Mining company sustainability reports to aid financial decision-making

Tushar Goel, Palak Jain, Ishan Verma, Lipika Dey, Shubham Paliwal

Innovation Labs, Tata Consultancy Services, News Delhi, India
(t.goel, jain.palak4, ishan.verma, lipika.dey, shubham.p3)@tcs.com

Abstract

Extracting information from financial documents like annual reports, sustainability reports or analyst reports plays an important role in investment decisions. Manual processing of these reports is time-consuming and tedious. In the past, pattern-based information extraction tools have been proposed to extract financial parameters from large documents. Rule-based approaches do not work for situations that need to extract information about actions or compliances along with targeted quantifiable information. In this paper, we present deep-learning based methodologies to retrieve information about sustainability practices from reports. Sustainability practices are becoming increasingly significant for investors to assess risk associated to a company. These reports have complex formats along with images and tables, due to which OCRs often fail to read the content correctly. We present methods to automatically detect text blocks from arbitrarily formatted PDF reports in a reliable way before using the OCR to read and index the content. This content is then searched for indicators that represent sustainability practices. The retrieved sentences are ranked based on their conceptual similarity to the indicators as well as quantitative content. Results show that the proposed methods retrieve sentences with high recall and precision and therefore substantially decrease human efforts to generate the right insights.

Introduction

Company reports, published by all public corporations, contain detailed information about their vision, operation, strategies and financial conditions. These reports have huge business implications, since several investor decisions are taken based on the reported information. Corporate banking divisions are also interested in gathering and analyzing this information, based on which they usually generate comparative portfolios of all their corporate clients, across different sectors. The insights extracted are used to generate appropriate advice both for the corporate clients as well as individuals interested in investing in these companies.

Along with detailed report about several business parameters and indicators like newly launched products and services, revenue, profit, future strategies etc. these reports also contain information about their Environmental, Social and Governance activities, collectively known as ESG factors which throw light on their social commitments. Performance indicators belonging to the ESG category are used in capital markets to determine how advanced the companies are with respect to sustainability practices. The three factors are described here in brief.

- **Environmental factor** examines how a business performs as a steward of natural environment focusing on practices around waste and pollution management, energy consumption, natural resource usage, reduction in greenhouse gas emission, addressing deforestation and climate change.
- **Social factor** looks at how the company treats people and concentrates on employee relations and diversity, working conditions, local communities, health, safety and conflicts.
- **Governance factor** examines the transparency around accounting methods and focuses on tax strategy, executive pay, donations and political lobbying etc. It also examines the company’s stand against corruption and bribery and commitment towards diversity.

It is found that, along with traditional business indicators like revenue, profit and market-share, ESG factors are playing increasingly significant roles in the investment process. Investors are choosing companies that are more socially responsible over the ones that do not have well-defined ESG policies. In (Bos 2014), author states that ESG integration in the mainstream investment process provides a mechanism to optimize risk return characteristics of a portfolio.

Though ESG factors currently play a crucial role in investment decisions, the related data is not easily available in a structured or semi-structured format like those available in 10K reports filed with SEC. The problem is compounded by the fact that the ESG reports follow no specific format or style and are heavy in graphics, info-graphics and text. Several person hours are currently devoted to this activity. Since these reports are usually voluminous, extracting relevant pieces of information to generate application-specific insights is a tedious and time-consuming task. It requires reading several pages which not only involves several person hours of effort but is also error-prone and subjective. Hence, partial or full automation of this task can greatly reduce the tedium and information curation time, leaving the human experts more time to validate and assimilate this information for subsequent decision-making.

In this paper, we present methods to automatically detect, retrieve and score relevant information about ESG indicators from textual content of company reports. We would like to emphasize that these types of information are text-heavy in nature and follows no specific pattern, thereby eliminating the use of simple pattern-based extraction methods. The sentences are often very long and complex making it difficult even for humans to comprehend without appropriate knowledge of domain and language. Simple search does not work well due to the many nuances of natural language.
We break the problem into a subset of tasks as follows:

(a). *Document Pre-processing* - The reports are assumed to be in PDF format. The text content from the PDF is first extracted using Google Tesseract (Smith 2007) which outputs text. We found that it does not always work perfectly for these reports, which does not follow a standard columnar structure. Since each page has multiple text blocks, the OCR often fails to detect the block boundaries correctly. We therefore introduce a pre-processing step, in which a block-detection algorithm is deployed to detect the block structures of any given page automatically. Thereby, the associated text obtained from the OCR is indexed and linked back to the block. The text output of each block is subjected to sentence detection using an NLP toolkit. It is observed that these steps improve the accuracy of the overall extraction.

