformer is a deep learning model that assigns a varied weight to the importance of various parts of the input data. The attention mechanism was proposed by Bahdanau, Cho, and Bengio (2015) and has been widely used in various areas of deep learning in recent years, for example in computer vision to capture perceptual fields on images, or in NLP to locate key tokens or features.

The Google Brain team released the Transformers model in 2017 (Vaswani et al., 2017) and it has since been the model of choice for a wide range of NLP problems (Wolf et al., 2020). This led to the development of systems such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and Generative Pre-trained Transformer (GPT) (Brown et al., 2020), which were trained using massive language datasets and could then be fine-tuned for specific downstream applications.

The attention mechanism is very similar to the human observation mechanism of external

things. When humans observe external things, they generally do not see things as a whole, but tend to selectively obtain certain important parts of the observed things according to their needs. Therefore, the attention mechanism can help the model to assign different weights to each part of the input and extract more critical and important information so that the model can make more accurate judgments without incurring greater computational and storage overheads, which is the reason why the Attention Mechanism is so widely used.

One of the most well-known transformer models is named BERT. BERT trains the transformer model using an unsupervised technique. One of the key features of this paradigm is the attention mechanism, which enables numerous focus points to be activated simultaneously for a single utterance rather than serial processing. BERT employs a two-step procedure to guarantee that the model understands the semantics of the utterance appropriately. The first stage is to mask 15% of an article's words and let the machine forecast the masked words based on context at the second stage. Transform models are originally trained by making omnidirectional predictions of the words that are hidden by masks.

Large-scale pre-trained models (PTMs) (Han et al., 2021), such as BERT and GPT, have become a key milestone in AI due to the time and resources required to train transformers. Pre-trained models are stored deep learning models trained on large datasets for particular NLP tasks. Through transfer learning (Zhuang et al., 2020), the rich information implicitly contained in enormous parameters may help a wide range of downstream NLP applications by fine-tuning the pretrained models. Time and resources may be saved by using PTMs instead of constructing models from scratch.

Transformers power lots of NLP applications, and progress in natural language processing has been fueled by improvements in both model design and model pretraining during the last several years (Wolf et al., 2020). The Transformer architecture has made it feasible to design larger-capacity models, and pretraining has made it possible to use this capacity efficiently for a broad range of tasks such as text classification, question answering, machine translation, sentiment analysis, and natural language inference. In the following subsections, we go through the three key transformer-enabled natural language processing applications used in this project.

### 2.5.1 Semantic Search

In this project, we use a vector-based search engine with pretrained Sentence Transformers (Reimers and Gurevych, 2019) and Faiss (Johnson, Douze, and Jégou, 2019)[7]. Traditional keyword-based search engines, such as Elasticsearch (Gormley and Tong, 2015), are user-friendly and effective in most situations. However, it cannot differentiate between words with several meanings in different contexts and has trouble with lengthy inquiries. We need to create a system that takes the context of the words into consideration in order to overcome these restrictions. In section 2.4.2, we introduce several modern language models that represent text as a numerical vector, and some representations may capture word context information. A vector-based or Semantic-based search engine indexes the vectors in high-dimensional vector spaces and calculates similarity between query vectors and indexed objects to rank relevant sentences or documents. Due to its focus on better understanding the content of the search question to improve the accuracy of the search results, semantic search may reveal synonyms as well as lexical matches. The semantic search schema used in this project for building the search engine is illustrated in Figure 2.2. We use a pretrained model from HuggingFace Model

[7]We use the implementation of https://github.com/kstathou/vector_engine

Hub for semantic search with Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), and more information related to the model's performance can be found in Table A.1 in the Appendix.

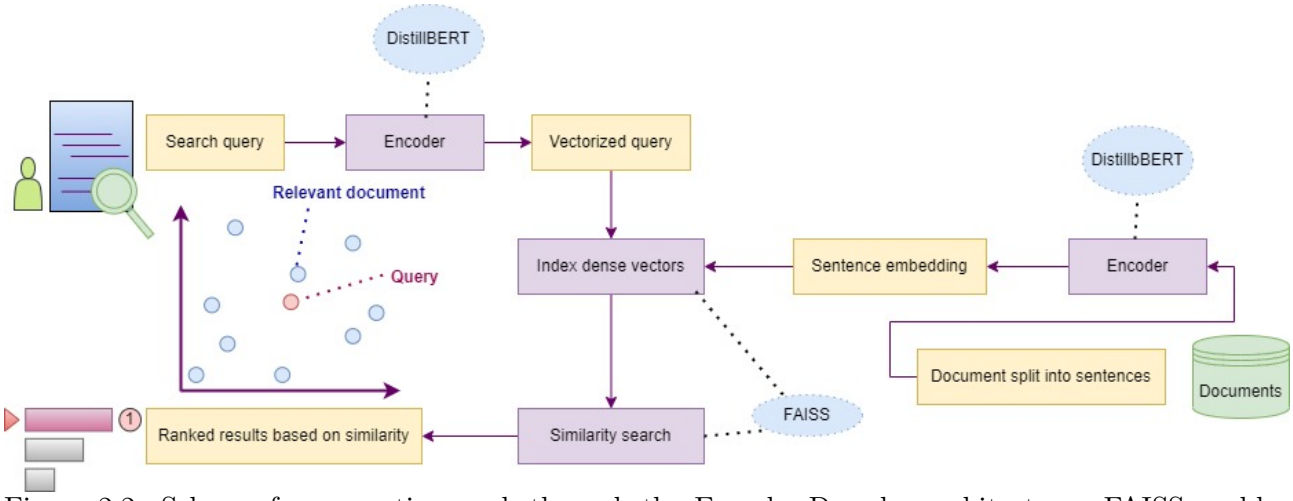



Figure 2.2: Schema for semantic search through the Encoder-Decoder architecture. FAISS enables rapid search for similar sentence embeddings.

### 2.5.2 Natural Language Inference

NLI is adopted in several industries, including banking, retail, and finance. It is often used to determine whether the end-user generated or retrieved result matches the hypothesis. Comparing the sentences it generates to a large corpus of reference texts allows NLI to mimic some degree of reasoning ability of a human auditor. As some of the content relating to a SDG does not indicate the activities taken by firms in support of that SDG, Natural Language Inference (NLI) (Bowman et al., 2015) is used to determine which firms are taking the most significant steps toward achieving each SDG. Given a premise and two sentences, the hypothesis and its implication, we may use NLI to infer a hypothesis. The hypothesis may be true, false, or irrelevant to the premise. True indicates entailment, false indicates contradiction, and indeterminate or unrelated indicates neutrality. This research analyzes sustainability reports using a NLI model constructed by fine-tuning BERT on the SNLI, MultiNLI, and Hans datasets (Gao, Colombo, and Wang, 2021). In this project, we adopt the BERT-based pretrained NLI model trained on the SNLI dataset implemented by Gao, Colombo, and Wang (2021)[8]. More information about the pretrained model used in this project for NLI can be found in Table A.2 in the Appendix.

### 2.5.3 Sentiment Analysis

A popular approach for collecting meaningful and subjective information from text-based data is sentiment analysis. Sentiment analysis is utilized in this project to find, analyze, and extract reactions, moods, or emotions from the news. For sentiment analysis, word embeddings are often used to model the text. However, as mentioned in Section 2.4.2, they ignore text sentiment and context and need a huge text corpus to train and generate precise vectors. These methods only create vectors for words in their vocabulary and disregard terms not in their vocabulary, resulting in information loss. As a result, contextualized embeddings are becoming more popular. For instance, BERT for sentiment analysis (Tabinda Kokab, Asghar, and Naz, 2022) explores the syntactic and semantic information along with the sentimental and

[8]The implementation is available here: https://github.com/yg211/bert_nli

contextual analysis of the data. Sentiment could be decomposed of several components such as polarity, magnitude and sentiment score. Sentiment polarity specifies sentiment's orientation. Sentiment magnitude measures emotional intensity. The sentiment score reflects the text's overall emotional tone.

## 2.6 Machine Learning on Graphs

First of all, we begin by introducing some fundamentals about graphs. A graph is a visual representation of the connection between entities. A knowledge graph is a logically coherent, connected graph that collectively comprises an interlinked set of facts that are machine readable. The set of facts is composed of RDF triples (semantic triples) of the form *[head, relation, tail]*. An example of a knowledge graph subgraph we pulled from Wikidata is shown in Figure 2.3 along with introducing some fundamental terms for knowledge graphs (A more completed version of the subgraph can be found in Appendix Figure A.1).

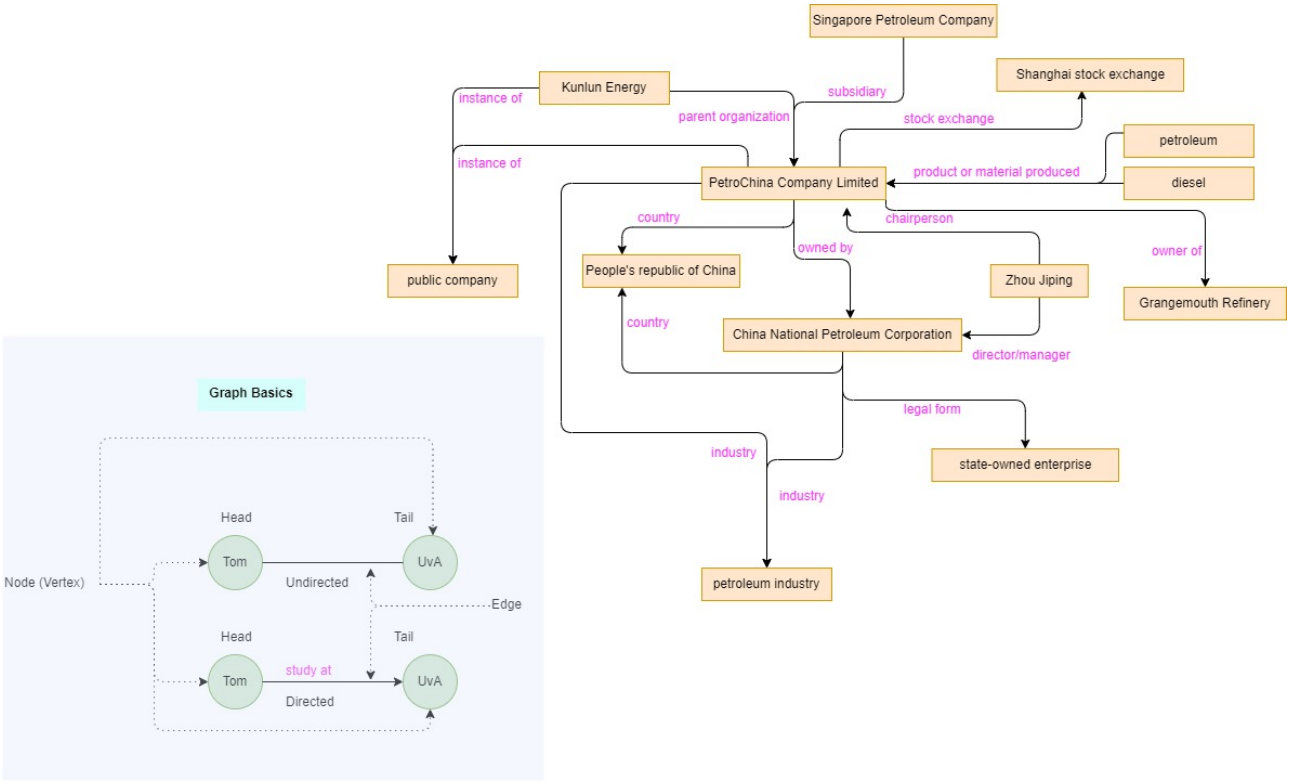



Figure 2.3: Knowledge graph example

Graphs allow us to depict the structure of complex systems in the real world. A graph consists of nodes linked by edges that describe relationships between the nodes. Deep learning on graphs (Zhang, Cui, and Zhu, 2022) has recently shown great promise in developing algorithms that are more accurate and more efficient for a wide variety of graph-related applications. Graph Neural Networks (GNNs) (Scarselli et al., 2009) are a class of deep learning algorithms designed to infer from graph data that can forecast graph-related tasks at the node, edge, and graph levels. Node-level tasks such as node classification and node clustering are employed in this project and will be discussed in further depth below.

Particularly, knowledge graphs (Ji et al., 2022) have a unique capacity to discriminate between various types of links as well as connections between nodes. Unlike in the majority of previous work, where unlabeled examples are connected in a graph with the implicit assumption

that similar nodes have similar labels, the ideal model in our case could customize which edge or relationship types in the knowledge graph are more important, as well as the node importance. The knowledge graph describes the relationships between companies and other entities, and that each company node in the knowledge graph can contain news, sustainability reports, and Wikipedia information, as well as a tabular format of data with both numerical and categorical characteristics.

### 2.6.1 Intuition of Network Economics for SDG

Network effects have been studied in the field of finance, such as portfolio diversification (Li and Wang, 2021) and portfolio selection (Peralta and Zareei, 2016). Network economics (Knieps, 2015) is the study of the network effect in business. The term "network economics" in the field of sustainability refers to the concept of diverse teams working together to develop environmentally friendly products and technologies, and together to build sustainable supply chains. For instance, the existing studies by Jouzdani and Govindan (2021) investigated the design of food supply chain networks to help achieve SDG goals.

Stamou (2021) and Palacios and Piedra (2019) discuss the potential application of knowledge graphs in achieving the Sustainable Development Goals (SDGs), given that knowledge graphs may include extensive structural information about organizations, such as supply chain activity and shareholder information. However, the purpose of the research (Palacios and Piedra, 2019) is to propose an endeavor to develop tools to aggregate and link the multiple resources relevant to the SDGs with the help of open data. Furthermore, rather than being company-specific, the research (Stamou, 2021) is undertaken at the level of countries. None of the studies suggest a knowledge graph-based solution or implementation that facilitates SDG alignment for companies.

### 2.6.2 Classification on graphs and knowledge graphs

Kipf and Welling (2017) mentions that there are two primary approaches for graph-based semi-supervised learning tasks, namely explicit graph Laplacian learning and graph embedding-based approaches. The prior work Graph Convolution Networks (GCN) (Kipf and Welling, 2017) is motivated by spectral method and further addresses semi-supervised classification using convolutional graph networks on graphs. The core mathematics behind GCN is explained in Section A.1 in the Appendix. Figure 2.4 depicts the architecture for employing GCN for classifying the nodes. The study includes an application based on a knowledge graph, but they preprocess the KG to a regular graph without taking edge properties into account. To address this issue, Schlichtkrull et al. (2018) devised RGCNs, which are a recent class of neural networks that operate on knowledge graphs. RGCN adjusts the way of message passing in GCN, and uses different neural network weights for different relation types in order to deal with multiple relation types. One of our interests is to figure out how to effectively use data from several sources to do weakly supervised classification tasks. On the one hand, a KG is a rich source of information due to the variety of nodes and edges it includes. On the other hand, external input signals such as the sustainability report and news may include critical information about a company's SDG score.

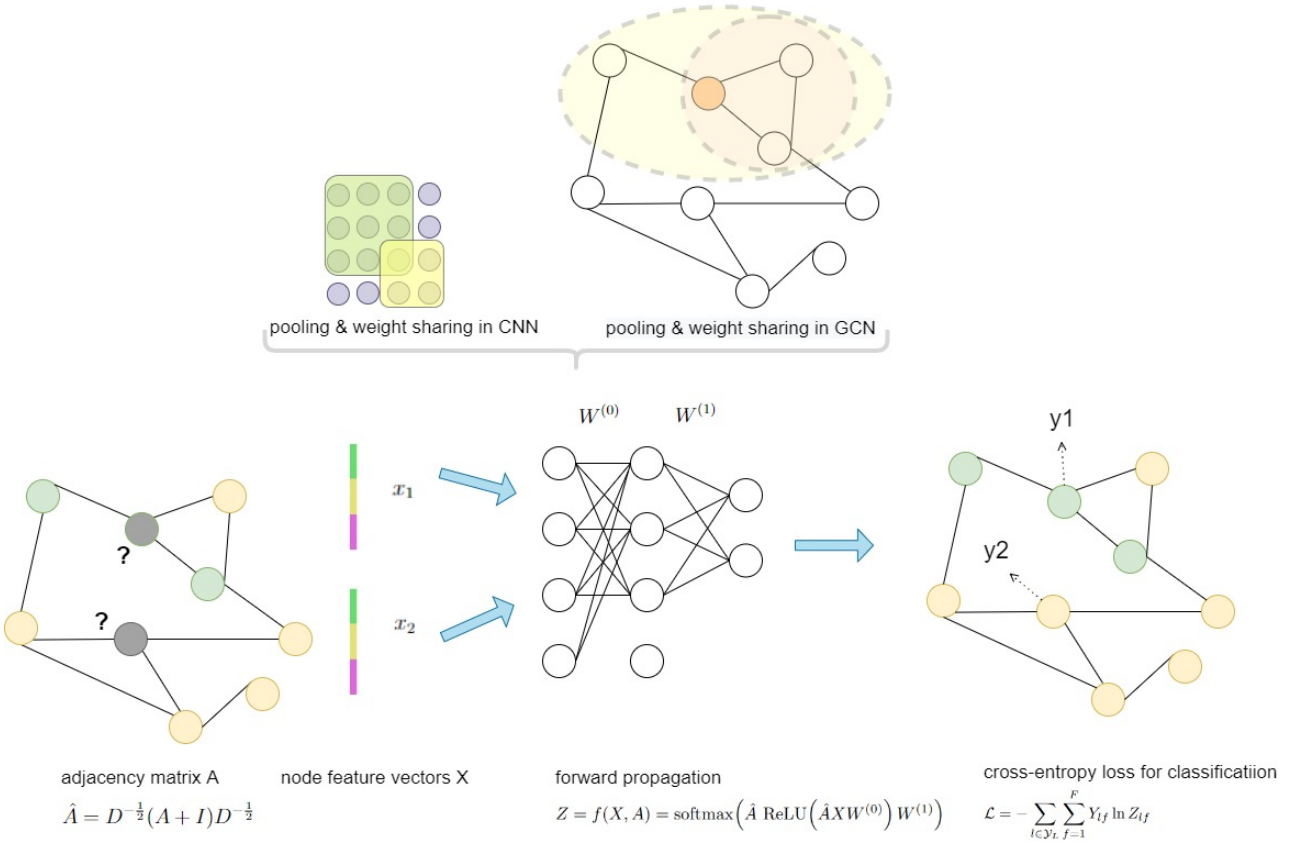



Figure 2.4: Classification with GCN (architecture)

### 2.6.3 Clustering on graphs

In order to utilize the graph structure, generative graph models are often used to detect community structures (Sun et al., 2019; Kolda et al., 2014; Yuan et al., 2022). Node clustering is a common method in graph analysis. Recent work on graph clustering may be categorized into probabilistic inference or spectral clustering methods. A nice review could be found in Balasubramanyan, Lin, and Cohen (2010). Probabilistic techniques propose a generative model for graphs, while spectral clustering and other graph partition algorithms leverage network properties to handle particular graph applications. In this project, we adopt the probabilistic approach to cluster companies with Graph Neural Networks (GNNs) as recent advances in GNN have been able to aid with the challenging graph clustering problem (Tsitsulin et al., 2020; Shchur and Günnemann, 2019).

### 2.6.4 Explaining classification on graphs

The pioneering work on generating entity-specific explanations for graph algorithms such as GNNExplainer (Ying et al., 2019) exists. This technique involves jointly training two neural network models to accomplish the prediction job and simultaneously learning to create explanations by selecting a subgraph of an input graph that has a significant association with the prediction. However, because this technique only considers typical GNN algorithms, it ignores the edge characteristics included in a knowledge graph. As previously stated, it is preferable to construct robust counterfactual explanations for our particular use case by explicitly modeling the common decision logic that exists in a knowledge graph in a way that resembles how humans intuitively grasp the logic in knowledge graphs.

# Chapter 3

# Dataset Construction & Preprocessing

The previous chapters provided an overview of the challenges, research questions, and relevant literature in the field of sustainability investing, as well as the NLP and graph machine learning techniques that we employ in this project. In this section, we first present a data engineering pipeline for creating a domain-specific dataset that can be used to measure companies' SDG performance. Then we demonstrate how we preprocess the text data and knowledge graphs so that we can identify essential information from the data for each SDG for different companies.

## 3.1 Pipeline for constructing sustainability dataset

In Section 2.2, we discuss the various sustainability datasets that are currently available, as well as their limitations. In Figure 1.2, we demonstrate that we collect data from a variety of sources in order to evaluate companies' contributions to the SDGs. Prior to gathering all of the information, the official websites of each company as well as the Wikipedia website are gathered. The company's official website is used to improve the pipeline of collecting sustainability reports. As the same firm may appear under several names in different data resources, the company's Wikipedia page is utilized as a common identifier of the same company across different databases. Multiple data sources spanning various aspects of sustainability are taken into account when trying to assess a company's sustainability performance[1]. In short, we collect the following data from several sources:

- The business descriptions of companies that include information about their products, services, and raw materials used in manufacturing. This information is available on their Wikipedia webpages.
- Sustainability reports provided by the companies. It encompasses all forms of sustainability reporting. As part of our efforts to assess how products from different companies contribute to each of the SDGs, we web scrape every company's sustainability report from their official website. However, it is critical to remember that green-washing occurs. Companies have an incentive to demonstrate only the positive aspects of their operations.
- News where companies are mentioned. A company's operational side of business is evaluated using information from the news, as news data may be used to document some of a company's most significant events regarding its operations. However, media coverage may be skewed (e.g. depending on who wrote the news and who published it). Therefore, we need to identify significant news for a company throughout the year.

[1] A brief view of our collected data can be viewed here: https://drive.google.com/file/d/1HPza7BKTGiS74Fn0-AiqnTGs7SKqE_WW/view?usp=sharing or in the code repository for readers to assess the quality of gathered data.

- Relations between companies and organizations are obtained using the Wikidata Knowledge Graph. It contains fundamental and structural information about a firm, such as industry, ownership, parent organization, supply chain activities (distributed by, operational area), and so on. It includes statements/claims describing how this corporation is connected to another company or entity in Wikipedia's Knowledge Graph, such as "A is an owner/instance/part of B,", "A is rewarded by B," and so on.

### 3.1.1 An accurate and fast approach for web-scraping the sustainability reports via Learning to Rank


Figure 3.1: Pipeline for web-scrapping the sustainability reports

As indicated in Section 1.3, one of the challenges to collecting sustainability reports is fragmented information sources: platforms such as UN Global Compact only cover a small number of companies in our interests as sustainability reporting is not sufficiently regulated at present. As a common practice, companies are most likely to provide their sustainability reports on their own websites. Therefore, we may recursively loop through each company's website to download all the files and detect which files contain sustainability relevant information. This method, however, is time-consuming because there could be hundreds or even thousands of links within the domain of a large corporation. In order to address this issue, we may conduct a Google/Bing search using the company's name and sustainability-related keywords in order to identify web-pages that may contain the company's sustainability report. However, because the web-pages ranked by Google/Bing could be quite noisy, we may restrict the Google/Bing search to only the official domains of each company. The overall procedure is depicted in Figure 3.1, and the detailed pipeline is explained as follows:

1. At the beginning, we retrieve each company's official website by using the Bing Web Search API: *search_term = company_name + "official website"*. The domain of the company is obtained through retrieving the result ranked first returned by the API.

2. After we obtain the domain for each company, we compose the query *search_term = "site:"+ company_domain + company_name + "sustainability environmental social governance report"* for the Bing Web Search API to look for the most relevant pages that could potentially contain sustainability reports of sustainability related information on each company's official website. The domain parameter is added to improve the search

accuracy and also as a restriction to make sure the search engine only considers the results listed on the company's official website.

3. Then we retrieve the top ten websites' urls returned by the API for each company. Optionally, an examination procedure could be performed to check if some keywords occur in the content of the websites in both steps to assure the quality of the retrieved results. In our implementation, we build a search engine to further rank the top 10 urls of the retrieved websites with keywords *recent sustainability reports 2022 2021 2020 2019* in order to obtain the most recent sustainability reports.The top three urls from our search engine are used for the next few phases in the process.

4. Following the acquisition of the websites of the companies that provide sustainability-related information, we search for files in pdf format on these pages and extract the URLs to download the reports. We extract the contents (text) of these websites as an alternate solution for sustainability reports if no pdf file can be found on the website[2].

