ASSIGNMENT

Create an LLM-based system that can automatically generate comprehensive ESG reports for companies and provide comparative analysis across industries.

[Divesh Yadav]	[Contact No.]
[2022mmb1376@iitrpr.ac.in]	[9811111492]

Table of Contents

1. Introduction

The project aims to develop an **LLM-based system** that automates the generation of **Environmental, Social, and Governance (ESG) reports** for companies, while also enabling comparative analysis across industries. The system involves fine-tuning a large language model to interpret unstructured financial and ESG data, extract key metrics, and generate detailed ESG reports. Additionally, a comparative analysis tool will be integrated, allowing users to compare the ESG performance of companies across industries through an interactive dashboard. The project emphasizes factual accuracy, data standardization, and identifying emerging ESG trends for reporting.

End Goal: The final product could be a software platform used by companies, investors, or regulators to generate ESG reports automatically, extract key metrics, and provide comparative insights across industries. The interactive dashboard would allow users to perform data-driven analysis on ESG metrics, identifying leaders, laggards, and emerging trends.

2. Related Research Papers

2.1. Advanced Unstructured Data Processing for ESG Reports: A Methodology for Structured Transformation and Enhanced Analysis

Link: “ https://arxiv.org/abs/2401.02992 ”

In the evolving field of corporate sustainability, analyzing unstructured Environmental, Social, and Governance (ESG) reports is a complex challenge due to their varied formats and intricate content. This study introduces an innovative methodology utilizing the "Unstructured Core Library", specifically tailored to address these challenges by transforming ESG reports into structured, analyzable formats. Our approach significantly advances the existing research by offering high-precision text cleaning, adept identification and extraction of text from images,

and standardization of tables within these reports. Emphasizing its capability to handle diverse data types, including text, images, and tables, the method adeptly manages the nuances of differing page layouts and report styles across industries. This research marks a substantial contribution to the fields of industrial ecology and corporate sustainability assessment, paving the way for the application of advanced NLP technologies and large language models in the analysis of corporate governance and sustainability. Our code is available at https://github.com/linancn/TianGong-AIUnstructure.git.

Innovative Methodology: The paper introduces the "Unstructured Core Library" to transform unstructured ESG reports into structured, analyzable formats.
High-Precision Techniques: It emphasizes text cleaning, identification and extraction from images, and table standardization for better data processing.
Handling Diverse Data: Capable of managing text, images, and tables from different industries and formats.

Contribution: This approach significantly advances corporate sustainability analysis using NLP and large language models (LLMs).

2.2. Building machine learning systems for automated ESG scoring

Link:“https://seco.risklab.ca/wp-content/uploads/2021/12/Building-Machine-Learning-Systems-for-ESG-Scoring.pdf ”

Key Takeaways:

1. The authors demonstrate the feasibility and advantages to applying state-of-the art Natural Language Processing (NLP) to identify ESG risks using social media data.
2. The authors discuss how advances in modern NLP can be leveraged to continuously build up algorithmic capabilities for processing ESG-relevant documents, by leveraging the capabilities of deep learning models for learning general representations of text data, which can then be applied across many tasks in the ESG domain.

3. The authors discuss results of NLP models can be used for creating aggregated ESG scores, as well design considerations for creating fully or semi-autonomous ESG scoring systems.

Proposed Approach: The paper introduces a method to convert unstructured text data into **ESG scores** using **NLP** techniques.
Deep Learning Application: The authors leverage **BERT**, a state-of-the-art NLP model, to assess the relevance and content of documents within the ESG context.
Practical Example: The approach is applied to **social media data** to demonstrate how it can enhance ESG scoring accuracy.
Relevance: This method aims to **automate ESG scoring** and **portfolio construction** using advanced NLP techniques.

2.3. ESG Data Collection with Adaptive AI

Link:"https://www.scitepress.org/Papers/2023/118445/118445.pdf "

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 longterm 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.

Sustainable Finance: Defined by the European Commission as incorporating ESG factors in investment decisions for long-term sustainable projects.