(b). *Relevant Sentence Retrieval* - Given a list of indicators in each of the three ESG categories, the most relevant content for each indicator is obtained using a novel sentence scoring algorithm, that exploits vector representation of words and phrases. Top ranked sentences are then retrieved for each indicator.

(c). *Quantitative Information extraction* - For each quantifiable indicator, the top ranked sentences are further subjected to a filtering process based on the contained quantitative information. Presently, the filtered set of sentences are verified by humans in the loop for extracting the relevant quantities to help in downstream financial decision-making tasks like computing risks for investment portfolios.

It may be noted that, though these reports also contain information presented as infographics and tables, we have limited the scope of the current work to analyzing text content only.

Rest of the paper is organized as follows. Section 2 presents a brief overview of the ESG indicators. Section 3 presents overview of the pre-processing method that obtains the textual content from the reports. Section 4 presents methods to detect sentences relevant to these indicators. Section 5 presents some results from experiments conducted. Section 6 presents a review of related work and highlights the key differentiating aspects of the proposed work. Finally, section 7 concludes with a summary of the open problems and future directions.

ESG Indicators – an overview

Table 1: Sample ESG Indicators

ESG Factor	**Indicator Name**	**Quantifiable**	**Unit**
Environment	Energy Efficiency	Yes	KW/ MJ / Watts
	GHG Emissions or Carbon intensity	Yes	Metric tonnes / %
	Waste by Unit Produced	Yes	Tonnes / Million pounds
	Environmental Policy mentions	No	
Social	Lost Time in accidents	May be	Hours
	EH&S compliance in formal agreements	No	
	Hours of training	Yes	Hours / Day
	Compliance with Child labor prohibition	No	
Governance	Adherence to Anti-corruption training	No	
	Amount of money allocated for Employee welfare	Yes	% / USD
	Percentage of Women Board members	Yes	%
	R&D expenses	Yes	USD

Each ESG factor is a collection of indicators, around each of which a company is expected to elaborate more on their practices corresponding to it. Risks around each of these factors are assessed based on quantitative information reported about the indicators in a specific category.

For example, environmental risks for a company is evaluated based on how much energy is consumed by the company, how it is reusing and recycling materials, its contribution towards reducing the impact global climate change and its compliance with government environmental regulations. Social indicators address a broad range of complex issues depending on where the company operates and what it does for these issues. These indicators look for compliance by a company on several issues related to human rights, labor rights in its operations or supply chain, commitment against proliferation of child labour and other sustainability practices related to land acquisition, community relocation or doing business in conflict-affected areas. Social indicators also include factors that ensure employee well-being and safety by looking at parameters related to presence of labor health and safety clauses in formal agreements, hours of training etc.

Regarding governance indicators, investors want assurances that companies avoid conflicts of interest in their choice of board members, do not use political contributions to obtain unduly favorable treatment and do not engage in illegal practices. Authors in (Bassen and Kovacs 2008) (Rahdari and Rostamy 2015) presented a comprehensive list of ESG indicators and their categories. For current work, we have taken a subset of these indicators in each category as shown in Table 1. It also states whether the expected information around each indicator is quantitative, qualitative or either. Table 2 presents extracts from different reports around increasing diversity in management, to highlight the variations encountered in text, thereby asserting the complexity of the information extraction task.

Table 2: Sample sentences illustrating complexity for Governance indicator “Percentage of Women Board Members”

Company Name	**Sentences in report stating their commitment to support Diversity in Management**
FedEx	Diversity Starts at the Top the FedEx Board of Directors includes twelve directors, four of whom are women and two who are African American.
BM Group	As per the General Meeting in 2017, the composition of the Supervisory Board is in line with this target, given that two out of six members are female.
Henkel	In 2018 we were again able to raise the proportion of women in management worldwide – to 34.7 percent at December 31, 2018.
KPN	The 2018 results show that gender diversity in senior management increased slightly in 2018 to 22% women from 20% in 2017.
AstraZeneca	Women comprise 50.1% of our global workforce, and there are currently five women on our Board (42%).

