


APPLICATION PAPER

Knowledge-driven neural models for extraction and analysis of sustainability events from the web

Abir Naskar, Tushar Goel, Vipul Chauhan, Ishan Verma, Tirthankar Dasgupta  and Lipika Dey

TCS Research, India.

Corresponding author: Tirthankar Dasgupta; Email: dasgupta.tirthankar@tcs.com

Received: 17 March 2023; **Revised:** 24 November 2023; **Accepted:** 16 September 2024

Keywords: Sustainability; ESG; Violation; Safety Incident; Classification

Abstract

Sustainability practices of a company reflect its commitments to the environment, societal good, and good governance. Institutional investors take these into account for decision-making purposes, since these factors are known to affect public opinion and thereby the stock indices of companies. Though sustainability score is usually derived from information available in self-published reports, News articles published by regulatory agencies and social media posts also contain critical information that may affect the image of a company. Language technologies have a critical role to play in the analytics process. In this paper, we present an event detection model for detecting sustainability-related incidents and violations from reports published by various monitoring and regulatory agencies. The proposed model uses a multi-tasking sequence labeling architecture that works with transformer-based document embeddings. We have created a large annotated corpus containing relevant articles published over three years (2015–2018) for training and evaluating the model. Knowledge about sustainability practices and reporting incidents using the Global Reporting Initiative (GRI) standards have been used for the above task. The proposed event detection model achieves high accuracy in detecting sustainability incidents and violations reported about an organization, as measured using cross-validation techniques. The model is thereafter applied to articles published from 2019 to 2022, and insights obtained through aggregated analysis of incidents identified from them are also presented in the paper. The proposed model is envisaged to play a significant role in sustainability monitoring by detecting organizational violations as soon as they are reported by regulatory agencies and thereby supplement the Environmental, Social, and Governance (ESG) scores issued by third-party agencies.

Impact Statement

The proposed work is envisaged to play a significant role in sustainability monitoring by detecting organizational violations as soon as they are reported by regulatory agencies and thereby supplement the Environmental, Social, and Governance (ESG) scores issued by third-party agencies. We hope such real-time sustainability monitoring and regulatory reporting definitely help in bringing down the number of incidents. Aggregate analysis like ours can also help other organizations to become cautious.

1. Introduction

The sustainability charter of an organization describes its commitment to environmental and societal good along with good governance (Tregidga et al., 2014; Garbie, 2015; Ahi, 2014). Organizations are awarded sustainability scores, that are computed based on their Environmental (E), Social (S), and Governance

(G) practices, collectively termed as ESG parameters (Lozano and von Haartman, 2018; et al., 2015; Fenwick, 2007). As awareness about sustainable practices is spreading across the world, it is becoming increasingly clear that better ESG performance can drive better investment outcomes too (Bang et al., 2023; Raghunandan and Rajgopal, 2022), since the reputation of an organization is squarely dependent on its sustainability program. Sustainability studies (ESG, 2018) indicate that around 69% of the companies that experienced a high or severe ESG incident¹, experienced an average market cap decline of 6% within the next ten days.

ESG scores are available readily from third-party sources like CRISIL², MSCI³, etc. who use proprietary methods to compute them. The scores are based on information gathered from multiple sources including self-published sustainability reports by organizations, organization websites, exchange filings, annual reports, investor presentations, Carbon Disclosure Project (CDP) filings, etc (Fiaschi et al., 2020; Yuan et al., 2022; Henisz et al., 2019). The scores may reflect the effect of performance over multiple years. While end users can obtain the scores, the reason for the scores is not readily available to them. One major ask of end users in recent times is to be provided with insights behind the scoring (Murphy and McGrath, 2013). Awareness about an organization's current activities, and not just the past, is also in demand. It is also important for end users like investment bankers interested in a company to obtain the views of the regulatory agencies since negative perception of these agencies can lead the company into trouble (Sinha et al., 2020; Geddes et al., 2018; Bouma et al., 2017; et al., 2023).

ESG regulatory bodies comprise of government authorities or independent agencies that exercise autonomous dominion over environmental, social, or governance-related activities of organizations falling within their area of judicial purview (Twinamatsiko and Kumar, 2022; Bătae et al., 2020; Alkaraan et al., 2022). For example, Occupational Safety and Health Administration (OSHA) monitors organizations in North America to ensure that they provide safe and healthful working conditions for workers by setting and enforcing standards. Similarly, Environmental Protection Agency (EPA) exercises efforts to reduce environmental risks. The agencies also announce rewards for positive actions. These events often shape public opinion about the involved organizations, and therefore provide important supplementary data for ESG performance assessment to end users. By choosing the correct set of monitoring agency websites, it can be ensured that the obtained information about an organization is correct and complete. Given the speed and volume of information flow, corporate risk assessment practices have been adopting natural language processing techniques for news analysis for quite some time now. End users use Google News to be informed about their objects of interest. There is a rising trend to build dedicated semantic search engines using named entities, events, and sentiments that are extracted from News automatically. Since events of interest can be widely varied, these systems are usually trained to track different kinds of events like supply-side risks, large-scale financial events (Murakami and Muraoka, 2022), terrorist attacks, market events (Bhadani et al., 2019; Mahajan et al., 2008), etc. The proposed work can be used to track all kinds of sustainability violations, which are not currently available. Since the model is trained to exploit linguistic structures of violations, it can track novel violations also very effectively.

Applications of language technologies for ESG assessment are gaining ground. Information extraction from various sources such as News articles (Sokolov et al., 2021; Sokolov et al., 2021; Gustavsson, 2022), scholarly articles (Perazzoli et al., 2022), company disclosure reports (Luccioni et al., 2020; Fischbach et al., 2022), business documents (Van Der Elst, 2022), and social media (Nugent et al., 2021; Caudron, 2022) have been reported. These works are primarily focused on the following aspects a) classifying text documents into either E, S, or G categories; b) identifying sentiment polarity of the documents; c) identifying the occurrences of a set of predefined types of ESG elements. However, none of the above works perform deep linguistic analysis of the textual contents to identify sustainability incidents based only on linguistic structure of text. The model presented in this paper is distinct from the earlier approaches since

¹ For the sake of simplicity, we consider the terms ESG event and incident as synonymous and therefore use both of them interchangeably.

² <https://www.crisil.com/>

³ <https://www.msci.com/>

it is capable of detecting novel and heretofore unknown incidents from reports based only on semantic and linguistic analysis of content.

Keeping in mind the above-mentioned requirements of end users, in this work we propose an event detection model that can identify sustainability-related insights and violations mentioned in regulatory reports. The proposed model uses a multi-tasking sequence labeling architecture that works with transformer-based document embeddings. Multi-task learning is a training paradigm in which machine learning models are trained with data from multiple tasks simultaneously, using shared representations to learn the common ideas between a collection of related tasks. These shared representations increase data efficiency and can potentially yield faster learning speed for related or downstream tasks. We have created a large annotated corpus containing relevant articles published by multiple regulatory agencies over three years (2015–2018) for training and evaluating the model. Care has been taken to ensure that all three domains are appropriately covered. The model has been thereafter applied to recent publications. Insights about incidents and violations obtained from these have been further categorized using indicators published by Global Reporting Initiative (GRI) standards (Weber et al., 2008; Polignano et al., 2022). Aggregate analysis of these extracted insights reveals interesting trends about sustainability practices.

The contributions of the paper are summarized as follows:

1. We present the concept of a screening system that can provide near-real-time insights about an organization's sustainability-related events including positive actions and violation incidents, if any, to supplement its third-party ESG scores, as required by end users. The system utilizes a deep neural architecture to process regulatory articles and detect incidents.
2. We present a multi-tasking neural architecture for detecting sustainability incidents, violation clauses, and other related parameters from regulatory articles. The proposed model exploits Sustainability-BERT (S-BERT), a BERT-based language model that was fine-tuned using sustainability News articles collected from the web. The model has been trained and evaluated using the above-mentioned corpus.
3. A training corpus of 900 regulatory articles under E, S, and G categories was manually annotated to mark risks or incidents, regulatory violations, and penalties, along with the target organization which was reported. These articles were published between 2015 and 2018. The annotation was done by multiple annotators using an incident knowledge schema, following a rigorous procedure to ensure acceptable inter-annotator agreement.
4. In this paper, we also introduce another corpus of 2,969 regulatory articles published between 2019 and 2022, which was processed using the above-mentioned model. The extracted incidents and violations are grouped using definitions proposed by GRI standards. We show how interesting insights can be derived about sustainability incidents observed over this period.

Rest of the paper is organized as follows: In [Section 2](#), we present a brief overview of sustainability indicators and reports and then move on to [Section 3](#) that provides details of a conceptual system for screening ESG activities. [Section 4](#) presents a knowledge-driven approach toward ESG information extraction. [Section 5](#) presents our proposed multi-tasking neural model for ESG knowledge extraction. [Section 6](#) presents a detailed evaluation of the proposed model with respect to the state of the arts. [Section 7](#) presents some interesting statistics in terms of quantifiable indicators used by the community. Related work done in the area of text mining for sustainability analytics is presented in [Section 8](#). Finally, [Section 9](#) concludes the paper.