Challenges in ESG Data: Financial institutions face difficulty in collecting ESG performance data due to the unstructured format and lack of a unified source.

AI-Based Method: The paper proposes an AI approach to address these challenges by collecting ESG data efficiently.

Implementation and Experiments: It discusses the implementation and results of experiments on real-world documents.

3. Medium Articles, Blogs, and Tutorials

Blog: Automated ESG Reporting Powered by Gen AI

Practical Implementations: This blog discusses how **Generative AI (Gen AI)** automates ESG report generation. It focuses on utilizing AI to process large volumes of ESG data and automate compliance reporting.

Technical How-To Guides: The blog outlines the use of NLP models and automation pipelines for real-time ESG metric extraction, making it relevant for our project's goal of automating report generation and analysis.

Gap Analysis: From analyzing blog posts on ESG report automation, key challenges often arise around **data standardization** across multiple ESG reporting frameworks (such as GRI, SASB). Many implementations lack robust mechanisms for harmonizing metrics, leading to inconsistencies in cross-company analysis. Another gap is the **limited explainability** of AI-generated reports, as current tools fail to provide clear reasoning behind specific ESG scores or metrics. These limitations can be addressed in our project by focusing on **improving standardization methods** and **enhancing transparency** in AI-generated reports, enabling more accurate cross-industry comparisons.

ESG Reporting Software:

https://www.ibm.com/products/envizi/esg-reporting?utm_content=SRCWW&p1=Search&p4=43700079220252664&p5=p&p9=58700008354102315&gclid=CjwKCAjwx4O4BhAnEiwA42SbVD23ppwvsvtIfp-rPkxWBhpXNhwvQKrZY3639m5xYdbVYd65BzsVCBoCr5AQAvD_BwE&gclsrc=aw.ds

IBM Envizi ESG Reporting Software is a platform designed to help organizations manage and streamline their Environmental, Social, and Governance (ESG) data collection, reporting, and analysis. It allows businesses to gather data from multiple sources, track ESG metrics, and generate compliance reports. The software integrates with various systems to automate data collection, ensuring accuracy and reducing manual work. It also provides insights into ESG performance, enabling users to monitor sustainability targets and meet regulatory requirements. The platform supports decision-making by offering data visualization and benchmarking capabilities.

Articles: “The Role of AI in ESG Reporting: A Technical Perspective**”**

Link:“https://www.linkedin.com/pulse/role-ai-esg-reporting-technical-perspective-momin-ali-u1cbf/ ”

Practical Implementations: This post emphasizes the use of **AI** to automate the ESG reporting process, focusing on Natural Language Processing (NLP) for unstructured data and AI models to streamline compliance and sustainability reports.

Technical How-To Guides: The article highlights leveraging **AI-based analytics** for real-time ESG metrics extraction, data visualization, and risk management, with examples of integrating AI models into existing frameworks.

Gap Analysis: One limitation noted is that while AI tools help, there's often a lack of standardization across various ESG frameworks. Addressing this in our project will require developing methods to harmonize data from different reporting structures for more consistent output.

4. Github Repositories:

Repository: LLaMA-Factory

- **Summary**: LLaMA-Factory is a framework designed for the efficient fine-tuning of large language models. It offers tools for streamlined model training, hyperparameter optimization, and integration with various datasets, enhancing adaptability for specific applications.
- **Relevance**: This repository is highly relevant for developing an LLM-based ESG reporting system, as it allows for efficient customization of models to better suit ESG-specific tasks.
- **Key Features**:

  ❖ Streamlined training processes with hyperparameter optimization tools.

  ❖ Support for integrating diverse datasets to enhance model performance.

- **Next Steps**: Utilize LLaMA-Factory to fine-tune models with ESG-related data, optimizing for specific reporting requirements

Repository: explosion/spaCy

- **Summary**: SpaCy is a popular NLP library used for named entity recognition (NER), dependency parsing, and text classification. It also supports training custom models for specific domains.
- **Relevance**: SpaCy can be employed for extracting key ESG metrics and entities from unstructured company filings, news articles, and sustainability reports. This will facilitate the creation of a knowledge graph representing ESG metrics.
- **Key Features**:
  - Robust entity extraction capabilities that can be adapted for ESG-specific entities.
  - Easily integrates with other Python libraries like Neo4j for building knowledge graphs.
  - Efficient processing for large text datasets, making it suitable for analyzing extensive ESG reports.
- **Next Steps**: Fine-tune SpaCy's NER model to detect and extract ESG-related entities like "carbon

emissions," "governance structure," or "social impact metrics."