Document pre-processing – sentence extraction from arbitrary content layout

As stated earlier, company reports are heavily unstructured with no specific format maintained even within a single document. Fig. 1 shows sample pages from two company reports to highlight the intra-document and inter-document diversities in format. The top two pages are from the report prepared by company BM Group and the bottom ones belong to the report of company FedEx. The layouts are complex and do not follow any specific template, thereby making the task of detecting and extracting coherent text very complex.

To read the text content from a PDF report, each page in the report is first converted into an image. Each page is then processed for obtaining the text content. Initially, Google’s Tesseract OCR was directly used on each page image to read the text content. It was observed that this tool reads the words correctly, however it often fails to parse the page structure properly, thereby leading to wrong sentences. Fig 2a shows an example of incorrect merging highlighted by the oval region. The text in two separate columns within the region are merged into a single continuous sentence, which is wrong. This error is extremely costly for our task, since it may lead to completely wrong information extraction. Also, Tesseract was not able to identify the tables correctly, hence it merged the text extracted from tables with other text in the page leading to incorrect sentences. Hence, we implemented a different pipe-line explained below, to obtain the content.

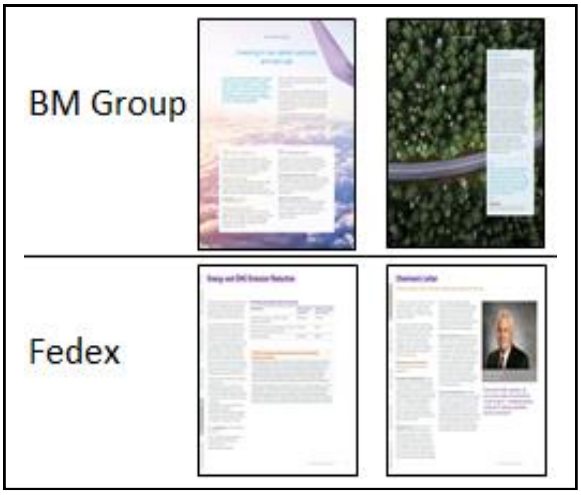
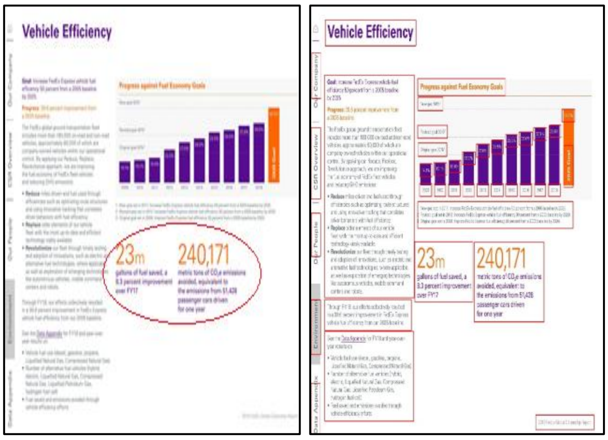


Fig. 1 Illustrating intra-report and inter-report format


Vehicle Efficiency
Progress against Fuel Economy Goals
23m
240,171
Vehicle Efficiency
Progress against Fuel Economy Goals
23m
240,171

(a) (b)

Fig. 2 (a) Encircled text incorrectly merged by Tesseract (b) Correct block detection using proposed algorithm.

To begin with, we first detect the table boundaries in each page image using the table detection method proposed in

TableNet (Paliwal 2019). The blocks enclosing tables are then removed from the page images. The resulting images are then processed to obtain proper textual blocks to eliminate incorrect merging of content, as explained below.

The block detection process starts with word boundary detection. For this, we used EAST (An Efficient and Accurate Scene Text Detector) (Zhou, et al. 2017) which uses deep learning to return the coordinates of the bounding boxes for each word. In the next step, a block detection algorithm is applied to merge neighboring boxes using a threshold. The threshold for each report is determined automatically using the approach explained below.

For every pixel in a page lying outside existing word bounding boxes, the boundary closest to it is determined. All distances are then normalized to values within 0 to 1, such that all pixels within and on a bounding box has value 0, and pixels furthest from any bounding box have value 1.