2. A Brief Overview of ESG Reporting

GRI is a global organization⁴, which sets the standards for sustainability practices across the globe. It provides a number of indicators in each of the E, S, and G categories, for which information has to be

⁴<https://database.globalreporting.org/>

provided by the organizations (Chiarini, 2017; Ching et al., 2014) GRI standards define a range of indicators falling in different categories.

- Series 200 defines economic indicators that are used to evaluate an organization's commitment to good governance practices including handling business ethics, intolerance toward child labor, etc. The indicators in this category are numbered as 2xx—for example, performance(201), anti-corruption(205), etc.
- Series 300 defines environmental indicators that are used to assess a company's contribution toward the environment using parameters like its commitment to reduction in carbon emission, water recycling, conservation tasks, etc. All indicators in this category are numbered as 3xx—for example, energy(302), biodiversity(304), emissions(305), etc.
- Series 400 defines social indicators like training and education(404), security practices(410), etc. These indicators assess an organization's commitment toward its employees' physical and mental well-being, training opportunities provided, etc.

An organization publishes its ESG activities in one or both of the following reports:

- Sustainability Report:** It mainly contains the non-financial data describing the journey of progress of organizational performance targets against specific economic, environmental, social, and governance goals along with a few metrics to establish its journey toward sustainable development;
- Integrated Report:** It contains both financial and non-financial data linked by a narrative explaining data linkage to future corporate profitability along with the company's vision on sustainable development.

However, these self-published reports only provide grossly positive information about the companies and not much about violation incidents. Analysts, regulators, and investors also use a number of other sources to track organizational activities that may contradict the principles of sustainability. Reports published by regulatory agencies are the most important sources of current information. These sources may contain information about ESG violations along with events that caused or are likely to lead to violations, along with the probable future impact of non-sustainable actions. The language of these reports is different from the reporting language.

The primary contribution of this paper is a linguistic model that can identify such violation incidents or events that may lead to future violations from reports and thereafter link them to GRI indicators. This helps in deriving insights in a more uniform way.

3. Monitoring Regulatory Websites for Screening ESG Activities—A Conceptual System

Figure 1 presents a schematic diagram of an enhanced sustainability screening system that is being currently conceptualized. This system collects information from websites of regulatory agencies on a regular basis and analyzes them to generate individual and aggregate statistics about sustainability incidents and violations. The aim is to enhance the visibility of sustainability practices of a company through constant monitoring of reliable web sources.

The present work uses regulatory reports published by three agencies namely OSHA⁵, EPA⁶, and Department of Justice (DoJ)⁷. Initially, a corpus of reports was used to create an annotated dataset. A neural model is then trained on the annotated dataset, validated, and then applied to detect such violation incidents from current reports. The detected events are mapped to corresponding GRI indicators and subjected to aggregate analysis for insight generation. Coming from authentic sites only, these insights are trusted and verifiable.

⁵ <https://www.osha.gov/>

⁶ <https://www.epa.gov/>

⁷ <https://www.justice.org/>

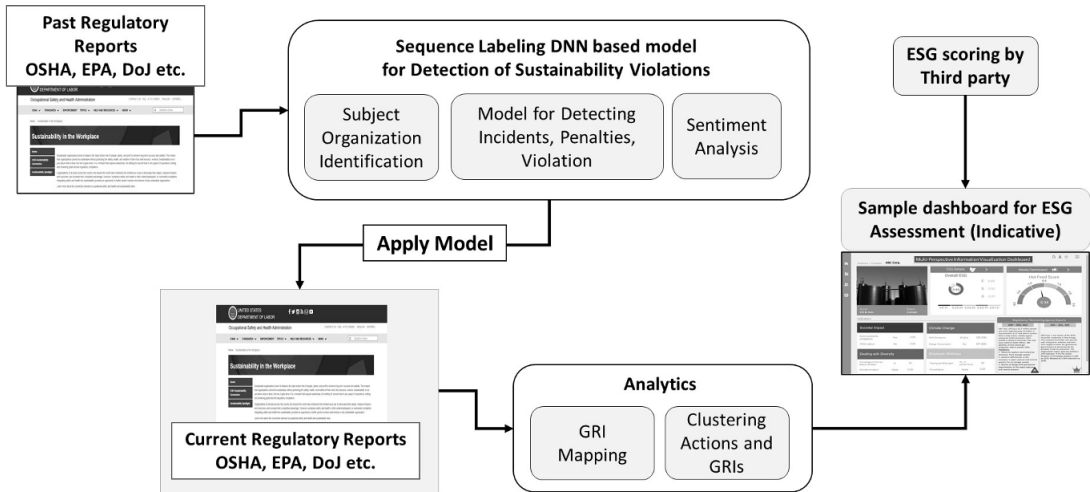


Figure 1. A schematic diagram of a unified sustainability information analytics.

For this work, we have crawled a total of 3,869 articles from the above-mentioned sites. While 900 among them have been used to train the model and evaluate it, the remaining articles have been used to derive insights about recent trends in the ESG domain. The crawlers continue to be active.

4. ESG Information Extraction—A Knowledge-driven Approach

A wide range of events starting from managing climate risk to building a diverse and inclusive workforce along with all aspects of governance can qualify as a potential ESG event. While each regulatory agency reports about events that primarily come under their purview, obviously a report can be quite mixed in nature. There is no accepted rigorous definition of what constitutes an event. For example, an incident like a factory accident that causes the death of employees thereby leading to a penalty, can be viewed as a collection of multiple events. Given the nature of the domain and the diversity of the reporting patterns, we propose a knowledge schema that elicits the different entities and their roles that are usually observed in a regulatory article and also captures the relationships among these entities. Figure 2 presents this knowledge schema. G represents the regulatory agency name while TO represents the target organization, which is the subject of the article. The other entities encountered are penalty (P), award mentions denoted by (A), location names denoted by L, and textual strings denoted by I and V, which represent incidents and violations. The relationships among a pair of entities may be expressed by different types of linguistic constructs, and hence established via multiple paths.

The cause–effect relationship between an incident and a violation is not very well defined. While in one case an accident may be reported as an incident caused by a violation of safety norms within a factory, in another instance an action like accepting a bribe may be reported as an incident that led to a violation of governance regulation. These reports often also contain information about the violated clause, the regulations or acts that were violated, the indicators that are affected, the penalty value imposed by a regulatory agency, extent of loss incurred due to the event, number of deaths or injuries, penalty paid, etc. While some of the entities like TO, L, or P can be detected using standard Named Entity Recognition (NER) tools, detecting incidents and violations is a non-trivial task. The proposed neural model is trained to do this. The above-mentioned knowledge graph is used to create an annotated dataset for the purpose. The detailed process of annotation is mentioned in the next subsections.

4.1. The Data Annotation Process

A training corpus of 900 regulatory articles comprising of around 10,800 sentences, under E, S, and G categories published between 2015 and 2018 by the U.S. Environmental Protection Agency (EPA),

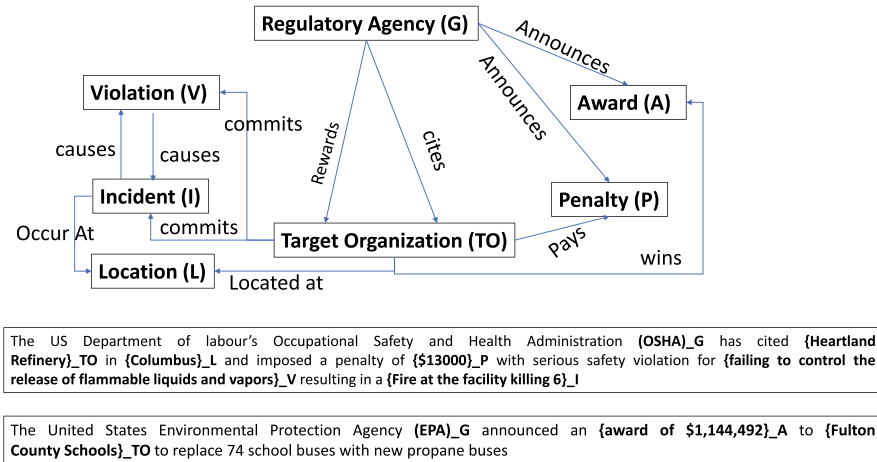


Figure 2. ESG incident knowledge schema illustrated along with example sentences from text containing relevant entities. Different events and entities in text are marked in bold.

Occupational Safety and Health Administration (OSHA)⁸, Equal Employment Opportunity Commission (EEOC)⁹, and the U.S. Department of Justice (DoJ)¹⁰ respectively were manually annotated to mark the various entities mentioned in the knowledge graph. The annotation was done by multiple annotators following a rigorous procedure to ensure acceptable inter-annotator agreement. The semantics of the nodes and relations mentioned in the knowledge graph were used to guide the annotators while identifying words, phrases, or portions of sentences that may represent an instance of a node in a document.

To help the annotators, each document is first processed using the Stanford NER (Manning et al., 2014) to obtain the organization names, locations, and currency values as named entities. This helps in quick localization of the first four elements, if present in the document. The annotators are domain experts who are knowledgeable about sustainability domain, familiar with incidents and violations.

