A Multi-Perspective Content Analysis Platform for Sustainability Assessment

Lipika Dey, Tirthankar Dasgupta, Abir Naskar, Tushar Goel, Ishan Verma, Vipul Chauhan, Uma M N and Rajkumar Pallikuth
TCS Research, India
(lipika.dey,dasgupta.tirthankar,abir.naskar,t.goel,ishan.verma,chauhan.vipul,uma.mn,raj.p1)@tcs.com

ABSTRACT

As awareness about sustainability increases, organizational initiative to remain compliant to sustainability development goals and disclosing data is also picking up. Consequently, there is an exponential rise in demand for sustainability assessment. Currently, a handful of rating agencies are collecting this data separately and analyzing them to come up with a score. This process however is not transparent and doesn't satisfy the need of all stakeholders. In this paper, we present how natural language processing techniques can be employed to extract the relevant information from the disclosures and other available reports. The proposed content analysis platform is powered by deep-neural language processing models that can extract, label and group contextually relevant information from multiple sources. These components complement and supplement each other to provide a holistic view of an organization. The platform itself is designed to provide access to underlying information at multiple levels of granularity, but can be linked to external risk analysis and decision making modules. The NLP models have been discussed in details along with sample results and evaluations.

KEYWORDS

sustainability assessment, violation detection, ESG information extraction, text analysis

ACM Reference Format:
Lipika Dey, Tirthankar Dasgupta, Abir Naskar, Tushar Goel, Ishan Verma, Vipul Chauhan, Uma M N and Rajkumar Pallikuth . 2022. A Multi-Perspective Content Analysis Platform for Sustainability Assessment. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With increasing awareness about sustainability issues, the future of finance clearly lies in socially responsible investing, environmental awareness, and championing for corporate ethics. Consequently, all stakeholders are increasingly demanding information about an organization's commitment towards the causes of Environment (E), Society (S) and good Governance (G)[2]. Not only the organizations themselves, Scope 3 compliance of an organization also demands sustainability compliance of all their vendors and partners [16]. Companies are therefore actively incorporating these goals in their financial planning, and disclosures are also on the rise. These reports contain information about the company's activities, agenda and performance in the areas of environmental, social, and corporate governance (ESG) loosely structured around the indicators defined by Global Resource Initiative (GRI) [8]. The objective of these indicators is to help in easing assessment, though not all of them are not strictly quantifiable. Disclosure regulations and standards are still emerging. Currently disclosure reports do not follow any specific format [1, 15, 18, 19, 21], thereby making the task of assessment difficult. Sustainability analysis, however cannot depend solely on these self-disclosures. Information provided by monitoring agencies, regulatory reports, alternative views coming from technology and industry analysts, News and social media can also contribute significantly towards the assessment [9, 11]. These sources provide information about new or undisclosed ESG violations, public sentiments, long or short term impact of certain choices, local biodiversity issues etc. all of which provide critical information for decision making.

The underlying information sources used for sustainability assessment is predominantly unstructured. Facts and figures are buried in reports and spread over a multitude of sources, all of which have to be aggregated and assimilated contextually. The volume and velocity of relevant information is also increasing at a furious rate. Insight generation from this heap of information requires Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques [7, 13]. In this paper, we present a content analysis framework powered by deep-neural language processing models for extracting contextually relevant information from a multitude of relevant sources. Given that the content here is diverse in nature, created with different intent from different perspectives, we propose knowledge-driven information processing for each type of content. The framework supports customizable sustainability assessment, by allowing an end-user to have a 360 degree view with information aggregated from diverse sources, and organized around pre-defined measurable factors aligned with ESG goals. Task specific reasoning mechanisms can also be built on top of this.

The motivation for the work comes from the fact that presently sustainability analysis mostly depends on sustainability scores awarded to an organization by a few designated third party rating organizations[8, 15, 18]. The scoring process is not transparent, and also not contextualized, as it uses the same lens for all industries. However, sustainability analysis is by itself a highly contextual task. While it is more important to look at the occupational health and safety standards to assess a healthcare company, for a financial company, more attention may be given towards prevention of

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

anti-money laundering and fraudulent practices. Similarly, for a retail company, the key focus may be towards scope 3 assessment of its manufactures and suppliers, while for a mining company one would like to rigorously assess its commitment towards maintaining biodiversity in its area of operations. Clearly, the risks associated with different industries arise from different sources. Though quantification of qualitative information is important, it is clear that sustainability assessment cannot rely on a single score. End-users also want access to organized information at multiple levels of granularity, get aggregated and fused views presented in a comprehensible way. The proposed framework is built to support guided access to an array of information along with interesting visualizations to help easy comprehension of information by end-users.

The rest of the paper is organized as follows. Section 2 provides an overview of the framework. Detailed design of different types of NLP applications that are designed for content analysis, are presented in subsequent sections. Evaluation of each individual module is presented alongside the models in each section. Section 6 presents how information is presented to end-users through sample visualizations.

2 SUSTAINABILITY CONTENT ANALYSIS - A FRAMEWORK FOR MULTI-SOURCE INFORMATION FUSION

Relevant content for sustainability assessment can be grouped under three categories - self-reported disclosures, reports published by regulatory agencies and consumer-generated content gathered from social media. Each type of content is generated from a unique perspective. Consequently the analytical objectives and techniques are also different.

