ESG Data Collection with Adaptive AI

Francesco Visalli[1][a], Antonio Patrizio[1], Antonio Lanza[1], Prospero Papaleo[1], Anupam Nautiyal[1], Mariella Pupo[1], Umberto Scilinguo[1], Ermelinda Oro[2][b] and Massimo Ruffolo[1][c]

[1]*altilia.ai, Piazza Vermicelli, c/o Technest, University of Calabria, Rende (CS), 87036, Italy*

[2]*High Performance Computing and Networking Institute of the National Research Council (ICAR-CNR),*
Via Pietro Bucci 8/9C, Rende (CS), 87036, Italy

Keywords: Intelligent Document Processing, Environmental Social and Governance (ESG), Sustainable Investment, Socially Responsible Investment (SRI), Artificial Intelligence, Natural Language Processing, Large Language Models, Computer Vision, Information Retrieval, Deep Learning, Knowledge Graph, Workflow, Hyperautomation.

Abstract: The European Commission defines the sustainable finance as the process of taking Environmental, Social and Governance (ESG) considerations into account when making investment decisions, leading to more long-term investments in sustainable economic activities and projects. Banks, and other financial institutions, are increasingly incorporating data about ESG performances, with particular reference to risks posed by climate change, into their credit and investment portfolios evaluation methods. However, collecting the data related to ESG performances of corporate and businesses is still a difficult task. There exist no single source from which we can extract all the data. Furthermore, most important ESG data is in unstructured format, hence collecting it poses many technological and methodological challenges. In this paper we propose a method that addresses the ESG data collection problem based on AI-based approaches. We also present the implementation of the proposed method and discuss some experiments carried out on real world documents.

1 INTRODUCTION

Environmental, Social, and Governance (ESG) pillars describe areas that characterize a sustainable, responsible, or ethical investment. ESG investing has evolved in recent years to meet the demands of investors and public authorities that wish to better incorporate long-term financial risks and opportunities into their investment decision-making processes.

There is currently a clear challenge with the quality and consistency of ESG data. It is important to have standardized data, but only fragmented information is available from multiple sources. Data sources can be divided into two main sub-groups: primary and secondary data sources. By primary data sources, we intend the self-reported ESG data (company websites, annual and sustainability reports, etc.), third-party ESG data (NGO/government websites and reports), and real-time ESG signals (news, social media, company reviews, and so on). Secondary data sources are ESG data vendors (or ESG data providers), whose job consists in manually collecting, systematizing, and analyzing ESG attributes obtained from a primary data source. Primary data sources lock ESG data within strongly unstructured texts and documents, while secondary data sources are slow, not timely, and provide limited subsets of manually built ESG data for subsets of businesses.

There's also a lack of a standard taxonomy. ESG factors continue to evolve, and the dictionary changes as investors move through sectors and industries. ESG factors are difficult to reproduce over time and across geographies, partly because of differences in data across regions and in how the recording of data has evolved. Moreover, there may be discrepancies between what certain factors are expected to do and what they end up doing.

Traditionally, data was collected by a human analysts also responsible for pre-processing and analyzing the data. This process requires a lot of human capital and is very time-consuming, and the chances

[a] https://orcid.org/0000-0002-6768-3921
[b] https://orcid.org/0000-0002-5529-1007
[c] https://orcid.org/0000-0002-4094-4810

Visalli, F., Patrizio, A., Lanza, A., Papaleo, P., Nautiyal, A., Pupo, M., Scilinguo, U., Oro, E. and Ruffolo, M.
ESG Data Collection with Adaptive AI.
DOI: 10.5220/0011844500003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 1*, pages 468-475
ISBN: 978-989-758-648-4; ISSN: 2184-4992

that human analysts can make some mistakes while performing these tasks are very high. Advances in AI have made it easier than ever before to automate complex tasks at incredible speeds and volumes, thus revolutionizing how companies work with data. Artificial Intelligence (AI) and Intelligent Document Processing (IDP) have the power to extract, filter and structure crucial data that is used by rating agencies, business analysts, investors, etc. at scale. IDP uses AI technologies such as computer vision and language models rooted in deep learning to classify, categorize, extract, and validate the relevant information extracted from a variety of document formats.