The rules followed by the annotators are as follows:

- Regulatory Agency (G):** This was programmatically done by using a list of regulatory agency names and their abbreviations, which was searched from among the NERs detected.
- Target Organization (TO):** Of the many organization names that may appear in a document and are detected by the NER earlier, the task during annotation is to identify and tag the violating or the award-winning organization as TO.
- Location (L):** Among possibly many locations appearing in a report, the task here is to mark the correct geographical location of the incident, in the absence of which, the location of the organization could be marked, if it appears in the document.
- Proposed Penalty (P):** The currency value that denotes the penalty that has been enforced upon the target organization by a governing body is marked.
- Incident (I):** Annotators have to tag phrases or sequences of words that collectively are indicative of an incident or an event.
- Guideline violations (V):** These are phrases or sets of words that collectively indicate non-compliance or failure to comply with guidelines or regulations.
- Award (A):** The award name and associated money value or citation phrases, are manually tagged during the annotation process.

⁸ <https://www.osha.gov/news/newsreleases/enforcement/>

⁹ <https://www.eeoc.gov/newsroom/search>

¹⁰ <https://www.justice.gov/news>

Table 1. Sample sustainability News texts with the respective annotated entities and events. Note that all the target organization names were intentionally masked by the token [ORGName] to maintain anonymity

BLOOMER, Wis. Underground construction contractor [ORGName] Inc./TO headquartered in Bloomer/L, has agreed to pay \$474,000/P in penalty as part of a settlement agreement with the OSHA addressing hazards/V cited during three inspections. After providing false documentation and making false representations/V claiming that previously cited hazards related to hydraulic/V presses had been corrected, [ORGName] LLC/TO has been issued 14 citations, with proposed fines totaling \$816,5003/P. Four violations cited also involve failing to train workers/V on how to properly stop machines before service and maintenance, which continuously exposed machine operators to laceration, amputation, burns, and having parts of the machine strike or crush them/V. Occupational Safety and Health Administration has cited [ORGName]/TO in Columbus/L with one alleged serious safety violation for failing to control the release of flammable liquids and vapors/V resulting in a July 17, 2010, fire at the facility. The U.S. Environmental Protection Agency (EPA) announced that Louisiana Department of [ORGName]/TO will receive \$5 million/A in Brownfield grant funding at Louisiana/L.

Six annotators took part in the annotation, with each expert annotating 400 documents using the Stanford simple manual annotation tool¹¹. This included 100 documents, which were sent to all the annotators to compute the inter-annotator agreement later. The average length of a document is around 23 sentences. The experts read each document and performed the following tasks:

1. **Task-1:** Label sentences of each document as **POSITIVE**—if the document reported events of winning awards or recognition for sustainability practices or policy announcements in favor of sustainability or **NEGATIVE**—if the document mentions events that violate ESG regulations or incidents reporting damages, loss of life, etc. or **NEUTRAL**—in case none of the above factors hold.
2. **Task-2:** This task had two components:
 - (a) From among the named entities, the target organization, penalty values, and locations were marked, if any, and
 - (b) Mark phrases in the text that indicate announcements, incidents, and violations.

At the end of the annotation, each word in the document is assigned a label TO, I, V, P, L, A, or None. Table 1 illustrates the annotation with some example News texts. For the sake of understanding, we have shown labels of only the phrases that belong to any one of the following classes {TO, I, V, P, L, A}. Table 4 presents the detailed overview of the dataset used to train and test our proposed models.

4.1.1. Computing annotator agreement

Using the annotations obtained for 100 common documents, we measured the inter-annotator agreement using the Fleiss Kappa (Fleiss et al., 1981) measure (κ). This is computed as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

The factor $1 - \bar{P}_e$ gives the degree of agreement that is attainable above chance, and $\bar{P} - \bar{P}_e$ gives the degree of agreement actually achieved above chance. It was observed that the inter-annotator score for

¹¹ <https://nlp.stanford.edu/software/>

Task-1 was 0.83, which is appreciably high. For Task-2, it was found to be 0.71. The scores are computed using word-label matches assigned by different annotators. The very high scores indicate that all experts were marking fairly uniformly and therefore, the expert annotated dataset is reliable to be used for training incident detection systems.

Out of the 20,700 sentences from 900 documents, around 7,500 sentences were found to contain words belonging to at least one type mentioned in the incident knowledge schema. Altogether, we obtained 3,000 violation phrases, 2,100 penalty phrases, 2,223 Target Organizations, 1,300 locations, 695 incidents, and 101 awards.

5. A Multi-tasking Neural Model for ESG Knowledge Extraction

In this section, we present the details of incident labeling architecture implemented using multi-task neural model. This model works on each sentence at a time to detect elements of interest that are defined in the incident knowledge schema. Multi-task learning utilizes the correlation between related tasks to improve classification by learning tasks in parallel. In the present work, the two related tasks are *task-1*: classifying a sentence into either *positive* or *negative* classes as discussed earlier, and *task-2*: labeling appropriate phrases in the text as per the incident knowledge schema. **Accordingly, for task-1 each sentence of the training data is labeled as either Positive (Award), Negative(Penalty), or Neutral whereas for task-2 each word/phrase of a sentence is labeled according to the following classes <TO,I,V,P,L,A>.**

Figure 3 presents a high-level view of the proposed model. It uses a cascaded CNN-BiLSTM layer for the combined tasks of sentence classification and sequence label prediction, using the fine-tuned S-BERT for creating the sequence embeddings. Thus, the task-(a) sequence classification is predicted as:

$$y^i = \text{softmax}(W_i * h_1 + b_i). \tag{1}$$

On the other hand, for task-(b) or the sequence labeling task, we feed the final hidden states of the *BERT – CNN – LSTM* network of the other tokens, h_2, \dots, h_T , into a softmax layer to classify over the sequence labels. The output of the model is represented as:

$$y_i^s = \text{softmax}(W^s * h_i + b_s), i \in (1 \dots N) \tag{2}$$

where h_i is the hidden state corresponding to the word w_i .

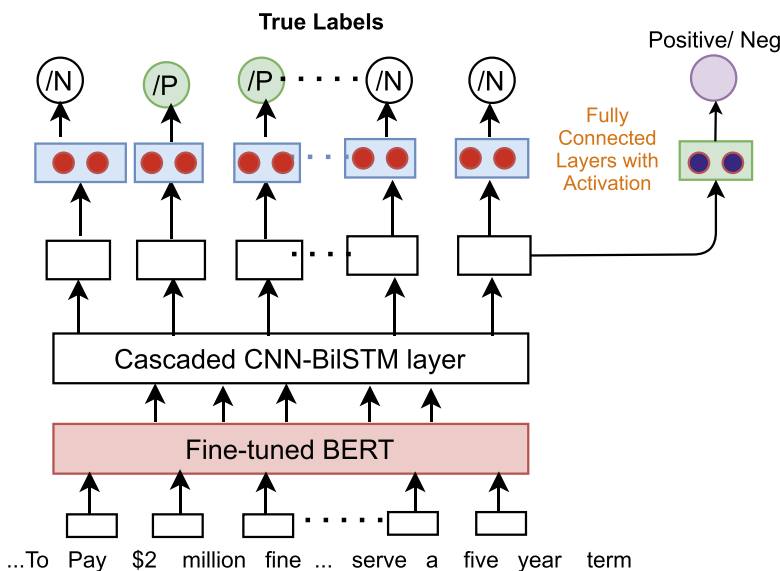


Figure 3. BERT-based multi-tasking network for ESG sequence classification and labeling.

To obtain the multi-tasking model for dual tasks of sequence classification and sequence labeling, the *BERT – CNN – BiLSTM* layers have been trained with two separate loss functions L_1 and L_2 . Where,

$$L_1(\theta) = - \sum_{t=1}^M \sum_{k=1}^K \bar{y}_t^k \log(y_t) \text{ and } L_2(\theta) = - \sum_{t=1}^N \sum_{j=1}^J \bar{q}_t^{i,j} \log(q_t^i)$$

q_t is the vector representation of the predicted output of the model for the input word w_t^i . K and J are the number of class labels for each task. The model is fine-tuned end-to-end via minimizing the cross-entropy loss.

We define the joint loss function using a linear combination of the loss functions of the two tasks as:

$$L_{joint}(\theta) = \lambda * L_1(\theta) + (1 - \lambda) * I_{[y_{sentence} == 1]} * L_2(\theta) \tag{3}$$

Where, λ is a hyper-parameter whose value is learned to control the contribution of losses of the individual tasks in the overall joint loss. $I_{[y_{sentence} == 1]}$ is an indicator function which activates the loss only when the corresponding sentence classification label is 1, since we do not want to back-propagate sequence labeling loss when the corresponding sequence classification label is 0.