Figure 1 presents the proposed multi-content analysis framework for processing different types of text data for sustainability assessment. The framework further supports plugging in of information visualization, drill down and customizable risk analysis modules for decision making. We now present an overview of the complementary and supplementary information components that can be extracted and used in a fusion framework.

(1) Self-disclosures - This category includes Sustainability and financial reports, company reports, 10K and SEC filings all prepared by the organizations themselves. These reports contain detailed information about objectives, goals, actions, achievements and outcomes for different activities undertaken under the Environment, Social and Governance (ESG) heads. Given a parameter for assessment, the task is to locate the relevant content within the report, extract information components like numbers, process details, implementation status or policies adopted. The report is most often a free-format PDF file containing a mixture of text and images. The techniques deployed include content location, indexing, information retrieval and information extraction. The methods are built on top of contextual embeddings of words and phrases, to take care of linguistic differences and synonyms.

(2) Regulatory reports - As ESG awareness grows, a number of regulatory and monitoring agencies are constantly engaged in framing new policies that support sustainable development for the entire planet, and also monitor the adherence to these principles. Regulatory reports contain information about organizational violations,the root causes identified, actions taken and penalties imposed. These reports not only provide a augmented information about an organization, but can also provide statistics about industry sectors, thus contributing significantly towards risk analysis. Extracting and compiling root causes from these reports can also help organizations to take corrective actions while formulating future practices and also effect policy changes, if required. The NLP methods employed for extracting information from regulatory reports is that of semantic role labeling, whereby portions of text are detected and labeled as violations, causes or effects. Collectively, the incidents and their causes mined from a large repository are used for deriving sector-level and regional-level insights. The information is also arranged into a causal graph that provide valuable insights for actionable intelligence. Using insights from regulatory reports is a novel aspect of the proposed framework, supported by novel methods for insight generation.

(3) Public view, sentiments and opinions - This is gathered from social media and News. The content is widely varied in nature and contain information about current events, pros and cons of emerging technologies, analyst views and consumer reactions. This content is streaming in nature, may contain previously unseen information, conjectures, sentiments and opinions, arguments and contradictions etc. The NLP methods employed for analysis of this type of content includes event detection, classification and sentiment analysis. At this point, the open source content is overlaid with the other content to enable decision makers have a 360 degree view for assessment.

The information components extracted are fed into a data fusion platform, where each piece of content is labeled with indicators defined by GRI. These indicators are further grouped under different heads for assessment purposes. Table 1 shows a possible grouping of the 32 indicators. Each indicator may belong to multiple groups also. A group score is derived based on aggregate information for the group. The scores can be numbers, fuzzy qualitative measures or a categorical letter grade. We propose that fusing information and insights derived from multiple sources can provide a wholesome view of sustainability performance. More details of the analytics modules are presented in the subsequent sections.

3 INFORMATION EXTRACTION FROM SUSTAINABILITY REPORTS

Sustainability reports are the main sources of information for computing ESG performance. The report contains facts and figures that contribute towards its score for each GRI indicator and consequently for a group. The NLP task we propose here is to automatically extract answers for a set of predefined questions formulated around the GRI indicators. The desired answer is either a numeric value for a measurable parameter like "percentage of reduction in green house gas emission" or some textual information, like (say) "about members of Audit Committee", for which the most word sequence that contains the answer has to be extracted.