### 3.1.2 Pipeline for extracting news from GDELT Global Entity Graph: Entity Linking through Wikipedia

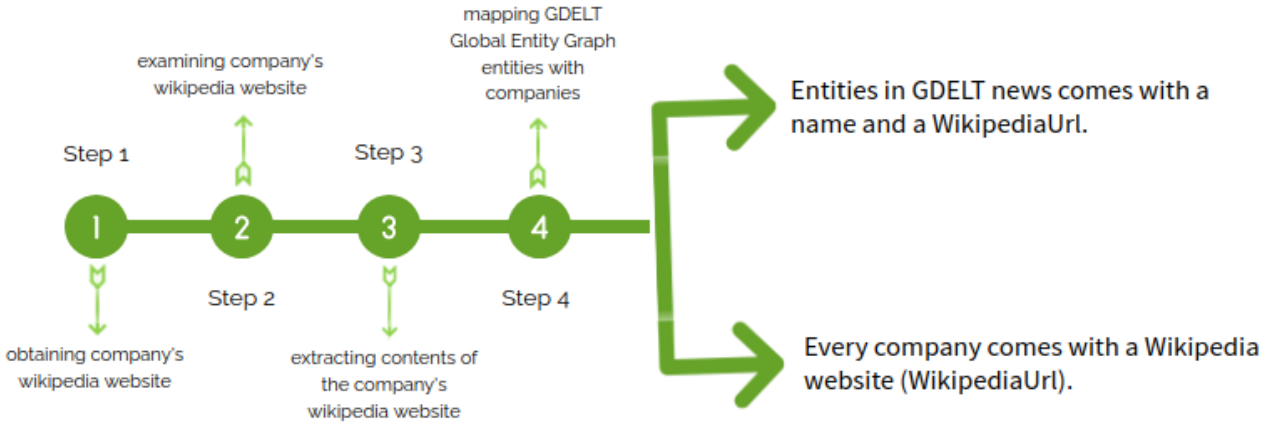



Figure 3.2: Pipeline for extracting company news.

In order to measure company's operations, news are retrieved from GDELT (Leetaru and Schrodt, 2013) for every company. GDELT is an open-source project that analyzes the world's news media in over 100 languages, in print, broadcast, and online formats. The GDELT Global Entity Graph [3] is a dataset of metadata entity annotations for the news in 11 languages where every piece of news contain a timestamp, an url to retrieve the news, the language of the news, the sentiment measures (sentiment score, polarity and magnitude of the sentiment) obtained through Google Natural Language Processing API, a list of distinct entities identified in the news and the number of mentions for each entities in the news. Every entitiy is identified through a WikipediaUrl and a Google id. We retrieve the news for the events happened in 2021 for a basket of companies of our interests.

As the Wikipedia website is used as a common identifier, we first obtain each company's Wikipedia website following the same procedure as in Section 3.1.1. Now that we have been

[2]The content (text) of the websites is also extracted for companies with reports on these websites.

[3]https://blog.gdeltproject.org/announcing-the-global-entity-graph-geg-and-a-new-11-billion-entity-dataset/

able to identify firms through the use of the *WikipediaUrl* as common identifier, we can filter the news in the GDELT Global Entity Graph database to ensure that only news pertaining to the companies of interest is retained. The overall procedure is illustrated in Figure 3.2.

Wikipedia is a powerful tool for entity recognition and linking from the text. We link companies from different databases or data resources using Wikipedia/Wikidata as linkers when a common identifier code is not available. In the GDELT database, they adopt the Google Natural Language API for entity recognition and sentiment analysis. There are also other tools which could be used for this purpose, such as DBpedia SPotlight (Mendes et al., 2011) or OpenTapioca (Delpeuch, 2020), which can annotate text with locations, organizations, and people from Wikidata. Additionally, we could obtain relations, statements, or claims from the Wikidata API, which could be utilized to form a knowledge graph in the following step.

### 3.1.3 Connecting the Dots: Constructing Knowledge Graphs

In order to represent the structural relationships that exist between the companies, we use Wikidata API to extract the claims for each company. The claim is a triple of the form *[head, relation, tail]*. The procedure of constructing knowledge graphs is as follows:

1. In the first round, we seek out all the relationships that each firm has with other entities that have a *Wikidata ID* in Wikidata.

2. In the second round, we look for the companies whose tails we obtained in the first round that are not companies of our interest. These tails will then be treated as heads and repeat the procedure in the first round in order to obtain more related entities in Wikidata knowledge graphs.

3. Iteratively repeating the procedure above twice more results in the final knowledge graph. The reason we repeat the procedure twice is that we believe it is adequate to consider companies reachable within 1-4 steps by connecting to shared entities.

The summary statistics of the constructed knowledge graph among 1852 companies of interest are shown in Table 3.1. Figure 3.3 shows the bar plot of the most frequently occurring 30 relations. Figure 3.4 shows the degree distribution of the nodes. All the relations from the extracted Wikidata knowledge graph is presented in Figure A.2 along with examples of triples in Figure A.3 in the Appendix.

We can see from Table 3.1 and Figure 3.4 that the major part of the extracted knowledge graph is very sparsely connected, while a small part of the graph is very densely connected. We use the directed knowledge graph as a starting point for further preprocessing this graph to obtain a company graph. Section 3.2.5 will provide the specifics.

Table 3.1: Original Graph Statistics

| | |
|---|---|
| Number of nodes | 74,840 |
| Number of edges | 160,994 |
| Number of types of relations | 610 |
| Average degree | 4.3024 |
| Density | 5.75E-05 |
| Clustering coefficient | 0.0587 |
| Triadic closure | 0.0074 |

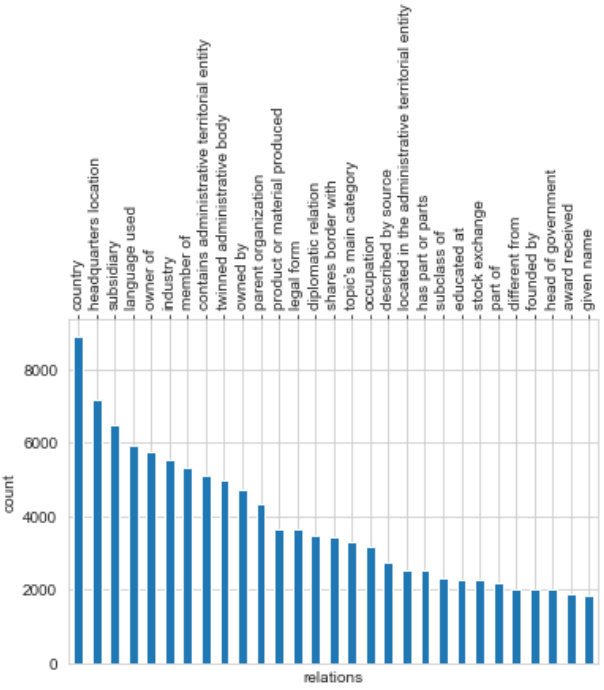



Figure 3.3: Histogram of top 30 relations

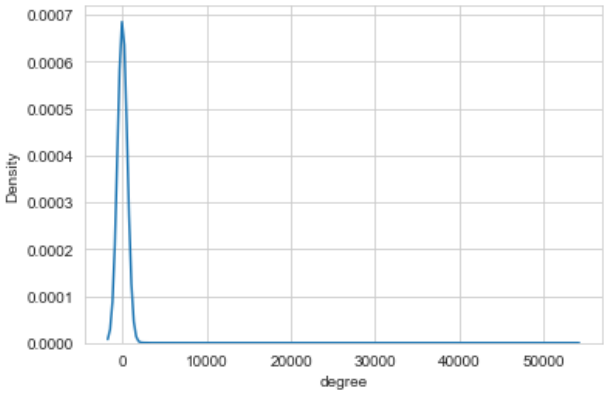



Figure 3.4: The node degree distribution

## 3.2 Data Preprocessing

### 3.2.1 Motivation

As for the sustainability reports, most companies' reports only contain a few sentences about each SDG, therefore modeling all the text with the methods discussed in Section 2.4.2 will result in poor classification performance because the relevant features (key words/phrases) only make up a very tiny percentage of total features in text representation models like BoW. Hence, we need to do some feature engineering beforehand to select the sentences relevant to each SDG in a company's sustainability reports.

In terms of Wikipedia descriptions, some well-known corporations have considerably more information spanning from history to controversies than others. The goal is to extract statements that characterize the company's business (products/services/materials). In terms of media coverage, some corporations such as Amazon, Microsoft and Tesla are frequently mentioned throughout the year. Filtering the news is essential in order to identify and distinguish the most important news for a company. The corporate nodes are fairly sparsely dispersed over the whole knowledge network. It would be preferable to create a denser corporate graph with just the sorts of connections that are relevant to our situation. For example, relationships characterize organizations' supply chain operations are more essential than their headquarters.

### 3.2.2 Sustainability Reports

There are two major motivations for analyzing the reports with machines: (1) machines are faster than humans in processing information; (2) machines are able to tackle some limitations of humans such as understanding different languages (multi-lingual models), or finding complex patterns or interactions. In this section, we would like machines to resemble how human analysts would read and analyze sustainability reports. In order to understand the context and reason behind the relations between the sentences, transformer-based methods are used to transfer a

series of the human procedures above to machine procedures as shown in Figure 3.5.

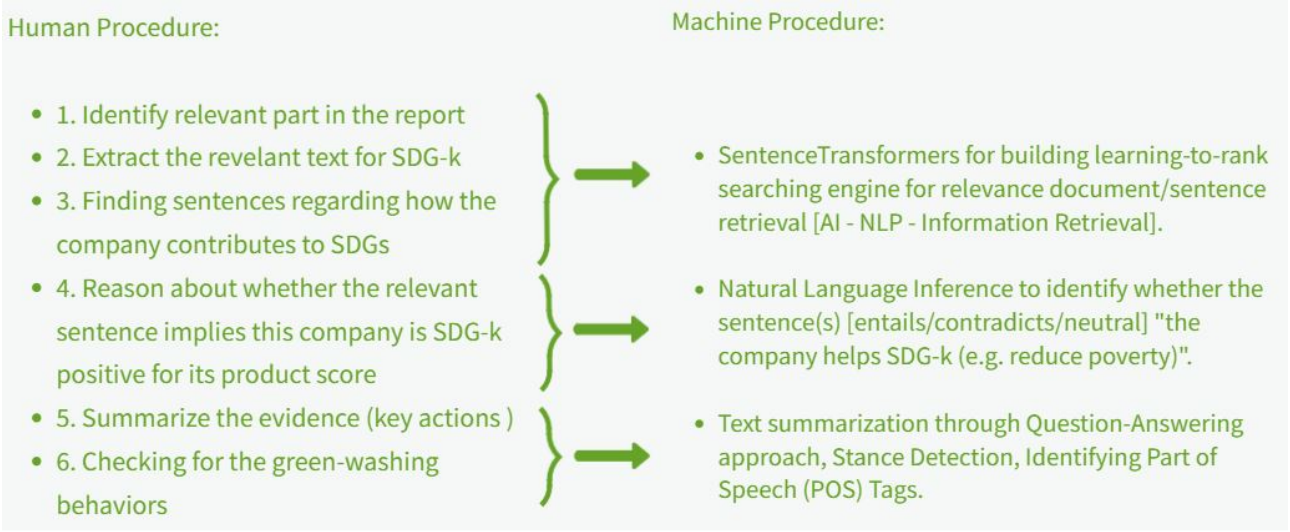



Figure 3.5: Pipeline for analyzing the sustainability reports

Table 3.2: Sample of summarized key actions for SDG 13: Climate Actions

| **company** | **resource** | **summarized key actions** |
|---|---|---|
| Cleanaway Waste Management Ltd | web | by reducing our greenhouse gas emissions , by the responsible management of our landfill gas , and by assisting our customers and the community in managing their waste impacts |
| Singapore Telecommunications Ltd | report | undertook a science based targets programme and engaged experts on developing science based targets |
| GlaxoSmithKline PLC | web | our climate strategy covers the full value chain of emissions reductions |
| Mapfre SA | web | protects the environment through public commitments |
| Carrefour SA | report | structured its climate action plan around three priority areas |
| Swiss Re AG | report | we use our existing processes and instruments to address climate - relat |
| Solvay SA | web | raising the bar |
| Crown Holdings Inc | web | drive climate action throughout our value chain |
| Enagas SA | report | through efficient use of energy |
| Cie Generale des Etablissements Michelin SCA | report | taking action both downstream from its operations to ght climate change , conserve natural protect objectives for 2050 to make all the production plants , supply chain operations and raw material and component inputs carbon neutral |
| NextEra Energy Inc | report | prepare our business to adapt to the effects of climate change |
| Reckitt Benckiser Group PLC | report | our business strategy sees the macro-trend of climate change as a key factor influencing both our development and society |
| McDonald's Corp | report | conducting and analyzing scenario modeling to understand the transition and physical risks |
| Prologis Inc | report | measuring our impact , reducing emissions wherever possible and investing in new technologies and offsets for emissions that , today , cannot be avoided |
| CF Industries Holdings Inc | report | leverage our unique capabilities to accelerate the worlds transition to clean energy |
| NEL ASA | report | incorporate sustainability into our strategic decision-making processes |
| SSE PLC | report | actively and positively advocates to drive accelerated climate action to achieve net zero |
| Catalent Inc | web | reducing our carbon footprint |
| Elia Group SA/NV | report | building our grid |
| Jubilant Pharmova Ltd | web | we reduced specific ghg emissions from our facilities year on year , by continuously working on reduction in energy consumption , waste generation , increase in renewable energy share and enhancing the carbon sinks by planting more trees |

The procedure for analyzing sustainability reports is as follows:

1. We extract sentences that contain the pronouns we/our/us/ourselves/business name in order to discover the sentences in which the firm is mentioned in their sustainability reports.

2. Then, a semantic search engine is developed utilizing a query comprised of the concatenated keywords[4] acquired from the Sustainable Development Solutions Network[5] to find relevant SDG statements in the sustainability report.

3. Next, we pick a threshold relevance score and filter out statements that are deemed less relevant to the given SDG by qualitatively looking at the ranked results.

4. The Natural Language Inference (NLI) method is used to determine if the relevant sentences suggest that the firm is assisting in the achievement of each SDG's goals.

5. Following that, we gathered all of the sentences for each firm that indicate that the company is contributing to the specific SDG. The evidence acquired in order to establish whether the company is contributing to a specific SDG will be used for the classification task in the next section. A sample of summarized key actions for SDG 13 is shown in Table 3.2.


Figure 3.6: Wordcloud for SDG 13: Climate Action evidence found for all companies

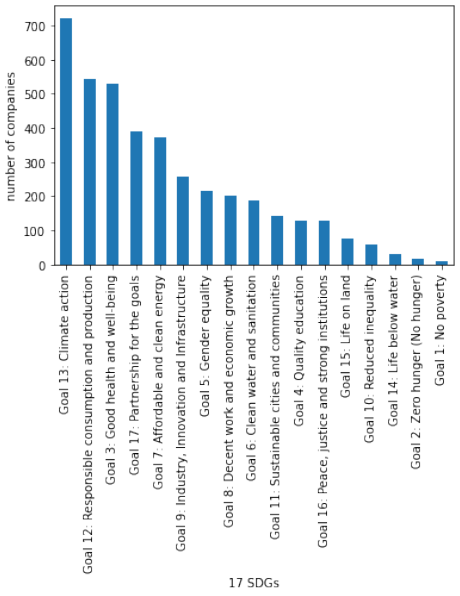



Figure 3.7: Number of companies with evidence found for 17 SDGs

In Figure 3.6, we generate a wordcloud of all the sentences with evidence discovered in companies' sustainability reports for SDG 13 (Climate Action). We can see that climate change, greenhouse gases, the environment, and sustainability are some of the most frequently occurring terms. Moreover, as shown in Figure 3.7, SDG 13 (Climate Action), SDG 12 (Responsible consumption and production), and SDG 3 (Good health and well-being) have the highest number of firms that have evidence to back up their claims.

Moreover, certain SDG topics may overlap, while others may have fewer commonalities. In order to find distinguished terms from the sentences with evidence and check how two categories

[4]The keywords list is available at this link: https://drive.google.com/file/d/1IJIUvS9tydCMxHE0T45OePGNCx7v9Q6r/view?usp=sharing

[5]https://www.unsdsn.org/about-us

of text are different from each other, we use the tool scattertext (Kessler, 2017) to visualize the difference between SDG 13 (Climate Action) and SDG 12 (Responsible consumption and production), and the difference between SDG 13 (Climate Action) and SDG 9 (Industry, Innovation and Infrastructure) in Figure 3.8 and Figure 3.9. The most associated terms for SDG 9 (Industry, Innovation and Infrastructure), SDG 12 (Responsible consumption and production), and SDG 13 (Climate Action) are illustrated in Table 3.3. Among the sentences with evidence, we can see from Figure 3.8 that both SDG 13 and SDG 12 regularly use terms like green house emissions, supply chain, and carbon footprint. This frequently occurs for for both SDG 13 (Climate Action) and SDG 12 (Responsible consumption and production). Nevertheless, as demonstrated in Table 3.3, the most associated terms are very distinct. However, there is no overlap between frequently occurring terms in SDG 13 (Climate Action) and SDG 9 (Industry, Innovation and Infrastructure) as shown in Figure 3.9.

Table 3.3: Most associated terms for SDG9, SDG12, SDG13

| **SDG9:Industry, Innovation and Infrastructure** | **SDG12:Responsible Consumption and Production** | **SDG13:Climate Action** |
|---|---|---|
| good health | raw materials | force on climate |
| sustainable cities | sustainable consumption | financial disclosures |
| economic growth | responsible sourcing | task force |
| decent work | responsible consumption | climate risk |
| resilient infrastructure | food waste | task force on climate |
| clean water | waste generation | scenario analysis |
| responsible consumption | hazardous waste | risk management |
| energy infrastructure | efficient use | cdp climate |
| greater metropolitan | waste management | global warming |
| back overview | product development | cdp climate change |
| social infrastructure | water use | paris agreement |
| quality education | water consumption | climate risks |
| community investment | circular economy | united nations |
| greater metropolitan area | zero waste | impacts of climate |

### 3.2.3 Wikipedia Product Information

The Wikipedia Product Information is retrieved through a search engine with a query made of "products materials service industry environment" similar to Step 2 in Section 3.2.2 for preprocessing sustainability reports. In Table 3.4 we provide a sample of collected Wikipedia product information.

### 3.2.4 Company News

The motivation for using news to assess SDGs is that news is an effective measure of companies' operational procedures. In order to retrieve valuable information from the news, the main challenge is related to how to identify important news for a company. Firstly, we need to determine the impactfulness of the news for the company. Secondly, we need to recognize the news describing the same events. In order to evaluate the impactfulness of the news, we create an aggregated sentiment score which takes sentiment, magnitude of sentiment, and the number of times the company is mentioned in this news into consideration (aggregated sentiment score = sentiment $\times$ magnitude of sentiment $\times$ number of times the company is mentioned in the news article). Threshold ($\sigma = 0.55$) filter based on the percentage similarity of news headers is used to filter out the news describing the same event. A comparison between the coverage of important news through our methods and through the MSCI SDG framework is made in Table 3.5 and Table 3.6 using Vodafone Group as an example. The primary difference is that MSCI has a propensity to select a piece of news in each nation that may be used to assess the operational side of businesses, while our technique is more based on media attention. The news items that MSCI selects frequently overlap with some of our selections of significant news, which implies that


Figure 3.8: Difference between SDG 12 and SDG 13. The figure is transposed.


Figure 3.9: Difference between SDG 9 and SDG 13. The figure is transposed.

Table 3.4: Wikipedia product information (sample)

| **company** | **product information** |
|---|---|
| SGL Carbon SE | It is one of the worlds leading manufacturers of products from 29 production sites around the globe (16 in Europe, 8 in North America and 5 in Asia), and a service network in over 100 countries, SGL Carbon is a globally operating company |
| Gerdau SA | These products are used in different sectors, such as industry, metallurgy, farming and livestock, civil construction, automotive industries, petrochemicals, railway and naval sectors, in addition to orthodontic, medical and food areas |
| Bridgestone Corp | Today, Bridgestone diversified operations encompass automotive components, industrial products, polyurethane foam products, construction materials, parts and materials for electronic equipment, bicycles and sporting goods |
| MTS Systems Corp | The companys products and services support customers in research and development and QAQC testing of products through the physical characterization of materials, such as ceramics, composites and steel |
| Berkshire Hathaway Inc | Moore formulates, manufactures, and sells architectural coatings that are available primarily in the United States and 2001, Berkshire acquired three additional building products companies |
| ANDRITZ AG | Xerium Technologies is a global manufacturer and supplier of machine clothing (forming fabrics, press felts, drying fabrics) and roll covers for paper, tissue, and board machines |
| 3D Systems Corp | Applications and industries 3D Systems products and services are used across industries to assist, either in part or in full, the design, manufacture andor marketing processes |
| International Paper Co | At the time of sale, Temple-Inlands corrugated packaging operation consisted of 7 mills and 59 converting facilities as well as the building products operation |
| Sasol Ltd | These products are used in the production process of numerous everyday products made worldwide and benefit the lives of millions of people around the world |
| Grasim Industries Ltd | This company supplies to sectors such as food, textiles, electrical and electronics, composites, leather, plastics and automobiles |

MSCI likely sorts news by country before picking the significant news in each country. However, it is unclear how counties are prioritized when it comes to gathering news for a company.

Table 3.5: GDELT most impactful news - Vodafone Group Example

| company | date | news header | aggregated sentiment score |
|---|---|---|---|
| Vodafone Group PLC | 2021-01-17 05:32:49+00:00 | Hey Teacher, Zoom those kids - Independent.ie | -105.3 |
| Vodafone Group PLC | 2021-05-06 11:02:32+00:00 | GRACE ON THE CASE: How do I get bad Vodafone credit marks removed? | -104.64 |
| Vodafone Group PLC | 2021-06-04 15:17:12+00:00 | March 26th and the aftermath - where next for the anti-cuts movement? | -79.02 |
| Vodafone Group PLC | 2021-01-29 00:32:37+00:00 | India must avoid the retrospective tax mess - Northlines | -63.9 |
| Vodafone Group PLC | 2021-05-04 14:33:03+00:00 | Vodafone Healthline Season 9: A Remarkable Journey Of Health Education And Life-Changing Surgeries | 210 |
| Vodafone Group PLC | 2021-05-03 16:18:20+00:00 | Vodafone and Google Cloud to Develop Industry-First Global Data Platform | 248.46 |

Table 3.6: MSCI most impactful news - Vodafone Group Example

| company | news header |
|---|---|
| Vodafone Group PLC | Egypt: National Telecom Regulatory Authority imposed EGP 7 million penalty over non-compliance with number portability regulation |
| Vodafone Group PLC | India: Tax issue over the acquisition of Hutchison-Essar |
| Vodafone Group PLC | UK: Employment Tribunal complaint over alleged racial discrimination, unfair dismissal and unauthorized wage deductions |
| Vodafone Group PLC | Portugal: Tax dispute over value added tax payable for early termination fees. |
| Vodafone Group PLC | Vodafone Italy SpA: EUR 773,000 AGCOM penalty over alleged lack of transparency regarding costs on unlimited internet service offer |
| Vodafone Group PLC | Italy: EUR 500,000 of AGCM penalty over alleged foreign IBAN discrimination |
| Vodafone Group PLC | Romania: NSAPDP penalty over alleged breach of data protection rules |
| Vodafone Group PLC | Spain: EUR 8.15 million AEPD penalty over alleged violation of data protection regulations in marketing campaigns |

### 3.2.5 Knowledge graph pre-processing

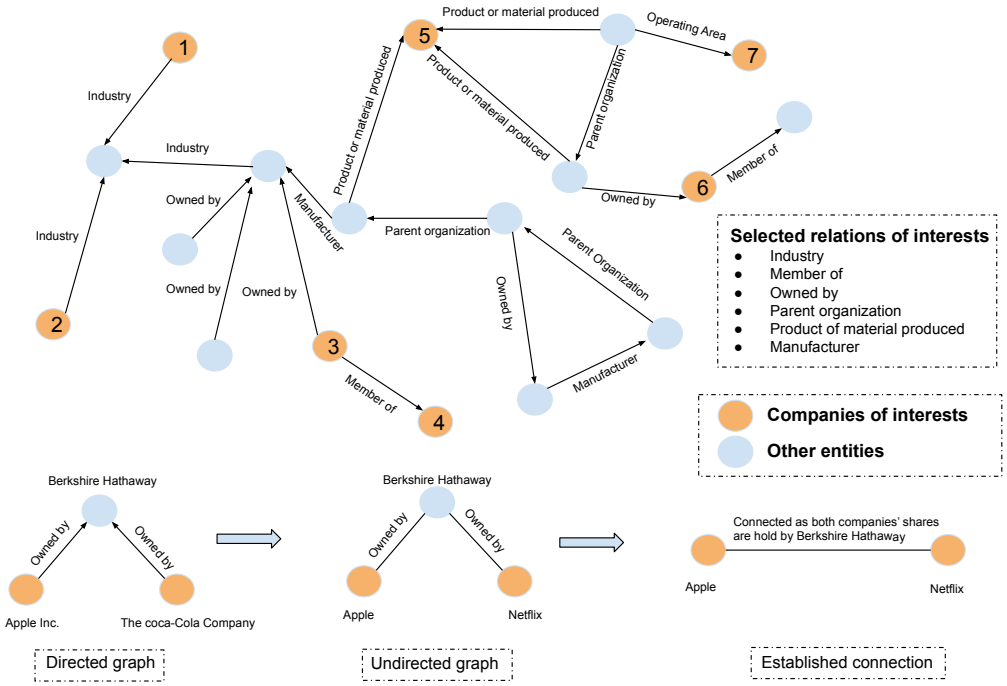



Figure 3.10: Schematic Diagram: Re-compose the original graph. The directed graph is transformed to an undirected graph, and only relations of our interest are used to compose the new company graph. The intuition for transforming a directed graph to an undirected graph is also shown at the bottom of the schema.