In this paper, we propose an ESG data collection method grounded on Altilia Intelligent Automation (AIA) an AI-based Intelligent Document Processing (IDP) platform. The proposed ESG data collection method allows gathering various documents and contents from disparate sources including annual reports, sustainability report, notes to financial statements, news, NGO reports, company websites obtained by web scraping approaches, etc. AIA platform is grounded on an hybrid and adaptive AI paradigms that makes use of deep learning algorithms for Computer Vision (CV) and Natural Language Processing (NLP). In particular, we combine large language models, Human-In-The-Loop AI techniques, continuous learning, knowledge representation methods to implement machine reading comprehension techniques that turn unstructured documents into structured data and provide answers to many different ESG related questions. The main contributions of this work are the following:

- We describe a new approach to automatically gather ESG related documents/contents, analyze them and extract ESG data to use for creating structured company profiles.
- We show results of some experiments, carried out on real world documents, that show how banks, wealth management agencies, rating agencies, investors, and business analysts, can use the proposed IDP approach to perform ESG data collection and extraction for every company regardless the size and the industry.

The rest of this paper is organized as follows: section 2 presents the related work in the area of AI powered approaches for ESG investing and ESG data collection; section 3 describes ESG data collection problems we address in the paper; section 4 describes the Altilia Intelligent Automation platform and the proposed ESG data collection method implemented by the platform; section 5 presents some experiments performed over a dataset we have built to show the depth and breadth of the proposed approach; finally, section 6 concludes the paper.

2 RELATED WORK

In this section we provide a short summary regarding papers related to ESG with AI.

ESG data collection has really gone mainstream because of the growing relevance that ESG rating is gaining in the investment community. There are a growing number of ESG rating agencies and reporting frameworks, all of which have evolved to improve the transparency and the consistency of the ESG information that firms are reporting publicly.

(Hughes et al., 2021) describes how traditionally, ESG ratings have been developed by human research analysts following proprietary methodologies to analyze company disclosures, articles, and industry research among other sources to identify the ESG credentials of a company. Process underpinning analyst-driven ESG research, imbued with subjectivity during data analysis and rating generation and how within the last few years, developments in Artificial Intelligence and Machine Learning have led to creation of a new type of ESG rating provider; one that analyzes the ESG risks and opportunities of companies by collecting (or “scraping”) and analyzing unstructured data from internet sources using AI.

(Macpherson et al., 2021) discusses how Artificial Intelligence and FinTech-powered ESG screening and analysis solutions have become “strategic enablers” that can address some of the inherent ESG information biases and potentially even ESG rating divergences arising from corporate self-reporting, and annualised, backward looking reporting of information. In this study they discussed about implications of regulatory and industry expectations around ESG data and frameworks management, and AI-backed solutions to better manage and align ESG information sources, e.g. for issuer and controversies screening.

(Lee et al., 2022) describes how to analyze ESG data through ML methods including regression, classification, and anomaly detection methods for the dataset to perform these experiments. Their main task is to classify whether investors conducted excellent or bad investments, detecting anomaly data to prevent an adversarial attack, predicting the revenue based on ESG funds, and classifying the sentences suggesting a straightforward method for predicting their ESG scores.

In (de Franco et al., 2020), an ML algorithm was developed to identify the patterns, between ESG profile and financial performances for companies. The ML algorithm maps region into high dimensional

ESG features. The aggregated predictions are converted into scores which are used to screen investments for stocks with positive scores. This Machine Learning algorithm nonlinearly links ESG features with financial performance. It is an efficient stock screening tool that outperforms classic strategies, which screen stocks based on their ESG ratings.