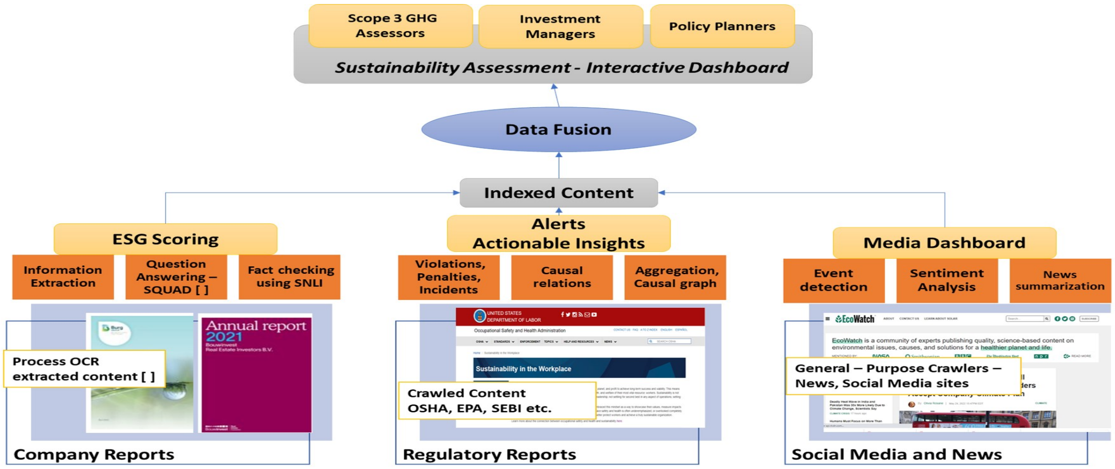


Figure 1: Proposed Framework

Category	GRI	Category	GRI
Climate Change	302_energy	Financial Performance	207_tax
Climate Change	301_materials	Financial Performance	417_marketing and labeling
Climate Change	305_emissions	Governance	415_public policy
Climate Change	307_environmental compliance	Planet Friendliness	303_water
Dealing with Diversity	405_diversity and equal opportunity	Planet Friendliness	304_biodiversity
Dealing with Diversity	406_Anti-discrimination	Planet Friendliness	306_effluents and waste
Dealing with Diversity	411_rights of indigenous people	Planet Friendliness	413_local communities
Economic Performance	201_Economic performance	Safe Work Environment	403_occupational health and safety
Economic Performance	202_Market presence	Safe Work Environment	410_security practices
Economic Performance	203_Indirect Economic impacts	Safe Work Environment	416_customer health and safety
Employee Wellness	402_labor management relations	Scope 3 GHG assessment	308_supplier_environmental_assessment
Employee Wellness	404_training and education	Scope 3 GHG assessment	414_supplier social assessment
Employee Wellness	407_freedom of association and collective bargaining	Societal Impact	401_employment
Employee Wellness	409_forced or compulsory labor	Societal Impact	408_child labor
Financial Performance	205_Anti-corruption	Societal Impact	412_human rights assessment
Financial Performance	206_Anti-competitive behavior	Societal Impact	419_Socio-economic compliance

Table 1: Grouping of the 32 GRI Indicators

In the proposed framework, the task is solved as a two-stage process. First, a sentence that is most similar to the question content is retrieved from the report, using cosine similarity between their stacked embeddings [7]. Next, the question and the sentence are both passed to the transformer based SQUAD architecture proposed in [6, 10], which extracts the exact sequence that comprises the answer. Table 2 presents a few sample questions and corresponding answers extracted from different reports.

This entire process has been evaluated against an erstwhile fully manual process. Generating answers for 30 questions for each company took a person to read a 150 - 200 page report, which on an average took 3 person days. With these methods deployed, the information extraction time reduced to 30 minutes per company, including manual verification and intervention, if required, which is a huge productivity gain. More questions also get added continuously. For the second task, answers to 65 questions from over 60 reports have been manually verified and accuracy of the SQUAD process is found to be 82.6%.

4 INFORMATION EXTRACTION FROM REGULATORY REPORTS

Regulatory and monitoring agencies regularly announce operational guidelines to be followed by organizations in order to remain compliant with respect to sustainable development goals. They also monitor their designated geographical regions to through inspections or otherwise, to locate violations, take appropriate action, mitigate litigation among different parties if they arise, and so on. Some of these agencies may also reward individuals and societies to encourage fair and sustainable practices. One of the oldest such organization is Occupational Safety and Health administration (OSHA) which has been operating in North America region for more than a decade now. OSHA has been actively pursuing the cause of safe and employee-friendly workplace. Analysis shows that the actions taken and issues identified by OSHA in the past, have helped in reducing the number of similar accidents in many regions [5]. Presently, Environment Protection Authority (EPA) is

Source of information	**Question**	**Relevant Sentence**	**Answer**
Essity Sustainability Report	How many women members are there in the board committee?	Five of the Board members are women, corresponding to 55% of the total number of AGM-elected Board members.	five
Essity Sustainability Report	what is the volume of water being discharged by the company?	The total volume of discharged water was 93 million cubic meters.	93 million cubic meters
Old ChangKee Limited Corporate Governance Report	Did the company disclose the orientation programmes for new directors?	All newly appointed Directors will undergo an orientation programme where the Director would be briefed on the Groupś history, strategic direction, governance practices, business and organisation structure as well as the expected duties and obligations of a director of a listed company, details of which are set out in a formal appointment letter provided to such newly appointed Director.	the director would be briefed on the groupś history, strategic direction, governance practices, business and organisation structure as well as the expected duties and obligations of a director of a listed company
STENGG Annual Report	Are all the audit committee members independent ?	The Audit Committee comprises three independent Directors, one of whom is also the Chairman of the Committee.	the audit committee comprises three independent directors
Signify Sustainability Report	How much total waste was recycled by the company ?	82% of total waste was recycled.	82%
Lexmark Sustainability Report	How much material is reused or recycled ?	97 percent or 7,861 metric tons of materials reclaimed from our customers' returned cartridges were reused or recycled.	97 percent or 7,861 metric tons
Lexmark Sustainability Report	Does the company pay any fines or sanctions for non-compliance with laws and regulations ?	Lexmark has not been subject to any significant fines or nonmonetary sanctions for noncompliance of laws and regulations related to accounting fraud, human rights, workplace discrimination, health and safety or corruption during this reporting period.	lexmark has not been subject to any significant fines or nonmonetary sanctions
Singapore airlines Annual Report	Does the company disclose the orientation programmes for new directors ?	The Company conducts orientation programmes for such new Board Directors, including site visits to the Company's main centres of operations such as the aircraft hangars and training facilities for cabin crew and pilots.	the company conducts orientation programmes for such new board directors

Table 2: Sample questions and corresponding answers extracted from different reports