Setting *t* as an initial threshold value, the algorithm repeatedly pushes pixels lying outside any box to a box that is closest to it. At one extreme, the entire page content will lie within one box. At another extreme, each word is encapsulated in a single box, which is also our starting condition. Let *t1* and *t2* denote two threshold values at which the above extremes will be obtained. It may be further noted that due to the variance in format, the optimal values of *t1* and *t2* may be different for different reports. The objective of the proposed algorithm is to determine the correct values of *t1* and *t2* for each report.

To determine the optimal values for a report, we start with the page that has maximum number of words in it. Let this number be denoted by N_d for document d. Algorithm Block-detection explains the steps for determining the thresholds for a given document.

Algorithm Block-detection

```
(1). For document d, set Maxblock = $N_d$ and Minblock = 1.
(2). Let Numblock denote the number of blocks in page, which is initially equal to $N_{d..}$ Set t = 0.
(3). While Numblock = $N_d$
    (3a). Set t = t + Δt.
    (3b). Each pixel lying within a distance *t* from a block, is assigned to that block.
    (3c). All block boundaries are recomputed.
    (3d). Update Numblock.
(4). Obtain t1 = t. t1 denotes the maximum threshold at which there is no merging of word blocks.
(5). Repeat until Numblock = 1
    (5a). Set t = t + Δt.
    (5b). Each pixel lying within a distance *t* from a block, is assigned to that block.
    (5c). All block boundaries are recomputed.
    (5d). Update Numblock.
(6). Obtain t2 = t. t2 denotes the minimum threshold at which the whole page is within a single block.
(7). Set final threshold for document d as
        $t = \beta_1 * t_1 + \beta_2 * t_{2,}$
     where $\beta_{1,} \beta_2$ are constants.
```

Currently β1 and β2 have been determined empirically through experiments and are found to give best results when set as 0.3 and 0.7 respectively. High weightage to t2 can be justified by the fact that for the present task, it is always better to get bigger continuous blocks as compared to multiple smaller text blocks. It is a vital step to preserve the semantics of sentences.

Once we get the text-blocks for a report, we apply OCR to each of these text-blocks to obtain the text. The text content of each block is further processed using NLTK (https://www.nltk.org/) sentence tokenizer from NLTK natural language toolkit to split the content into its constituting sentences. Each sentence is then indexed according to its block-id and page-id, which helps to locate the sentence at the time of retrieval. Fig. 2 (b) shows the block detection output. After block detection, OCR on individual blocks eliminates the merging problem.

Retrieving Relevant Sentences for ESG Indicators

We now explain how relevant sentences for each ESG indicator are identified and ranked, based on their similarity to the indicators. The problem can be seen, as akin to a search problem, where the blocks form a repository and the indicators are like search queries, though the challenge is more complex due to the following reasons:

(a). The matching has to happen at more granular levels of words and phrases and has to be more precise, since the ultimate target is a sentence rather than a document.

(b). Each indicator name is a conceptual representation of a set of elements. For example, for the indicator "energy consumption" – information bearing sentences may actually be reporting about usage of elements like gasoline/ petrol, diesel, Jet fuel, Bio Diesel, coal, LPG etc. Hence simple lexical or semantic similarities do not work.

Vector representation of words created from large corpora can be utilized to exploit conceptual similarity; we propose the use of Glove (Global Vectors) (Pennington, Socher and Manning 2014) for computing sentence similarity to an indicator in a novel way. Glove constitutes a single "word embedding" representation for each word or phrase in the vocabulary that is learnt by aggregating global word-word cooccurrence in a known corpus.

In our work, we created vector representation of the indicators using 100-dimensional Glove embedding for each word. These representations are then used to compute indicator-sentence similarity.

Generating indicator representation - Since indicators are usually multi-word phrases, we have used a stacked embedding representation. An indicator constituting 'w' words has a 'w+1' size stack representation where embeddings from 1 to wth layers correspond to the words in the indicator and the (w+1)th embedding is obtained by taking tuple-wise average of the constituent word embeddings.