(Gupta et al., 2021) provides a framework for conducting statistical analysis and leveraging ML techniques to gauge the importance of ESG parameters for investment decisions and how they affect financial performance of firms. For companies with the best ESG ratings, “return on equity” was found to be greater than rest of the companies. While using linear and random forest regression models, prediction accuracy of growth variables “profit margin” and “return on assets” increased when ESG data was used along with financial data as input. Companies having the highest “profit margins” were the ones having the best ESG ratings.

(Schultz and Tropmann-Frick, 2020) have developed a method for detecting unusual journal entries within individual financial accounts using autoencoder neural networks. A manually tagged list of entries is compared with identified journal entries. In the comparison, all analyzed financial accounts showed high F-scores and high recall.

(Rony et al., 2022) propose Climate Bot a machine reading comprehension system for question answering over documents that provides answers related to climate changes. The proposed Climate Bot makes available an interface for users to ask questions in natural language and get answers from reliable data sources. All the papers discussed above describe general AI methods adopted in the ESG investment practice. To the best of our knowledge, no work describes the application of advanced adaptive AI techniques to the extraction of information from ESG sources such as sustainability reports, annual reports, websites, and so on. This paper reflects one of the components in our platform where we provide the possibility to ask question related to ESG letting our AI-based platform to extract structured and pre-processed information before presenting it to the end user.

3 ESG DATA COLLECTION PROBLEMS

In this section we discuss how ESG information can be spread across different sources of information and the importance of mixing information that comes from heterogeneous sources.

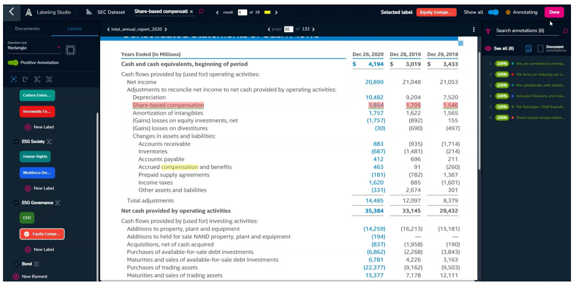
3.1 ESG Data Sources

ESG information can be disclosed through different types of documents such as non-financial and sustainability reports, as well as financial statements, annual and management reports. One of the main issues is the scanty availability of ESG data for small-cap or mid-cap companies. Large-cap companies, due to the availability of more resources and the fact that ESG practices are becoming mandatory for this type of businesses, provide ESG data through annual reports, sustainability reports, media reports, social media, news, etc. But for small-cap and mid-cap companies, due to limited resources and lack of standardization in ESG data disclosure, it’s hard to find ESG-related data.

There are two types of data sources: primary data source and secondary data source. The primary data source is available through the company website, annual report, proxy report, sustainability reports, corporate social report (CSR), news, social media, company reviews. This kind of data is typically available for free, and can in turn be classified into self-reported ESG data (all the data that the company itself discloses), third-party ESG data (such as NGO or government websites and reports) and real-time ESG signals (such as news, social media, company reviews etc.). The problem with self-reported ESG data is that it can be biased as companies can choose what type of information to disclose. On the other hand, if we rely only on real-time signals it could happen that, for example, a competitor spreads a misleading news about a company. So it is important to mix both self-reported data and real-time signals. A major challenge is the format of data available because every company has its way to present ESG information. Companies in different countries have different standards to disclose data. Due, to a lack of standards the data is present in different structures, for example, it could be present in text form, complex tables, or multifaceted data points. Another big challenge is that there exists no specific place to find all the ESG-related information for the particular company, all the data is fragmented and available in multiple places.

In secondary data source there are ESG data vendors or ESG data providers whose job consists in collecting, systematizing, and analyzing environmental, social, and governance attributes obtained from a primary data source. The final product of this type of ESG information providers are reports that contain analysis of a given company/sector equipped with some ESG scores.