4.1. Existing Solutions

Repository: Neo4j Graph Data Science

- **Summary**: Neo4j’s Graph Data Science (GDS) library focuses on advanced graph algorithms for machine learning. It provides tools for running graph analytics, machine learning workflows, and graph-based data integration.
- **Technology Stack**: Built using **Java** for the core, **Python** for integrations, and **Cypher** for querying graph databases.
- **Code Structure**: The repository includes APIs for graph analytics, node embedding models, and integrations with Neo4j databases
- **Relevance**: For this project, Neo4j can be used to store ESG metrics and their relationships in a knowledge graph. This will allow more structured queries and ensure factual consistency in generated ESG reports..

- **Next Steps**: Utilize Neo4j GDS for building knowledge graphs from ESG data and applying graph-based analytics to discover insights from ESG reports.

Repository: Elasticsearch

- **Summary**: Elasticsearch is a distributed search and analytics engine used for full-text searches and complex queries across large datasets.
- **Technology Stack**: Built primarily in **Java** with REST APIs, it leverages **Lucene** for indexing and supports integrations with Kibana for visualization.
- **Code Structure**: The repository follows a modular design with core components handling indexing, queries, and cluster management, alongside REST API endpoints for communication.
- **Relevance**: Elasticsearch can be integrated into the ESG reporting system to retrieve and analyze ESG data from various sources like regulatory filings and news articles in real-time. Its powerful search and filtering capabilities will allow for industry-wide ESG comparisons.
- **Next Steps**: Use Elasticsearch for indexing ESG reports, enabling efficient querying, and powering interactive dashboards for ESG metric comparisons.

5. Existing Tools and APIs:

Best Financial Market Data for LLMs : https://medium.datadriveninvestor.com/best-financial-market-data-for-llms-cacad36f359a

Refinitiv Eikon API :

The Refinitiv Eikon API is a powerful tool offering access to a vast array of financial data and analytics. Designed for professional use, it provides comprehensive coverage of global markets, including equities, commodities, foreign exchange, and fixed income. The API integrates seamlessly with the Eikon desktop application, allowing for advanced data manipulation and analysis.

Key Features: Refinitiv Eikon API delivers real-time market data, historical data, and in-depth analytics. Users can access detailed financial statements, economic indicators, news, and research reports. The API also supports extensive data visualization tools and has robust integration capabilities with various programming environments such as Python, R, and Excel.

Pricing: Refinitiv Eikon operates on a subscription-based model, with pricing tailored to the specific needs and scale of the user:

Basic Plan: $22,000 per year, providing access to real-time data, historical data, and basic analytics.

Advanced Plan: $24,000 per year, including additional features such as advanced analytics, news feeds, and in-depth research reports.

Premium Plan: Custom pricing, offering unlimited API calls, extensive customer support, and access to exclusive data sets and analytics tools.

The pricing includes a set number of API calls, with additional calls available at an extra cost. The plans are designed to cater to different levels of data needs, from individual analysts to large financial institutions.

Advantages: The Refinitiv Eikon API offers unparalleled data accuracy and depth, making it a preferred choice for professional financial analysts and institutions. Its integration with the Eikon desktop application enhances its functionality, allowing for seamless data manipulation and analysis. The extensive range of data, combined with powerful analytics and visualization tools, sets it apart from other financial APIs.

5.1. Libraries and Frameworks:

To develop an LLM-based system for generating ESG reports and enabling comparative analysis, here’s a breakdown of key libraries:

1. SpaCy

Use: Excellent for Named Entity Recognition (NER) and natural language processing (NLP) tasks.