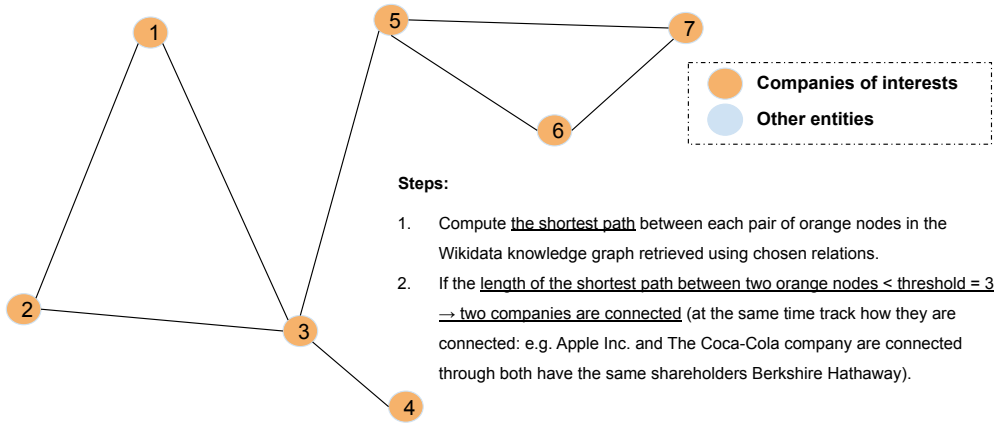



Figure 3.11: Schematic Diagram: Simplify the knowledge graph to a graph with only company nodes. We call the new graph company graph, where we only consider companies of interest that can be reachable within two steps in the undirected knowledge graph of selected relations from the previous step (Figure 3.10).

Table 3.7: Company graph statistics

| | All Relations | Selected Relations |
|---|---|---|
| Number of nodes | 1,783 | 1,426 |
| Number of edges | 1,335,313 | 35,890 |
| Number of components | 6 | 78 |
| Average degree | 1,497.8273 | 50.3366 |
| Density | 0.8405 | 0.0353 |
| Clustering coefficient | 0.9484 | 0.6884 |
| Triadic closure | 0.9504 | 0.7354 |

The extracted Wikidata Knowledge Graph with companies will be preprocessed before conducting graph classification or clustering tasks. The fundamental concept behind this endeavor is that we recreate a corporate graph using the initial knowledge graph extracted from Wikidata. As shown in Figure 3.10, we begin by selecting relationships of interest, then convert the original directed knowledge graph into an undirected graph and create a direct link between two company nodes. The reason for converting the graph into an undirected one is illustrated at the bottom of Figure 3.10 where two entities are connected through different types of relations to a common entity and we are only concerned with whether they are connected rather than the direction of information flow. For instance, if A provides environmentally friendly products to B, B may be affected by A by scoring higher in sustainability ratings, but over time, A will likely receive more orders from B and other companies, which encourages it to continue providing sustainable materials, resulting in A receiving higher sustainability ratings.

Table 3.8: All relations

| distance | count | percentage |
|---|---|---|
| 2 | 1,335,044 | 84.04% |
| 3 | 209,481 | 13.19% |
| 4 | 36,981 | 2.33% |
| 5 | 5,732 | 0.36% |
| 6 | 1,128 | 0.07% |
| 1 | 269 | 0.02% |
| 7 | 17 | 0.00% |
| 8 | 1 | 0.00% |

Table 3.9: Selected relations

| distance | count | percentage |
|---|---|---|
| 6 | 261,146 | 0.259935 |
| 4 | 239,699 | 0.238588 |
| 5 | 235,029 | 0.233940 |
| 7 | 113,603 | 0.113076 |
| 8 | 54,839 | 0.054585 |
| 3 | 39,242 | 0.039060 |
| 2 | 35,714 | 0.035548 |
| 9 | 19,065 | 0.018977 |
| 10 | 4,820 | 0.004798 |
| 11 | 1,153 | 0.001148 |
| 1 | 176 | 0.000175 |
| 12 | 148 | 0.000147 |
| 13 | 23 | 0.000023 |

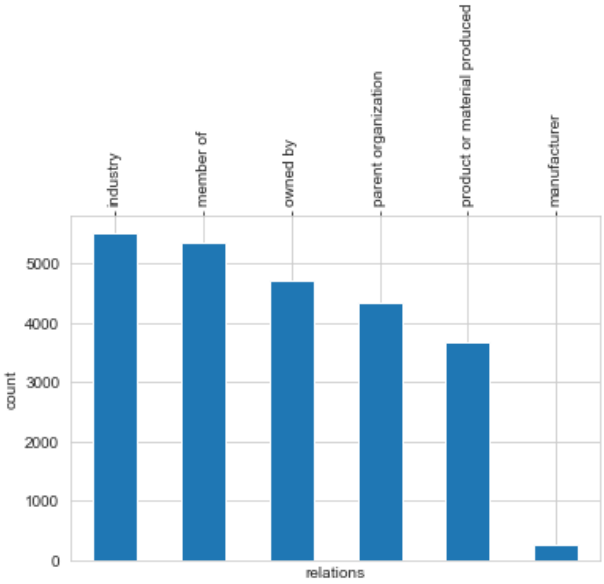



Figure 3.12: Histogram of selected relations

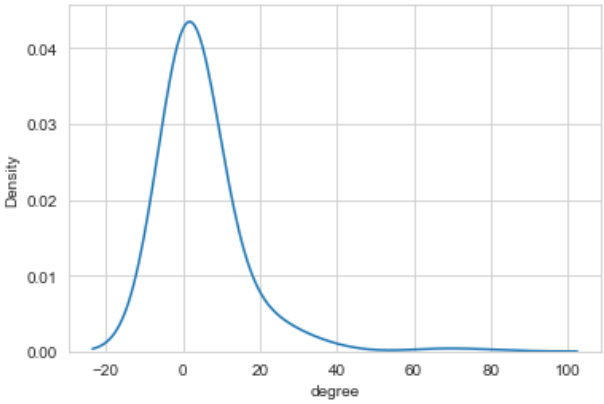



Figure 3.13: The node degree distribution of selected relations

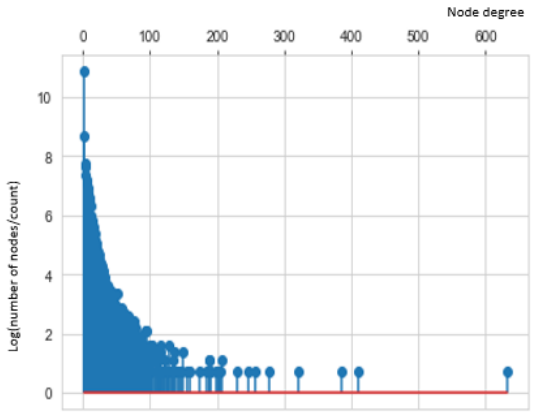



Figure 3.14: Lollipop chart of original graph

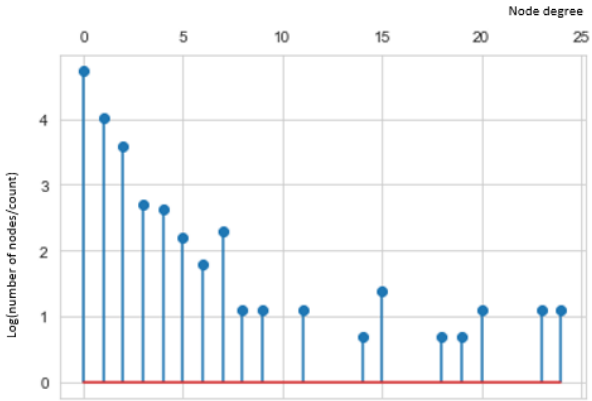



Figure 3.15: Lollipop chart of company graph

The detailed preprocessing procedure is as follows:

1. For each pair of companies, we calculate the minimum distance between them in the

graph as shown in Table 3.8. We can see that there are 0.02% of pairs of companies that are reachable within 1 step (directly connected), and 84.04% of companies that are reachable within 2 steps in the original extracted knowledge graph.

2. We apply a threshold to the minimum distance matrix between companies: if two companies are reachable within 2 steps ($\sigma_{\text{dist}} < 3$), we draw an edge between them. In this way, we construct a new company graph as shown in Figure 3.11. The reason we pick 2 as the threshold is that there are other companies in the knowledge graph that are not of our interests, and it is good to take them into consideration as they could be direct suppliers or shareholders of companies in our interests (for classification). The connections between the companies get weaker when the threshold is raised.

3. Additionally, we can see from the node distribution plot in Figure 3.4 and the statistics of this new company graph shown in column 1 of Table 3.7 that the company graph is very densely connected when all types of relations are considered.

4. However, not all types of relations are highly relevant for describing a company's supply chain activities, therefore we select a subset of relations of interests (*industry, member of, owned by, parent organization, product or material produced, manufacturer*) as shown in Figure 3.12. There are two reasons for selecting these relations: (1) Shareholder information and supply chain activities[6] are important factors to take into account when evaluating companies' SDG performance. (2) These chosen relations are among the most frequently occurring types of relations in the knowledge graph and can therefore cover the majority of the connections of interests. The graph statistics of this new company graph with selected relations are shown in column 2 of Table 3.7, and the minimum distance between each pair of companies is shown in Table 3.9. We can see in node distribution plot Figure 3.13 that this new graph is less dense than the original extracted graph. Additionally, Figure 3.14 and Figure 3.15 depict a comparison of node degree counts between the original graph and the corporate graph.

[6]This article from the United Nations discusses the importance of the supply chain for SDGs: https://www.unglobalcompact.org/take-action/leadership/integrate-sustainability/roadmap/supply-chain

# Chapter 4

# Experiments & Results

In this section, we utilize the data we collected and preprocessed from the previous section to classify companies' SDG scores, then examine the results to see if our data can accurately predict the SDG scores of existing frameworks. We further analyze the results to determine which data elements are most important in predicting SDG product and operation scores, as well as whether network structure information could enhance the classification performance. Finally, we propose a novel approach for generating SDG scores that incorporates data from many sources and provides explanations for the resulting SDG scores.

## 4.1 A classification-based analysis of the existing SDG frameworks

### 4.1.1 Motivation

#### General motivation

In the previous section, we applied modern NLP techniques and recent NLP breakthroughs in deep bidirectional transformer models to extract information from Wikipedia and sustainability reports, as well as preprocess news and Wikidata knowledge graphs to find evidence of companies contributing to SDGs from both product and operational perspectives. In this section, we take a closer look at two industry-leading SDG frameworks from MSCI and RSAM to see how companies are assessed in terms of their contributions to 17 SDGs. We want to determine how effectively the information we obtained can predict the MSCI SDG score so that we may extend the model to forecast a wider selection of stocks. To conclude, our main objective is to determine whether the open source data we gathered can accurately predict the SDG scores provided by MSCI and RSAM, and to scale up the measurement of firms' contributions to SDGs by combining NLP and machine learning algorithms for classification.

#### Motivation for using graph classification algorithm

Numerous studies (Zhang, Cui, and Zhu, 2018) concentrate on the development of deep learning algorithms for graph-structured data, with a particular emphasis on Graph Neural Networks (GNNs). It is essential to comprehend when graph algorithms are applicable and under what conditions they are beneficial. 2708 scientific papers in the citation network have been included in the Cora dataset, and they've been divided into seven different categories. Figure 4.1 shows the visualization of the Cora dataset (Bodnar, Cangea, and Liò, 2021). We apply both Graph Convolution Networks (GCN) and Balanced Random Forest (BRF) to the cora datasets and vary the fractions used for constructing the training set. Figure 4.2 shows the results of com-

paring the accuracy of the two approaches with various training ratios. When using GCN, it is essential to keep in mind that it employs a sampling strategy that extracts at least one sample from each class. We can see from Figure 4.2 that more training samples raise the performance of BRF from 0.3 to 0.8, while GCN stays at 0.8 consistently. GCN achieves an accuracy of 0.80 with just 0.01 percent of the data utilized for the training set, but BRF only achieves an accuracy of 0.30. However, when additional data is added to the training set, GCN's performance no longer improves. Consequently, GCN seems to be most effective by leveraging the network structure when there is insufficient training data. However, when adequate training examples are available, the network topology does not seem to provide added value. In addition, the structure of this citation network itself offers crucial information for classifying publications (nodes) in this network. In contrast, if the network architecture lacks information relevant to the classification objective, it may not improve classification results. Similarly, GCN and rGCN are used on the company graph and the extracted original Wikidata knowledge graph to see whether the connections have any effect on forecasting MSCI and RSAM SDG scores.

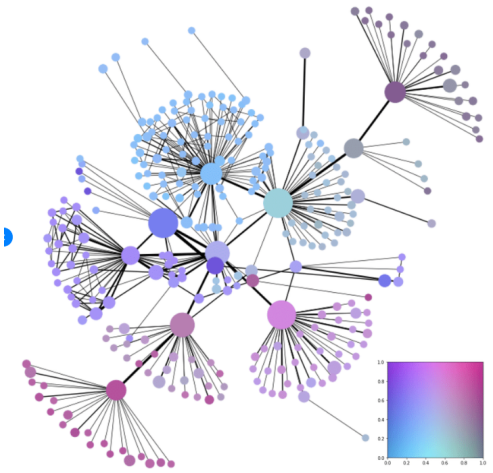


Figure 4.1: Visualization of the Cora dataset: same-colored nodes have the same class. The figure is taken from Bodnar, Cangea, and Liò (2021).

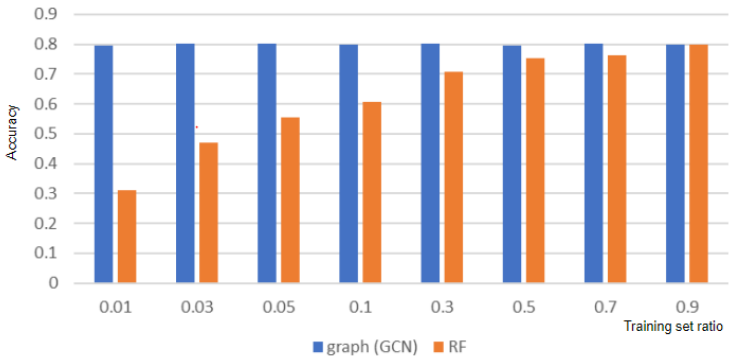



Figure 4.2: Accuracy compared between GCN and RF with different training set ratios.

### 4.1.2 Data: Summary Statistics

We classify MSCI product scores, operation scores, net alignment scores, and RobecoSAM (RSAM) net alignment scores. MSCI scores have 5 categories: Strongly Misaligned, Misaligned, Neutral, Aligned and Strongly Aligned. RSAM's SDG ratings vary from -3 to 3 on the scale. According to the distribution of 17 SDG scores in Figure 4.3, we can observe that the majority of firms have a neutral score, while the other classes have unbalanced data. The distribution of MSCI sectors of the companies is shown in Figure 4.4. The MSCI sector classification is used as a benchmark model. Some SDGs are simpler to classify than others, as seen in Figure 4.5. We can also see from Figure 4.5 that RSAM's classification outcomes using model Balanced Random Forest (BRF) for SDG 8 (Decent Work and Economic Growth), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 10 (Reduced Inequality) seem to yield a much higher F1-micro score than that of MSCI when considering sectors alone.

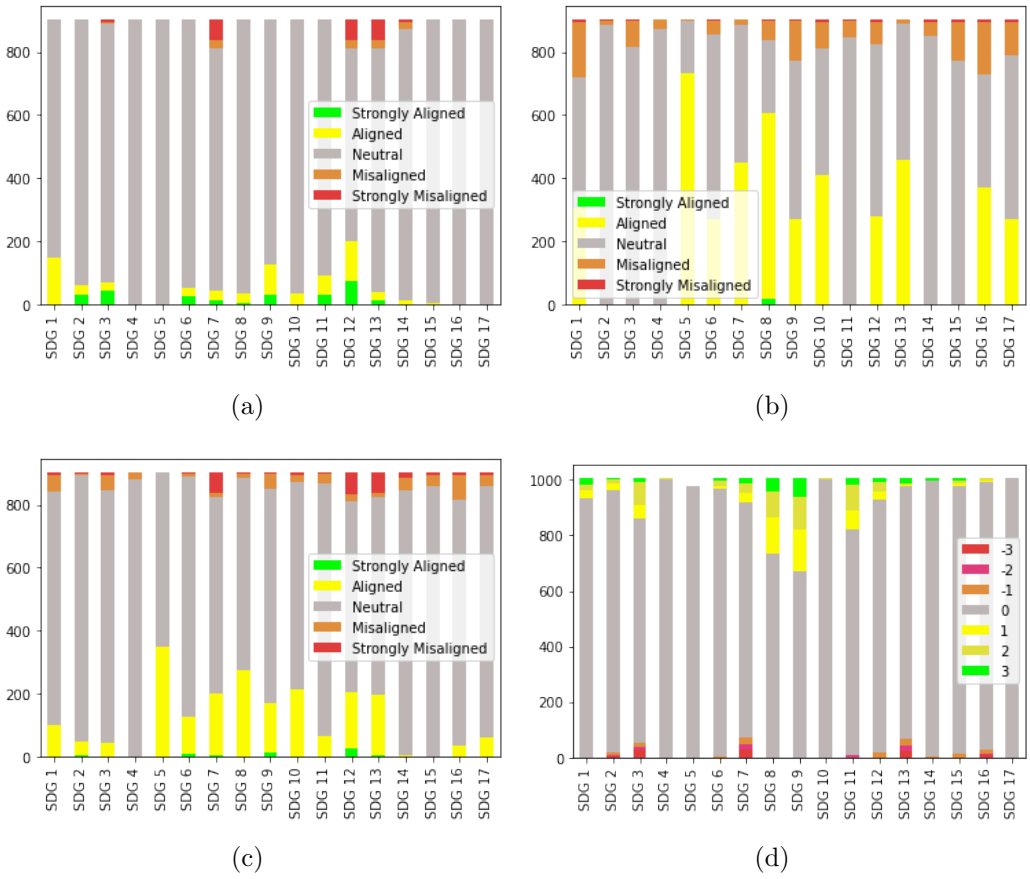



Figure 4.3: (a) Distribution of MSCI product scores (b) Distribution of MSCI operation scores (c) Distribution of MSCI net alignment scores (d) Distribution of RSAM net alignment scores

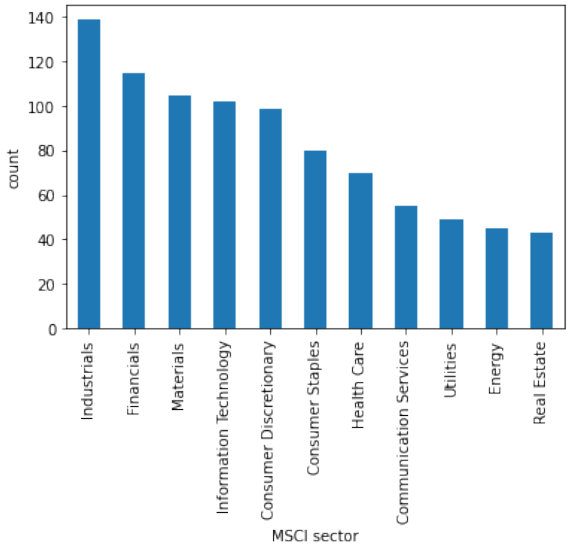



Figure 4.4: Histogram of MSCI sectors for all the companies used in classification

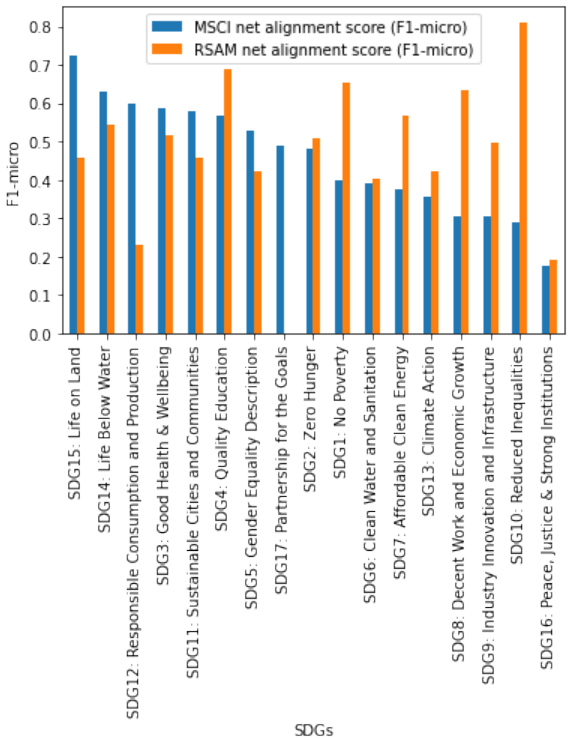



Figure 4.5: Sector classification using Balanced Random Forest (BRF) for MSCI and RSAM net alignment score (measured by F1 micro)

### 4.1.3 Experiment Design

The three major models involved in classifying SDG scores are Balanced Random Forest (BRF), Graph Convolution Networks (GCN) and Relational Graph Convolution Networks (RGCN). BRF is introduced in Section 2.3. As part of the model's incorporation of network structure, we introduce deep learning on graphs for classification purposes in Section 2.6.2. Regarding the use of BRF, we use all the preprocessed information in Section 3.2 as features to predict the SDG scores. In order to classify companies using network structures with GCN and RGCN, the company graph and the original extracted knowledge graph are utilized in conjunction with all the BRF model features.

Multiple groups of experiments are undertaken. The features of the models are shown in Table 4.1, and the details of all the experiments are listed in Table 4.2. The goal of the BRF experimentation groups is to determine if the obtained and preprocessed data can reliably predict the SDG scores of MSCI and RSAM. The purpose of the sets of experiments using graph algorithms is to determine whether the graph containing extra supply chain activities and shareholders' data might enhance the performance of BRF.

There are several issues to consider when it comes to model evaluation. Firstly, applying 5-fold cross-validation directly to the data for evaluating the performance of imbalanced classifiers is inappropriate. Hence, we adopt oversampling of minority classes in order to split data into 5-folds uniformly to enhance the performance of BRF. However, the oversampling method does not function on graphs, and the imbalanced data problem may not pose a threat to applying graph algorithms like GCN, which only requires training splits to have labeled examples from each class.