The problem is that nowadays as ESG is becoming a hot topic and financial institutions and firms



Years Ended (In Millions)	Dec 26, 2020	Dec 28, 2019	Dec 29, 2018
Cash and cash equivalents, beginning of period	$ 4,194	$ 3,019	$ 3,433
Cash flows provided by (used for) operating activities:			
Net income	20,899	21,048	21,053
Adjustments to reconcile net income to net cash provided by operating activities:			
Depreciation	10,482	9,204	7,520
Share-based compensation	1,854	1,705	1,546
Amortization of intangibles	1,757	1,622	1,565
(Gains) losses on equity investments, net	(1,757)	(892)	155
(Gains) losses on divestitures	(30)	(690)	(497)
Changes in assets and liabilities:			
Accounts receivable	883	(935)	(1,714)
Inventories	(687)	(1,481)	(214)
Accounts payable	412	696	211
Accrued compensation and benefits	463	91	(260)
Prepaid supply agreements	(181)	(782)	1,367
Income taxes	1,620	885	(1,601)
Other assets and liabilities	(331)	2,674	301
Total adjustments	14,485	12,097	8,379
Net cash provided by operating activities	35,384	33,145	29,432
Cash flows provided by (used for) investing activities:			
Additions to property, plant and equipment	(14,259)	(16,213)	(15,181)
Additions to held for sale NAND property, plant and equipment	(194)	—	—
Acquisitions, net of cash acquired	(837)	(1,958)	(190)
Purchases of available-for-sale debt investments	(6,862)	(2,268)	(3,843)
Maturities and sales of available-for-sale debt investments	6,781	4,226	3,163
Purchases of trading assets	(22,377)	(9,162)	(9,503)
Maturities and sales of trading assets	15,377	7,178	12,111

Figure 1: Point-and-click document annotation.

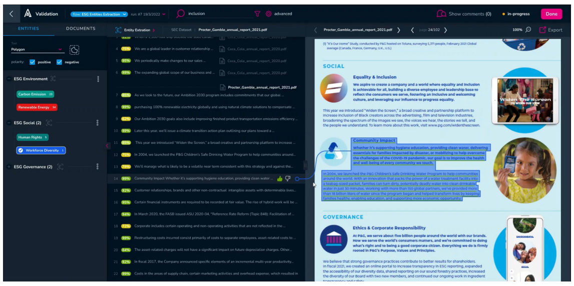


Figure 2: The visual inspection and review of a datapoint recognized within a document.

are heavily investing in getting ESG data and reports for rating companies. But primary data sources lock ESG data within strongly unstructured texts and documents, while secondary data sources are slow, not timely, provide limited sub-sets of manually built ESG data for subsets of businesses and, sometime, cannot be trusted due to a lack of transparency.

4 ALTILIA INTELLIGENT AUTOMATION PLATFORM

Altilia Intelligent Automation (AIA) is a platform designed with the unique goal to democratize the use of AI for Intelligent Document Processing (IDP). The platform enables the automation of business processes that require the understanding of complex documents and unstructured data sources. It makes use of AI techniques that allow the adaptation of its AI models to real-world changes by exploiting the human feedback. The AIA platform gives enterprises a no-code/low-code interface, in a cloud-based environment, allowing business domain experts to transfer their knowledge into algorithms by training AI models (Figure 1), combining models to create AI skills, and use AI skills within workflows to automate the extraction of relevant ESG data. The platform allows processing and understanding complex and visually rich documents, with variable and non-standardized layouts. This is crucial to automate processes that require the extraction and/or comparison of data buried in long in-depth reports (e.g. financial and annual reports).

4.1 ESG Data Collection by Altilia Intelligent Automation Platform

In order to allow a better understanding of the ESG data collection workflow built by using the AIA platform, we consider a running example based on the collection of datapoints from notes to financial statements, sustainability reports, and web sites. Datapoints considered in the running example are the following: (i) `Company Description` - it is a string describing what the company concretely does; (ii) `Ateco` - it is an entity representing the Italian code for the industry which the company belongs to; (iii) `Green Product` - it is a string that contains the name and/or the description of a product obtained by following organic procedures or low environmental impact; (iv) `Non-Renewable Energy` - it is a string describing how much of non-renewable energy the company uses for the production; (v) `Efficiency Initiatives` - it is a string that describes company initiatives to reduce the environmental impact of the production; (vi) `Environmental Certifications` - it is a an entity representing the name of the environmental certifications obtained by the company (for example ISO 14001).