Best Fit: Extracting ESG-related entities (e.g., "carbon emissions", "corporate governance") from text.

Pros: Fast, scalable, customizable.

2. Hugging Face Transformers

Use: Fine-tuning pre-trained models like GPT, T5 for text generation.

Best Fit: ESG report generation using domain-specific data.

Pros: Large repository of models, easy fine-tuning.

3. Neo4j

Use: Knowledge graph construction and management.

Best Fit: Structuring relationships between ESG data points and companies, enabling advanced querying.

Pros: Ideal for graph-based relationships and real-time queries.

4. D3.js

Use: Data visualization, especially for dashboards.

Best Fit: Interactive dashboard creation for comparing ESG performance across industries.

Pros: Great for custom, dynamic visualizations.

Best Combination: Hugging Face for fine-tuning the LLM for ESG report generation, SpaCy for extracting ESG-specific entities, Neo4j for building a knowledge graph for comparison, and D3.js for visualizing the ESG metrics on an interactive dashboard.

5.2. Open-Source Tools vs Proprietary Solutions:

By using open-source tools (e.g., Hugging Face Transformers, SpaCy, Neo4j, and D3.js), it’s possible to replicate many features offered by proprietary solutions while enabling more innovation. Custom fine-tuning of language models, building knowledge graphs for ESG metrics, and generating interactive dashboards can extend capabilities beyond existing proprietary tools. Moreover, open-source solutions offer better flexibility for adapting to emerging ESG trends and generating company-specific insights, making them ideal for scalable, innovative solutions without the high costs of proprietary systems.

Key Advantages of Open-Source Solutions:

- **Customization:** Open-source frameworks allow developers to tailor models and processes to specific needs, such as fine-tuning LLMs for ESG-specific reports.
- **Cost-Effectiveness:** Proprietary solutions like Refinitiv and Bloomberg come with high subscription fees, while open-source tools are often free or have lower costs for deployment and development.
- **Scalability and Flexibility:** With open-source tools, it's easier to integrate multiple data sources and adjust algorithms to suit evolving ESG standards and requirements.
- **Challenges:** Replicating the reliability, data coverage, and support that proprietary tools provide can be challenging. Integrating real-time data sources, managing large datasets, and ensuring compliance with reporting frameworks

are critical aspects where proprietary tools have an advantage. However, with proper development, it’s possible to build open-source systems that can innovate and surpass the capabilities of closed systems.

6. Data Sources and Datasets:

6.1. Publicly Available Datasets

To train, test, and validate the LLM for generating ESG reports, curating relevant datasets is crucial. Key datasets include:

Financial Data: From sources like World Bank, Refinitiv, or Yahoo Finance.

ESG Metrics: Sources include MSCI ESG Ratings, Sustainalytics, and Refinitiv ESG Database, offering company-level ESG scores.

Company Regulatory Filings: SEC EDGAR or Global Reporting Initiative (GRI) databases provide access to sustainability reports and financial disclosures.

These datasets help in training the model for financial analysis and ESG scoring.

6.2. Synthetic Data Generation

If access to real-world data is limited due to privacy or availability constraints, synthetic data generation techniques can be employed. Generative models like GANs (Generative Adversarial Networks) can produce synthetic financial and ESG data that mirrors real-world data distributions. This allows the model to be trained on diverse data without breaching confidentiality regulations. Techniques such as:

Statistical Data Simulation: Using probabilistic distributions of existing data.

Data Augmentation: Modifying available data to create additional training instances

6.3. Data Licensing

Carefully managing dataset licensing is critical to avoid legal or financial risks. Common licensing issues arise with proprietary datasets (e.g., Bloomberg, Refinitiv) or subscription-based ESG databases. Publicly available datasets from sources like the World Bank are typically open for use but may have restrictions on commercial exploitation.

For datasets that are commercially available, ensure compliance with licensing terms by reviewing:

Permitted Use: Whether the data can be used for training AI models or for commercial purposes.

Attribution Requirements: Some datasets require acknowledgment or referencing.

Choosing appropriate datasets while ensuring legal compliance will allow for smoother implementation and scalability of your system.