Table 4.1: Features of the models

| Feature | Feature names |
|---|---|
| (1) | MSCI Sector (one-hot encoded) |
| (2) | Wikipedia product information (BoW) |
| (3) | Sustainability report SDG-k relevant evidence (BoW) |
| (4) | Selected news headers (BoW) |
| (5) | Selected news sentiment statistics (mean, median, max, min, variance, different percentiles) |

Table 4.2: Details of experiments

| Classification | BRF | Graph |
|---|---|---|
| **Algorithm** | Balanced Random Forest | GCN & RGCN |
| **Sampling method** | Oversampling the minority | None |
| **Validation method** | 5-fold cross-validation repeated for 3 times | Cross-validation: Sampling 0.6/0.2/0.2 percentage from each class to compose the training/validation/test set |
| **Evaluation** | F-micro, F-macro | F-micro, F-macro |
| **Experiements** | Group 1: Feature (1)(2)(3) → forecast MSCI product score | Group 1: GCN + Feature (1)(2)(3)(4)(5) |
| | Group 2: Feature (1)(2)(4)(5) → forecast MSCI product score | Group 2: BRF + Feature (1)(2)(3)(4)(5) |
| | Group 3: Feature (1)(2)(3)(4)(5) → forecast MSCI net alignment score | Group 3: BRF + Feature (1) |
| | Group 4: Feature (1)(2)(3)(4)(5) → forecast RSAM net alignment score | |

### 4.1.4 Evaluations

This section contains the results of the experiments reported in Table 4.2. Firstly, we model all the pre-processed data as features in a tabular fashion through balanced random forest. We demonstrate which attributes are most relevant to making forecasts of product, operation, and overall scores. Then, we utilize GCN to model the network topology using the same features

as BRF. This stage aims at determining whether the network topology may improve classification performance. Furthermore, two additional experiments are conducted for the ablation study. In the first experiment, we compare the classification performance of a GCN model that employs all the features to the classification performance of a model that uses no features (featureless approach). The motivation for applying GCN is discussed in Section 4.1.1: Figure 4.2 indicates that GCN seems to be most effective when there is a very limited number of samples for the training set. Therefore, we run a second experiment with a decreased training ratio to determine the contribution of network topology on the classification performance.

As for evaluating the classification performance, we introduce F1 micro, F1 macro and micro-average Precision Recall Curve as evaluation methods in Section 2.3.3. F1 micro would prioritize making the class with the biggest sample size (neutral class in our case) as accurate as possible, while F1 macro is preferable when the minority classes are most valued. Furthermore, the micro-average Precision-Recall Curve is used to evaluate the multi-class classification with respect to each class. A large area under the curve implies both good recall and precision.

**Classification with BRF**

In terms of forecasting the product scores of MSCI, the columns involving Wikipedia extracted product information have the highest F1 micro or macro score as shown in Table 4.3 and Table 4.4. When compared to a benchmark model that just employs MSCI sector information, F1 micro gains by 10% to 40% as shown in column (1). The increase is due to the inclusion of Wikipedia product information. For instance, the F1 micro climbs from 0.438 to 0.885, and the macro rises from 0.331 to 0.601 when Wikipedia product information is provided for forecasting SDG 9 (Industry, Innovation, and Infrastructure). For MSCI SDG product scores, F1-micro has risen by 25.4%, 21.9% and 41.4% after we included all the new information we acquired to estimate BRF (as shown in Table 4.3 and Table 4.5) while F1-macro has increased 23.2%, 11.6% and 42.7% respectively (as shown in Table 4.4 and Table 4.6) for forecasting MSCI product, operation and net alignment scores compared to the baseline model in column (1) of each table which only utilizes MSCI sector information. When it comes to predicting the RSAM net alignment to SDGs, F1-micro and F1-macro scores have risen by an average of 20.3% and 30.6% (as shown in Table 4.5 and Table 4.6).

Incorporating Wikipedia product information seems to be the most helpful for predicting MSCI product scores. For predicting MSCI operation scores, the use of sentiment features seems to aid the minority class in SDG1, SDG4, and SDG8 in achieving improved performance. For forecasting MSCI net alignment scores, including Wikipedia information is still the most effective as net alignment scores aggregate product and operation scores, although additionally integrating all characteristics seems to provide a little higher F1-micro score. Some SDGs seem to receive the highest F1-macro score when news headlines or sentiment measurement features are included. For forecasting RSAM net alignment scores, including sentiment measures seems to increase both F1 micro and macro scores for SDG6 and SDG13, whereas including news headlines might aid the minority class in achieving improved F1-macro performance for SDG11. For net alignment score in column (4) for both MSCI and RSAM in Table 4.5 and Table 4.6, the row with the bolded and underlined number indicates that incorporating Wikipedia product information is most beneficial to forecast this SDG. For net alignment score in column (5) for MSCI or column (6) for RSAM, the row with bolded number which is not underlined suggests that this SDG benefits from all features.

Additionally, Figure 4.6 and Figure 4.7 show the Micro-average PR curve along with iso-f1

curves for each class in terms of forecasting MSCI and RSAM SDG7 net alignment scores. A large area under the curve is indicative of both high recall and high accuracy, where high precision corresponds to a low false positive rate and high recall corresponds to a low false negative rate. We could calculate the average precision scores by utilizing the Area Under the Precision-Recall Curve (PR AUC) for each recall threshold. For MSCI, classes *Strong Misaligned*, *Strong Aligned* and *Neutral* yield the best average precision (AP) above 0.8, while RSAM classes *-2* and *0* achieve the best AP. However, if we set the recall threshold at 0.4, classes *Misaligned* and *Strongly Aligned* from MSCI and classes *-2* and *3* from RSAM would have the same size area under the curve (AUC).

Table 4.3: F1-micro scores of forecasting MSCI product score and operation score with BRF

| **SDGs** | **MSCI Product score** | | | | | | | **MSCI Operation score** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **1** | **2** | **3** | **4** | **5** |
| SDG1: No Poverty | 0.772 | 0.825 | 0.168 | **0.854** | 0.772 | 0.822 | 0.847 | 0.391 | 0.441 | 0.474 | 0.417 | **0.476** |
| SDG2: Zero Hunger | 0.809 | 0.934 | 0.930 | **0.938** | 0.822 | 0.933 | 0.936 | 0.799 | 0.978 | 0.976 | **0.978** | **0.978** |
| SDG3: Good Health and Well-being | 0.622 | 0.965 | 0.852 | **0.966** | 0.849 | 0.962 | 0.966 | 0.735 | 0.956 | 0.955 | 0.956 | **0.961** |
| SDG4: Quality Education | 0.891 | **0.997** | 0.831 | **0.997** | 0.911 | **0.997** | **0.997** | 0.779 | 0.963 | 0.965 | **0.967** | 0.966 |
| SDG5: Gender Equality | 0.763 | **0.996** | 0.363 | **0.996** | 0.889 | **0.996** | **0.996** | 0.581 | 0.791 | 0.848 | 0.850 | **0.858** |
| SDG6: Clean Water and Sanitation | 0.581 | 0.942 | 0.911 | 0.948 | 0.717 | 0.949 | **0.950** | 0.454 | 0.646 | 0.683 | 0.706 | **0.721** |
| SDG7: Affordable and Clean Energy | 0.556 | 0.912 | 0.826 | **0.938** | 0.724 | 0.918 | 0.935 | 0.460 | 0.637 | 0.672 | 0.646 | **0.688** |
| SDG8: Decent Work and Economic Growth | 0.739 | 0.950 | 0.354 | **0.961** | 0.812 | 0.961 | 0.959 | 0.290 | 0.602 | 0.675 | 0.673 | **0.687** |
| SDG9: Industry, Innovation and Infrastructure | 0.438 | 0.881 | 0.727 | **0.885** | 0.627 | 0.880 | 0.883 | 0.438 | 0.583 | 0.599 | **0.647** | 0.627 |
| SDG10: Reduced Inequality | 0.829 | 0.945 | 0.243 | 0.953 | 0.843 | 0.944 | **0.954** | 0.389 | 0.477 | 0.461 | 0.475 | **0.490** |
| SDG11: Sustainable Cities and Communities | 0.682 | 0.897 | 0.881 | 0.913 | 0.709 | 0.899 | **0.913** | 0.788 | 0.938 | 0.943 | 0.935 | **0.946** |
| SDG12: Responsible Consumption and Production | 0.510 | 0.878 | 0.528 | **0.895** | 0.667 | 0.874 | 0.892 | 0.482 | 0.813 | 0.826 | 0.832 | **0.840** |
| SDG13: Climate Action | 0.606 | 0.970 | 0.795 | **0.974** | 0.891 | 0.969 | 0.971 | 0.606 | 0.957 | 0.961 | 0.958 | **0.966** |
| SDG14: Life Below Water | 0.692 | **0.945** | 0.031 | 0.945 | 0.684 | 0.945 | 0.945 | 0.713 | 0.937 | 0.938 | 0.932 | **0.941** |
| SDG15: Life On Land | 0.850 | **0.997** | 0.078 | **0.997** | 0.852 | **0.997** | **0.997** | 0.788 | 0.833 | 0.846 | 0.862 | **0.873** |
| SDG16: Peace, Justice, and Strong Institutions | | | | | | | | 0.386 | 0.426 | 0.493 | 0.434 | **0.508** |
| SDG17: Partnerships for the Goals | | | | | | | | 0.429 | 0.687 | 0.676 | 0.701 | **0.704** |
| Average | 0.689 | 0.936 | 0.568 | 0.944 | 0.785 | 0.936 | 0.943 | 0.559 | 0.745 | 0.764 | 0.763 | 0.778 |

MSCI product score: The columns are (1) MSCI sector (2) Wikipedia extracted product information (3) report evidence (4) MSCI sector + Wikipedia extracted product information (5) MSCI sector + report evidence (6) Wikipedia extracted product information + report evidence (7) All features included. MSCI operation score: The columns are (1) MSCI sector (2) Preprocessed GDELT news headers (3) Preprocessed GDELT news sentiment features (4) Wikipedia extracted product information (5) All features included. The highest value in each row is **bold**. If multiple highest values occur in the same row, these values are underlined.

Table 4.4: F1-macro scores of forecasting MSCI product score and operation score with BRF

| SDGs | MSCI Product score | | | | | | | MSCI Operation score | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **1** | **2** | **3** | **4** | **5** |
| SDG1: No Poverty | 0.491 | 0.535 | 0.133 | **0.669** | 0.491 | 0.529 | 0.640 | 0.327 | 0.318 | **0.372** | 0.263 | 0.361 |
| SDG2: Zero Hunger | 0.458 | 0.498 | 0.442 | **0.555** | 0.496 | 0.479 | 0.523 | **0.314** | 0.297 | 0.296 | 0.297 | 0.297 |
| SDG3: Good Health and Well-being | 0.306 | 0.818 | 0.232 | **0.821** | 0.312 | 0.809 | 0.816 | 0.417 | 0.752 | 0.745 | 0.752 | **0.757** |
| SDG4: Quality Education | 0.493 | **0.699** | 0.628 | **0.699** | 0.495 | **0.699** | **0.699** | 0.493 | 0.491 | **0.545** | 0.492 | 0.516 |
| SDG5: Gender Equality | 0.441 | **0.499** | 0.270 | **0.499** | 0.489 | **0.499** | **0.499** | 0.347 | 0.457 | 0.446 | 0.455 | **0.464** |
| SDG6: Clean Water and Sanitation | 0.369 | 0.696 | 0.370 | 0.708 | 0.368 | **0.713** | 0.711 | 0.320 | 0.484 | 0.524 | 0.509 | **0.524** |
| SDG7: Affordable and Clean Energy | 0.376 | 0.733 | 0.285 | **0.771** | 0.422 | 0.723 | 0.758 | 0.385 | 0.488 | 0.524 | 0.540 | **0.548** |
| SDG8: Decent Work and Economic Growth | 0.409 | 0.418 | 0.186 | **0.461** | 0.441 | 0.420 | 0.423 | 0.200 | 0.316 | **0.385** | 0.325 | 0.369 |
| SDG9: Industry, Innovation and Infrastructure | 0.331 | 0.600 | 0.300 | **0.601** | 0.424 | 0.567 | 0.586 | 0.385 | 0.455 | 0.464 | 0.455 | **0.477** |
| SDG10: Reduced Inequality | 0.340 | 0.557 | 0.123 | 0.592 | 0.352 | 0.564 | **0.620** | 0.292 | 0.310 | 0.330 | 0.266 | **0.333** |
| SDG11: Sustainable Cities and Communities | 0.450 | 0.554 | 0.406 | 0.616 | 0.462 | 0.559 | **0.621** | **0.443** | 0.378 | 0.418 | 0.375 | 0.408 |
| SDG12: Responsible Consumption and Production | 0.429 | 0.789 | 0.199 | **0.814** | 0.427 | 0.773 | 0.798 | 0.397 | 0.686 | 0.696 | 0.667 | **0.706** |
| SDG13: Climate Action | 0.392 | **0.908** | 0.245 | 0.892 | 0.450 | 0.872 | 0.881 | 0.493 | 0.962 | 0.964 | 0.962 | **0.969** |
| SDG14: Life Below Water | 0.234 | 0.221 | 0.012 | 0.221 | **0.238** | 0.204 | 0.208 | **0.373** | 0.349 | 0.359 | 0.349 | 0.348 |
| SDG15: Life On Land | 0.476 | 0.699 | 0.072 | **0.699** | 0.477 | **0.699** | **0.699** | **0.518** | 0.394 | 0.338 | 0.408 | 0.381 |
| SDG16: Peace, Justice, and Strong Institutions | | | | | | | | 0.317 | 0.321 | 0.367 | 0.301 | **0.379** |
| SDG17: Partnerships for the Goals | | | | | | | | 0.343 | 0.488 | 0.481 | 0.476 | **0.499** |
| Average | 0.400 | 0.615 | 0.260 | 0.641 | 0.423 | 0.607 | 0.632 | 0.374 | 0.467 | 0.485 | 0.464 | 0.490 |

MSCI product score: The columns are (1) MSCI sector (2) Wikipedia extracted product information (3) report evidence (4) MSCI sector + Wikipedia extracted product information (5) MSCI sector + report evidence (6) Wikipedia extracted product information + report evidence (7) All features included. MSCI operation score: The columns are (1) MSCI sector (2) Preprocessed GDELT news headers (3) Preprocessed GDELT news sentiment features (4) Wikipedia extracted product information (5) All features included. The highest value in each row is **bold**. If multiple highest values occur in the same row, these values are underlined.

Table 4.5: F1-micro scores of forecasting MSCI and RSAM net alignment score with BRF

| SDGs | MSCI Net Alignment score | | | | | RSAM Net Alignment score | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **1** | **2** | **3** | **4** | **5** | **6** |
| SDG1: No Poverty | 0.398 | 0.729 | 0.806 | **0.817** | 0.816 | 0.654 | 0.832 | 0.928 | 0.929 | 0.023 | **0.929** |
| SDG2: Zero Hunger | 0.481 | 0.861 | 0.936 | **0.937** | **0.937** | 0.508 | 0.863 | 0.936 | **0.939** | 0.080 | 0.939 |
| SDG3: Good Health and Well-being | 0.587 | 0.951 | 0.959 | 0.956 | **0.961** | 0.519 | 0.879 | 0.918 | 0.924 | 0.532 | **0.928** |
| SDG4: Quality Education | 0.569 | 0.972 | 0.973 | **0.976** | 0.975 | 0.689 | **0.994** | 0.994 | **0.994** | 0.142 | 0.994 |
| SDG5: Gender Equality | 0.527 | 0.687 | 0.715 | 0.711 | **0.730** | 0.421 | **0.983** | 0.983 | **0.983** | 0.300 | 0.983 |
| SDG6: Clean Water and Sanitation | 0.392 | 0.819 | 0.870 | **0.878** | 0.877 | 0.403 | 0.919 | **0.975** | 0.975 | 0.942 | 0.975 |
| SDG7: Affordable and Clean Energy | 0.375 | 0.738 | 0.762 | 0.813 | **0.821** | 0.567 | 0.839 | 0.895 | 0.896 | 0.820 | **0.900** |
| SDG8: Decent Work and Economic Growth | 0.307 | 0.665 | 0.687 | **0.734** | 0.730 | 0.634 | 0.686 | 0.746 | 0.774 | 0.162 | **0.783** |
| SDG9: Industry, Innovation and Infrastructure | 0.304 | 0.758 | 0.787 | **0.803** | 0.802 | 0.499 | 0.647 | 0.708 | **0.749** | 0.157 | 0.747 |
| SDG10: Reduced Inequality | 0.291 | 0.633 | 0.698 | 0.728 | **0.734** | 0.811 | 0.911 | 0.993 | **0.994** | 0.241 | **0.994** |
| SDG11: Sustainable Cities and Communities | 0.578 | 0.892 | 0.903 | 0.898 | **0.907** | 0.459 | 0.763 | 0.824 | 0.830 | 0.134 | **0.831** |
| SDG12: Responsible Consumption and Production | 0.598 | 0.864 | 0.860 | 0.884 | **0.890** | 0.231 | 0.938 | 0.954 | 0.949 | 0.636 | **0.955** |
| SDG13: Climate Action | 0.356 | 0.965 | 0.977 | 0.973 | **0.982** | 0.421 | 0.990 | **0.990** | 0.981 | 0.835 | 0.990 |
| SDG14: Life Below Water | 0.630 | 0.924 | 0.930 | 0.925 | **0.933** | 0.544 | **0.982** | 0.981 | **0.982** | 0.035 | **0.982** |
| SDG15: Life On Land | 0.724 | 0.947 | 0.945 | 0.939 | **0.950** | 0.458 | 0.861 | 0.957 | **0.958** | 0.035 | **0.958** |
| SDG16: Peace, Justice, and Strong Institutions | 0.177 | 0.809 | 0.861 | 0.861 | **0.864** | 0.193 | 0.893 | 0.954 | **0.955** | 0.136 | **0.955** |
| SDG17: Partnerships for the Goals | 0.490 | 0.888 | 0.914 | 0.921 | **0.923** | | | | | | |
| Average | 0.458 | 0.829 | 0.858 | 0.868 | 0.872 | 0.501 | 0.874 | 0.921 | 0.926 | 0.325 | 0.928 |

MSCI Net Alignment score: The columns are (1) MSCI sector (2) Preprocessed GDELT news headers (3) Preprocessed GDELT news sentiment features (4) Wikipedia extracted product information (5) All features included. RSAM Net Alignment score: The columns are (1) MSCI sector (2) Preprocessed GDELT news headers (3) Preprocessed GDELT news sentiment features (4) Wikipedia extracted product information (5) Report evidence (6) All features included. The highest value in each row is **bold**. If multiple highest values occur in the same row, these values are underlined.

Table 4.6: F1-macro scores of forecasting MSCI and RSAM net alignment score with BRF

| SDGs | MSCI Net Alignment score | | | | | RSAM Net Alignment score | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 6 |
| SDG1: No Poverty | **0.291** | 0.275 | 0.268 | 0.254 | 0.272 | **0.235** | 0.224 | 0.231 | 0.231 | 0.013 | 0.231 |
| SDG2: Zero Hunger | 0.156 | **0.237** | 0.222 | 0.223 | 0.223 | 0.128 | 0.184 | 0.203 | **0.206** | 0.060 | 0.205 |
| SDG3: Good Health and Well-being | 0.376 | 0.822 | 0.845 | 0.838 | **0.850** | 0.229 | 0.749 | 0.816 | 0.824 | 0.147 | **0.833** |
| SDG4: Quality Education | 0.254 | 0.427 | **0.466** | 0.428 | 0.428 | 0.280 | **0.465** | 0.443 | **0.465** | 0.105 | **0.465** |
| SDG5: Gender Equality | 0.354 | 0.644 | 0.647 | 0.647 | **0.671** | 0.093 | 0.833 | 0.810 | **0.839** | 0.104 | 0.811 |
| SDG6: Clean Water and Sanitation | 0.183 | 0.416 | 0.435 | **0.439** | 0.436 | 0.148 | 0.596 | **0.716** | 0.713 | 0.202 | 0.713 |
| SDG7: Affordable and Clean Energy | 0.341 | 0.691 | 0.696 | 0.761 | **0.775** | 0.263 | 0.579 | 0.638 | 0.648 | 0.217 | **0.659** |
| SDG8: Decent Work and Economic Growth | 0.186 | 0.330 | **0.363** | 0.331 | 0.332 | 0.283 | 0.373 | 0.370 | 0.409 | 0.114 | **0.415** |
| SDG9: Industry, Innovation and Infrastructure | 0.202 | 0.397 | 0.400 | **0.403** | 0.399 | 0.239 | 0.442 | 0.429 | **0.495** | 0.137 | 0.494 |
| SDG10: Reduced Inequality | 0.184 | 0.275 | **0.306** | 0.264 | 0.281 | 0.343 | 0.318 | 0.465 | **0.465** | 0.097 | **0.465** |
| SDG11: Sustainable Cities and Communities | 0.344 | **0.388** | 0.377 | 0.374 | 0.382 | 0.153 | **0.442** | 0.420 | 0.413 | 0.075 | 0.425 |
| SDG12: Responsible Consumption and Production | 0.473 | 0.746 | 0.739 | 0.796 | **0.803** | 0.121 | 0.602 | 0.651 | 0.640 | 0.198 | 0.652 |
| SDG13: Climate Action | 0.345 | 0.958 | 0.966 | 0.959 | **0.971** | 0.192 | 0.945 | **0.945** | 0.924 | 0.208 | 0.919 |
| SDG14: Life Below Water | **0.301** | 0.261 | 0.263 | 0.266 | 0.264 | 0.145 | 0.264 | 0.264 | **0.264** | 0.016 | **0.264** |
| SDG15: Life On Land | **0.376** | 0.324 | 0.329 | 0.323 | 0.325 | 0.114 | 0.174 | 0.176 | **0.176** | 0.020 | **0.176** |
| SDG16: Peace, Justice, and Strong Institutions | 0.145 | 0.299 | **0.314** | 0.266 | 0.303 | 0.067 | 0.171 | 0.203 | **0.205** | 0.037 | **0.205** |
| SDG17: Partnerships for the Goals | 0.260 | 0.470 | 0.506 | 0.515 | **0.519** | | | | | | |
| Average | 0.281 | 0.468 | 0.479 | 0.476 | 0.484 | 0.190 | 0.460 | 0.486 | 0.495 | 0.109 | 0.496 |

MSCI Net Alignment score: The columns are (1) MSCI sector (2) Preprocessed GDELT news headers (3) Preprocessed GDELT news sentiment features (4) Wikipedia extracted product information (5) All features included. RSAM Net Alignment score: The columns are (1) MSCI sector (2) Preprocessed GDELT news headers (3) Preprocessed GDELT news sentiment features (4) Wikipedia extracted product information (5) Report evidence (6) All features included. The highest value in each row is **bold**. If multiple highest values occur in the same row, these values are underlined.

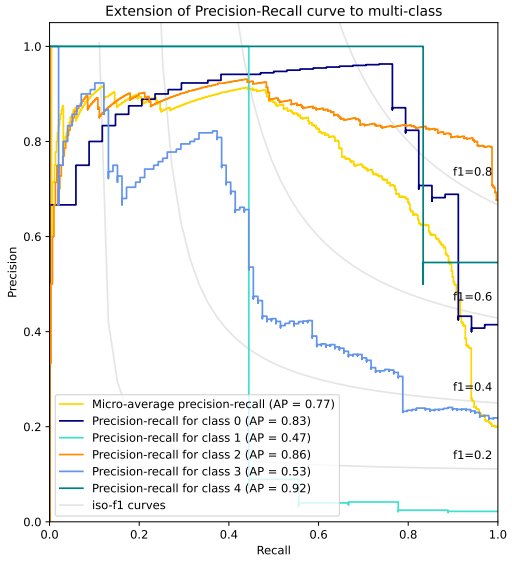



Figure 4.6: Micro-average Precision Recall curve for MSCI SDG7 net alignment score. Class 0-4: Strongly Misaligned, Misaligned, Neutral, Aligned, Strongly Aligned

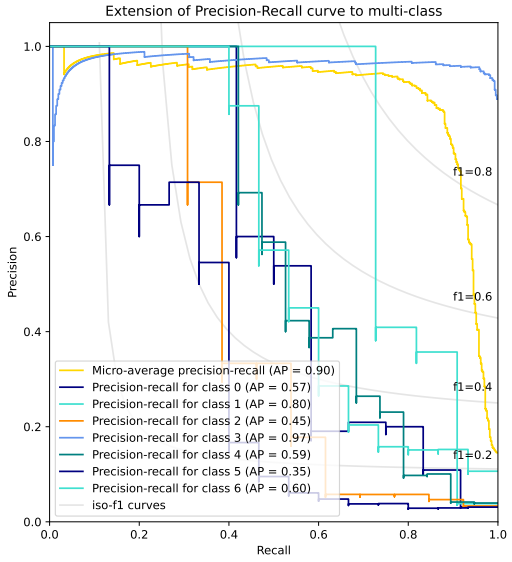



Figure 4.7: Precision Recall curve for RSAM SDG7 net alignment score. Class 0-6: -3, -2, -1, 0, 1, 2, 3

### Explaining the text classifier results

In this section, we use LIME to construct explanations for text classification utilizing Wikipedia product information modelled by BoW with bigrams to predict MSCI product scores and

RSAM net alignment scores[1]. Only nouns are retained for generating explanations. Following classification, LIME generates feature importance plots as shown in Figure A.4, A.5, A.6, A.7 that allow users to find keywords associated with each class. Here, we can clearly see a considerable preference for businesses that generate power or electricity over those that generate oil or gas (petrol) for MSCI. In contrast to this, RSAM tends to see businesses that deal with gas (fuel) and wind energy as good, while those that deal with oil (refinery), lubricants, and plastic are considered bad. Additionally, an example producing the individual explanation is shown in Figure 4.8 for forecasting SDG7 (Affordable and Clean Energy) of RSAM. The key words *wind* and *energy* contribute significantly to Vestas Wind Systems A/S's positive score.