5.1. Fine-tuning the BERT language model

We use the *BERT_{base}* model that has been built over the BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia. The *BERT_{base}* was then fine-tuned with a portion of the corpus using over-sampling, to create S-BERT. The fine-tuning of the *BERT_{base}* over the sustainability corpus enforces the BERT model to be both generalized as well as focused over the domain data. A labeled document is broken into multiple smaller chunks, such that each chunk can be fed as a unit to *BERT_{base}* to create its corresponding vector. Each chunk is associated with a label that is the same as its parent document. A classification task is now defined with these chunks during which the basic BERT model is fine-tuned while training. This model is designed as a fully connected layer over the BERT base model, with softmax as the activation function. Training was done with learning rate set to 2×10^{-5} using the Adam optimizer (Kingma and Ba, 2014). The model is fine-tuned for few epochs (3-4) only to avoid over-

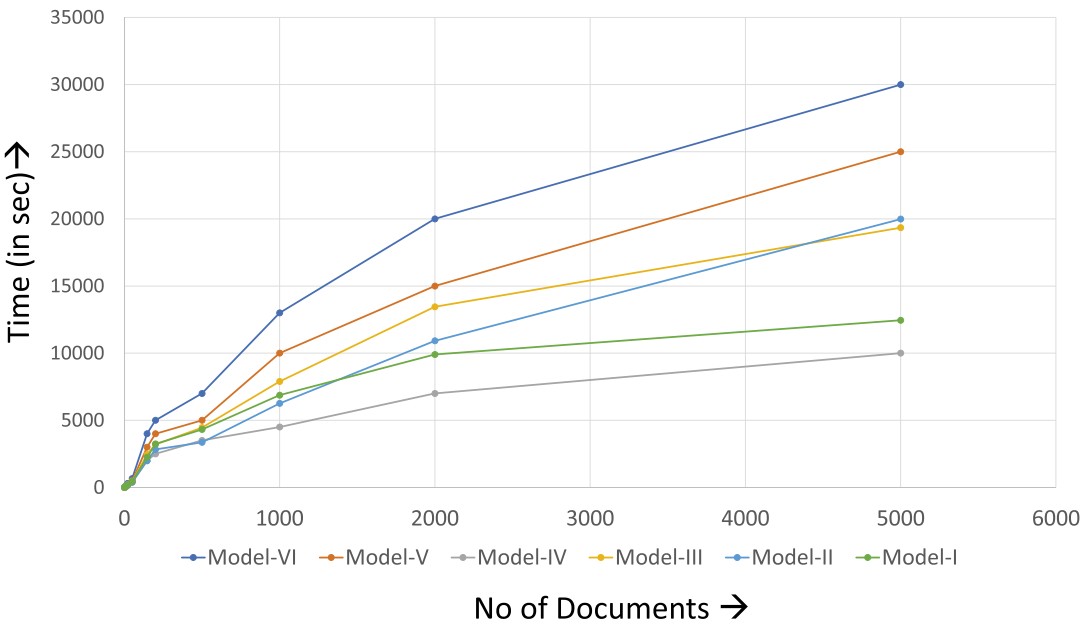


Figure 4. A comprehensive running time comparison between different models.

fitting. The chunk representations are saved from the [CLS] token embeddings created during the process. The fine-tuned BERT model, S-BERT, is subsequently used for document and incident recognition tasks. Figure 4 depicts a comprehensive runtime analysis of the proposed model along with all the baseline systems. The runtime evaluation is performed over a training document of 8,917 documents comprising of 222,925 sentences over an epoch of 10. All the above testing results have been obtained from the Tesla T4 GPU, with NVIDIA-SMI 460.32.03, Driver Version: 460.32.03, and CUDA Version: 11.2 hardware configurations.

6. Evaluation and Results

6.1. Evaluating the multi-task network for ESG knowledge extraction

Next, we present results for the multi-objective sequence labeling model for the ESG knowledge extraction task. Table 2 presents the precision, recall, and F1 scores of classifying sustainability events as awards or violations. We have obtained a highest F1 score of 0.91 with a high precision of 0.93. However, we observed that the single task BERT-CNN-BiLSTM performs better than the proposed Multi-Class network in terms of recall. A possible hypothesis behind this anomaly is the fact that given the size of the training data, incorporating the CNN-BiLSTM layer is causing the models to overfit and thus affecting the recall value. This hypothesis is further supported by the fact that the recalls are consistent throughout all the models having the CNN-BiLSTM layer except the case where the layers are removed in the single-task model.

Table 3 presents the accuracy of subsequent labeling of word sequences within a sentence by their respective categories—*TO*, *L*, *I*, *V*, *P*, *A*—as described in Section 4. For both the cases, the performance of the proposed multi-objective architecture has been compared with a number of baseline state-of-the-art models designed with single objective functions. It was observed that for almost all the categories, the Multi-task BERT-CNN-BiLSTM model significantly outperforms the baseline models. For example, in the *Target Organization* class, it was found that the Multi-task BERT-CNN-BiLSTM model significantly reduces the false negative score and achieves a high true positive score thereby achieving a high precision and recall. In general, an F-measure of 0.89 with a precision of 0.87 and recall of 0.92 was achieved. For the class *Incident*, F-measure of 0.82 with a precision of 0.79 and recall of 0.85 was achieved. A high F1 score of 0.90 was achieved for *Penalty*. The target organizations were detected correctly 89% of the times. In the remaining cases, either wrong organizations were detected or missed out altogether. Detailed analysis reveals that for violation and incident phrases majority of the sub-sequences are detected correctly. The errors occur due to a portion of the sequence not detected correctly.

Apart from evaluating the classifiers, we also evaluated the performance of the violation to GRI indices mapping module. We have randomly chosen a sample dataset comprising 100 data points for analysis. Two expert evaluators were engaged to perform manual verification of the mapping between GRIs and corresponding violation phrases within the dataset. The validation process yielded a commendable

Table 2. Results reporting accuracy of classifying sustainability events as awards or violations

Model No.	Model Name	Sequence Classification		
		P	R	F1
I.	Single task CNN-BiLSTM	0.83	0.89	0.86
II.	Single task PreTrained BERT	0.85	0.92	0.88
III.	Single-task-BERT-CNN-BiLSTM	0.85	0.89	0.88
IV.	Multi-task-CNN-BiLSTM	0.83	0.87	0.85
V.	Multi-task-BERT	0.84	0.90	0.89
VI.	Multi-task BERT-CNN-BiLSTM	0.93	0.90	0.91

Table 3. Results reporting the accuracy of sustainability event and entity extraction task

Model		Sequence Labeling																	
No.	TO			I			V			L			P			A			
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	
I.	0.76	0.78	0.77	0.71	0.67	0.69	0.77	0.78	0.77	0.72	0.77	0.74	0.77	0.8	0.78	0.69	0.73	0.71	
II.	0.79	0.82	0.80	0.69	0.74	0.71	0.79	0.77	0.78	0.82	0.85	0.83	0.8	0.85	0.82	0.73	0.82	0.77	
III.	0.80	0.87	0.83	0.71	0.75	0.73	0.76	0.86	0.8	0.79	0.89	0.83	0.83	0.91	0.87	0.80	0.87	0.83	
IV	0.80	0.79	0.86	0.81	0.86	0.86	0.79	0.85	0.81	0.88	0.86	0.80	0.80	0.84	0.87	0.81	0.84	0.82	
V	0.88	0.80	0.79	0.84	0.81	0.89	0.81	0.87	0.83	0.88	0.83	0.89	0.81	0.84	0.89	0.80	0.89	0.85	
VI.	0.81	0.89	0.85	0.79	0.85	0.82	0.82	0.92	0.87	0.86	0.92	0.89	0.88	0.93	0.90	0.86	0.89	0.87	

accuracy rate of 92.6%. However, in certain instances, the mapping process encountered challenges, leading to instances where accurate mappings were not achieved. For instance, a violation phrase “failing to provide proper lead-based paint disclosure to buyers and renters of homes built” was maximally mapped with GRI “Child labor (408)”. Similarly, another violation, involving “establish programmatic capability to better compete for federal brownfield cleanup grants” demonstrated a maximal match with GRI “Human Rights Assessment (412)”. This discrepancy in mappings may be attributed to the absence of specific GRIs in the provided list that ideally should have corresponded to these particular violations. This underscores the potential need for the inclusion of additional GRIs to comprehensively capture the diverse spectrum of violations encountered in the dataset.

6.2. Analysis of performance

Extracting ESG knowledge components from documents is a challenging problem due to the complexity and variability of language that also includes many legal terms and constructs. With a few examples, we highlight below the aspects in which our model is doing good, as well as those aspects that need further improvement.

(a) Language Diversity is handled quite effectively: Linguistic constructs used for reporting incidents and violations in texts published across different domains and data sources vary quite a lot. The following example sentences highlight the differences exhibited across three different sources while reporting violations and penalties. The ordering of concepts is quite different in the three.

1. *Worker’s Permanent Injury Results in \$75,000 Fine for London-Area Company.* Source: osha.gov
2. *The U.S. Environmental Protection Agency (EPA) has announced settlements of \$18,000 with [ORGName], Inc.* Source: epa.org
3. *The [ORGName], has agreed to forfeit \$500 million to the United States in connection with a conspiracy.* Source: justice.org

The model successfully identifies the penalty amounts correctly from all the constructs with correct resolution of associated words like “*results in \$75,000 fine*”, “*announced settlements of \$18,000*”, and “*Agreed to forfeit \$500 million*”.

(b) Violations are learned as a time and source invariant phenomenon: Though the nature of violations change over time and sources as new regulations and compliance factors are introduced, it was observed that the proposed models are capable of recognizing new types of violations without additional training.