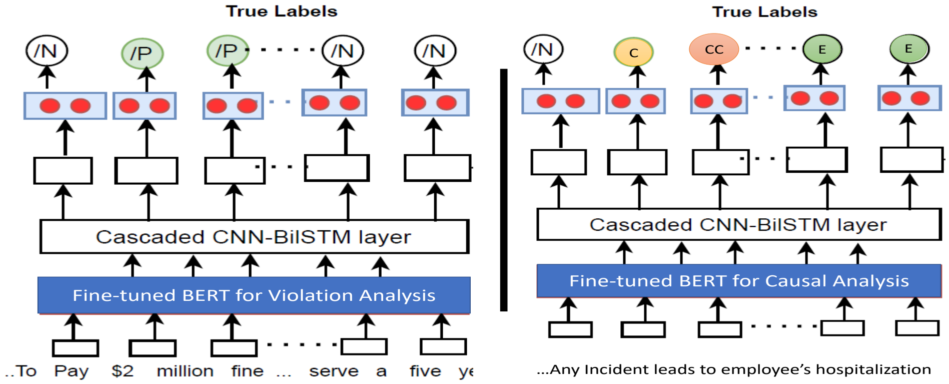

True Labels
/N
/P
/P
/N
/N
Cascaded CNN-BiLSTM layer
Fine-tuned BERT for Violation Analysis
..To Pay $2 million fine ... serve a five ye
True Labels
/N
C
CC
E
E
Cascaded CNN-BiLSTM layer
Fine-tuned BERT for Causal Analysis
...Any Incident leads to employee's hospitalization

Figure 2: Sequence labeling models for analyzing regulatory reports

The U.S. Environmental Protection Agency (EPA) announced the court approval of a settlement agreement with an Ottumwa, Iowa, man and his two companies for [***alleged violations of the federal Clean Water Act (CWA), including the illegal construction of a recreational vehicle campground***]$_{Violation}$ in Ottumwa on the Des Moines River. The Clean Water Act seeks to protect the nation's water resources, said David Cozad, director of EPA Region 7's Enforcement and Compliance Assurance Division. [***Placing unauthorized fill material into rivers and wetlands can degrade watershed health***]$_{Cause}$, [***create loss of wildlife habitat, and deprive downstream landowners and the public from the use and enjoyment of public waters***]$_{Effect}$. According to EPA, Russell Kirk and his companies, [***Ottumwa Northshore LLC***]$_{ViolatingOrganization}$ and ***[Breaking Gate LLC]***$_{ViolatingOrganization}$, filled in approximately 5 acres of protected wetlands and conducted unauthorized bank stabilization along approximately 2,000 feet of the Des Moines River without first obtaining a required CWA permit.

Table 3: Sample output from the two sequence labeling models depicting mentions of violation, violating organization names, causes and effects of the violation.

playing a similar role in tracking environmental violations as well as documentation of best practices, in the USA. Similarly, The Securities and Exchange Board of India (SEBI) is a regulatory body for securities and commodity market in India. Singapore Exchange also has dedicated sites to report organizational sustainability issues. Since a self-disclosure report is published at periodic intervals, the regulatory reports provide authentic information about the current state of sustainability practices for a company. Following is a sample sentence from a regulatory report published by Environment Protection Authority (EPA) - *ORANGE COUNTY - The U.S. Environmental Protection Agency (EPA) has announced settlements with Basin Marine, Inc. and Balboa Boatyard of California, Inc., to resolve Clean Water Act violations for discharging contaminants into Newport Bay. The violations at both facilities related to regulations preventing the discharge of pollutants through stormwater as well as the failure to comply with California's industrial stormwater permit.*

The first task for regulatory report analysis is to extract the names of the violating organization (O), violation incident (V), penalty amounts (P) from the report. In the current framework, this is modeled as a sequence labeling task, which is solved using

a transformer based neural network architecture. The model is trained to recognize the semantic role of each word in the sentence and subsequently mark it with a label O, V, P or none (N). It is implemented using a CNN-BiLSTM sequence classification layer over BERT. The overall architecture of the model is depicted in Figure 2. The model works on one sentence at a time to determine appropriate labels for words in the text. The learning task is formally represented as follows:
During training, the model is given a sentence x_t from the training data, in which its $i - th$ word, w_t^i is accompanied by a label $\bar{q}_t^i$, resulting in a sequence of labels $\bar{y_t}$. The task is to learn $\bar{y_t}$ for x_t. $\bar{q_t^i}$ is represented using the one-hot representation of the ground-truth class labels, which are V, O, P or none.

Each input sentence is passed to the BERT encoder containing 12 transformer blocks, 12 self-attention heads, and a hidden layer of 768 units [17]. BERT tokenizes an input sentence and usually generates its contextual representation as a single vector, termed as CLS. BERT can take in an input sentence of no more than 512 tokens at a time, which is enough to represent a sentence. To generate the sequence labels for each token, along with the CLS vector, we also take the token level representations from the hidden states of BERT, and then pass them through an additional classification network. This network is designed using a CNN layer followed by a BiLSTM network. While BERT creates token level representation of each word baked by the influence of other words in the sentence without taking their relative positions into account, the CNN layer adds more information by dealing with neighbouring words or n-grams at a time. The BiLSTM layer on top of learns even more structural information about a sentence, by looking at it from both ends simultaneously. The many-layered representation of a sentence is found to be ideal for learning to recognize information components correctly from arbitrarily complex sentences. The final hidden states of the $BERT - CNN - BiLSTM$ network of all the tokens are passed into a softmax layer to classify over the sequence labels. The output of the model is represented as: $q_t^i = softmax(W * h_i + b), i \in (1...N)$ where h_i is the hidden state corresponding to the word w_i. The loss functions (L) for the sequence labeling task is defined as:

$$L(\theta) = -\sum_{t=1}^{N}\sum_{j=1}^{J}\bar{q}_t^{i,j} log(q_t^i)$$

The second module, i.e the cause-effect sequence labeling module is also implemented using an identical design employing a $BERT - CNN - BiLSTM$ network model, The only difference is in the output layer, as the softmax classifier is now required to label each token as cause (C), effect (E), causal connector (CC) or none. Obviously a cause or an effect can be spread over multiple consecutive words. While the first model learns to exploit the linguistic characteristics of text like the presence of named entities in the vicinity of typical verbs like *fined, penalized etc.*, the second model learns the relative dependence of words within a sentence, bound by causal connectors like *'caused by, due to etc.* Table 3 illustrates how the violation component detection and causal relation extraction model works, by depicting sample input sentences and the final labels awarded to each word by the respective models.