Retrieving Relevant sentences for indicators – To retrieve relevant sentences for each indicator, we compute significance of indicator words in a page by using their term frequencies and distribution across pages. The indicator words have different presence in different pages. The significance for each indicator word $I_{w,}$ denoted by $\boldsymbol{\sigma}$ **(I_w),** determines the importance of a word for a page in the document. Treating each page as an individual document, I_w is computed as follows:

$$\sigma\ (I_w) = TF(I_w) * IPF(I_w)$$

where TF(I_w) is the frequency of indicator word (I_w) and $IPF(I_w)$ is the inverse page frequency of the indicator word (I_w) , calculated as :

$$IPF(I_w) = log\frac{N}{1+N_{I_w}}$$

Here, N denotes the total number of pages in a document and N_{I_w} is the number of pages containing indicator word I_w.

Each word in each sentence as well as each word in an indicator now has a stacked representation as well as a significance value associated to it, where the significance varies from page to page. The relevance of a sentence with respect to a given indicator is computed as a weighted function of cosine-similarity between the contained words.

Let n denote the number of words in a sentence, $S_{w(j)}$ denote the jth word of the sentence S and $sim(I_w, S_{w(j)})$ denote the cosine similarity between an indicator word I_w and a sentence word S_w. For each indicator word I_w, $maxSim$ function captures the maximum similarity between the word and the words of a sentence and is calculated as:

$$maxSim(I_w) = \max[sim(I_w\ , S_{w(j)})] \qquad j = 1\ to\ n$$

The final similarity score between the indicator and a sentence is calculated by multiplying the tf-ipf weights of indicator words to the corresponding *maxSim* score as given below.

$$\text{Score} = \frac{\sum_{i=1}^{a}[\ (\sigma(I_{w(i)}) * maxSim(I_{w(i)})] + maxSim(I_{w+1}))}{a+1}$$

where a is the number of words in indicator I , I_{w+1} is the (w+1)th embedding of indicator I.

It has been observed that few indicator words, which are rare in the repository, tend to have higher significance value. This high significance value can lead to higher score for a sentence even if the maxSim of these words are low. In order to reduce the effect of these rare but non-relevant indicator words on the final score, the significance σ (I_w) of an indicator word I_w is set to 1 if its maxSim score is less than 0.5.

Each sentence is then ranked for each indicator based on its similarity score with that indicator. Higher the score, better the rank. To extract ESG information from a document, the sentences may be explored in decreasing order of relevance.

Quantitative Information Extraction from ESG reports

We reiterate that our goal in this work is to sentences having more often than not, quantitative information about the indicators to aid the investment decisions. To achieve that, we define a set of sentence prioritization steps that can be applied to further reduce the set of possible sentences to be explored. These steps are designed to assess the potential of a conceptually relevant sentence to provide the quantitative information. The set of relevant sentences retrieved in the previous step are assigned a default priority of 0. The sequence of prioritization steps applied are as follows.

a) *Step 1*: Here we check for presence of quantifiable information in the sentence. Quantifiable information in a text can be identified using Named Entity Recognition. A named entity is a sequence of words that identifies some real-world entity like organization, person, place, number, date etc. In our case, we have utilized numeric named entities like money, percentage, cardinal, date, quantity and ordinal numbers. We have used SpaCy's Named Entity Recognizer (NER)[1] to locate and identify numeric named entities from the relevant sentences. Sentences containing quantifiable information are given priority 1.

b) *Step 2*: In this step, sentences with priority 1 are further screened with respect to presence of date named entity. Sentences other than those containing only date named entities but no other quantitative information are assigned priority 2. This ensures sentences having only numeric values and sentences having both date and numeric values are given higher priority. For example, the sentence, "*During 2018, training efforts focused on investigation and fraud reporting to ensure designated team members have the knowledge and resources to investigate potential fraud using standard rules of practice"*, will have priority 1 assigned by step 1 but it's priority will remain same after step 2 due to absence of any other numeric information than date.

[1] https://spacy.io

c) *Step 3*: Sentences with priority 2 are checked for presence of futuristic information. In this step, priority of sentences having mentions of future dates is reduced to 1. For an instance, after step 2, "*Investment in Communities Invest $200 million in the world by 2025"* – will get a priority of 2 but after current step its priority will be decreased to 1 due to presence of futuristic information.

The final ranking of retrieved sentences is obtained by first ordering them as per their priority and then within the same priority level by their respective scores.