In the following we explain step-by-step how we leveraged the AIA platform to build the end-to-end workflow, having the general structure depicted in figure 3, that allows collecting the ESG datapoints we have defined above.

Gathering. This step aims at gathering the documents to use for the ESG data collection. In our running example we used three different types of documents. Annual and sustainability reports were manually gathered, while company websites were automatically gathered by executing a data sources connector capable of applying web crawling and web scraping techniques. Such a connector works in two phases. In the first phase, it crawls the Web on the base of a list of companies where each company is equipped by its set of firmographic data such as: company name, vat number, headquarter address, etc. The scope of the Web crawling phase is to provide in output the websites of the companies in the input list. In the second phase, the connector applies web scraping techniques to gather contents of the web pages in the website. The output of this phase is a set of documents, one for each web page, ready to be ingested in the platform.

Ingestion. In this step the AIA platform applies document analysis and indexing methods. AIA takes in

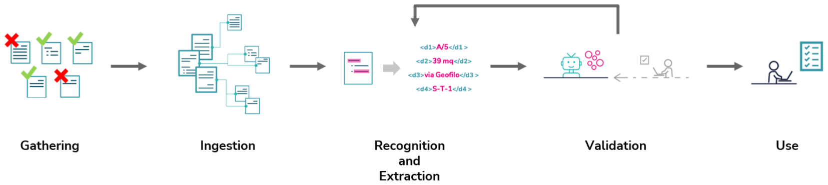


Figure 3: The ESG data collection process with Altilia Intelligent Automation platform.

input documents in PDF format and produces in output the Altilia Spatial Document Object Model (ASDOM), a spatial document format (patented by Altilia). More in details, document ingestion consists of two main phases: document analysis and document indexing.

In the first phase the platform applies: (i) intelligent optical character recognition (iOCR) algorithms to documents in image format; (ii) language detection and page orientation detection algorithms to enable better layout and texts recognition within the documents; (iii) document layout analysis and recognition algorithms to extract main layout elements. Furthermore, the OCR error correction technique defined in the paper (Nguyen et al., 2021) is applied. The output of this first phase is the document content turned into the ASDOM format. ASDOM allows turning document contents in machine readable format representing, in combined way, both textual contents and document layout elements such as: text paragraphs; tables and their sub elements like cells rows, columns, row headers, and column headers; text columns; charts; images; page headers and footers. ASDOM plays a twofold role, it supports document searching, filtering and visualization, and it simplifies Machine Learning based document processing workflows. In particular, ASDOM enables further document processing by machine learning models aimed at performing the concrete ESG data collection.

In the second phase the platform stores and indexes the ASDOM within the Altilia Knowledge Base allowing to jointly query and retrieve document contents and layout elements. More in detail, the platform applies text embedding algorithms like word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), BERT (Devlin et al., 2018), etc. to produce dense vector representations of the text contained in the documents. Dense vectors produced in this phase are linked with the layout elements in the ASDOM in order to create the final document representation stored, indexed, and managed by the Altilia Knowledge Base.

Recognition and Extraction. This step takes as input documents stored in ASDOM format within the knowledge base and applies deep learning algorithms to recognize and extract datapoints for which the algorithms have been trained to. More in detail, the datapoint recognition and extraction step works in two main phases: text retrieval and text reading.

Text retrieval phase is needed because we deal with documents that can count hundred of pages, hence to accurately extract a specific datapoint, we have first to identify elements of the layout that are candidate to contain the datapoint under analysis. We adopt two different methods to execute the retrieving phase. The first method is based on BM25-like algorithms (Kim and Gil, 2019; Amati,), the second makes use of dense passage retrieval search like the one discussed in (Karpukhin et al., 2020). BM25-like algorithms are based on queries composed by keywords and do not take into account the semantics of the datapoints. While dense passage retrieval allows retrieving text passages by queries composed of piece of texts that express the semantics of the datapoints.