(c). Target organizations are not always identified correctly: In around 7% of the cases, we have observed that incorrect target organization detection can be attributed to the structural complexity of the underlying text. For example, in the statement “[ORGName], which was contracted by [ORGName1] to provide steam fitting services at the site, was cited for six serious violations, with 31,500 USD in penalties.”, the system fails to identify the *target organization* “[ORGName]”. This is possibly due to the long dependency distance between the *violation* event and the *penalty* entities, which the algorithm is unable to make use of.

7. Deriving Insights from Regulatory Reports

We now present some interesting insights about commonly occurring ESG incidents and violations obtained from the articles published between 2019 and 2022, using the model presented earlier.

Mapping incidents and violations to GRI indicators: The extracted ESG incidents and violations derived from the regulatory articles are mapped to GRI indicators introduced in Section 2. In order to map incidents and violation phrases to GRI indicators, we have utilized the GRI indicator definition to obtain better matches. The violation phrases are compared with GRI indicator definition using Universal Sentence Encoder (USE) (Cer et al., 2018) embeddings. Since each sentence may contain multiple

Table 4. Data Statistics

Domain	Source	URL	Sustainability	
			Corpus Size	Timeline
Environment	U.S. Env. Protection Agency,	epa.gov	785	2019–2022
Social	Occupational safety and health administrator (OSHA), EEOC, USA	osha.org, eeoc.org	1133	2020–2022
Governance	Securities and Exchange Commission, U.S. Dept. of Justice,	sec.gov, justice.gov	1051	2019–2021

incidents or violations, only violation phrases are used to compute similarity and not the whole sentence. Also, GRI indicator definition provides comprehensive explanation of the indicator in text form. We have used the entire sentence to compare against the phrases. Indicator with the highest cosine similarity between USE embeddings and above a minimum similarity threshold is mapped to a violation phrase under consideration.

Table 5 presents a few sample sentences with the extracted incidents and violations marked in bold, along with their GRI mappings. The samples illustrate that the proposed model is powerful enough to mine a wide range of discriminatory incidents related to gender, age, color, etc.

Once the incidents and violations are mapped to a fixed set of GRI indicators, these can be used for further analytics at aggregate and individual levels. A single report usually maps to multiple GRI indicators, since usually a single incident, say for example, an accident at a mining site leading to loss of life is linked to multiple lapses like unsafe workplace, inadequate training, environmental hazards, etc. This is illustrated with the following example:

Sample sentence: “*The alleged violations included failure to manage and contain hazardous wastes; failure to comply with air emission limits; failure to comply with chemical accident prevention safety requirements; and failure to timely report use of certain toxic chemicals.*”

GRI indicators mapped to: i) Effluents and waste (306) ii) Emissions (305) iii) Occupational health and safety (403).

Figure 5 shows trends of violations observed for GRI indicators obtained from reports published between the years 2020 and 2022. Count of each indicator over a time period is calculated from the number of documents that contain it. A steep rise can be seen in 2022 for violations in the categories of corruption (205), biodiversity (304), discrimination in workplace (406), diversity and equal opportunity (405), public policy (415), and customer healthcare (416). Figure 5 also shows rising trends for environmental indicators like biodiversity (304), water resources (303), etc. Violations related to economy like indirect economic impacts (308) and socio-economic conditions (414) are observed only in the year 2022. It is observed that “lack of adequate training” consistently co-occurs with a large number of incidents across various categories. Clearly that is an area of improvement for all organizations.

Figure 6 shows the distribution of the sectors across which discrimination (406) was observed, most of which were in Healthcare. Manual examination of these discriminatory violations revealed that these could be further grouped into a few categories. We used a key-phrase-based approach to group them under the following categories: *sexual, racial, religious, pregnancy, retaliation, national origin, age, and disability*. These key-phrases are primarily motivated from the EEOC Prohibited Employment Policies/Practices guidelines¹². Figure 7 shows the distribution of incidents reported in these areas from

¹² <https://www.eeoc.gov/prohibited-employment-policiespractices>

Table 5. Sample violation events picked up from News articles of different categories and are mapped across GRI Standards. Note that the target organization names were masked by the token [ORGName] to maintain anonymity.

GRI Standards with name	Sample violation phrases picked up by Sequence Labeling Algorithm mapped to GRI's
406_Non-discrimination	<ul style="list-style-type: none"> • ORGName, a leading pet waste removal company, will pay \$40,000 and provide significant equitable relief to resolve a federal pregnancy and disability discrimination lawsuit.—2020 • ORGName tolerated a work environment hostile to female and African-American employees in [ORGName]'s Denver and Nashville offices. EEOC alleged that African-American employees were referred to as “lazy” had stress balls thrown at them, and were subjected to racially demeaning cartoons.—2019
305_Emission	<ul style="list-style-type: none"> • The settlement addresses [ORGName]'s failure to capture and control air emissions from storage vessels and to comply with associated inspection, record keeping, and reporting requirements.—2021 • The alleged violations included failure to manage and contain hazardous wastes; failure to comply with air emission limits; failure to comply with chemical accident prevention safety requirements; and failure to timely report use of certain toxic chemicals.—2021
306_Effluents and Waste	<ul style="list-style-type: none"> • The case stems from several transformer spills at locations in Massachusetts and Connecticut, involving improper manifesting of PCB remediation waste, improper storage of a PCB transformer, and improper disposal of PCBs.—2019, • Those violations included discharges of pollutants primarily chlorides and sodium in excess of its permit, failure to properly monitor and maintain records, and failure to adequately operate and maintain its wastewater treatment system.—2020
404_Training and Education	<ul style="list-style-type: none"> • The serious violations, with \$ 9,680 in penalties, were cited for failing to provide workers with eye and face protection; to ensure that the manufacturer did not exceed its safe operating pressure for hoses, pipes, and valves; and not providing a training program for workers exposed to fall hazards.—2019 • The company has also failed to initiate and maintain an accident prevention program, a hazard cited by OSHA.—2019
415_Public Policy	<ul style="list-style-type: none"> • The SEC's order finds that [ORGName] overcharged more than 149,000 advisory clients because it failed to adopt and implement compliance policies and procedures reasonably designed to ensure that clients were billed accurately according to the terms of their advisory agreements.—2021 • On September 27, the U.S. Attorney's Office for the Eastern District of Pennsylvania obtained a grand jury indictment of [ORGName] charging him with sixteen counts of mail fraud, eighteen counts of wire fraud, one count of bank fraud, one count of making false statements to the government, three counts of filing false tax returns, one count of impeding the administration of the revenue laws, fifteen counts of money laundering, and twenty-seven counts of money laundering to promote an unlawful activity.—2021

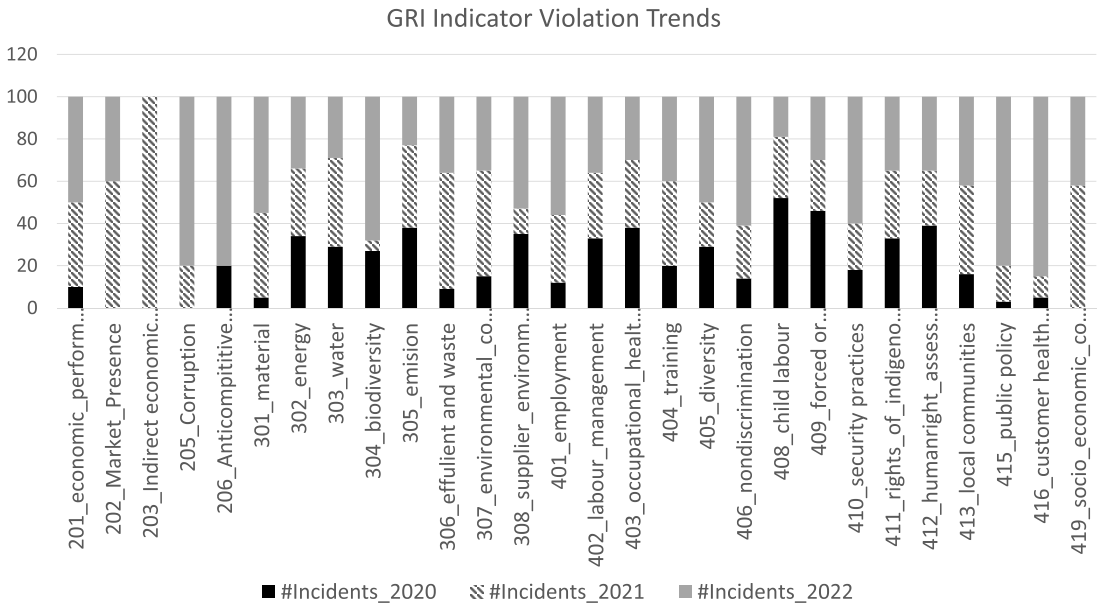


Figure 5. Violation trends by volume for 28 GRI indicators (2020–2022).