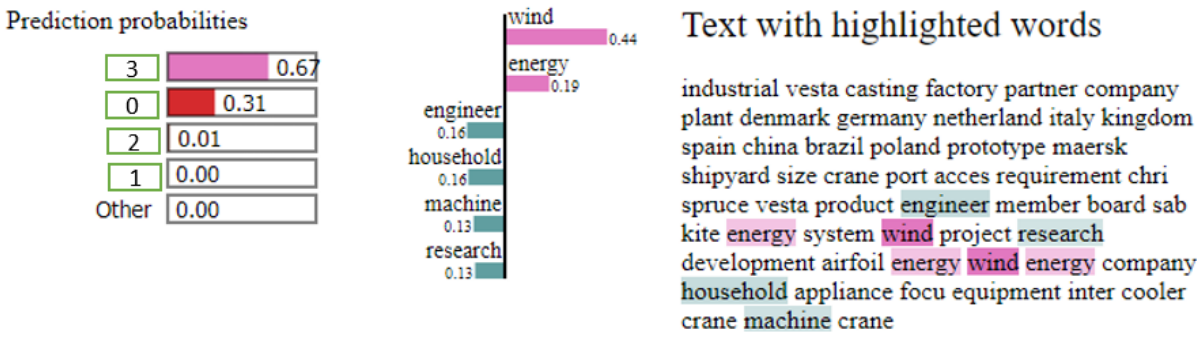


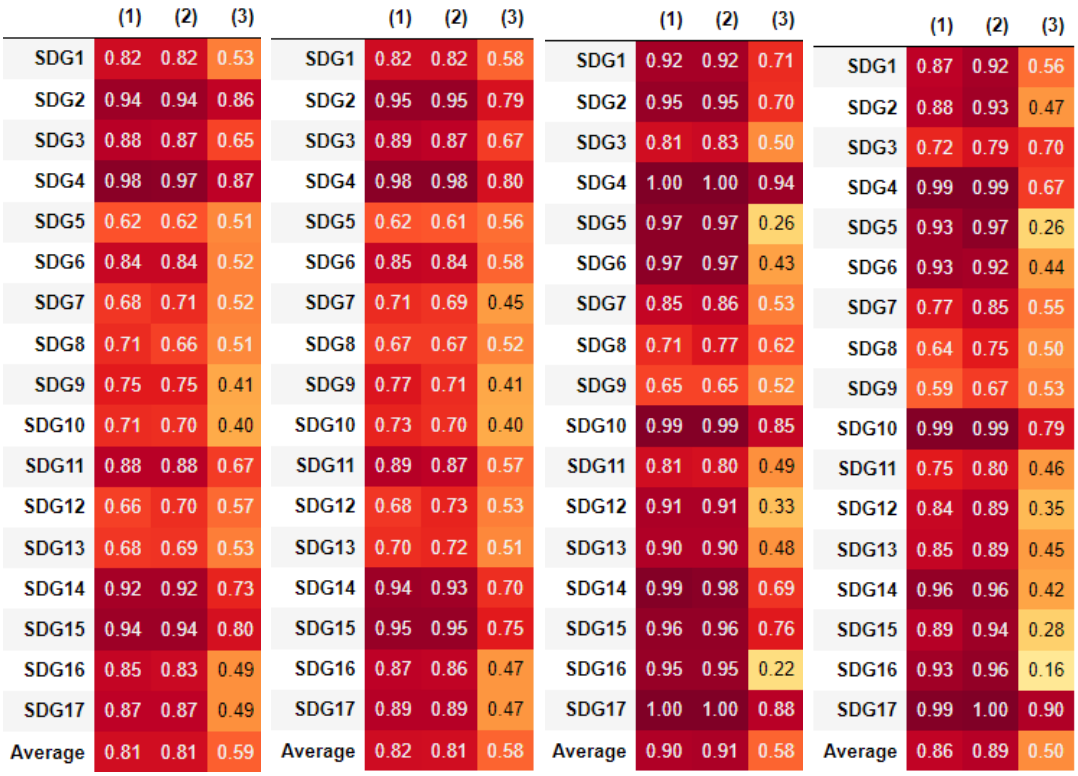

Figure 4.8: Explaining the text classifier results for Vestas Wind Systems A/S (RSAM SDG7)

### Utilizing the network structure

GCN and RGCN are used to predict net alignment scores in order to see whether the network structure improves classification performance. The relevant hyperparameters for GCN and RGCN are reported in Table A.3 in the Appendix. GCN is applied to the company graph, whereas RGCN is used on the extracted Wikidata knowledge graph. What is noteworthy is that, the training, evaluation, and test sets are created in a different way compared to running BRF. In Section 4.1.3, we briefly explain the differences in sampling and evaluation methods for BRF and GCN/RGCN. The splits are different. If we were to apply the split used for training GCN to BRF, it would result in a poorer classification score for BRF in comparison to the BRF with the oversampling approach. We can see that F1 micro scores in Table 4.5 (where BRF is applied after oversampling the minority classes) are higher than column (2) in Figure 4.9 (where BRF is applied using the same splits as GCN). It is essential that we have training samples for each class in order to perform classification on graphs. As shown in Figure 4.9[2], the values/colors in columns (1) and (2) do not vary significantly for F1 micro score and are much higher than the baseline model in column (3), indicating that neither GCN nor RGCN give much extra information beyond BRF that might enhance F1 micro score. Figure 4.10 (a) demonstrates that, compared to BRF, the F1-macro score improves by an average of 3% when GCN is used. The intriguing phenomenon is that column (3) has a darker color than columns (1) and (2) for Figure 4.10 (a) and (b): the baseline model (BRF with MSCI sector information) in column (3) seems to provide the highest F1-macro score, contrary to what was discovered in the previous section when BRF was used. This is a consequence of the sampling procedures used to build the training set. Samples from each minority class shall appear at least once in the training set, indicating that the low variety of features in minority classes is what contributes to their high F1-macro scores. However, this tendency is not seen in Figure

[1]The explanations for forecasting all SDGs are inside of folder *explanations* in the code repository.
[2]The color is normalized to the range of 0-1.

4.10 (c)(d), indicating that the RSAM likely has a greater variety of characteristics for the minority classes.



| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.82 | 0.82 | 0.53 |
| SDG2 | 0.94 | 0.94 | 0.86 |
| SDG3 | 0.88 | 0.87 | 0.65 |
| SDG4 | 0.98 | 0.97 | 0.87 |
| SDG5 | 0.62 | 0.62 | 0.51 |
| SDG6 | 0.84 | 0.84 | 0.52 |
| SDG7 | 0.68 | 0.71 | 0.52 |
| SDG8 | 0.71 | 0.66 | 0.51 |
| SDG9 | 0.75 | 0.75 | 0.41 |
| SDG10 | 0.71 | 0.70 | 0.40 |
| SDG11 | 0.88 | 0.88 | 0.67 |
| SDG12 | 0.66 | 0.70 | 0.57 |
| SDG13 | 0.68 | 0.69 | 0.53 |
| SDG14 | 0.92 | 0.92 | 0.73 |
| SDG15 | 0.94 | 0.94 | 0.80 |
| SDG16 | 0.85 | 0.83 | 0.49 |
| SDG17 | 0.87 | 0.87 | 0.49 |
| Average | 0.81 | 0.81 | 0.59 |

(a)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.82 | 0.82 | 0.58 |
| SDG2 | 0.95 | 0.95 | 0.79 |
| SDG3 | 0.89 | 0.87 | 0.67 |
| SDG4 | 0.98 | 0.98 | 0.80 |
| SDG5 | 0.62 | 0.61 | 0.56 |
| SDG6 | 0.85 | 0.84 | 0.58 |
| SDG7 | 0.71 | 0.69 | 0.45 |
| SDG8 | 0.67 | 0.67 | 0.52 |
| SDG9 | 0.77 | 0.71 | 0.41 |
| SDG10 | 0.73 | 0.70 | 0.40 |
| SDG11 | 0.89 | 0.87 | 0.57 |
| SDG12 | 0.68 | 0.73 | 0.53 |
| SDG13 | 0.70 | 0.72 | 0.51 |
| SDG14 | 0.94 | 0.93 | 0.70 |
| SDG15 | 0.95 | 0.95 | 0.75 |
| SDG16 | 0.87 | 0.86 | 0.47 |
| SDG17 | 0.89 | 0.89 | 0.47 |
| Average | 0.82 | 0.81 | 0.58 |

(b)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.92 | 0.92 | 0.71 |
| SDG2 | 0.95 | 0.95 | 0.70 |
| SDG3 | 0.81 | 0.83 | 0.50 |
| SDG4 | 1.00 | 1.00 | 0.94 |
| SDG5 | 0.97 | 0.97 | 0.26 |
| SDG6 | 0.97 | 0.97 | 0.43 |
| SDG7 | 0.85 | 0.86 | 0.53 |
| SDG8 | 0.71 | 0.77 | 0.62 |
| SDG9 | 0.65 | 0.65 | 0.52 |
| SDG10 | 0.99 | 0.99 | 0.85 |
| SDG11 | 0.81 | 0.80 | 0.49 |
| SDG12 | 0.91 | 0.91 | 0.33 |
| SDG13 | 0.90 | 0.90 | 0.48 |
| SDG14 | 0.99 | 0.98 | 0.69 |
| SDG15 | 0.96 | 0.96 | 0.76 |
| SDG16 | 0.95 | 0.95 | 0.22 |
| SDG17 | 1.00 | 1.00 | 0.88 |
| Average | 0.90 | 0.91 | 0.58 |

(c)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.87 | 0.92 | 0.56 |
| SDG2 | 0.88 | 0.93 | 0.47 |
| SDG3 | 0.72 | 0.79 | 0.70 |
| SDG4 | 0.99 | 0.99 | 0.67 |
| SDG5 | 0.93 | 0.97 | 0.26 |
| SDG6 | 0.93 | 0.92 | 0.44 |
| SDG7 | 0.77 | 0.85 | 0.55 |
| SDG8 | 0.64 | 0.75 | 0.50 |
| SDG9 | 0.59 | 0.67 | 0.53 |
| SDG10 | 0.99 | 0.99 | 0.79 |
| SDG11 | 0.75 | 0.80 | 0.46 |
| SDG12 | 0.84 | 0.89 | 0.35 |
| SDG13 | 0.85 | 0.89 | 0.45 |
| SDG14 | 0.96 | 0.96 | 0.42 |
| SDG15 | 0.89 | 0.94 | 0.28 |
| SDG16 | 0.93 | 0.96 | 0.16 |
| SDG17 | 0.99 | 1.00 | 0.90 |
| Average | 0.86 | 0.89 | 0.50 |

(d)

Figure 4.9: F1 micro scores of forecasts. Columns: (1) corresponding graph algorithm (GCN/RGCN) with all features (2) BRF with all features using the same training validation test split for running graph algorithms (3) BRF with only MSCI sector as feature using the same training validation test split for running graph algorithms. Captions: (a) GCN + MSCI net alignment score (b) RGCN + MSCI net alignment score (c) GCN + RSAM net alignment score (d) RGCN + RSAM net alignment score

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.31 | 0.29 | 0.34 |
| SDG2 | 0.22 | 0.19 | 0.31 |
| SDG3 | 0.39 | 0.23 | 0.51 |
| SDG4 | 0.54 | 0.49 | 0.48 |
| SDG5 | 0.59 | 0.55 | 0.34 |
| SDG6 | 0.22 | 0.19 | 0.26 |
| SDG7 | 0.31 | 0.29 | 0.44 |
| SDG8 | 0.22 | 0.25 | 0.34 |
| SDG9 | 0.20 | 0.21 | 0.24 |
| SDG10 | 0.21 | 0.25 | 0.23 |
| SDG11 | 0.34 | 0.26 | 0.38 |
| SDG12 | 0.35 | 0.31 | 0.50 |
| SDG13 | 0.32 | 0.32 | 0.42 |
| SDG14 | 0.29 | 0.24 | 0.32 |
| SDG15 | 0.39 | 0.37 | 0.38 |
| SDG16 | 0.25 | 0.30 | 0.27 |
| SDG17 | 0.27 | 0.23 | 0.25 |
| Average | 0.32 | 0.29 | 0.35 |

(a)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.31 | 0.27 | 0.35 |
| SDG2 | 0.31 | 0.19 | 0.34 |
| SDG3 | 0.32 | 0.23 | 0.39 |
| SDG4 | 0.53 | 0.49 | 0.33 |
| SDG5 | 0.61 | 0.57 | 0.34 |
| SDG6 | 0.29 | 0.19 | 0.25 |
| SDG7 | 0.25 | 0.31 | 0.38 |
| SDG8 | 0.24 | 0.20 | 0.37 |
| SDG9 | 0.27 | 0.18 | 0.28 |
| SDG10 | 0.46 | 0.25 | 0.25 |
| SDG11 | 0.28 | 0.24 | 0.39 |
| SDG12 | 0.25 | 0.22 | 0.41 |
| SDG13 | 0.28 | 0.27 | 0.45 |
| SDG14 | 0.42 | 0.24 | 0.30 |
| SDG15 | 0.48 | 0.33 | 0.43 |
| SDG16 | 0.34 | 0.28 | 0.26 |
| SDG17 | 0.33 | 0.27 | 0.24 |
| Average | 0.35 | 0.28 | 0.34 |

(b)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.26 | 0.30 | 0.23 |
| SDG2 | 0.16 | 0.16 | 0.11 |
| SDG3 | 0.22 | 0.25 | 0.21 |
| SDG4 | 0.50 | 0.50 | 0.47 |
| SDG5 | 0.25 | 0.25 | 0.07 |
| SDG6 | 0.24 | 0.16 | 0.20 |
| SDG7 | 0.15 | 0.16 | 0.24 |
| SDG8 | 0.31 | 0.25 | 0.21 |
| SDG9 | 0.25 | 0.29 | 0.23 |
| SDG10 | 0.50 | 0.50 | 0.37 |
| SDG11 | 0.18 | 0.19 | 0.13 |
| SDG12 | 0.24 | 0.16 | 0.12 |
| SDG13 | 0.17 | 0.16 | 0.22 |
| SDG14 | 0.30 | 0.25 | 0.18 |
| SDG15 | 0.16 | 0.16 | 0.16 |
| SDG16 | 0.16 | 0.16 | 0.10 |
| SDG17 | 1.00 | 1.00 | 0.48 |
| Average | 0.30 | 0.29 | 0.22 |

(c)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.26 | 0.28 | 0.22 |
| SDG2 | 0.15 | 0.14 | 0.12 |
| SDG3 | 0.22 | 0.25 | 0.28 |
| SDG4 | 0.33 | 0.33 | 0.29 |
| SDG5 | 0.16 | 0.25 | 0.05 |
| SDG6 | 0.19 | 0.16 | 0.13 |
| SDG7 | 0.15 | 0.16 | 0.25 |
| SDG8 | 0.22 | 0.26 | 0.26 |
| SDG9 | 0.17 | 0.25 | 0.23 |
| SDG10 | 0.50 | 0.50 | 0.32 |
| SDG11 | 0.20 | 0.16 | 0.18 |
| SDG12 | 0.21 | 0.17 | 0.14 |
| SDG13 | 0.20 | 0.14 | 0.15 |
| SDG14 | 0.20 | 0.20 | 0.11 |
| SDG15 | 0.18 | 0.16 | 0.09 |
| SDG16 | 0.18 | 0.20 | 0.09 |
| SDG17 | 0.50 | 1.00 | 0.47 |
| Average | 0.24 | 0.27 | 0.20 |

(d)

Figure 4.10: F1 macro scores of forecasts. Columns: (1) corresponding graph algorithm (GCN/RGCN) with all features (2) BRF with all features using the same training validation test split for running graph algorithms (3) BRF with only MSCI sector as feature using the same training validation test split for running graph algorithms. Captions: (a) GCN + MSCI net alignment score (b) RGCN + MSCI net alignment score (c) GCN + RSAM net alignment score (d) RGCN + RSAM net alignment score

### 4.1.5 Ablation study for classification on graphs

In this section, we perform two further sets of experiments: one employs the featureless strategy for GCN, while the other employs a reduced training ratio for GCN. If we compare columns (1) and (2) in Figure 4.11 (a) with Figure 4.9 (a), and columns (1) and (2) in Figure 4.11 (b) to Figure 4.10 (a), we can observe a slight decrease in the F1 micro and macro scores, despite the fact that the difference between the featureless approach and the featured approach appears to be quite small.

We demonstrated that when the training ratio is low, GCN outperforms RF on Figure 4.2 provided that the network topology is advantageous for predicting the target variables. Comparing column (1) in Figure 4.11(c)(d) to column (1) in Figure 4.9 (a) and in Figure 4.10 (a), the difference between the average F1-micro and F1-macro scores is between 0 and 1 percent, further validating that the network topology does not include crucial information that may be leveraged to enhance classification performance.

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.81 | 0.81 | 0.61 |
| SDG2 | 0.94 | 0.94 | 0.80 |
| SDG3 | 0.88 | 0.88 | 0.64 |
| SDG4 | 0.97 | 0.97 | 0.87 |
| SDG5 | 0.60 | 0.59 | 0.52 |
| SDG6 | 0.84 | 0.84 | 0.52 |
| SDG7 | 0.68 | 0.68 | 0.47 |
| SDG8 | 0.66 | 0.66 | 0.52 |
| SDG9 | 0.75 | 0.75 | 0.42 |
| SDG10 | 0.71 | 0.71 | 0.50 |
| SDG11 | 0.88 | 0.88 | 0.61 |
| SDG12 | 0.70 | 0.67 | 0.62 |
| SDG13 | 0.68 | 0.68 | 0.44 |
| SDG14 | 0.92 | 0.92 | 0.73 |
| SDG15 | 0.94 | 0.94 | 0.80 |
| SDG16 | 0.85 | 0.85 | 0.39 |
| SDG17 | 0.87 | 0.87 | 0.40 |
| Average | 0.81 | 0.80 | 0.58 |

(a)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.22 | 0.22 | 0.37 |
| SDG2 | 0.19 | 0.19 | 0.34 |
| SDG3 | 0.23 | 0.23 | 0.47 |
| SDG4 | 0.49 | 0.49 | 0.52 |
| SDG5 | 0.39 | 0.37 | 0.34 |
| SDG6 | 0.18 | 0.18 | 0.27 |
| SDG7 | 0.16 | 0.16 | 0.42 |
| SDG8 | 0.20 | 0.20 | 0.39 |
| SDG9 | 0.17 | 0.17 | 0.29 |
| SDG10 | 0.21 | 0.21 | 0.29 |
| SDG11 | 0.23 | 0.23 | 0.39 |
| SDG12 | 0.32 | 0.16 | 0.49 |
| SDG13 | 0.30 | 0.16 | 0.44 |
| SDG14 | 0.24 | 0.24 | 0.32 |
| SDG15 | 0.32 | 0.32 | 0.44 |
| SDG16 | 0.23 | 0.23 | 0.29 |
| SDG17 | 0.23 | 0.23 | 0.26 |
| Average | 0.26 | 0.24 | 0.37 |

(b)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.80 | 0.77 | 0.70 |
| SDG2 | 0.94 | 0.94 | 0.96 |
| SDG3 | 0.88 | 0.87 | 0.77 |
| SDG4 | 0.98 | 0.98 | 0.96 |
| SDG5 | 0.60 | 0.59 | 0.50 |
| SDG6 | 0.84 | 0.84 | 0.72 |
| SDG7 | 0.68 | 0.66 | 0.61 |
| SDG8 | 0.68 | 0.58 | 0.60 |
| SDG9 | 0.74 | 0.66 | 0.60 |
| SDG10 | 0.70 | 0.63 | 0.58 |
| SDG11 | 0.89 | 0.88 | 0.81 |
| SDG12 | 0.70 | 0.65 | 0.65 |
| SDG13 | 0.68 | 0.63 | 0.56 |
| SDG14 | 0.92 | 0.93 | 0.85 |
| SDG15 | 0.94 | 0.94 | 0.91 |
| SDG16 | 0.85 | 0.84 | 0.76 |
| SDG17 | 0.87 | 0.87 | 0.64 |
| Average | 0.81 | 0.78 | 0.72 |

(c)

| | (1) | (2) | (3) |
|---|---|---|---|
| SDG1 | 0.27 | 0.31 | 0.34 |
| SDG2 | 0.26 | 0.20 | 0.32 |
| SDG3 | 0.34 | 0.23 | 0.53 |
| SDG4 | 0.54 | 0.49 | 0.56 |
| SDG5 | 0.36 | 0.37 | 0.34 |
| SDG6 | 0.23 | 0.21 | 0.24 |
| SDG7 | 0.29 | 0.22 | 0.40 |
| SDG8 | 0.21 | 0.25 | 0.27 |
| SDG9 | 0.25 | 0.20 | 0.28 |
| SDG10 | 0.23 | 0.24 | 0.34 |
| SDG11 | 0.32 | 0.23 | 0.30 |
| SDG12 | 0.35 | 0.22 | 0.40 |
| SDG13 | 0.32 | 0.22 | 0.41 |
| SDG14 | 0.28 | 0.24 | 0.32 |
| SDG15 | 0.40 | 0.36 | 0.47 |
| SDG16 | 0.34 | 0.28 | 0.37 |
| SDG17 | 0.26 | 0.23 | 0.27 |
| Average | 0.31 | 0.26 | 0.36 |

(d)

Figure 4.11: F1 scores of forecasts. Columns: (1) corresponding graph algorithm (GCN/RGCN) with all features (2) BRF with all features using the same training validation test split for running graph algorithms (3) BRF with only MSCI sector as feature using the same training validation test split for running graph algorithms. Captions: (a) GCN + MSCI + featureless (F1-micro) (b) GCN + MSCI + featureless (F1-macro) (c) GCN + MSCI + training ratio modified from 0.6 to 0.1 (F1-micro) (d) GCN + MSCI + training ratio modified from 0.6 to 0.1 (F1-macro)

## 4.2 Generating SDGs Scores & Producing Explanations

### 4.2.1 Motivation

There are 2 main objectives for generating SDG scores: (1) The primary reason for being able to predict MSCI and RSAM scores was to be able to estimate a score for which these providers do not publish one. (2) The objective of creating new SDG ratings and providing explanations for those scores is to enhance the present MSCI SDG framework. It may be in one's interest to know the difference between predicting SDG scores and generating SDG scores. In our context, generating SDG scores refers to modifying existing SDG ratings to include additional data (company's extra features and network structure), whereas predicting SDG scores refers to generating missing SDG values for companies without a label. In reality, we must first understand the factors MSCI considers when producing SDG scores. In the previous section (Section 4.1), we tested the accuracy with which we predict the SDG scores using the publicly available information we gathered online, and we found that neither MSCI nor RSAM seem to take the graph structure into account or fully use the report evidence and news we unearthed. We also discovered that the straightforward logic for constructing their frameworks makes it feasible to predict these scores with reasonable accuracy. However, because MSCI and RSAM primarily focus on the products from which the company generates the majority of its revenue and the general impactful news that has occurred to the company, their evaluation is rather two-dimensional and does not effectively use all of the information that could be deemed rele-

vant or even important for constructing the SDG framework. Therefore, we see an opportunity to enhance their framework by including more dimensions, such as using the additional data we have gathered like Wikidata knowledge graphs, which might represent the relationships (e.g., certain supply chain activities) between the organizations.