The text reading phase depends on the complexity of the documents in input and on how much difficult are the datapoints. For very easy datapoints we use a pure syntactic approach base on Altilia Spatial Information Extraction Language (ASIEL) a rule-based spatial language, that has the expressiveness of a context free grammar enriched by a spatial algebra, exploiting the ASDOM to recognize and extract datapoints and objects from documents. For more complex and semantic datapoints, that need sense disambiguation, we use different types of LLMs (Large Language Models) and machine reading and comprehension algorithms to perform NLP task such as: token sequence classification, text classification, entity extraction, and question answering. In both cases we get very high accuracy and extraction performances. For example, for the Ateco datapoint we use ASIEL because extraction at token level is preferred, while for the datapoint Company Description it is required an extraction at sentence level, and because company descriptions can be very different each other, we need powerful text classification capabilities provided by LLMs.

The ground truth dataset is used to develop both the the text retriever and the text reader as discussed in paper (Ni et al., 2018). The fine-tuning of deep learning algorithms is performed by using Altilia Models a platform tool that assists in fine tuning LLMs. After fine-tuning the text retriever and the text reader algorithms, we combine them into an AI skill by means of the platform module named Altilia Skills. The AI skill is then used to configure a Workflow by the platform tool Altilia Workflow to make everything ready for the final datapoint collection process.

Validation. In this step, when accuracy values of extracted data are under given thresholds the platform sends a warning to the user that can inspect the data by the Altilia Reviews tool. This way users can check, validate and/or correct each single datapoint.

Review step (Figure 2) implements a Human-In-The-Loop AI approach (Wu et al., 2021) because after each validation action performed by domain experts, validated data becomes feedback stored into the ground truth dataset that can be used for the "continuous retraining/fine-tuning" of the AI models embedded in a workflow.

The continuous retraining takes place on demand or can be scheduled to be automatically executed by the platform. This way models improve while using the platform. It is worthwhile nothing that the human feedback continuously enrich the ground truth dataset with new examples coming from workflow executions, hence the platform can automatically maintain AI models up-to-date ensuring that AI models don't degrade over time, avoiding the data drift (model drift) phenomenon (Ackerman et al., 2020).

The validation step, and more in general the Altilia Reviews tool, helps in visualizing collected data for data quality check (Sheth and Thirunarayan, 2021) and explainable AI (Došilović et al., 2018) purposes. In particular, because we face long and complex documents such as annual and sustainability reports, a tool that allows exploring results of deep learning models and provide feedback is critical.

Our experience has taught us that such documents are difficult to label, many annotations are lost. In the context of ESG data we performed at least two rounds of Human-In-The-Loop interactions for each datapoint, each time the model brought up a large amount of "false" false positives (e.g. correct examples marked as "false positive" because escaped the human eye during the annotation phase). Hence, we extended the ground truth dataset improving significantly the final accuracy of the model.

Use. This last step of the workflow allows exporting extracted data towards third-party databases, applications, and tools by Connectors that are software artifacts allowing the platform interoperability. Connectors and platform APIs allow accessing the data stored in the Altilia Knowledge Base and export them in different formats such as XML, RDF, JSON, etc. Output connectors can directly feed external systems and applications or interacting with other RPA, CRM, CMS, ERP systems, etc.

5 EXPERIMENTS

In this section we describe experiments carried out in order to train AI Skills for the recognition of ESG datapoints. First we present the dataset and the annotations within it. Then, we describe in depth the experiments and the obtained results.

5.1 Dataset

We focused on three type of unstructured data sources: annual reports, sustainability reports and company websites. In particular, the numbers of documents took into account in the experiments are: 322 annual reports, 185 sustainability reports and 2495 web pages. Experiments involve around 2000 companies spread over various industry and with revenues ranging from 0-2 million to >1 billion.