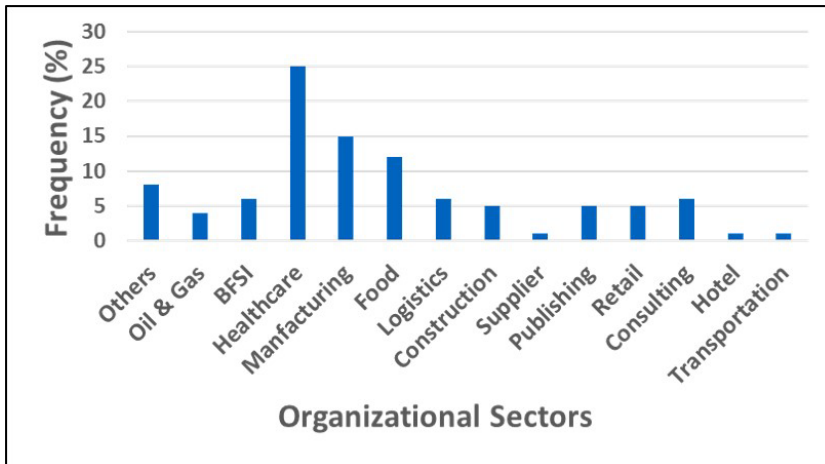


Figure 6. Distribution of Incidents of discrimination (GRI 406) across organizational sectors for the Year 2020.

2020 to 2022. We observe that the highest number of discriminatory violations reported were related to incidents involving *disability, gender, and retaliations*.

Interestingly, Figure 5 shows no significant change corresponding to occupational health and safety (403) incidents observed over the last three years. We believe that this can be attributed to the fact that occupational health and safety has been actively monitored globally for the last five years, which has helped in the reduction of violations in this area. Awareness about environment, economic impact, and some other social factors are fairly recent phenomena. These results clearly establish that sustainability monitoring and regulatory reporting definitely help in bringing down the number of incidents. Aggregate analysis like ours, can also help other organizations to become cautious. In the next section, we show how these insights can be clustered to obtain intelligence at a higher level of granularity.

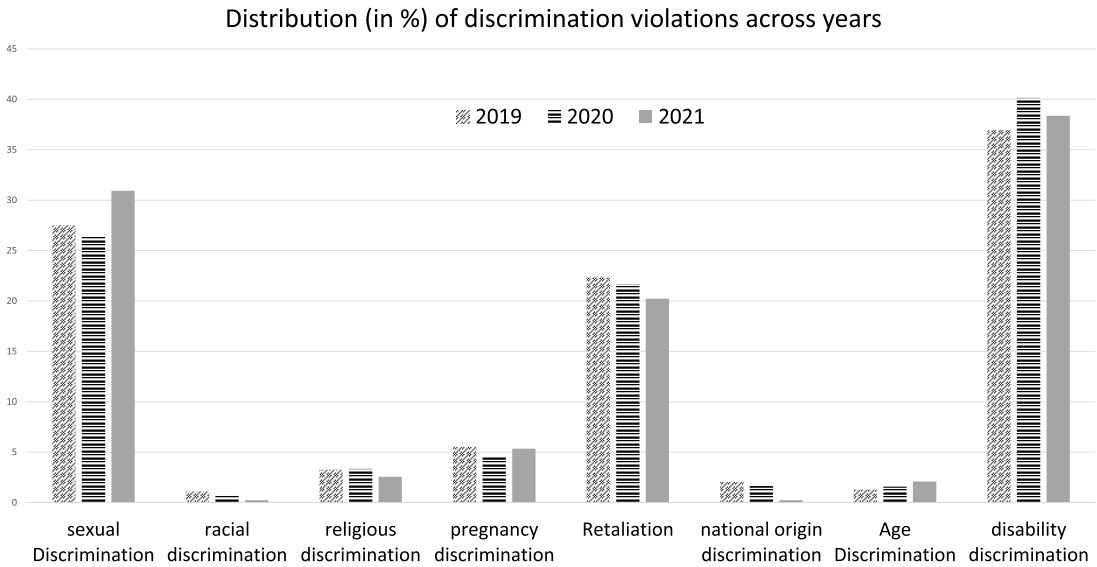


Figure 7. Distribution (in %) of discrimination violations across years.

7.1. Clustering incident and violation phrases—Obtaining another view

As is obvious with text reports, there are many incidents which are not exactly identical but similar or related. While some of these incidents could be mapped easily to GRI indicators, the mapping was not possible for some violation and incident phrases. This could be either due to a conceptual gap between the definition of the indicator and the incident or due to the fact that certain phrases were not covered by any GRI. Hence, to derive additional insights, the incidents and violations extracted from each set of articles belonging to the E, S, and G categories were clustered separately. For each category, embeddings for the sequences labeled as violation and incident by the model were created using the USE (Cer et al., 2018). These vectors were then clustered using the K-means clustering algorithm (Aggarwal and Zhai, 2012), with cosine similarity as the underlying distance measure. We have used the popular Elbow method (Joshi and Nalwade, 2013) for selecting the optimal number of clusters, k , for each set.

Figure 8 depicts the different clusters obtained from environmental violations mined from the corpus, along with their percentage occurrences. The clusters reveal that many organizations are operating without the necessary certifications required to ensure clean air and clean water at their premises. Results indicated that since certification requirements are region-specific, the violations differed a lot in their descriptions. These could not be mapped to the relevant GRI indicator, which in most cases is defined by very high-level specifications only. Figure 9 depicts percentages for social violations encountered in the corpus. It can be again observed that “lack of training” co-occurs with most of the incidents, indicating that this is the most frequent violation leading to unpleasant workplace incidents. It should be noted that the violations are not mutually exclusive.

8. Related Work

The rising awareness about sustainability practices among consumers, investors, and regulators is forcing organizations to pay attention toward ESG factors, as these are known to affect financial outcomes (Tarmuji et al., 2016). A number of researchers around the world have reported results on how ESG factors are likely to affect financial performance of organizations. One of the earliest works in the area, reported in (Shahi et al., 2014), presented an automated solution for scoring corporate sustainability data reported. A variety of classifiers including decision trees and neural networks were used to predict the sustainability score. Other notable works reported in the area of impact assessment are

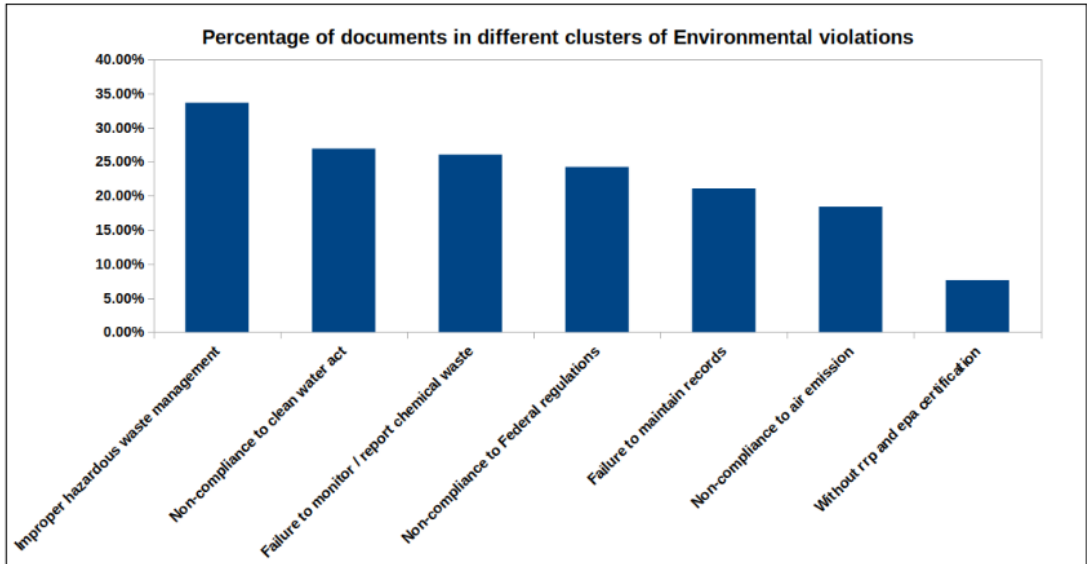


Figure 8. Distribution of violation clusters in environmental sector.

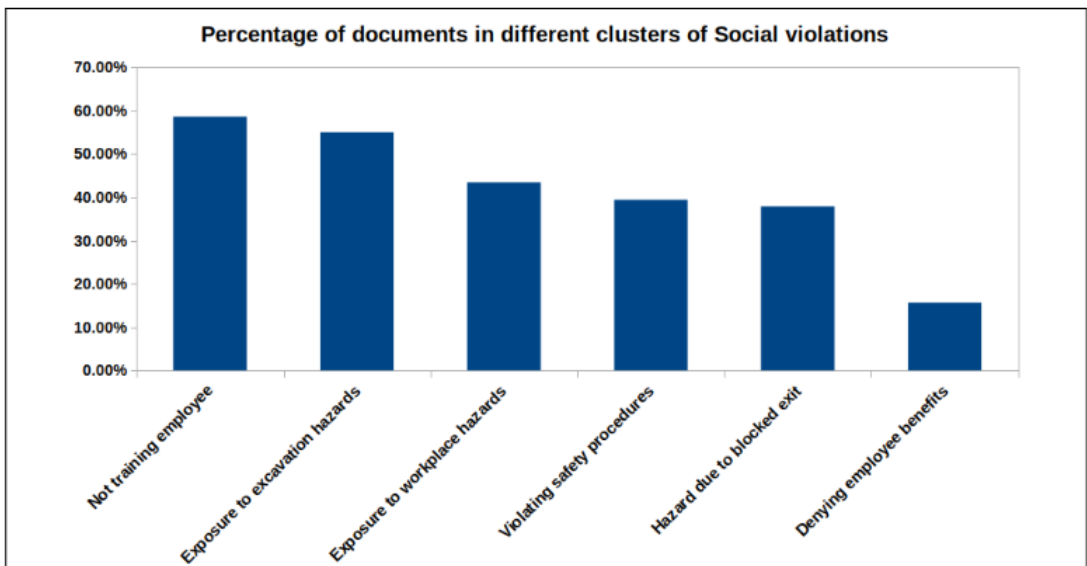


Figure 9. Distribution of violation clusters in occupational health and safety sector.