4.1 Dataset - Training and Evaluation

For the violation extraction task, we have collected and manually annotated a subset of 900 documents, 300 each for the Environmental, Social and Governance domains, all published between the time-period of 2015-2017, by EPA, OSHA and Gov.org respectively. The average length of a document is around 23 sentences. Six annotators took part in the annotation using the Stanford simple manual annotation tool[1]. The experts read each document and performed the following tasks : (a). From among the named entities, marked them as the target organization (O) and penalty values (P). (b). Mark phrases in the text that cover incidents and / or violation (V). At the end of annotation, each word in the document is either assigned a label O, V, P or None.

For the cause-effect extraction task, we took a few openly available cause-effect datasets that were shared for a challenge, CEREX 2020, during an international conference FIRE 2020[14]. It contains around 4500 causal sentences, some of which were released by a challenge task during SEMEVAL 2010 Task-8 and others were newly introduced. In this dataset, words in a sentence are annotated with labels cause (C), effect (E), causal connector (CC) or none.

The pre-trained BERT model provides a powerful context dependent sentence representation and can be used for various target tasks through the fine-tuning procedure. Fine-tuning the pre-trained model with training data from different domains is known to improve the performance of language processing tasks. Further, we set the early stopping of fine-tuning to 800 steps in order to prevent over-fitting. We use a batch size 32, a maximum sequence length of 128, and a learning rate of $2 * 10^5$ for fine-tuning this model. Each model was trained using 70%, validated with 10% and evaluated using the remaining 20% of the corresponding annotated corpus.

Table 4 (a) presents the accuracy of labeling word-sequences within a sentence by their respective categories - *O, V, P* as described earlier. It was found that the BERT-CNNBiLSTM model yields high recall for all the labels. Detailed analysis shows that the model significantly reduces the false negative scores and achieves a high true positive score, thereby achieving high precision and recall values for all the three classes. The highest F-measure of 0.92 was obtained for violations, followed by In general, an F-Measure of 0.90 was obtained for penalty, followed by 0.88 for violations and 0.86 for organizations. Organizations are named entities. They do not exhibit linguistic characteristics, hence the result.

Table 4 (b) presents the results for the cause-effect extraction task. In this case also, the BERT-CNNBiLSTM model is found to achieve a reasonable high F-measure across all the three class labels namely Cause(C), Effect (E) and Connectives (CC). The comparatively lower scores are due to less data availability for training. The nature of the sentences here are also more complex than those used for training.

The trained model is thereafter deployed to work on new reports that are crawled daily from a multitude of designated sources, over and above the sources that were used to gather training articles mentioned earlier. It works quite well in detecting actual violations from any News article. An example of a violation captured from a recent new article is shown in figure 6. It can be further extended to capture possible violation events also. We have observed that

[1] https://nlp.stanford.edu/software/

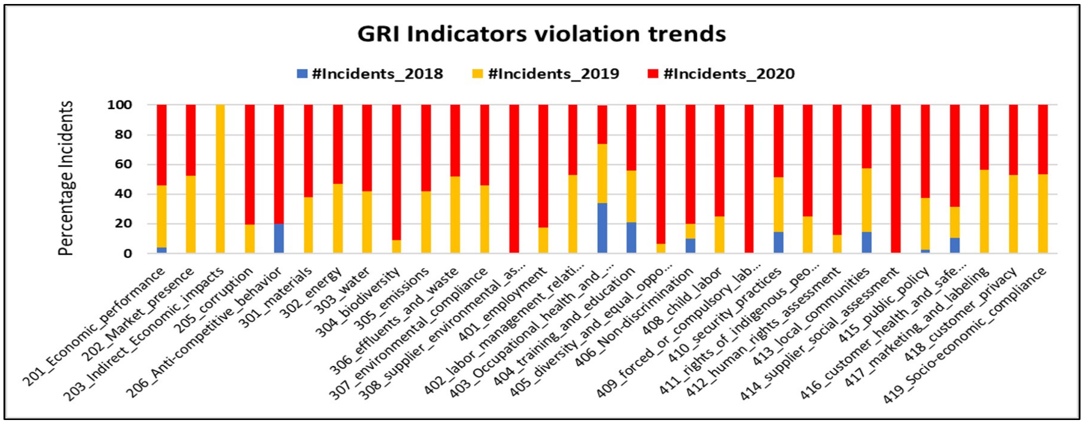


Figure 3: GRI indicators violation trends

(a) Violation Identification				(b) Cause-Effect Extraction			
Class	P	R	F1	Class	P	R	F1
Vio. Org.	0.83	0.89	0.86	Cause	0.73	0.77	0.75
Violation	0.85	0.92	0.88	Effect	0.77	0.79	0.78
Penalty	0.85	0.89	0.90	Connective	0.81	0.82	0.81

Table 4: Results reporting Precision(P), Recall (R) and F1-Score (F1) of (a). sustainability events (*violation, violating organization* and *penalty*) and (b). Causal relation extraction (as *cause, effect* and *causal connectives.*

several News articles mention about events which may lead to a violation, which are more like conjectures rather than facts. The present model may be fine tune to handle those also, however using these event information would need an added verification task.