Table 1 shows the number of annotations for each datapoint (rows) in each documents type (column). The ⟨datapoint, document⟩ pairs, shown in bold in the table, are the ones with a sufficient number of annotations to continue with the experimental phase.

Table 1: Number of annotations for each couple ⟨datapoint, document⟩.

	Sustainability Report	Annual Report	Web pages
Company Description	**242**	**592**	**645**
Ateco	2	**577**	-
Green Product	**371**	15	**146**
Non-Renewable Energy	**192**	1	-
Efficiency Initiatives	**274**	21	19
Environmental Certifications	**457**	**68**	**173**

5.2 Experimental Settings

In order to extract the 5 datapoints of interest, we created eight AI skills. Each skill listed here is capable of extracting a datapoint from a specific document layout: (i) `Company Description in text` composed by a syntactic retriever and neural reader; (ii) `Company Description in table` composed by a syntactic retriever and a neural reader; (iii) `Ateco` composed by a syntactic retriever and a syntactic reader; (iv) `Green Product` composed by a neural retriever and neural reader; (v) `Non-Renewable Energy in text` composed by a neural retriever; (vi) `Non-Renewable Energy in table` composed by a syntactic retriever; (vii) `Efficiency Initiatives` composed by a neural retriever and a neural reader; (viii) `Environmental`

Table 2: Results of applying AI skills to documents of interest.

Datapoint	Document Type	Precision Lenient (%)	Recall Lenient (%)	F1 Lenient (%)
Company Description - text	Annual Report	61.00	73.61	63.32
Company Description - table	Annual Report	97.00	99.00	97.99
Ateco	Annual Report	99.97	100.00	99.98
Green Product	Sustainability Report	77.50	96.50	84.30
Green Product	Web Page	35.71	100.00	52.63
Non-Renewable Energy - text	Sustainability Report	55.56	73.61	63.62
Non-Renewable Energy - table	Sustainability Report	71.00	89.00	78.99
Efficiency Initiatives	Sustainability Report	50.00	60.00	54.00
Environmental Certifications	Annual Report	40.02	98.31	57.89
Environmental Certifications	Sustainability Report	26.26	99.56	41.56
Environmental Certifications	Web Page	33.12	92.31	48.75

Certifications composed by a syntactic retriever and a syntactic reader. As regards the Company Description and Non-Renewable Energy datapoints, we created two different AI skills for handling data that can be found in tables and in text paragraphs. When we talk about syntactic retriever we refer to BM25-like algorithms (Kim and Gil, 2019; Amati,), instead when we say syntactic reader we refer to rules written in ASIEL. Finally, when we talk about neural retriever and neural reader we always refer to LLMs and Transformers-based algorithms (Vaswani et al., 2017).

All results presented in the next section have been obtained by 5-fold cross validation. Since some of the data are long text paragraphs we decided to use an F1-lenient as evaluation metric where true positives are defined as follow: span prediction ∩ span annotation AND at least 30% of words in common.

5.3 Results and Discussion

Table 2 shows the obtained results.

In general, we have obtained a good recall on all datapoints. It means that both syntactic and neural retrieval perform well. The worst - but still acceptable - datapoint in terms of recall is Efficiency Initiatives (60.00%), this is due to the fact that to capture this datapoint is very complex because of its semantic. The precision of some datapoints is instead below an acceptable threshold and will be the subject of future works.

When analyzing the results of these experiments, we have to take into consideration that we have worked with a view to having AI skills that generalizes on all document categories. It means, for example, that the same model leveraged for extracting Green Product from sustainability reports has been leveraged also for extracting the same datapoint from web pages. It explains why, in the first case, we have a precision of 77.50% and in the second case just 35.71%. These models are still far from generalizing on different document categories.