Experiments and Evaluations

We conducted experiments on various publicly available sustainability reports provided by Global Reporting Initiative (GRI) organization (https://database.globalreporting.org/). These reports are available in pdf format only. A set of 50 reports have been taken for evaluating the performance of the proposed algorithm.

In order to evaluate the performance of the proposed algorithm, we have selected a set of 18 ESG indicators comprising of 11 quantifiable indicators and 7 non-quantifiable indicators. We have chosen a Mean Average precision (MAP) (Christopher, Prabhakar and Hinrich 2008) and Recall@k as the two metrics for evaluation.

MAP is one of the standard evaluation measures that provides a single-value for the quality of the retrieval system. If the set of relevant documents for an information need $q_j \in$ set of all queries Q is $\{d_1, d_2, ..., d_{mj}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k. then MAP is calculated as follows

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{K=1}^{m_j}Precision\left(R_{jk}\right)$$

Here in our case we are retrieving sentences instead of documents. Precision here is the fraction of retrieved sentences that are relevant and when a relevant sentence is not retrieved at all, the precision value in the above equation is taken to be 0.

Recall@k can be defined as the ratio of number of retrieved sentences @k that are relevant to the total number of relevant sentences. In our experiments, we observed that there are on average only 2-3 relevant sentences that contain the required information for each indicator in a report.

Table 3 shows sample retrieved sentences for a few indicators for FedEx. Table 4 and Table 5 show MAP and Recall values at different k for various quantifiable and non-quantifiable indicators respectively. The average values shown for individual indicators are calculated across all documents. It can be observed that for most of the quantifiable indicators, MAP value reduces as k increases from 5 to 20 which shows that we are retrieving most of the relevant sentences in the top 5 positions itself. Lower MAP value for indicators like 'investment in technology' is because most of the companies do not report the relevant quantitative information for these indicators in textual content. Rather, they use infographics to report their observations for such indicators and in this work; we are only analyzing the textual information.

Also, not all indicators are present in all reports; as we are considering diversity within companies. For example, FedEx is a transportation company, so it does not report any information related to 'water withdrawn by source'.

It can be observed from the tables that we are getting good recall scores, which is continuously increasing as we increase the number of retrieved sentences. We have observed that for most of the quantifiable indicators, Recall@10 is 1 which means that we are retrieving all the relevant sentences within top 10 retrieved sentences.

Table 3: Top 3 Retrieved sentences for company FedEx for some indicators

Indicator	Retrieved Sentences	Correct
Total Hazardous Waste Generated	In FY18, 78 percent of the solid waste generated in our operations was sent to recyclers, an increase of 7 percentage points over the previous year.	Yes
	We also diverted close to 22,500 metric tons of wood waste materials, including pallets, away from landfills in FY18.	Yes
	In FY18, one Prevention through Design effort recommended engineering controls to reduce potential hazards between load lane staging conveyors and large- package conveyors in hubs and stations.	No
GHG emissions or Carbon intensity	Including just our Scope 1 and 2 emissions as part of this intensity comparison would have shown a greater reduction of approximately 43 percent for the same period.	Yes
	From FYO9 through FY18, we've decreased CO, emissions intensity (on a revenue basis) by about 37 percent.	Yes
	Thanks to these collective efforts, we decreased CO, emissions intensity (on a revenue basis) by about 37 percent from FYO9 through FY18, a period when our revenue grew by 84.5 percent.	Yes
Hours of training	For management and non-management training hours, see the Data Appendix.	No
	Team members participated in an average of 19 hours of formal training in FY18, many of them through our online FedEx Learning Center.	Yes
	For example, we continued working with U.S. government leadership on unique approaches for identifying, training and hiring opportunity youth the 5 million young Americans outside the workforce.	No