From another perspective, we try to mimic humans, but the data-driven and algorithmic-driven procedures have the ability to scale, and they beat humans in terms of speed. In the end, we intend to create a more a more transparent, configurable, and scalable framework for a large number of companies: by transparent, firstly it implies the data used to build the framework is fully public (accessible to everyone). Secondly, by being transparent, we are able to explain the process, and even offer individual-specific explanations for classification scores. By being customizable, the new framework could maintain the essential logic of rating firms based on MSCI, RSAM, or other key sustainability SDG framework providers while it may be tailored to the needs of the user, such as they could select the relations of interests in the graph and the attributes (e.g., evidence from sustainability reports, important company news, and Wikipedia product information) they want the model to take into consideration.

### 4.2.2 Algorithm: Generating SDG Scores

In this section, we propose an algorithm for generating SDG sores as illustrated in the Schema shown in Figure 4.12. The ultimate objective is to include more data dimensions (such as supplementary news, reports, and network structure) while preserving the underlying logic of how MSCI evaluates the products of firms. We firstly develop a heuristic method to convey the concept as shown in Figure A.8 in the Appendix. In order to achieve this, we firstly perform a graph cluster algorithm (Shchur and Günnemann, 2019) to cluster companies in the company graph into 50 clusters. The relevant hyperparameters to run this algorithm are reported in Table A.4 in the Appendix. The clustering algorithm uses node characteristics and network topologies to divide companies into 50 groups. In the second stage, we label each firm in the cluster with its MSCI label and then take the mean of all labels for that cluster as the new label for all companies in that cluster. Thus, the labels of all the companies inside the cluster are updated to the same new label. If the new label is positive, we mark companies inside of this cluster as positive; if the new label is negative, we mark as negative; neutral otherwise. If a company lacks an initial MSIC label, its label is predicted using GCN.

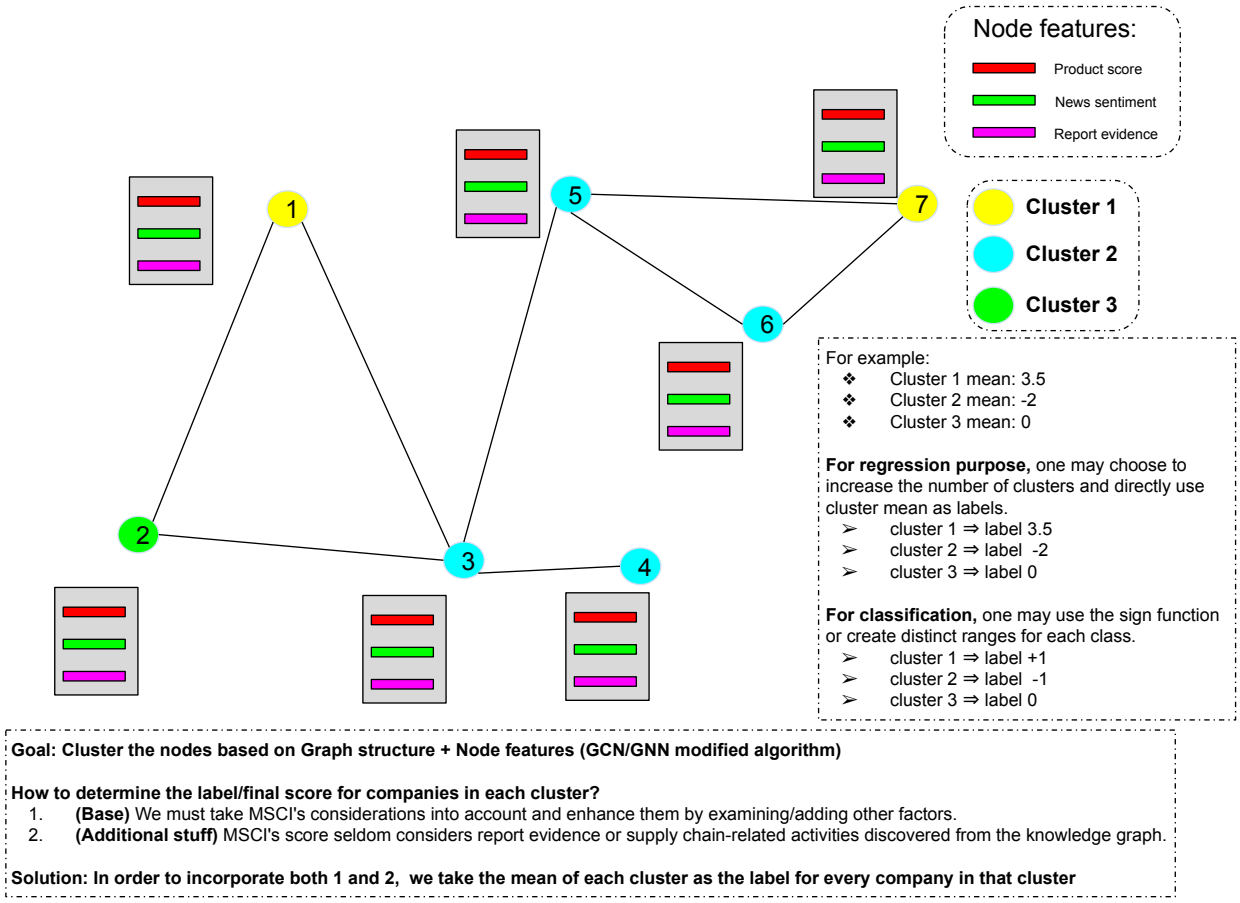



Figure 4.12: Schematic Diagram: Generating new SDG scores through graph clustering method. New SDG scores are generated by first using the graph clustering method to incorporate new characteristics of each company and the network structure utilized to convey shareholder information and supply chain activity, followed by the aggregation of MSCI scores for companies within the same cluster.

### 4.2.3 Algorithm: GNNExplainer for producing individual specific explanations

In order to explain the classification results obtained for every company of interest, GNNExplainer is adopted for generating explanations for classification results directly. The explainer produces explanations in two dimensions: (1) the most important neighbors and connections (subgraph) to the node of interest (2) the most important attribute (feature) that drives classification results for this node. An example is illustrated in the schema shown in Figure 4.13. In addition, two case examples below are offered to illustrate how the explanations are formed.

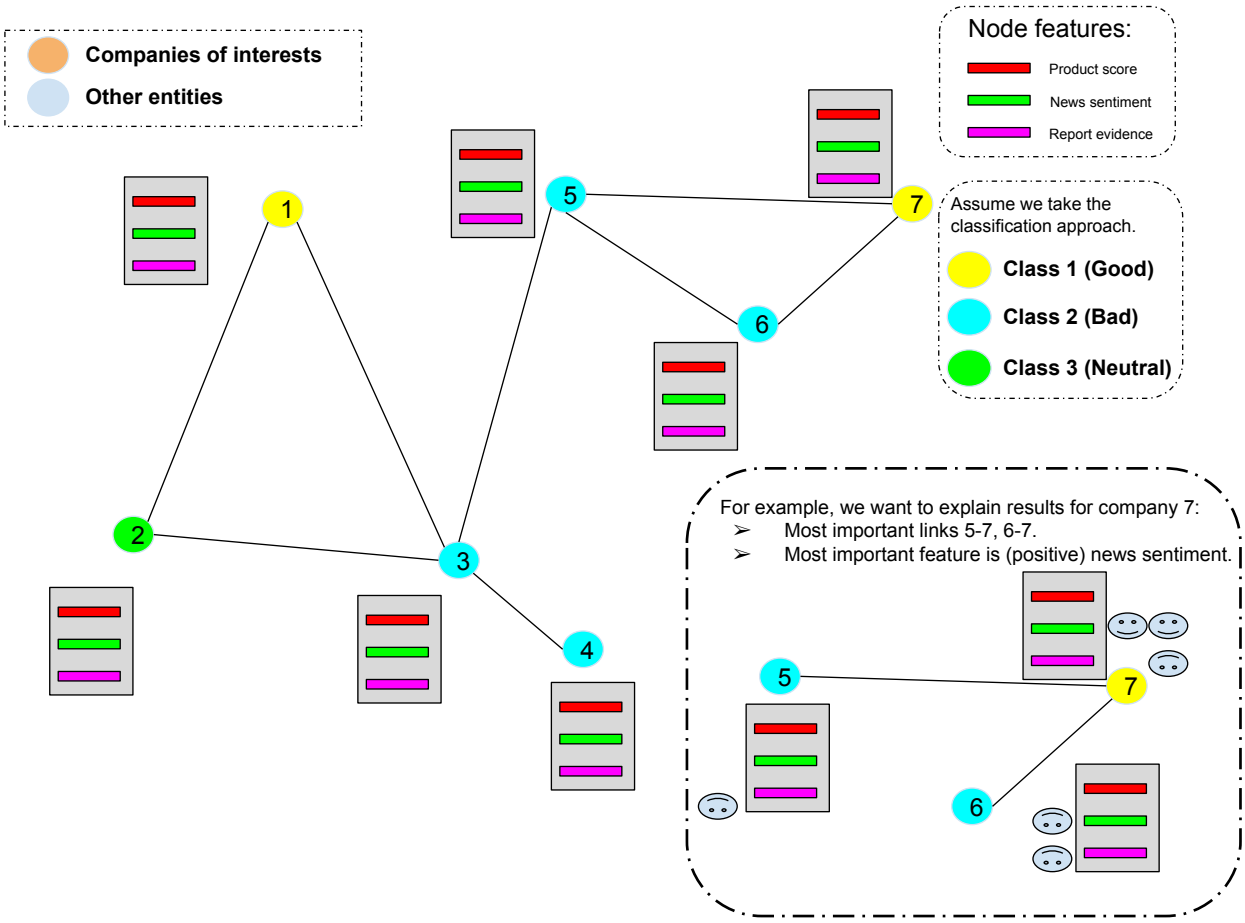
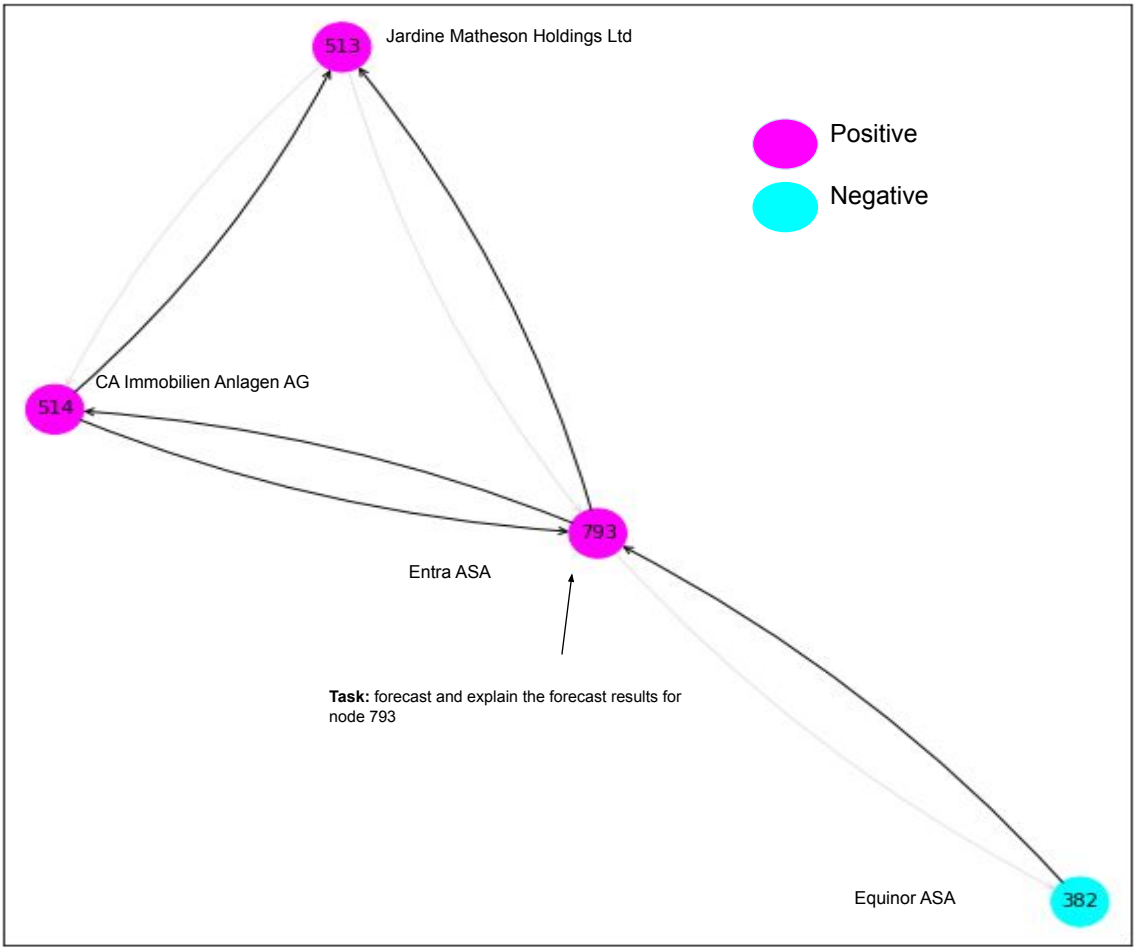



Figure 4.13: Schematic Diagram: Explain the classification results with GNNExplainer. The explanation is produced in two dimensions: (1) the most relevant subgraph for determining this node's forecast. (2) the most important features shared by the subgraph's nodes for drawing the prediction.

### Case Studies

In Figure 4.14 and Figure 4.15, we supplied two examples to illustrate how to explain the classification scores of firms of interest. For instance, as shown in Figure 4.14, we want to forecast the SDG score for node 793 (and node 793's predicted label is *positive*). The most essential subgraph (part of company graph) for determining the label of this node is shown in the picture along with the two tables that describe (1) the basic information (node id, Wikidata id, MSCI sector information and label) about all the nodes in this subgraph. (2) The most significant factors that lead to the favorable prediction of node 793. The leftmost column in the second table provides the most significant factor that determines this node's forecast (e.g., the feature that distinguishes this node's forecast from those of its neighbors).



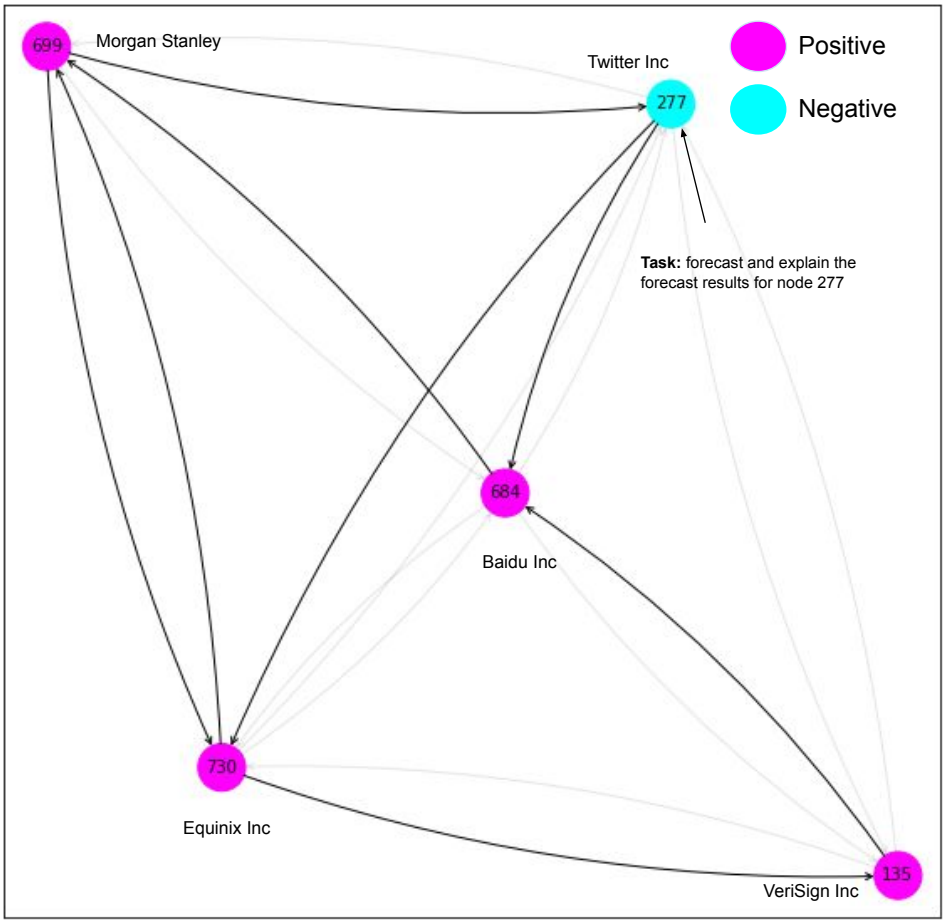

Basic Information

| index | company | wikidata id | GICS Industry | label |
|---|---|---|---|---|
| 382 | Equinor ASA | Q1776022 | Oil, Gas & Consumable Fuels | negative |
| 513 | Jardine Matheson Holdings Ltd | Q1683452 | Industrial Conglomerates | positive |
| 514 | CA Immobilien Anlagen AG | Q1022921 | Real Estate Management & Development | positive |
| 793 | Entra ASA | Q5380649 | Real Estate Management & Development | positive |

Features (left to right ranked by importance)

| index | company | report evidence | product score | news sentiment |
|---|---|---|---|---|
| 382 | Equinor ASA | 15.00 | -10.00 | -30.64 |
| 513 | Jardine Matheson Holdings Ltd | 0.00 | -3.00 | 0.34 |
| 514 | CA Immobilien Anlagen AG | 16.00 | 0.00 | 14.93 |
| 793 | Entra ASA | 13.00 | 0.00 | -10.65 |

Figure 4.14: GNNExplainer Example 1: explaining the classification results for the company *Entra ASA*. The figure illustrates the most relevant subgraph for getting the prediction of the company *Entra ASA*. The first table below the figure provides the basic information for every company in the subgraph. The second table ranks the most important features (from left to right) that these companies share in order to draw the final prediction.




| Basic Information | | | | |
|---|---|---|---|---|
| **node index** | **company name** | **wikidata id** | **GICS Industry** | **label** |
| 135 | VeriSign Inc | Q734338 | IT Services | positive |
| 277 | Twitter Inc | Q918 | Interactive Media & Services | negative |
| 684 | Baidu Inc | Q14772 | Interactive Media & Services | positive |
| 699 | Morgan Stanley | Q334204 | Capital Markets | positive |
| 730 | Equinix Inc | Q851641 | Equity Real Estate Investment Trusts (REITs) | positive |

| Features (left to right ranked by importance) | | | | |
|---|---|---|---|---|
| **index** | **company** | **news sentiment** | **product score** | **report evidence** |
| 135 | VeriSign Inc | -1.19 | 0 | 0 |
| 277 | Twitter Inc | -164.18 | 0 | 8 |
| 684 | Baidu Inc | 22.52 | 0 | 0 |
| 699 | Morgan Stanley | 1.59 | 0 | 15 |
| 730 | Equinix Inc | 192.76 | 0 | 0 |

Figure 4.15: GNNExplainer Example 2: explaining the classification results for the company *Twitter Inc.* The figure illustrates the most relevant subgraph for getting the prediction of the company *Twitter Inc.* The first table below the figure provides the basic information for every company in the subgraph. The second table ranks the most important features (from left to right) that these companies share in order to draw the final prediction.

# Chapter 5

# Discussion & Future Work

## 5.1 Conclusion

In this dissertation, we demonstrate how to incorporate and model alternate data while evaluating current SDG frameworks and constructing new SDG frameworks. We begin by constructing a data engineering pipeline for extracting sustainability reports, Wikipedia pages, corporate news, and the Wikidata knowledge network of firms with which they are related, and then we link all the companies in various databases via a unique Wikipedia identifier. The sustainability report, Wikipedia product information, and MSCI sector information are used to evaluate the firm's product side, while company news and Wikidata knowledge graphs are used to evaluate the company's operational side. Then, we preprocessed the sustainability reports further to identify significant evidence in the reports that indicates the firm is actively contributing to a certain SDG, and we extracted essential information from the companies' Wikipedia pages that describe their core products, businesses, and services. The company news is retrieved from GDELT Global Entity Graph, where a company appearing in the news could be identified though Wikipedia identifier. The repetitious news are filtered out, and the important news of a company are selected through a proposed aggregated sentiment measure.

After preprocessing all the acquired data, we attempted to reverse-engineer the current SDG framework using classification to see how well our data could predict their scores. We discovered that product information derived from Wikipedia pages might be used as a complement to MSCI sector information to better characterize a company's business, thus providing an efficient method for predicting product scores. In addition, we illustrated how MSCI and RSAM evaluate products differently by explaining the text classifier via LIME. On the operational side, we discovered that news headlines and sentiments may assist in predicting operation scores. However, our news filtering methodology and data sources seem to be distinct from MSCI's. Incorporating all the information we gathered results in a considerable improvement in the F1 micro and macro scores compared to the baseline models, which only employ MSCI sector information. Additionally, we discovered that using graphs in classification enhances the performance of minority classes: it does not increase F1 micro scores, but it does improve F1 macro scores. However, the improvement is modest, suggesting that neither MSCI nor RSAM adequately leverage the graph structure. Finally, on the basis of the current frameworks, we suggest a new SDG framework. This new framework makes use of the characteristics of firms with a network structure and provides customized explanations for the automated scores.

### 5.1.1 Answering the research question

**RQ1: How can a company's contribution to sustainability be measured? How does the present SDG framework classify companies in terms of the SDGs? What is the most important factor that present SDG frameworks consider when classifying companies?**

A company's contribution to sustainability could be measured by looking at both the product, operational side, shareholders, and supply chain related activities of the company. The product side is mainly associated with the core products of the company. The operational side includes important news that is associated with this company. The present SDG frameworks from RobecoSAM and MSCI classify companies' contributions to sustainability mainly through assessment of the company's performance from both product and operational side. As for the product side, including Wikipedia product information significantly improves the classification performance compared to that of only using MSCI sector information. The events from the news also aid in predicting SDG scores, although the news we consider important do not fully overlap with that of MSCI. Overall, the company's core business and products are the most crucial components that the current SDG frameworks take into account, after which they further change the existing scores depending on the operational side of the business. The Wikidata knowledge graph contains Wikipedia product information, but it seems that RobecoSAM and MSCI SDG scores do not make use of the extra shareholder information and supply chain related activities present in the network.

**RQ2: How to evaluate the company's sustainability performance when it has not been evaluated by human experts? What causes firms' SDG ratings to differ from one another?**

The SDG scores of each company are based on the fact that they offer a variety of products and services, as well as operate in a variety of ways. Different SDG score providers place varying emphasis on various aspects of the firm's products and operations, as well as on news about the company that they deem important. For instance, we found that within the same class of SDG scores, RobecoSAM had a greater variety of firm products than MSCI. A company's sustainability performance may be evaluated by looking at its products and core business or services to see whether they contribute positively or badly to 17 SDG goals, and then adjusting this score by looking at the company's news. If a company's data is out of sample, we could use the network to make a sensible forecast of its SDG score. The network is typically useful to label a company when its information is out of sample, which indicates there is no similar company to this company in the training set. In this case, we could find companies similar to this company or companies that are connected to this company through the graph with relations of our interests.

**RQ3: How can the SDG scores be automated? How can a new framework be created that preserves the most significant aspects of the present framework while still being adaptable enough to accommodate new data? How can a framework be designed to improve transparency, accountability, and coherence?**

The SDG scores could be automated by studying the patterns in the attributes of existing data via classification algorithms. If there is a very limited amount of data with similar features, we could still infer an automated scores using the network structure. In order to handle both scenarios, we first cluster all companies in the company graph according to their features and connections. For companies in the same cluster, we take the average of their MSCI SDG (product) scores as the final automated score for the company. In this way, we are able to

improve the MSCI SDG scores by incorporating the network structure and additional company features of interest while still preserving the most significant aspects of their framework. In order to improve the transparency, accountability, and coherence of the framework, the framework is fully data-driven and is designed to be customizable with the ability to intuitively explain the individual automated score. The explanation consists of three parts: by looking at what the most important neighbors are linked to this company; how they are linked; and what the most important attributes this company shares with its most important neighbors.