(Zhao et al., 2018; Yoon et al., 2018; KPMG, 2017; Velte, 2017; Sultana et al., 2018). We are not going into the depths of the analysis procedure, since our focus is on gathering relevant information about sustainability incidents and not on the analysis process itself.

Being a fairly new area, there is very little work in the area of sustainability information extraction from text documents. In a more recent work that is closer to the proposed work, Raman et. al (Raman et al., 2020) proposed a transfer learning framework for profiling ESG-related language in business corpora and detect historical trends in ESG discussions by analyzing the transcripts of corporate earning calls using contextual embeddings. The authors developed a classification model that categorizes the relevance of a text sentence to ESG using fine-tuned language models built on small corporate sustainability reports

dataset and evaluated their results on the Earnings Call Transcript dataset. Their analysis indicates that in the last 5 years, nearly 15% of the discussions during earning calls pertained to ESG, implying that ESG factors are integral to business strategy. A similar effort was reported in (Goel et al., 2020), in which a methodology to automatically detect, retrieve, and score relevant information about ESG indicators from a company's sustainability reports was presented. The relevant information was mined from the textual content using conceptual similarity of the sentences to the indicators, as well as quantitative information present in them. The above-mentioned works focused on information extraction about reported parameters with the explicit aim of reducing the manual effort required to read and extract information from reports. In (Lee and Kim, 2023), authors proposed a 4-class BERT-based classifier that can discriminate ESG information. The training data was manually constructed. Authors also conducted three application experiments to verify the usability of their proposed model. First, in the cross-sectoral experiment, the results confirmed that the ESG classifier showed a significant performance accuracy for sectors not included in the training data. Second, in multi-source adaptation experiment involving extracting ESG-related information from multi-source text data, the performance was qualitatively verified. Finally, in the data augmentation experiment, the performance was verified on an ESG dataset that was additionally constructed through the pseudo-labeling technique.

Efforts toward analyzing publicly available data for detecting ESG factors were reported in (Nugent et al., 2020), where the authors have introduced a domain-specific BERT language model which has been further pre-trained using a corpus of financial and business News articles. They have advocated that fine-tuning their proposed model for multi-class ESG controversy classification task results in an improvement in performance compared to the general domain BERT model. They have also proposed the use of back translation approaches to deal with limited dataset problem. The classifier works at the document level and does not extract the actual incidents. Moreover, they have also limited their scope to Reuters News only which has specific style guidelines and sacrificing stylistic diversity by using only a single source whereas we have collected the data from a variety of sources. A very interesting study by Guo, (Guo, 2020), reports the use of information extracted from ESG News to predict future volatility of stock returns. The model was trained using pairs of numerical News representation and the volatility of the corresponding companies to predict the future volatility of the companies mentioned in newly arrived News. They found that investors considered ESG News as a relevant factor which significantly impacts the future return and risk of companies. ESG News is segregated from general financial News using a hand-crafted ESG vocabulary created by domain experts. No language models were used for the purpose, thus limiting the capability of the model in detecting new and unforeseen events. In (Pasch and Ehnes, 2022), authors showed a novel solution to fine-tune transformer-based models for the ESG domain. An ESG sentiment model was created by combining ESG ratings with text documents from annual reports. It has been reported that the model outperformed traditional text classifiers in predicting the ESG behavior of companies by up to 11 percentage points. The work also demonstrated practical applications of the ESG sentiment models by predicting individual sentences and by tracking ESG-related news coverage over time.

In (Murakami and Muraoka, 2022), the authors have proposed modeling sustainability report scoring sequences using an attractor network. The authors have used the Attractor Neural Network (ANN) model to predict the sustainability reporting scores within the GRI for a set of global companies while literature (Pasch and Ehnes, 2022) presents sustainability reporting based on GRI standards within an organization in Romania. Here, the authors analyzed the content of standardizing sustainability reporting in Romania and the content of sustainability reports. They studied the occurrence of GRI indexes specified in the company's sustainability reports. On the other hand, we are mapping violation incidents or phrases with GRI indicators to derive additional insights.

Recently, there is a growing interest in applying multi-task networks to representation learning (Liu et al., 2019; Collobert et al., 2011; Liu et al., 2015; Luong et al., 2015; Xu et al., 2018; Wan et al., 2021). This is primarily due to the fact that multi-task learning frameworks provide an effective way of leveraging supervised data from many related tasks (Liu et al., 2019). We argue that multi-task learning and pre-trained language models are complementary techniques that can be combined to improve the learning of text representations to boost the performance of various language processing tasks. Although,

a lot of recent attempts have been made in applying deep multi-task learning frameworks for information extraction (Xu et al., 2018; Wan et al., 2021), however, to the best of our knowledge no such model has been employed to extract sustainability insights from the incoming stream of articles. The proposed framework can help in gathering large-scale insights about sustainability practices in general, as well as provide real-time knowledge for sustainability scoring.

9. Conclusion

In this paper, we have proposed computational models for extraction and curation of sustainability related incidents and violations from digitally published regulatory reports. We have written crawlers to gather a large number of such articles from multiple reliable sources. A portion of this corpus has been manually annotated to train and evaluate a deep neural network architecture for automated extraction and curation of sustainability incidents and violations. Knowledge about sustainability events, violations, awards, and penalties were used for the annotation task. The model is multi-tasking in nature. It simultaneously classifies a sentence as positive, negative, or neutral and also labels portions of the sentence as incidents, violations, or awards. The proposed multi-task network has been extensively evaluated with respect to some of the state-of-the-art baseline models. We observed that for almost all the defined tasks the proposed model surpasses the baseline models. The incident and violation phrases are thereafter mapped to GRI indices, from which insights can be generated at various granularities.

This work has focused on using language technologies for analyzing web content to extract sustainability-related information. Since the web contains updates about all sustainability-related incidents reported by regulatory agencies, this technology can help supplement sustainability scores with real-time information. One possible area of extension is to include these elements in an impact assessment framework for the target organizations. The methods can also be extended to work with other sources of information like News, reports, social media, etc. with appropriate frameworks in place for verification of the information. Another interesting area that emerges is the ability to utilize the insights to obtain predictive intelligence for policy makers and business leaders. Organizations can take preemptive actions based on aggregate insights for their own sectors. They can also use the knowledge for assessment of their vendors or partners.

Author contribution. Abir Naskar, Tushar Goel, Vipul Chauhan, Ishan Verma, Tirthankar Dasgupta and Lipika Dey these authors contributed equally to this work.

Competing interest. The authors have no relevant financial or non-financial interests to disclose.

Data availability. All the data collected for training, testing, and further analytics purposes have been collected from open regulatory news articles under Environmental (E), Social(S), and Governance (G) categories published between 2015 and 2022 by the following sources:

- Environmental—EPA¹³
- Social—OSHA¹⁴ and EEOC¹⁵
- Governance—DoJ¹⁶

References

- Aggarwal CC and Zhai C (2012) *A Survey of Text Clustering Algorithms*. Springer, pp. 77–128.
- Ahi P (2014) *Sustainability analysis and assessment in the supply chain*. D. Tech Dissertation. Toronto: Ryerson University .
- Alkaraan F, Albitar K, Hussainey K and Venkatesh V (2022) Corporate transformation toward industry 4.0 and financial performance: The influence of environmental, social, and governance (ESG). *Technological Forecasting and Social Change* 175, 121423.