In contrast, the syntactic retriever performs very well on pipelines that involve extracting data from tables (Company Description and Non-Renewable Energy: 97.00% and 71.00% of recall, respectively). This is due to the fact that tables, by their nature, report information in the form of keywords.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we discussed how to build automatic workflows that collect environmental, social, and governance (ESG) data from different types of unstructured data sources and the challenges associated with ESG data collection. We extensively discussed how the Altilia Intelligent Automation platform and its modules are used to teach create AI Skills and workflows that automate ESG data collection from documents having different formats and layouts like: pdf documents, image documents, flat text, table structures, web pages etc. We presented experiments carried our to extract five datapoints leveraging eight AI skills making use of different data extraction techniques, based on a retriever-reader pipeline. Information of interest was extracted from different element of the layout such as text paragraphs and tables.

ESG data extraction capabilities of the AIA platform can be further improved, enhancing its document layout analysis and recognition algorithms. As future work we are extending the AIA platform to enable complete machine reading and comprehension methods providing full question answering features that allow extracting data from documents in a conversational manner.

Finally, we are working on extending the initial set of datapoints in order to cover the ESG taxonomy approved by the European Parliament that allows us to meet the ESG data collection needs of banks and other players in the European financial service arena.

ACKNOWLEDGEMENTS

This paper has been supported by the following projects:

- “ESG - Alternative data in credit management” realized in the context of the first call for project of the Fintech Milano Hub[1];
- “Validated Question Answering” n. F/190114/01/X44 - CUP: B28I20000040005 PON “I&C” 2014-2020 FESR - And for sustainable growth - Sustainable manufacturing DM 05.03.2018 - DD 20/11/2018, art. 38, 47 e 48 D.P.R. n. 445 of 28/12/2000.

REFERENCES

Ackerman, S., Farchi, E., Raz, O., Zalmanovici, M., and Dube, P. (2020). Detection of data drift and outliers affecting machine learning model performance over time.

Amati, G.

de Franco, C., Geissler, C., Margot, V., and Monnier, B. (2020). Esg investments: Filtering versus machine learning approaches.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215.

Gupta, A., Sharma, U., and Gupta, S. K. (2021). The role of esg in sustainable development: An analysis through the lens of machine learning. In *2021 IEEE International Humanitarian Technology Conference (IHTC)*, pages 1–5.

Hughes, A., Urban, M. A., and Wójcik, D. (2021). Alternative esg ratings: How technological innovation is reshaping sustainable investment. *Sustainability*, 13(6).

Karpukhin, V., Oguz, B., Min, S., Wu, L., Edunov, S., Chen, D., and Yih, W. (2020). Dense passage retrieval for open-domain question answering. *CoRR*, abs/2004.04906.

Kim, S.-W. and Gil, J.-M. (2019). Research paper classification systems based on tf-idf and lda schemes. *Human-centric Computing and Information Sciences*, 9(1):30.

Lee, O., Joo, H., Choi, H., and Cheon, M. (2022). Proposing an integrated approach to analyzing esg data via machine learning and deep learning algorithms. *Sustainability*, 14(14).

Macpherson, M., Gasperini, A., and Bosco, M. (2021). Artificial intelligence and fintech technologies for esg data and analysis (february 15, 2021).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Nguyen, Q.-D., Le, D.-A., Phan, N.-M., and Zelinka, I. (2021). Ocr error correction using correction patterns and self-organizing migrating algorithm. *Pattern Analysis and Applications*, 24(2):701–721.

Ni, J., Zhu, C., Chen, W., and McAuley, J. (2018). Learning to attend on essential terms: An enhanced retriever-reader model for open-domain question answering.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rony, M. R. A. H., Zuo, Y., Kovriguina, L., Teucher, R., and Lehmann, J. (2022). Climate bot: A machine reading comprehension system for climate change question answering. *IJCAI*.

Schultz, M. and Tropmann-Frick, M. (2020). Autoencoder neural networks versus external auditors: Detecting unusual journal entries in financial statement audits.

Sheth, A. P. and Thirunarayan, K. (2021). The inescapable duality of data and knowledge. *CoRR*, abs/2103.13520.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2021). A survey of human-in-the-loop for machine learning. *CoRR*, abs/2108.00941.

[1] https://www.bancaditalia.it/focus/milano-hub/