Table 4 MAP@k and Recall@k for Quantifiable Indicators across all reports

Indicator	MAP @5	MAP @10	MAP @15	MAP @20	Recall@1	Recall@3	Recall@5	Recall@8	Recall@10
Energy Efficiency	0.41	0.39	0.39	0.37	0.10	0.60	0.85	0.89	0.96
Energy saved through initiatives	0.33	0.34	0.34	0.34	0.16	0.53	0.77	0.89	0.96
GHG emissions or Carbon intensity	0.45	0.42	0.41	0.41	0.08	0.36	0.60	0.89	0.97
Use of Recycled Material	0.48	0.46	0.46	0.46	0.55	0.77	0.88	0.95	1
Lost time in accidents	0.29	0.27	0.27	0.27	0.25	0.58	0.83	0.83	0.91
Hours of training	0.50	0.53	0.53	0.53	0.42	0.81	0.87	0.97	1
Amount of money allocated for employee welfare	0.43	0.43	0.43	0.43	0.4	0.6	1	1	1
Percentage of women board members	0.72	0.72	0.72	0.72	0.38	1	1	1	1
R&D Expenses	0.26	0.26	0.26	0.26	0.23	0.86	1	1	1
Water withdrawn by source	0.83	0.83	0.83	0.83	0.8	1	1	1	1
Waste by unit produced	0.48	0.48	0.48	0.48	0.19	0.58	0.66	0.95	1
Overall	**0.47**	**0.47**	**0.46**	**0.46**	**0.32**	**0.7**	**0.86**	**0.94**	**0.98**

Table 5 MAP@k and Recall@k for Non-quantifiable Indicators across all reports

Indicator	MAP@5	MAP@10	MAP@15	MAP@20	Recall@1	Recall@3	Recall@5	Recall@8	Recall@10
Environmental policy mentions	0.71	0.7	0.67	0.66	0.31	0.59	0.69	0.72	0.81
EH&S compliance in formal agreements	0.62	0.63	0.61	0.61	0.17	0.55	0.71	0.76	0.82
Compliance with Child labour prohibition	0.65	0.65	0.65	0.65	0.75	0.83	1	1	1
Adherence to Human rights policies	0.32	0.32	0.33	0.33	0.33	0.66	0.75	0.75	0.75
Supplier compliance to code of conduct	0.84	0.84	0.83	0.83	0.34	0.78	0.93	0.97	0.97
Adherence to Anti-corruption training	0.5	0.5	0.5	0.5	0.5	1	1	1	1
Mentions about Significant fines and sanctions	0.6	0.54	0.53	0.53	0.25	0.83	0.83	0.83	0.91
Overall	**0.6**	**0.59**	**0.59**	**0.58**	**0.38**	**0.75**	**0.84**	**0.86**	**0.89**

Related Work

Company sustainability reports have recently gained a lot of traction in proving information on factors capable of influencing business decision making. (Epstein 2018) A number of approaches have been proposed to exploit the unstructured information and retrieve relevant text for multiple domains such as biomedical, legal, scientific literature and finance (Chieze, Farzindar and Lapalme 2010) (Costantino and Coletti 2008) (Milward, et al. 2005) (Subramaniam, et al. 2003).

Karanikas (Karanikas, Tjortjis and Theodoulidis 2000) introduced TextMiner, a technique to mine information from financial documents using terms and events extracted from each document to find domain specific features and then used them for data extraction. Saggion (Saggion, et al. 2007) proposed an ontology-based information extraction and merging techniques to identify key pieces of knowledge from multiple sources to aid Business intelligence. The goal was to create a company intelligence tool for gathering corporate information with a utility towards business decision making.

Piskorski (Piskorski and Yangarber 2013) presented an overview of role of information extraction in finding factual information in free text. Facts in their work are defined as structured objects, such as database records, which can capture a real-world entity, event, occurrence, or state, with its arguments or actors. Information was typically sought in a target setting, e.g., corporate mergers and acquisitions. Authors in (Faloutsos and Oard 1998) discussed how to capture semantic information using natural language processing in information extraction techniques. Yuan et al. (Yuan, Liu and Yu 2006) described a pattern matching algorithm based on a tree model for extracting information from PDF files by parsing the PDF files and transforming the text into semi-structured information by injecting additional information tags. Kitamori (Kitamori, Sakai and Sakaji 2017) proposed a deep learning technique for extraction of sentences concerning business performance and economic forecast from summaries of financial statements. In their method, they first, selected sentences relating to prediction by using clue expressions. Second, the selected sentences were automatically classified into two prediction classes.

The closest to our work is financial entity extraction through candidate phrases (Kumar, et al. 2016). The authors developed a system that tags and extracts financial concepts along with corresponding numeric monetary values using machine learning and NLP techniques. However, financial terms are well defined as compared to ESG indicators and financial reports are in text and semi-structured format while ESG reports are unstructured in nature.