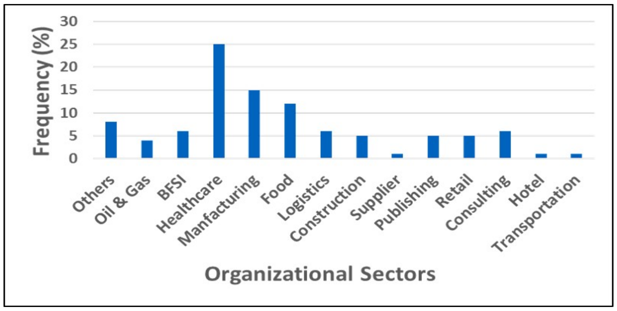


Figure 4: Distribution of Incidents of discrimination (GRI 406) across organizational sectors for the Year 2020

4.2 Aggregate Analysis of Violation Incidents

In order to show how aggregated statistics can be generated, the incidents extracted from the corpus was mapped to the GRI indicators, using cosine-similarity between the vector representations of both the indicator and the incidents. The sectors of all the extracted organization names were looked up from external sources.

Figure 3 shows trends of all 32 GRI violations obtained from reports published between years 2018 to 2020. A steep rise can be seen in 2020 for violations related to corruption (205), biodiversity (304), child labor (408), discrimination in workplace (406), diversity and equal opportunity (405), and human rights (412). Rising trends are also observed for environmental indicators like biodiversity (304), water resources (303) etc. Violations related to supplier environment assessment (308) and supplier social assessment (414) are observed only in the year 2020. Interestingly there is no significant change corresponding to occupational health and safety (403) observed over last three years. We believe all these observations can be attributed to the fact that occupational health and safety monitoring has been actively monitored globally for last ten years, while awareness about environment and some other social factors are fairly recent phenomena. This indicates that sustainability monitoring and reporting itself can help in bringing down the numbers of incidents in future, as other organizations become cautious. One significant observation is 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 4 shows the distribution of the sectors across which discrimination, indicated by presence of text related to GRI label 406, was observed, most of which were in Healthcare.

4.3 Aggregation of causal knowledge

We now show how the cause-effect pairs extracted from the corpus was aggregated to generate a causal knowledge graph. Though there was a high degree of partial or full semantic similarity among the causes or effects extracted, they could not be used straight away. Hence, we applied the K-means clustering algorithm to group conceptually similar elements together. Clustering of the causes and effects were done independently. Each cause and effect sub-sequence was first converted to a 512 dimensional vector using the Universal Sentence Encoder [4]. The K-Means [3] clustering algorithm provided by the Sklearn library[2] was implemented with Euclidean distance as the distance metric. The number of clusters is chosen using the Elbow method [12]. To obtain semantic interpretation of the clusters, a combination of frequent uni-gram, bi-gram and tri-grams of each cluster is used. Assuming that there are *m* cause

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

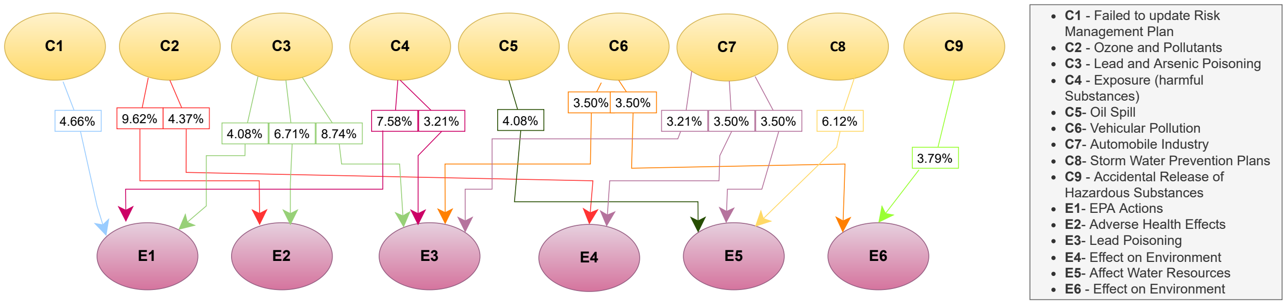

C1
C2
C3
C4
C5
C6
C7
C8
C9
4.66%
9.62%
4.37%
4.08%
6.71%
8.74%
7.58%
3.21%
4.08%
3.50%
3.50%
3.21%
3.50%
3.50%
6.12%
3.79%
E1
E2
E3
E4
E5
E6
C1 - Failed to update Risk Management Plan
C2 - Ozone and Pollutants
C3 - Lead and Arsenic Poisoning
C4 - Exposure (harmful Substances)
C5- Oil Spill
C6- Vehicular Pollution
C7- Automobile Industry
C8- Storm Water Prevention Plans
C9 - Accidental Release of Hazardous Substances
E1- EPA Actions
E2- Adverse Health Effects
E3- Lead Poisoning
E4- Effect on Environment
E5- Affect Water Resources
E6 - Effect on Environment

Figure 5: Cause Effect cluster relation from Environmental Violations. The number inside the edge denotes the percentage of documents containing the representative relation.

clusters and n effect clusters obtained from a repository, there are $m * n$ combinations of causal relations possible. For each of these cluster pairs, the percentage of documents containing a representative relation from the cluster is computed. These pairs are then used to present consolidated insights.