## 5.2 Machine Learning in Impact Investing: industrial concerns v.s. academic concerns

There are several AI applications in the financial services industry. A few percentage points of increase may be crucial in disciplines such as computer vision, where applications like autonomous driving demand great accuracy. The goal of machine learning in finance should be to address actual financial issues, such as integrating alternative investment data, identifying fraud, recognizing patterns in time series data, etc., in order to have an influence on the industry. In this research, many NLP approaches are combined with classification algorithms and generative models to examine SDG frameworks and develop a new SDG framework.

AI researchers and financial specialists seem to approach financial issues from distinct vantage points. Financial experts frequently find that AI researchers apply machine learning techniques to a financial problem without understanding the underlying financial intuitions, whereas AI researchers may find that financial experts could have adopted more comprehensive machine learning techniques to solve the problems. Another widespread disagreement among financial professionals is that AI is seen as a black box since consumers cannot often explain or grasp why an AI model recommends or forecasts a certain result. This is frequently an issue in areas where model transparency is highly valued, such as the trading industry. In addition, several areas of finance have stringent restrictions requiring the use of AI without the inclusion of prejudice or discrimination. The significance of fairness and accountability in AI cannot be overstated. Aside from this, another problem is that the advantages of AI, such as its ability to iteratively improve with new data, may not always result in more accurate financial forecasts due to the fact that the past does not always repeat itself and valuable patterns change with time. In the area of quantitative trading, for instance, machine learning models might easily overfit, and training with less (more recent) data could be preferable to training with all data. For instance, for financial time series data, the high noise-to-signal ratio might cause machine learning models to fail. Hence, dimension reduction is often the first step in analyzing such data in finance. A black swan event, such as the financial crisis or COVID-19, might also influence the ML models' tendencies to deviate due to an anomaly in the data. The absence of data on these unforeseen occurrences diminishes the predicted accuracy and performance of ML models. Despite AI's technical and computational capabilities, many applications in the world of finance need a human-in-the-loop strategy for improved accuracy and continuous development. In the next section, we will go into further depth on this topic.

## 5.3 Discussion on the design of human-in-the-loop learning for continuous update and improvement

Due to the difficulties of developing high-quality labels, and the fact that many labels are very subjective and abstract, industrial applications frequently require data-centered research, which

is anticipated to play a critical and significant role once model-centered research has run its course. Manually labeling only a subset of the data enables businesses to conserve valuable resources, such as time and human capital. In other words, the suggested system should be capable of learning well from a limited number of labeled samples as analysts only manually score a portion of companies regarding their sustainability performance. However, the score generated by multiple analysts is subjective, as the technique is opaque and may vary significantly amongst experts. Therefore, the labels created by analysts are subjective, and not all labels can accurately reflect a company's sustainability category. In general, it is required that the model be overseen or supervised by the internal logic of human analysts while also being resilient to human mistakes during the labeling procedure. To put it another way, at a high level, "sustainability" is really an abstract concept. As a result, we want a robust weakly supervised classification model that is capable of learning from human concepts while also being resilient to human errors. The weakly supervision concept originated from Mahajan et al. (2018), and the current weakly supervised methods include noisy supervision and high-level supervision. Additionally, analysts in financial firms may wish to verify the quality of the classification output before rejecting or accepting the algorithm's score. This procedure enables analysts to constantly enhance the model's quality iteratively, which is why creating interactive AI systems for industrial applications that enable human-in-the-loop learning is critical[1] (active learning).

## 5.4 Copyrights of web data and its implications on machine learning

Due to exemptions for scientific research under the Copyright Law of the United States[2] and Europe Union[3], our work falls under fair use[4] of web resources, since we do not utlize personal data but only gather data that companies have made publicly available on the internet. Furthermore, most of the AI approaches are data-driven hence advancements in artificial intelligence relies on accessible data to train AI systems. This highlights a fundamental intellectual property challenge when training data includes copyrighted content. Especially for web-scrapping: text, photos, and videos are regularly utilized as ML training data without the owner's permission. Although use of some public domain data or Creative Commons-licensed data like Wikipedia appears to be a simple solution to circumvent copyright issues, it is still of one's interest to know whether scrapping data from the web is legal and whether including copyrighted information during AI model training constitute copyright infringements. Hence, we need tools and resources to understand copyright law and its relevance to bias in machine learning algorithms.

## 5.5 Future Work

Although our research has extensively addressed the subjects linked to investigating and extending SDG frameworks, there is still room for future improvement. First of all, our dataset could also be used to empower other research in the field of sustainability. Some concrete examples are illustrated in Figure A.9. For instance, the news and sustainability reports could be used together to perform a walk-talk analysis to detect companies' green washing or SDG

[1] https://labelyourdata.com/articles/human-in-the-loop-in-machine-learning

[2] https://www.copyright.gov/title17/92chap1.html#107

[3] https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0029:en:HTML

[4] https://guides.nyu.edu/fairuse#:~:text=Fair%20use%20allows%20limited%20use,research%20and%20scholarship%2C%20and%20teaching.

washing behaviors[5]. The pipeline we utilize to gather our data could also be inspiring for other researchers who need to collect a large amount of company-specific or entity-specific data from different corners of the web. Additionally, researchers in machine learning might create platforms and tools for getting legal counsel on the copyrights of the data (on the web) and how it could potentially introduce bias into ML algorithms. Secondly, future research could develop human-in-the-loop machine learning algorithms that could modify and update abstract labels in real time. Finally, Explainable Artificial Intelligence (XAI) should help people make informed decisions, especially in sectors like finance. In this paper, we use LIME for explaining text classifiers and GNNExplainer for explaining graph algorithms. GNNExplainer is applied to the company graph (which is not a knowledge graph) created from the original knowledge graph because it cannot deal with the edge properties present in a knowledge graph and it also produces an explanation from a less logical computational graph. It is preferable to construct robust counterfactual explanations for our particular use case by explicitly modeling the common decision logic that exists in the original knowledge graph in a way that resembles how humans intuitively grasp the logic in knowledge graphs.

[5] https://www.emeraldgrouppublishing.com/opinion-and-blog/csr-greenwashing-general-sdg-washing-potential-threat-sdg-implementation

# Bibliography


[1] Simran Arora et al. "Contextual Embeddings: When Are They Worth It?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 2650–2663. DOI: 10.18653/v1/2020.acl-main.236. URL: https://aclanthology.org/2020.acl-main.236.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473 (2015).

[3] Ramnath Balasubramanyan, Frank Lin, and William W. Cohen. "Node Clustering in Graphs: An Empirical Study". In: (Dec. 2010). DOI: 10.1184/R1/6475925.v1. URL: https://kilthub.cmu.edu/articles/journal_contribution/Node_Clustering_in_Graphs_An_Empirical_Study/6475925.

[4] Florian Berg, Julian F. Kölbel, and Roberto Rigobon. "Aggregate Confusion: The Divergence of ESG Ratings". In: *Corporate Governance & Finance eJournal* (2019).

[5] Federico Bianchi, Silvia Terragni, and Dirk Hovy. "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 759–766. DOI: 10.18653/v1/2021.acl-short.96. URL: https://aclanthology.org/2021.acl-short.96.

[6] Cristian Bodnar, Cătălina Cangea, and Pietro Liò. "Deep Graph Mapper: Seeing Graphs Through the Neural Lens". In: *Frontiers in Big Data* 4 (2021). ISSN: 2624-909X. DOI: 10.3389/fdata.2021.680535. URL: https://www.frontiersin.org/article/10.3389/fdata.2021.680535.

[7] Richard J. Bolton and David J. Hand. "Statistical Fraud Detection: A Review". In: *Statistical Science* 17.3 (2002), pp. 235–249. ISSN: 08834237. URL: http://www.jstor.org/stable/3182781 (visited on 05/04/2022).

[8] Samuel R. Bowman et al. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://aclanthology.org/D15-1075.

[9] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[10] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly Detection: A Survey". In: *ACM Comput. Surv.* 41.3 (2009). ISSN: 0360-0300. DOI: 10.1145/1541880.1541882. URL: https://doi.org/10.1145/1541880.1541882.

[11] C Chao, A Liaw, and L Breiman. “Using random forest to learn imbalanced data”. In: *Berkeley* 110 (2004), pp. 1–12.

[12] N V Chawla et al. “SMOTE: Synthetic minority over-sampling technique”. In: *Journal of Artificial Intelligence Research* 16 (June 2002), pp. 321–357.

[13] Mike Chen et al. “NLP for SDGs: Measuring Corporate Alignment with the Sustainable Development Goals”. In: *The Journal of Impact and ESG Investing* (2021). ISSN: 2693-1982. DOI: 10.3905/jesg.2021.1.035. eprint: https://jesg.pm-research.com/content/early/2021/12/12/jesg.2021.1.035.full.pdf. URL: https://jesg.pm-research.com/content/early/2021/12/12/jesg.2021.1.035.

[14] Tom Corringham et al. “BERT Classification of Paris Agreement Climate Action Plans”. In: *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*. 2021. URL: https://www.climatechange.ai/papers/icml2021/45.

[15] TÃmea CzvetkÃ³ et al. “The intertwining of world news with Sustainable Development Goals: An effective monitoring tool”. In: *Heliyon* 7.2 (2021), e06174. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2021.e06174. URL: https://www.sciencedirect.com/science/article/pii/S2405844021002796.

[16] Antonin Delpeuch. “OpenTapioca: Lightweight Entity Linking for Wikidata”. In: *Proceedings of the 1st Wikidata Workshop (Wikidata 2020) co-located with 19th International Semantic Web Conference(OPub 2020), Virtual Conference, November 2-6, 2020*. Ed. by Lucie-Aimée Kaffee et al. Vol. 2773. CEUR Workshop Proceedings. CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2773/paper-02.pdf.

[17] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423. URL: https://doi.org/10.18653/v1/n19-1423.

[18] Robert G. Eccles, Ioannis Ioannou, and George Serafeim. “The Impact of Corporate Sustainability on Organizational Processes and Performance”. In: *Management Science* 60.11 (2014), pp. 2835–2857. DOI: 10.1287/mnsc.2014.1984. eprint: https://doi.org/10.1287/mnsc.2014.1984. URL: https://doi.org/10.1287/mnsc.2014.1984.

[19] George Forman and Martin Scholz. “Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement”. In: *SIGKDD Explor. Newsl.* 12.1 (2010), 49–57. ISSN: 1931-0145. DOI: 10.1145/1882471.1882479. URL: https://doi.org/10.1145/1882471.1882479.

[20] David Friederich et al. “Automated Identification of Climate Risk Disclosures in Annual Corporate Reports”. In: *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*. 2021. URL: https://www.climatechange.ai/papers/icml2021/25.

[21] Yang Gao, Nicolò Colombo, and Wei Wang. “Adapting by Pruning: A Case Study on BERT”. In: *CoRR* abs/2105.03343 (2021). arXiv: 2105.03343. URL: https://arxiv.org/abs/2105.03343.

[22] Clinton Gormley and Zachary Tong. *Elasticsearch: The Definitive Guide*. 1st. O’Reilly Media, Inc., 2015. ISBN: 1449358543.

[23] Guo Haixiang et al. “Learning from class-imbalanced data: Review of methods and applications”. In: *Expert Systems with Applications* 73 (2017), pp. 220–239. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2016.12.035. URL: https://www.sciencedirect.com/science/article/pii/S0957417416307175.

[24] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. “Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning”. In: *Lecture Notes in Computer Science*. Lecture notes in computer science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 878–887.

[25] Xu Han et al. “Pre-trained models: Past, present and future”. In: *AI Open* 2 (2021), pp. 225–250. ISSN: 2666-6510. DOI: https://doi.org/10.1016/j.aiopen.2021.08.002. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000231.

[26] Zellig S. Harris. “Distributional Structure”. In: *¡i¿WORD¡/i¿* 10.2-3 (1954), pp. 146–162. DOI: 10.1080/00437956.1954.11659520. eprint: https://doi.org/10.1080/00437956.1954.11659520. URL: https://doi.org/10.1080/00437956.1954.11659520.

[27] Haibo He et al. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Hong Kong, China: IEEE, June 2008.

[28] Tin Kam Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[29] Jin Huang and C.X. Ling. “Using AUC and accuracy in evaluating learning algorithms”. In: *IEEE Transactions on Knowledge and Data Engineering* 17.3 (2005), pp. 299–310. DOI: 10.1109/TKDE.2005.50.

[30] Shaoxiong Ji et al. “A Survey on Knowledge Graphs: Representation, Acquisition, and Applications”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2 (2022), pp. 494–514. DOI: 10.1109/TNNLS.2021.3070843.

[31] Jeff Johnson, Matthijs Douze, and Hervé Jégou. “Billion-scale similarity search with GPUs”. In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547.

[32] Javid Jouzdani and Kannan Govindan. “On the sustainable perishable food supply chain network design: A dairy products case to achieve sustainable development goals”. In: *Journal of Cleaner Production* 278 (2021), p. 123060. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2020.123060. URL: https://www.sciencedirect.com/science/article/pii/S095965262033105X.

[33] Jens Keilwagen, Ivo Grosse, and Jan Grau. “Area under Precision-Recall Curves for Weighted and Unweighted Data”. In: *PLOS ONE* 9.3 (Mar. 2014), pp. 1–13. DOI: 10.1371/journal.pone.0092209. URL: https://doi.org/10.1371/journal.pone.0092209.

[34] Jason Kessler. “Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ”. In: *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 85–90. URL: https://aclanthology.org/P17-4015.

[35] Elham Kheradmand et al. “A NLP-based Analysis of Alignment of Organizations’ Climate-Related Risk Disclosures with Material Risks and Metrics”. In: *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*. 2021. URL: https://www.climatechange.ai/papers/neurips2021/69.

[36] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[37] Günter Knieps. *Network Economics*. Springer Texts in Business and Economics 978-3-319-11695-2. Springer, 2015. ISBN: ARRAY(0x4691e030). DOI: 10.1007/978-3-319-11695-2. URL: https://ideas.repec.org/b/spr/sptbec/978-3-319-11695-2.html.

[38] Tamara G. Kolda et al. "A Scalable Generative Graph Model with Community Structure". In: *SIAM Journal on Scientific Computing* 36.5 (2014), pp. C424–C452. DOI: 10.1137/130914218. eprint: https://doi.org/10.1137/130914218. URL: https://doi.org/10.1137/130914218.

[39] M Kubat and S Matwin. "Addressing the curse of imbalanced training sets: One-sided selection". In: *Proceedings of the 14th International Conference on Machine Learning* 97 (1997), pp. 179–186.

[40] Agostina J. Larrazabal et al. "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis". In: *Proceedings of the National Academy of Sciences* 117.23 (2020), pp. 12592–12594. DOI: 10.1073/pnas.1919012117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1919012117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1919012117.

[41] Quoc Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, 2014, pp. 1188–1196. URL: https://proceedings.mlr.press/v32/le14.html.

[42] Kalev Leetaru and Philip A. Schrodt. "GDELT: Global data on events, location, and tone". In: *ISA Annual Convention* (2013). URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.686.6605.

[43] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.

[44] Shouwei Li and Chao Wang. "Network structure, portfolio diversification and systemic risk". In: *Journal of Management Science and Engineering* 6.2 (2021), pp. 235–245. ISSN: 2096-2320. DOI: https://doi.org/10.1016/j.jmse.2021.06.006. URL: https://www.sciencedirect.com/science/article/pii/S2096232021000433.

[45] Qi Liu, Matt J. Kusner, and Phil Blunsom. "A Survey on Contextual Embeddings". In: *ArXiv* abs/2003.07278 (2020).

[46] Sasha Luccioni, Emi Baylor, and Nicolas Duchene. "Analyzing Sustainability Reports Using Natural Language Processing". In: *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*. 2020. URL: https://www.climatechange.ai/papers/neurips2020/31.

[47] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[48] Dhruv Mahajan et al. "Exploring the Limits of Weakly Supervised Pretraining: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II". In: Sept. 2018, pp. 185–201. ISBN: 978-3-030-01215-1. DOI: 10.1007/978-3-030-01216-8_12.

[49] I Mani and J Zhang. ""kNN approach to unbalanced data distributions: A case study involving information extraction". In: *Proceedings of the Workshop on Learning from Imbalanced Data Sets*. 2003, pp. 1–7.

[50] Massimo Melucci. "Vector-Space Model". In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M. TAMER ÖZSU. Boston, MA: Springer US, 2009, pp. 3259–3263. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_918. URL: https://doi.org/10.1007/978-0-387-39940-9_918.

[51] G Menardi and N Torelli. "Training and assessing classification rules with unbalanced data"". In: *Data Mining and Knowledge Discovery* 28 (2014), pp. 92–122.

[52] Pablo Mendes et al. "DBpedia Spotlight: Shedding Light on the Web of Documents." In: *In the Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*. 2011.

[53] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781 (2013). URL: http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781.

[54] Prakamya Mishra and Rohan Mittal. "NeuralNERE: Neural Named Entity Relationship Extraction for End-to-End Climate Change Knowledge Graph Construction". In: *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*. 2021. URL: https://www.climatechange.ai/papers/icml2021/76.

[55] Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. "Borderline over-sampling for imbalanced data classification". In: *In Proceedings of the 5th International Workshop on computational Intelligence and Applications* 3.1 (2011), p. 4.

[56] José Eduardo Eguiguren Palacios and Nelson Piedra. "Connecting Open Data and Sustainable Development Goals using a Semantic Knowledge Graph Approach". In: *ONTOBRAS*. 2019.

[57] Harshita Patel et al. "A review on classification of imbalanced data for wireless sensor networks". In: *International Journal of Distributed Sensor Networks* 16.4 (2020), p. 155014772091640. DOI: 10.1177/1550147720916404. URL: https://doi.org/10.1177%2F1550147720916404.

[58] Gustavo Peralta and Abalfazl Zareei. "A network approach to portfolio selection". In: *Journal of Empirical Finance* 38 (2016), pp. 157–180. ISSN: 0927-5398. DOI: https://doi.org/10.1016/j.jempfin.2016.06.003. URL: https://www.sciencedirect.com/science/article/pii/S0927539816300603.

[59] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.

[60] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[61] Yutaka Sasaki. "The truth of the F-measure". In: *Teach Tutor Mater* (Jan. 2007).

[62] Franco Scarselli et al. "The Graph Neural Network Model". In: *IEEE Transactions on Neural Networks* 20.1 (2009), pp. 61–80. DOI: 10.1109/TNN.2008.2005605.

[63] Michael Schlichtkrull et al. "Modeling Relational Data with Graph Convolutional Networks". In: *The Semantic Web*. Ed. by Aldo Gangemi et al. Cham: Springer International Publishing, 2018, pp. 593–607. ISBN: 978-3-319-93417-4.

[64] Oleksandr Shchur and Stephan Günnemann. "Overlapping Community Detection with Graph Neural Networks". In: *CoRR* abs/1909.12201 (2019). arXiv: 1909.12201. URL: http://arxiv.org/abs/1909.12201.

[65] Filomena Stamou. "Data Science for Social Good: development of a Knowledge Graph targeted to Sustainable Development Goals". In: (Dec. 2021).

[66] Fan-Yun Sun et al. "VGraph: A Generative Model for Joint Community Detection and Node Representation Learning". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[67] Yanmin Sun, Andrew K. C. Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review". In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.04 (2009), pp. 687–719. DOI: 10.1142/S0218001409007326. eprint: https://doi.org/10.1142/S0218001409007326. URL: https://doi.org/10.1142/S0218001409007326.

[68] Sayyida Tabinda Kokab, Sohail Asghar, and Shehneela Naz. "Transformer-based deep learning models for the sentiment analysis of social media data". In: *Array* (2022), p. 100157. ISSN: 2590-0056. DOI: https://doi.org/10.1016/j.array.2022.100157. URL: https://www.sciencedirect.com/science/article/pii/S2590005622000224.

[69] I Tomek. "Two modifications of CNN". In: *IEEE Transactions on Systems, Man, and Cybernetics* 6 (1976), pp. 769–772.

[70] Anton Tsitsulin et al. "Graph Clustering with Graph Neural Networks". In: *CoRR* abs/2006.16904 (2020). arXiv: 2006.16904. URL: https://arxiv.org/abs/2006.16904.

[71] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[72] Ricardo Vinuesa et al. "The role of artificial intelligence in achieving the Sustainable Development Goals". In: *Nature Communications* 11.1 (Jan. 2020). DOI: 10.1038/s41467-019-14108-y. URL: https://doi.org/10.1038/s41467-019-14108-y.

[73] Shirui Wang, Wen'an Zhou, and Chao Jiang. "A survey of word embeddings based on deep learning". In: *Computing* 102 (2019), pp. 717–740.

[74] Nicolas Webersinke et al. "ClimateBERT: A Pretrained Language Model for Climate-Related Text". In: *arXiv preprint arXiv:2110.12010* (2021).

[75] Dennis L Wilson. "Asymptotic properties of nearest neighbor rules using edited data". In: *IEEE Trans. Syst. Man Cybern.* SMC-2.3 (July 1972), pp. 408–421.

[76] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

[77] Pengyi Yang et al. “Sample Subset Optimization for Classifying Imbalanced Biological Data”. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Joshua Zhexue Huang, Longbing Cao, and Jaideep Srivastava. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 333–344. ISBN: 978-3-642-20847-8.

[78] Rex Ying et al. “GNNExplainer: Generating Explanations for Graph Neural Networks”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: NIPS’19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019.

[79] Shunjie Yuan et al. “Community Detection with Graph Neural Network using Markov Stability”. In: *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 2022, pp. 437–442. DOI: 10.1109/ICAIIC54071.2022.9722614.

[80] Ziwei Zhang, Peng Cui, and Wenwu Zhu. *Deep Learning on Graphs: A Survey*. cite arxiv:1812.04202Comment: 15 pages, 10 figures. 2018. URL: http://arxiv.org/abs/1812.04202.

[81] Ziwei Zhang, Peng Cui, and Wenwu Zhu. “Deep Learning on Graphs: A Survey”. In: *IEEE Transactions on Knowledge and Data Engineering* 34.1 (2022), pp. 249–270. DOI: 10.1109/TKDE.2020.2981333.

[82] Fuzhen Zhuang et al. “A Comprehensive Survey on Transfer Learning”. In: *Proceedings of the IEEE* PP (July 2020), pp. 1–34. DOI: 10.1109/JPROC.2020.3004555.

# Appendix A

# Optional appendix

## A.1 Core mathematical details of GCN shortly explained

We need to understand fast approximating convolutions on graphs and spectral graph convolutions in order to gain a better understanding on how GCN works. GCN considers one-hop neighborhood then aggregate projected features before scaling and summing up to get a new representation. A is the adjacency matrix, and after add self-connection we have $\hat{A} = A + I$; D is the degree matrix, and similarly we have $\hat{D} = D + I$. Suppose we have N nodes and D features for each node, $H^{(l)} \in R \times D$ is the feature matrix for the node at layer $l$; $W^{(l)}$ is the linear projection layer. Then the layer-wise propagation rule could be presented as $H^{l+1} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$. Notice that $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$ is a fix term used for the normalization purpose. The purpose of normalization is to avoid exploding or diminishing vanishing gradient problems. The rest of the terms are projected feature vectors for every node in the graph. The sigma activation function will accumulate/aggregate the features and scale them using the fix term, then they will be summed up to get a new representation. For a node in the graph, it will accumulate all the vectors for all the nodes that connected to this node. As a matter of fact, its final representation will contain some portions of all these features vectors.