Conclusion and Future Work

In this work, we have presented methodology to extract ESG indicator information from sustainability reports that are primarily in non-standard format. The reports are first converted to image and then a block detection algorithm is applied to identify text blocks properly in order to facilitate correct OCR output. Subsequently, a word vector-based similarity scheme is applied to retrieve relevant sentences for each indicator. Our results show promising outputs as we are getting most of the relevant sentences in top 5 retrieval.

In future, we are working towards application of machine learning methods once sufficient training data is available. Also, identifying information from infographics presented in the reports is another future work dimension.

References

Bassen, Alexander, and Ana Maria Masha Kovacs. 2008. "Environmental, social and governance key performance indicators from a capital market perspective." *Zeitschrift für Wirtschafts-und Unternehmensethik* 182-192.

Bos, Jeroen. 2014. "Using ESG Factors for Equity Valuation." *CFA Institute Magazine, Volume 25 Issue 6.*

Chieze, Emmanuel, Atefeh Farzindar, and Guy Lapalme. 2010. "An automatic system for summarization and information extraction of legal information." In *Semantic Processing of Legal Texts,* 216-234. Springer.

Christopher, D. Manning, Raghavan Prabhakar, and Schütze Hinrich. 2008. "Introduction to information retrieval." *An Introduction To Information Retrieval* 151: 5.

Costantino, Marco, and Paolo Coletti. 2008. *Information extraction in finance.* Vol. 8. Wit Press.

Epstein, Marc J. 2018. *Making sustainability work: Best practices in managing and measuring corporate social, environmental and economic impacts.* Routledge.

Faloutsos, Christos, and Douglas W. Oard. 1998. "A survey of information retrieval and filtering methods." Tech. rep.

Karanikas, Haralampos, Christos Tjortjis, and Babis Theodoulidis. 2000. "An approach to text mining using information extraction." *Proceedings of Workshop of Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference.*

Kitamori, Shiori, Hiroyuki Sakai, and Hiroki Sakaji. 2017. "Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning." *2017 IEEE Symposium Series on Computational Intelligence (SSCI).* 1-7.

Kumar, Aman, Hassan Alam, Tina Werner, and Manan Vyas. 2016. "Experiments in Candidate Phrase Selection for Financial Named Entity Extraction-A Demo." *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations.* 45-48.

Milward, David, Marcus Bjäreland, William Hayes, Michelle Maxwell, Lisa Öberg, Nick Tilford, James Thomas, Roger Hale, Sylvia Knight, and Julie Barnes. 2005. "Ontology-based interactive information extraction from scientific abstracts." *International Journal of Genomics* (Hindawi) 6: 67-71.

Paliwal, Shubham and Sharma, Monika and Vig, Lovekesh. 2019. "TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images." *International Conference on Document Analysis and Recognition.* Sydney, Australia.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).* 1532-1543.

Piskorski, Jakub, and Roman Yangarber. 2013. "Information extraction: Past, present and future." In *Multi-source, multilingual information extraction and summarization*, 23-49. Springer.

Rahdari, Amir Hossein, and Ali Asghar Anvary Rostamy. 2015. "Designing a general set of sustainability indicators at the corporate level." *Journal of Cleaner Production* (Elsevier) 108: 757-771.

Saggion, Horacio, Adam Funk, Diana Maynard, and Kalina Bontcheva. 2007. "Ontology-based information extraction for business intelligence." In *The Semantic Web*, 843-856. Springer.

Smith, Ray. 2007. "An overview of the Tesseract OCR engine." *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007).* 629-633.

Subramaniam, L. Venkata, Sougata Mukherjea, Pankaj Kankar, Biplav Srivastava, Vishal S. Batra, Pasumarti V. Kamesam, and Ravi Kothari. 2003. "Information extraction from biomedical literature: methodology, evaluation and an application." *Proceedings of the twelfth international conference on Information and knowledge management.* 410-417.

Yuan, Fang, Bo Liu, and Ge Yu. 2006. "A study on information extraction from PDF files." In *Advances in Machine Learning and Cybernetics*, 258-267. Springer.

Zhou, Xinyu, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. "EAST: an efficient and accurate scene text detector." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 5551-5560.