Frequent violations related to Occupational safety and Health	% of reports
Inadequate Training leading to accidents	56
exposing employees to trenching hazards	46
Failure to comply with OSHA's fall protection standard	38
Blocked exit leading to burns, injuries and other hazards	31
Exposure to toxic chemicals and high heat	17
Fatal injuries, amputation	16
Violation of OSHA guidelines for coronavirus	7
Frequent causal violation Incidents from Financial Sector	**% of reports**
Discriminatory practices in financial institutions	63
Broker - dealer and Portfolio manager misconducts	53.6
banking wilful engagement in acts and practices that create or maintain inappropriate influence by investment banking over research analysts	45.7
Inappropriate selling of Mutual Fund shares	43.9
Failing to disclose details of transactions and purchases to clients	39.3
Wilful software manipulation; use of electronic chat to send data; Undetected coding errors	34.8

Table 5: Frequent Workplace violations mined from OSHA and Financial Regulatory Reports

Figure 5 presents part of a causal graph generated from a collection of 343 articles gathered from the site of United States Environment Protection Agency (EPA)[3] published between 2018 to 2020. The cause nodes on the top show frequently observed issues like lead poisoning, oil spill, ozone and pollutant release etc. that lead to health hazards as well as pose dangers to the environment. The nodes on top left illustrate that very often, violations occur due to procedural errors like not updating risk management practices, leading to legal action and penalty.

Following are some example sentences from occupational safety and financial regulatory reports which have cause, effect or both cause and effect.

- The company faces $281,108 in fines.
- The roofing contractor faces $19,890 in penalties.
- In addition, the SEC fined four of the firms for violating the record-keeping requirements concerning business-related internal e-mail communications during the period July 1999 through June 2001.

Since the effects were most often financial penalty, we only present a few causal incidents from these domains in Table 5.

5 OPEN-SOURCE DOCUMENT ANALYSIS

For open source content analysis, universal crawlers are deployed to gather content and process them using the language models proposed in earlier sections. Sentences relevant to a GRI indicator are determined using the similarity measures defined in section 3. The violation extraction module of section 4 is also run over the articles. Additionally, open source documents are also subjected to sentiment analysis. News articles are presented as summaries generated using PEGASUS [20], an abstractive content summarizer. The summaries are presented with links to the source, along with the violations extracted, if any. Positive scoring articles are also summarized and presented.

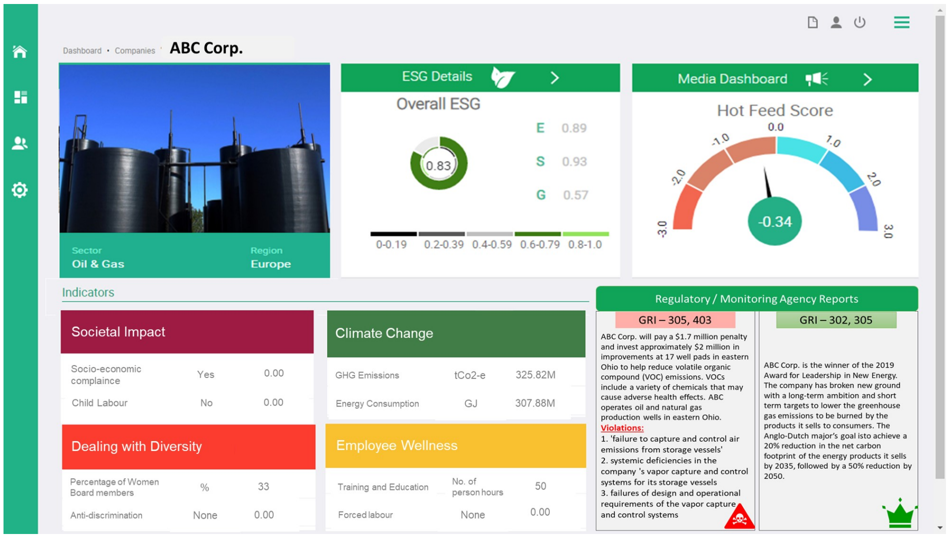
6 REPORTING AND VISUALIZATION

Figure 6 presents a sample visualization screen for a company, whose name has been anonymized to ABC. The different blocks on the screen represent different categories of information, at different levels of granularity. The bottom left blocks show ESG scores for groups, based on individual GRI information extracted from the reports, using the question answering mechanism explained in section 4. The media dashboard shows social media sentiments for company ABC, for sustainability-related posts. The block below shows summaries of positive and negative regulatory agency reports, along with the violations detected and GRI indicators that they were mapped to. One can also drill down to see sector-level performance for Oil and Gas sector and relative performance of ABC.