When it comes to spectral graph convolutions, we have the Laplacian matrix $L = D - A$ and we want to find the eigen-decomposition of L which will result in eigenvalues and corresponding eigenvectors. The spectral convolutions on a graph is defined as $g_\theta * x = U \cdot g_\theta \cdot U^T \cdot x$ where $x \in R^N$ is a signal (a scalar for every node), $g_\theta$ is a filter, and U is a matrix of eigenvectors for $L = I_N - \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}} = U\Lambda U$ where $\Lambda$ is a diagonal matrix with eigenvalues. It might be interesting to notice that $U^T x$ is actually the graph Fourier transformation of x. The meaning of this step that we project the signal into the new basis defined by the Laplacian eigenvectors. In this way, we do not learn the W matrix as previously. Instead, we try to learn the filter $g_\theta$. However, this raises the problem that multiplication of the eigenvector matrix U is expensive ($O(N^2)$) and calculating the eigen-decomposition of L in order to get U is expensive for large graphs ($O(N^3)$). Therefore, the GCN paper suggests to use a method which use Chebyshev polynomial approximation recursively. However, they only consider 1-hop neighborhood because they expect this could prevent overfitting on the local neighborhood. Renormalization trick is also applied: after adding up self-connections $I_N + D^{-\frac{1}{2}}\hat{A}D^{-\frac{1}{2}}$ we have $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$, and we can get the convolved signal matrix $Z = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}x\theta$. This results in a reduced complexity as the final expression does not have to compute eigenvalues or eigenvectors.

## A.2 Pretrained models information

Table A.1: Pretrained model information for semantic search

| Pre-trained model | Sentence embeddings | Semantic search | Avg performance | Speed | Model Size |
|---|---|---|---|---|---|
| all-mpnet-base-v2 | 69.57 | 57.02 | 63.30 | 2800 | 420MB |

Information obtained from https://www.sbert.net/docs/pretrained_models.html

Table A.2: Pretrained model information for BERT-base model fine-tuned for NLI task

| Evaluation | Contradiction | Entail | Neutral |
|---|---|---|---|
| Precision | 0.88 | 0.90 | 0.81 |
| Recall | 0.88 | 0.87 | 0.84 |
| F1 | 0.88 | 0.88 | 0.82 |

Information obtained from https://github.com/yg211/bert_nli

## A.3 Algorithm details (Hyperparameters)

Table A.3: Hyperparameters of GCN and RGCN for classification

| Hyperparameters | GCN | RGCN |
|---|---|---|
| optimizer | Adam | Adam |
| learning rate | 0.01 | 0.01 |
| weight decay | 5.00E-03 | 5.00E-04 |
| epoch | 5000 | 5000 |
| number of layers | 2 graph convolution layers | 2 fast relational graph convolution layers |
| hidden layer size | 16 | 16 |
| loss function | softmax | softmax |
| activation function | RELU | RELU |
| drop out ratio | 0.5 | FALSE |

Table A.4: Hyperparameters of GNN for clustering

| **GNN (clustering) hyperparameters** | |
|---|---|
| K | 50 |
| hidden layer size | 128 |
| weight decay | 1.00E-02 |
| drop out ratio | 0.5 |
| batch normalization | TRUE |
| learning rate | 1.00E-03 |
| number of epochs | 2000 |
| balance loss | TRUE |
| stochastic loss (not full batch training) | TRUE |
| batch size | 200 |
| normalize node features | TRUE |
| early stopping | TRUE |
| optimizer | Adam |

## A.4 Knowledge graph related supplementary


Figure A.1: Knowledge graph example (full)

| | | | | | |
|---|---|---|---|---|---|
| stock exchange | instance of | headquarters location | country | location of formation | industry |
| product or material produced | category for employees of the organization | owner of | uses | member of | different from |
| replaces | on focus list of Wikimedia project | parent organization | follows | employer | partnership with |
| country of citizenship | located in the administrative territorial entity | board member | said to be the same as | topic's main template | supervisory board member |
| brand | affiliation | sport | topic's main Wikimedia portal | business division | described by source |
| work location | occupation | copyright representative | operating area | location | creator |
| followed by | significant person | replaced by | language of work or name | subject has role | cuisine |
| official app | depicted by | notable work | programming language | operator | author |
| stock market index | main regulatory text | sex or gender | given name | family name | educated at |
| business model | has facility | member of sports team | nominated for | platform | online access status |
| official language | father | mother | instrument | languages spoken, written or signed | distribution format |
| copyright status | copyright license | operating system | input method | language used | is a list of |
| subsidiary | named after | part of | topic's main category | award received | legal form |
| founded by | chief executive officer | archives at | owned by | chairperson | director / manager |
| field of work | investor | has works in the collection | significant event | typically sells | history of topic |
| subclass of | applies to jurisdiction | external auditor | has part | has quality | distributed by |
| permanent duplicated item | motto | Wikimedia outline | country of origin | copyright status as a creator | participant in |
| currency | genre | item operated | office held by head of the orga | located on street | track gauge |
| mascot | record label | house publication | foundational text | open data portal | affiliated worker organisation |
| side effect | manufacturer | official podcast | chief operating officer | official color(s) | separated from |
| place of birth | has list | writing system | represents | flag | category for maps |
| convicted of | merged into | main subject | origin of the watercourse | mouth of the watercourse | website account on |
| location of creation | participant | category of associated people | publisher | developer | prohibits |

Figure A.2: All the relations in the initial extracted Wikidata knowledge graph

| wikidata_id_start_name | property_name | wikidata_id_end_name |
|---|---|---|
| 3D Systems | stock exchange | New York Stock Exchange |
| 3D Systems | instance of | business |
| 3D Systems | instance of | public company |
| 3D Systems | headquarters location | Rock Hill |
| 3D Systems | country | United States of America |
| 3D Systems | location of formation | Valencia |
| 3D Systems | industry | 3D printing |
| 3D Systems | industry | computer-aided design |
| 3D Systems | subsidiary | Z Corporation |
| 3M | named after | Minnesota |
| 3M | named after | mining |
| 3M | stock exchange | New York Stock Exchange |
| 3M | stock exchange | Tokyo Stock Exchange |
| 3M | instance of | conglomerate |
| 3M | instance of | brand |
| 3M | instance of | public company |
| 3M | part of | Dow Jones Industrial Average |
| 3M | part of | S&P 500 |
| 3M | headquarters location | Maplewood |
| 3M | headquarters location | St. Paul |
| 3M | headquarters location | Two Harbors |
| 3M | headquarters location | Saint Paul |
| 3M | headquarters location | Minnesota |
| 3M | topic's main category | Category:3M |
| 3M | award received | National Medal of Technology and Innovation |
| 3M | award received | Silver Anvil Awards |
| 3M | award received | Silver Anvil Awards |
| 3M | country | United States of America |
| 3M | subsidiary | 3M Innovative Properties |
| 3M | subsidiary | 3M (Canada) |
| 3M | subsidiary | 3M (Israel) |
| 3M | subsidiary | 3M (France) |
| 3M | subsidiary | 3M (United Kingdom) |
| 3M | subsidiary | 3M (Germany) |
| 3M | legal form | public company |

Figure A.3: Examples of triples in the initial extracted Wikidata knowledge graph

## A.5 Feature importance plot of text classifiers

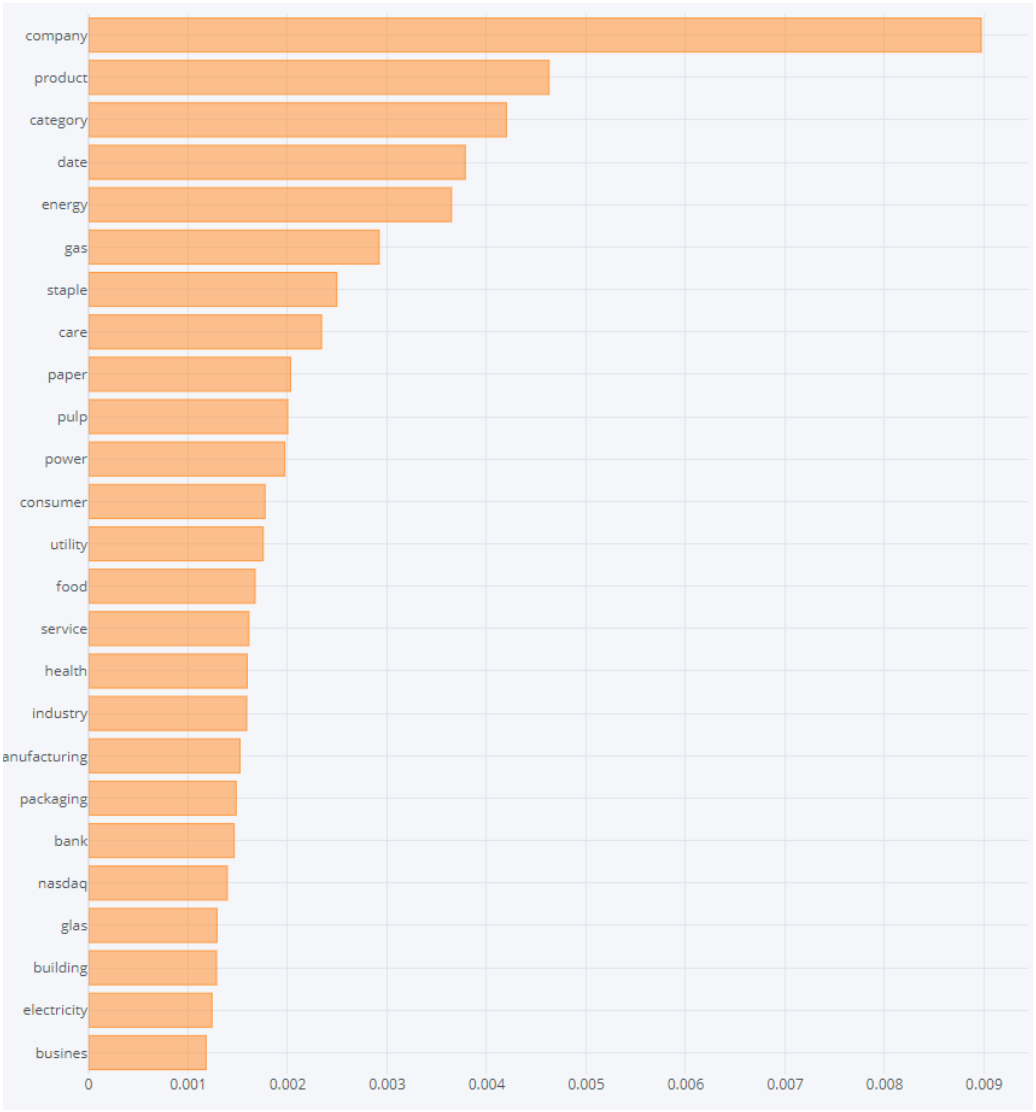



Figure A.4: Aggregate importance split by class SDG 7 RSAM

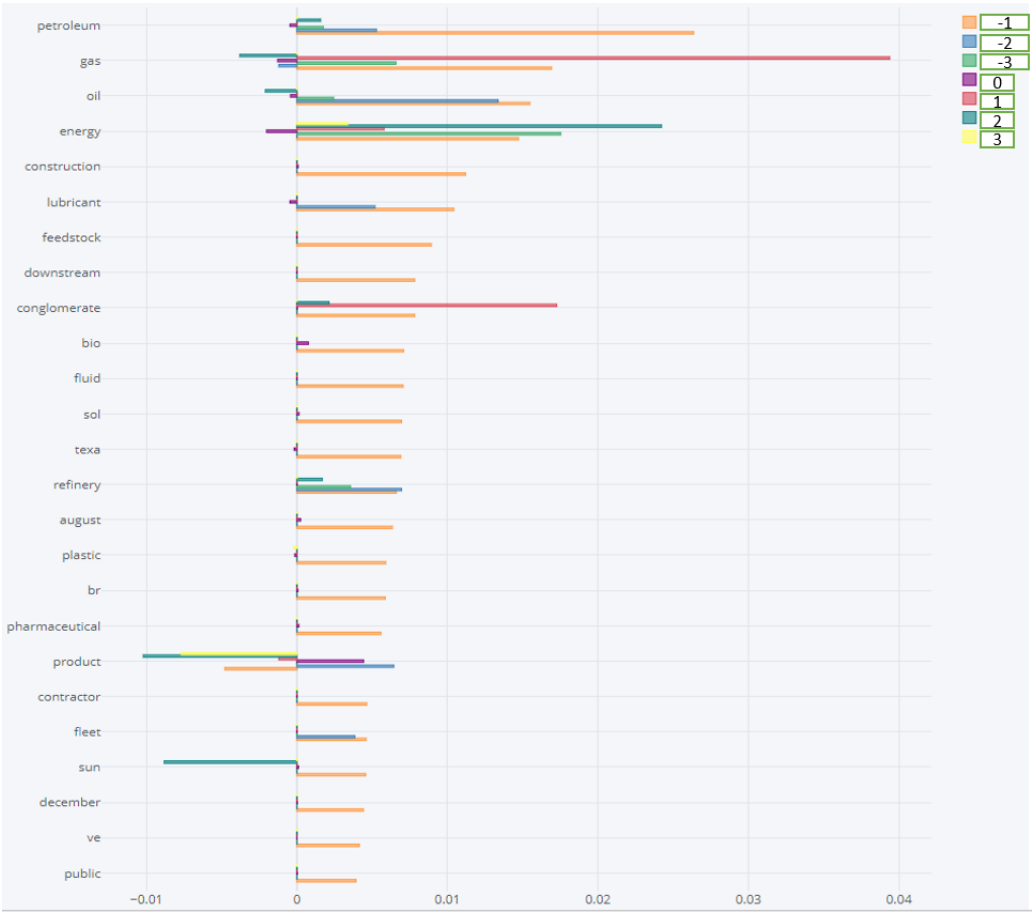


Figure A.5: Aggregate importance SDG 7 RSAM

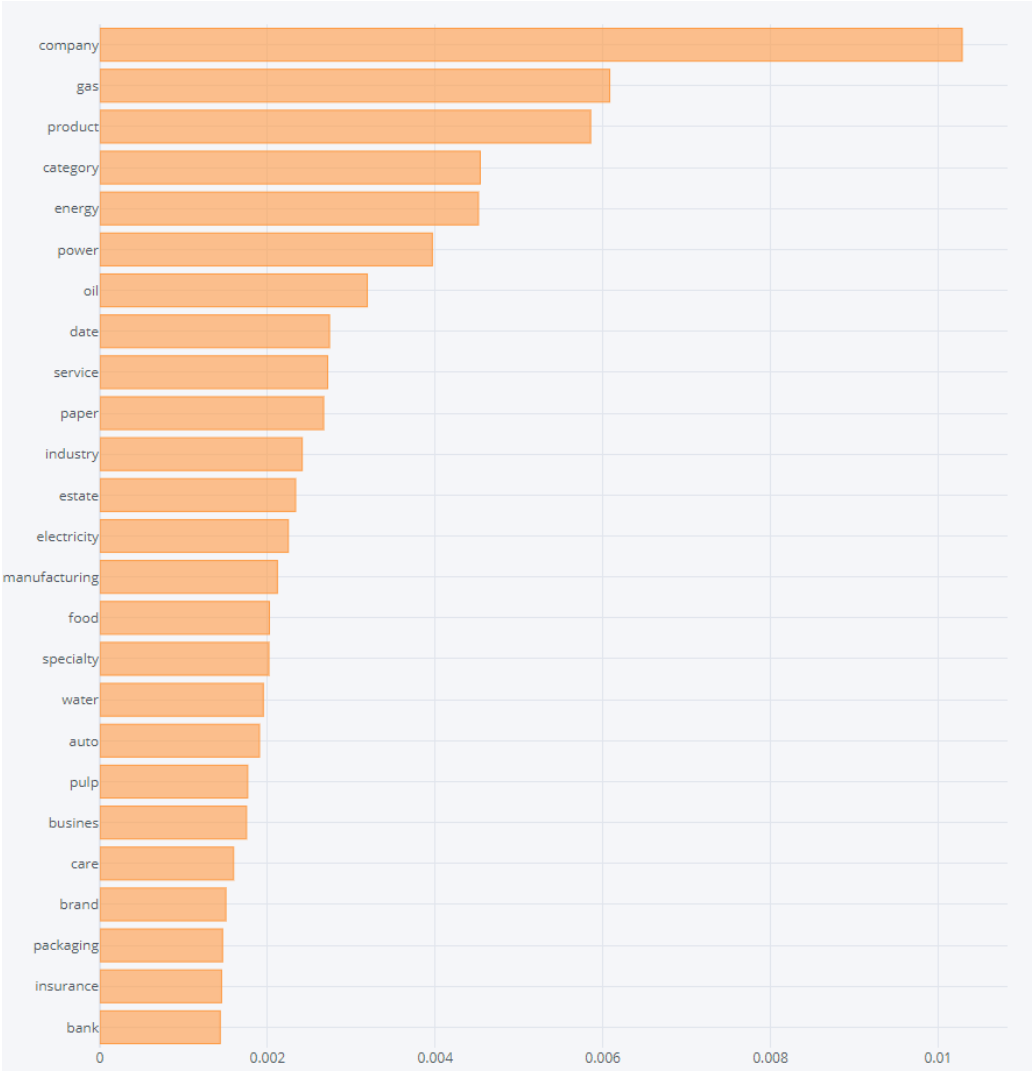


Figure A.6: Aggregate importance split by class SDG 7 MSCI

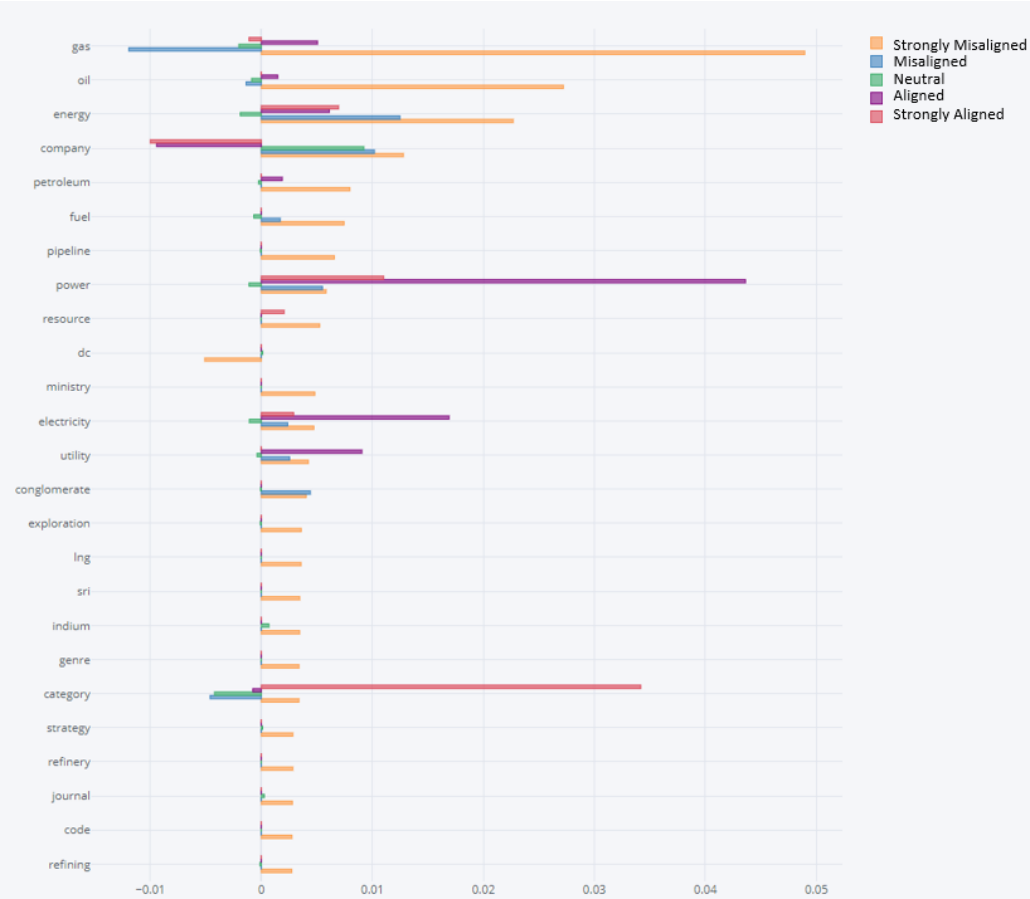


Figure A.7: Aggregate importance SDG 7 MSCI

## A.6 Schema for heuristically generating scores

### GENERATING SCORES - HEURISTIC APPROACH

- Product score (Base)
  - Wiki + Sector → predict product score (what we did before)
- Report Evidence
  - Found or not found
  - If found, base score +1, 0 otherwise
- News
  - Average sentiment of most impactful good&bad news
  - If avg_sentiment > 0 → base score + 1; avg_sentiment < 0 → base score -1; unchanged otherwise
- Modified score = modified by report evidence and news
- Connections
  - Knowledge Graph, if a group of companies are connected, for company in this group:
    - Get the average score of modified scores in this group
    - If company A's modified score > avg_modified_score_of_group & sign(A's modified score) != sign(avg_modified_score_of_group): adjust A's modified score up or down
    - E.g. If A is above group average, adjust down; adjust up otherwise.

Figure A.8: Concepts for generating scores: heuristic way of generating scores. The code and generated scores are available in the code repository.

## A.7 Other

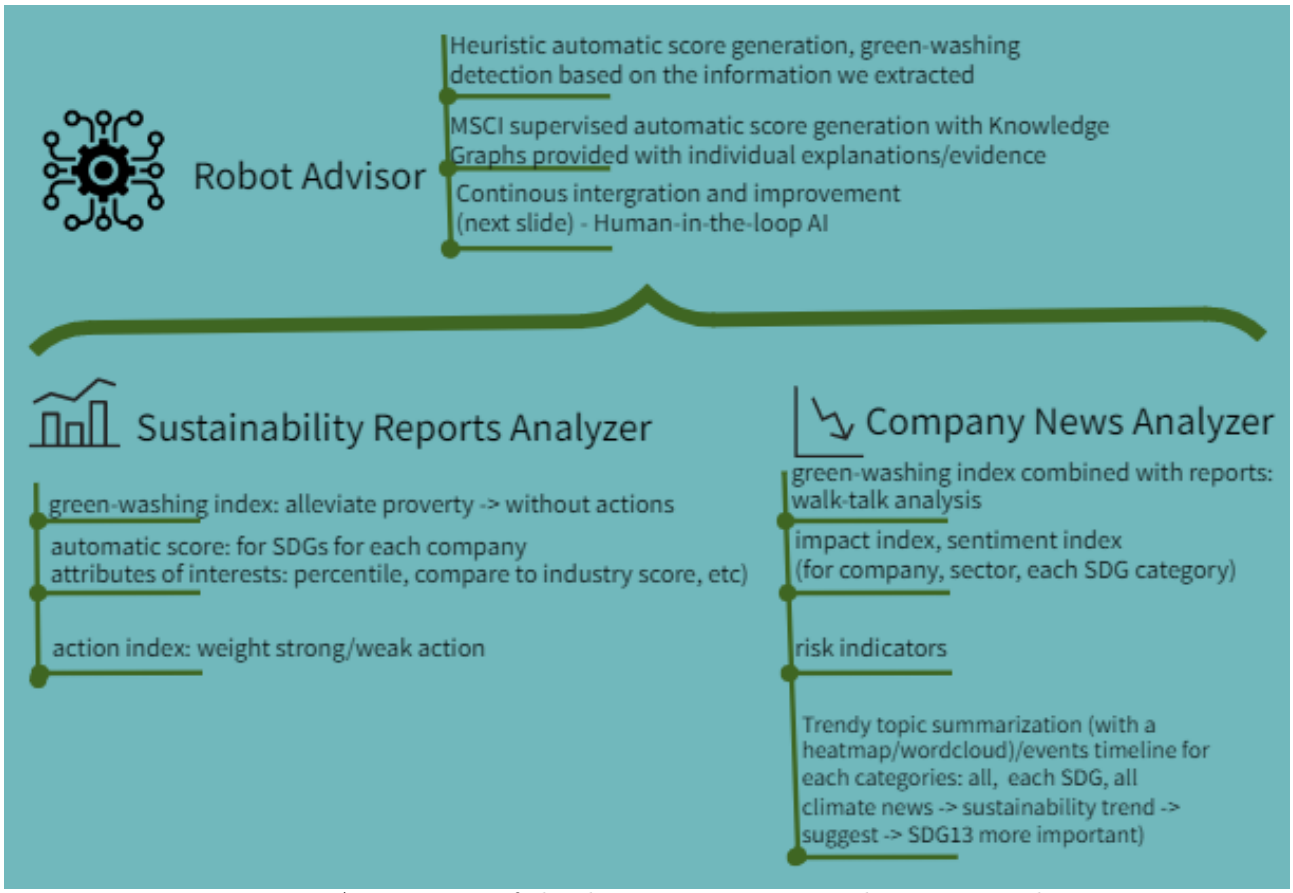



Figure A.9: Usages of the dataset to empower relevant research