¹³ <https://www.epa.gov/newsreleases/search/newsreleaseslanguage/en/subject/compliance-and-enforcement-226191>

¹⁴ <https://www.osha.gov/news/newsreleases/enforcement/>

¹⁵ <https://www.eeoc.gov/newsroom/search>

¹⁶ <https://www.justice.gov/news>

- Bang J, Ryu D and Yu J** (2023) ESG controversies and investor trading behavior in the Korean market. *Finance Research Letters*, 54, 103750. <https://www.sciencedirect.com/science/article/pii/S154461232300123X>. <https://doi.org/10.1016/j.frl.2023.103750>.
- Bătae OM, Dragomir VD and Feleagă L** (2020) Environmental, social, governance (ESG), and financial performance of European Banks. *Accounting and Management Information Systems* 19(3), 480–501.
- Bhadani S, Verma I, Dey L and Bitetta V** et al. (2019) Mining financial risk events from news and assessing their impact on stocks. In Bitetta V et al (eds.), *Mining Data for Financial Applications–14th ECML PKDD Workshop, MIDAS 2019, Würzburg, Germany, September 16, 2019, Revised Selected Papers*, Vol. 11985 of *Lecture Notes in Computer Science*, 85–100 (Springer, 2019). https://doi.org/10.1007/978-3-030-37720-5_7.
- Bouma JJ, Jeucken M and Klinkers L** (2017) *Sustainable Banking: The Greening of Finance* (Routledge).
- Caudron E** (2022) *Measuring ESG Performance: A Text Mining Approach*. Master's thesis, Louvain School of Management, Université catholique de Louvain, 2022. Prom.: Vrins, Frédéric. Available at <http://hdl.handle.net/2078.1/thesis:35409>.
- Cer D** et al. (2018) Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Chiariini A** (2017) Environmental policies for evaluating suppliers' performance based on GRI indicators. *Business Strategy and the Environment* 26(1), 98–111.
- Ching HY, Gerab F and Toste TH** (2014) Scoring sustainability reports using GRI indicators: A study based on ISE and ftse4good price indexes. *Journal of Management Research* 6(3), 27.
- Cojoianu TF** et al. (2023) The city never sleeps: But when will investment banks wake up to the climate crisis? *Regional Studies* 57(2), 268–286.
- Collobert R** et al. (2011) Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug), 2493–2537.
- ESG incidents and value destruction: insights from the sustainability incidents integration study** (2018).
- Fenwick T** (2007) Developing organizational practices of ecological sustainability: A learning perspective. *Leadership & Organization Development Journal*, 28(7).
- Fiaschi D, Giuliani E, Nieri F and Salvati N** (2020) How bad is your company? Measuring corporate wrongdoing beyond the magic of ESG metrics. *Business Horizons* 63(3), 287–299.
- Fischbach J** et al. (2022) Automatic esg assessment of companies by mining and evaluating media coverage data: NLP approach and tool. *arXiv preprint arXiv:2212.06540*.
- Fleiss JL, Levin B, Paik MC** et al. (1981) The measurement of interrater agreement. *Statistical Methods for Rates and Proportions* 2(212–236), 22–23.
- Garbie IH** (2015) Sustainability awareness in industrial organizations. *Procedia Cirp* 26, 64–69.
- Geddes A, Schmidt TS and Steffen B** (2018) The multiple roles of state investment banks in low-carbon energy finance: An analysis of Australia, the UK and GERMANY. *Energy policy* 115, 158–170.
- Goel T, Jain P, Verma I, Dey L and Paliwal S** (2020) *Mining Company Sustainability Reports to Aid Financial Decision-Making. Proc. of AAI Workshop on Know. Disc. from Unstructured Data in Fin. Services*.
- Guo T** (2020) Esg2risk: A deep learning framework from esg news to stock volatility prediction. Available at SSRN 3593885.
- Gustavsson V** (2022) *Zero-shot Text Classification of Sustainability Factors on Websites: Detecting Environmental, Social and Corporate Governance factors (ESG) with Natural Language Processing (NLP)*. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS).
- Henisz W, Koller T and Nuttall R** (2019) Five ways that ESG creates value. *McKinsey Quarterly*, (4), 1–12.
- Joshi KD and Nalwade P** (2013) Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing* 2(7), 219–223.
- Kingma DP and Ba J** (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- KPMG survey of corporate responsibility reporting** (2017). Available at <https://home.kpmg/xx/en/home/campaigns/2017/10/survey-of-corporate-responsibility-reporting-2017.html>.
- Lee J and Kim M** (2023) Esg information extraction with cross-sectoral and multi-source adaptation based on domain-tuned language models. *Expert Systems with Applications* 221, 119726. Available at <https://www.sciencedirect.com/science/article/pii/S0957417423002270>. <https://doi.org/10.1016/j.eswa.2023.119726>.
- Liu X, He P, Chen W and Gao J** (2019) Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.
- Liu X** et al. (2015) Representation learning using multi-task deep neural networks for semantic classification and information retrieval.
- Lozano R and von Haartman R** (2018) Reinforcing the holistic perspective of sustainability: Analysis of the importance of sustainability drivers in organizations. *Corporate Social Responsibility and Environmental Management* 25(4), 508–522.
- Luccioni S, Baylor E and Duchene N** (2020) Analyzing sustainability reports using natural language processing. *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*. Available at <https://www.climatechange.ai/papers/neurips2020/31>.
- Luong M-T, Le QV, Sutskever I, Vinyals O and Kaiser L** (2015) Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.
- Mahajan A, Dey L and Haque SKM** (2008) *Mining Financial News for Major Events and their Impacts on the Market*. IEEE Computer Society, pp. 423–426 <https://doi.org/10.1109/WIIAT.2008.309>.

- Manning CD et al. (2014) *The Stanford CoreNlp Natural Language Processing Toolkit*, pp. 55–60.
- Murakami S and Muraoka S (2022) Exploring the potential of internet news for supply risk assessment of metals. *Sustainability* 14(1). Available at <https://www.mdpi.com/2071-1050/14/1/409>. <https://doi.org/10.3390/su14010409>.
- Murphy D and McGrath D (2013) ESG reporting—class actions, deterrence, and avoidance. *Sustainability Accounting, Management and Policy Journal* 4(2), 216–235.
- Nugent T, Stelea N and Leidner JL (2020) Detecting ESG topics using domain-specific language models and data augmentation approaches. *arXiv preprint arXiv:2010.08319*.
- Nugent T, Stelea N and Leidner JL (2021) *Detecting Environmental, Social and Governance (ESG) Topics using Domain-Specific Language Models and Data Augmentation*. Springer, pp.157–169.
- Pasch S and Ehnes D (2022) *NLP for Responsible Finance: Fine-tuning Transformer-Based Models for ESG*. IEEE, pp. 3532–3536.
- Perazzoli S, Joshi A, Ajayan S and de Santana Neto JP (2022) Evaluating environmental, social, and governance (esg) from a systemic perspective: An analysis supported by natural language processing. *Social, and Governance (ESG) from a Systemic Perspective: An Analysis Supported by Natural Language Processing* (October 11, 2022).
- Polignano M et al. (2022) *An NLP Approach for the Analysis of Global Reporting Initiative Indexes from Corporate Sustainability Reports*, 1–8.
- Raghuandan A and Rajgopal S (2022) Do ESG funds make stakeholder-friendly investments? *Review of Accounting Studies* 27(3), 822–863.
- Raman N, Bang G and Nourbakhsh A (2020) Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction* 2(4), 453–468.
- Shahi AM, Issac B and Modapothala JR (2014) Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications* 13(01), 1450006 .
- Sinha R, Datta M and Ziolo M (2020) *ESG Awareness and Perception in Sustainable Business Decisions: Perspectives of Indian Investment Bankers Vis-a-vis selected European Financial Counterparts*. Springer, pp. 261–276.
- Sokolov A, Caverly K, Mostovoy J, Fahoum T and Seco L (2021) Weak supervision and Black-Litterman for automated ESG portfolio construction. *The Journal of Financial Data Science* 3(3), 129–138 .
- Sokolov A, Mostovoy J, Ding J and Seco L (2021) Building machine learning systems for automated ESG scoring. *The Journal of Impact and ESG Investing* 1(3), 39–50.
- Sultana S, Zulkifli N and Zainal D (2018) Environmental, social and governance (ESG) and investment decision in Bangladesh. *Sustainability* 10(6), 1831.
- Tarmuji I, Maelah R and Tarmuji NH (2016) The impact of environmental, social and governance practices (ESG) on economic performance: Evidence from ESG score. *International Journal of Trade, Economics and Finance* 7(3), 67.
- Tregidga H, Milne M and Kearins K (2014) (Re) presenting ‘sustainable organizations. *Accounting, Organizations and Society* 39(6), 477–494.
- Twinamatsiko E and Kumar D (2022) *Incorporating ESG in Decision Making for Responsible and Sustainable Investments using Machine Learning*. IEEE, pp. 1328–1334.
- Van Der Elst J (2022) *Extracting ESG Data from Business Documents*. Master’s thesis, Ecole polytechnique de Louvain, Université Catholique de Louvain, 2021. Nijssen, Siegfried. Available at <http://hdl.handle.net/2078.1/thesis:30732>.
- Velte P (2017) Does ESG performance have an impact on financial performance? Evidence from Germany. *Journal of Global Responsibility*, 8(2), 169–178.
- Wan C et al. (2021) Multi-task sequence learning for performance prediction and KPI mining in database management system. *Information Sciences* 568, 1–12.
- Waring TM et al. (2015) A multilevel evolutionary framework for sustainability analysis. *Ecology and Society* 20(2), 34–49.
- Weber O, Koellner T, Habegger D, Steffensen H and Ohnemus P (2008) The relation between the GRI indicators and the financial performance of firms. *Progress in Industrial Ecology, an International Journal* 5(3), 236–254.
- Xu Y, Liu X, Shen Y, Liu J and Gao J (2018) Multi-task learning for machine reading comprehension. *arXiv preprint arXiv:1809.06963*.
- Yoon B, Lee JH and Byun R (2018) Does ESG performance enhance firm value? Evidence from Korea. *Sustainability* 10(10), 3635.
- Yuan X, Li Z, Xu J and Shang L (2022) ESG disclosure and corporate financial irregularities—evidence from Chinese listed firms. *Journal of Cleaner Production* 332, 129992.
- Zhao C et al. (2018) ESG and corporate financial performance: Empirical evidence from China’s listed power generation companies. *Sustainability* 10(8), 2607.