7 CONCLUSION

This paper presents a content analysis framework for information extraction from sustainability related documents. We have presented a detailed analysis of different kinds of documents and their content, which can present complementary and supplementary

[3] https://www.epa.gov/newsreleases/search



Sector: Oil & Gas
Region: Europe

Societal Impact		
Socio-economic complaince	Yes	0.00
Child Labour	No	0.00

Climate Change		
GHG Emissions	tCo2-e	325.82M
Energy Consumption	GJ	307.88M

Dealing with Diversity		
Percentage of Women Board members	%	33
Anti-discrimination	None	0.00

Employee Wellness		
Training and Education	No. of person hours	50
Forced labour	None	0.00

GRI – 305, 403

ABC Corp. will pay a $1.7 million penalty and invest approximately $2 million in improvements at 17 well pads in eastern Ohio to help reduce volatile organic compound (VOC) emissions. VOCs include a variety of chemicals that may cause adverse health effects. ABC operates oil and natural gas production wells in eastern Ohio.

Violations:
1. 'failure to capture and control air emissions from storage vessels'
2. systemic deficiencies in the company 's vapor capture and control systems for its storage vessels
3. failures of design and operational requirements of the vapor capture and control systems

GRI – 302, 305

ABC Corp. is the winner of the 2019 Award for Leadership in New Energy. The company has broken new ground with a long-term ambition and short term targets to lower the greenhouse gas emissions to be burned by the products it sells to consumers. The Anglo-Dutch major's goal isto achieve a 20% reduction in the net carbon footprint of the energy products it sells by 2035, followed by a 50% reduction by 2050.

Figure 6: Sample screenshot presenting fused information for sustainability assessment

information to get a holistic picture about a company's sustainability practices. The platform is powered by deep neural language processing models that help in automated extraction, semantic categorization and aggregation of contextually relevant information components like indicator level performance for ESG activities, violations and penalties etc. This allows end users to obtain a comprehensive view, yet empowers them to sift through large volumes of content in a systematic way.

The platform is currently being extended to create a hyper-linked collection that can also provide weighted opinionated views especially of contradictory opinions around technology and policies. This would provide a different lens altogether for viewing sustainability practices. Controversies abound in this domain be it around electrical vehicle technologies or excessive use of packaging content due to increase in online shopping practices, though the later is supposed to be contributing towards lower use of fuel by consumers. Ability to access all the contextually relevant information while assessing performance is by itself an important task. Incorporating some of these information for quantitative assessment is a challenging but interesting task. Extending the platform to support knowledge-driven reasoning is our next goal.

REFERENCES

[1] 2017. KPMG Survey of Corporate Responsibility Reporting 2017. https://home.kpmg/xx/en/home/campaigns/2017/10/survey-of-corporate-responsibility-reporting-2017.html

[2] 2018. ESG incidents and value destruction: insights from the Sustainalytics incidents integration study.

[3] Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining text data*. Springer, 77–128.

[4] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[5] Tirthankar Dasgupta, Abir Naskar, Rupsa Saha, and Lipika Dey. 2018. Extraction and visualization of occupational health and safety related information from open web. In *2018 IEEE/WIC/ACM international conference on web intelligence (WI)*. IEEE, 434–439.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Tushar Goel, Palak Jain, Ishan Verma, Lipika Dey, and Shubham Paliwal. 2020. Mining company sustainability reports to aid financial decision-making. In *Proc. of AAAI Workshop on Know. Disc. from Unstructured Data in Fin. Services*.

[8] Tian Guo. 2020. ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction. *Available at SSRN 3593885* (2020).

[9] Tim Nugent, Nicole Stelea, and Jochen L Leidner. 2020. Detecting ESG topics using domain-specific language models and data augmentation approaches. *arXiv preprint arXiv:2010.08319* (2020).

[10] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[11] Natraj Raman, Grace Bang, and Armineh Nourbakhsh. 2020. Mapping ESG Trends by Distant Supervision of Neural Language Models. *Machine Learning and Knowledge Extraction* 2, 4 (2020), 453–468.

[12] Stan Salvador and Philip Chan. 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE international conference on tools with artificial intelligence*. IEEE, 576–584.

[13] Amir Mohammad Shahi, Biju Issac, and Jashua Rajesh Modapothala. 2014. Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications* 13, 01 (2014), 1450006.

[14] Manjira Sinha, Tirthankar Dasgupta, and Lipika Dey. 2020. CEREX@FIRE-2020: Overview of the Shared Task on Cause-Effect Relation Extraction. In *Forum for Information Retrieval Evaluation (FIRE 2020)*. Association for Computing Machinery, 18–20.

[15] Sayema Sultana, Norhayah Zulkifli, and Dalilawati Zainal. 2018. Environmental, social and governance (ESG) and investment decision in Bangladesh. *Sustainability* 10, 6 (2018), 1831.

[16] Indarawati Tarmuji, Ruhanita Maelah, and Nor Habibah Tarmuji. 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 (2016), 67.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[18] Patrick Velte. 2017. Does ESG performance have an impact on financial performance? Evidence from Germany. *Journal of Global Responsibility* (2017).

[19] Bohyun Yoon, Jeong Hwan Lee, and Ryan Byun. 2018. Does ESG performance enhance firm value? Evidence from Korea. *Sustainability* 10, 10 (2018), 3635.

[20] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, 11328–11339.

[21] Changhong Zhao, Yu Guo, Jiahai Yuan, Mengya Wu, Daiyu Li, Yiou Zhou, and Jiangang Kang. 2018. ESG and corporate financial performance: Empirical evidence from China's listed power generation companies. *Sustainability* 10, 8 (2018), 2607.