SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

EcoLens: An AI-Powered ESG and Carbon Footprint Analyzer

Dr. Meera Narvekar[1], Neeharika Bhaide[2], Nilay Rathod[3], Parth Das[4], Pooja Divekar[5]

[1]Professor and HoD (Computer Engineering), Department/ Institution: Computer Engineering/ Dwarkadas Jivanlal Sanghvi College of Engineering, City/State: Mumbai/ Maharashtra, E-mail: Meera.narvekar@djsce.ac.in, ORCID ID:0000-0003-4602-4094

[2]Student, Department/ Institution: Computer Engineering/ Dwarkadas Jivanlal Sanghvi College of Engineering, City/State: Mumbai/ Maharashtra, E-mail: bneeha20@gmail.com, ORCID ID: https://orcid.org/0009-0006-7722-3749

[3]Student, Department/ Institution: Computer Engineering/ Dwarkadas Jivanlal Sanghvi College of Engineering, City/State: Mumbai/ Maharashtra, E-mail: nilayrathod1303@gmail.com, ORCID ID: https://orcid.org/0009-0007-6783-9678

[4]Student, Department/ Institution: Computer Engineering/ Dwarkadas Jivanlal Sanghvi College of Engineering, City/State: Mumbai/ Maharashtra, E-mail: dasparthhh24@gmail.com, ORCID ID: https://orcid.org/0009-0009-1879-4922

[5]Student, Department/ Institution: Computer Engineering/ Dwarkadas Jivanlal Sanghvi College of Engineering, City/State: Mumbai/ Maharashtra, E-mail: divekarpooja.456@gmail.com, ORCID ID: https://orcid.org/0009-0006-2697-2061

ABSTRACT

ESG metrics are now widely regarded as key measures of a company's long-term viability, its risk, and its corporate governance. Traditionally, ESG metrics were primarily derived from disclosures that are sufficiently static in nature (i.e., self-reported and operating on a retrospective basis), but also lack cooperation, consistency, and conviction. Similarly, the inability to utilize carbon emissions data in the context of broader corporate sustainability metrics often leads to an inaccurate assessment of emissions levels. In this paper, we will outline the creation of EcoLens, an AI-powered analytics platform which can provide companies with actionable insights into their carbon emissions and ESG scores. The EcoLens system combines both structured data (e.g., corporate financials) and unstructured data (e.g., sustainability reports, regulatory filings, news alerts) to generate real-time reports on ESG ratings, critical carbon emissions information, and interpretable indicators of corporate risk using advanced an-alytical techniques such as Natural Language Processing (NLP), machine learning, and Explainable AI. The implementation of EcoLens will provide a level of transparency, timeliness, and informed decision-making that will ultimately improve the quality of reporting for all stakeholders involved with a company, including investors, regulators, and the companies themselves.

Keywords: ESG Analytics, Carbon Footprint, Machine Learning, Natural Language Processing, Explainable AI, Sus-tainability

I. Introduction

With the increasing emphasis on corporate accountability, the role of Environmental, Social and Governance (ESG) criteria in the development of business strategy and investment decision making has become a primary area of focus for many businesses [8]. Countries around the globe are also beginning to implement more stringent legal and regulatory frameworks mandating that companies provide investors with comprehen-sive sustainability disclosure [3]. The increase in attention towards ESG criteria has been paralleled by a greater desire on the part of investors for credible and auditable evidence that the companies in which they invest are conducting business in a responsible manner.

There are, however, still many limitations surrounding the evaluation and disclosure of ESG performance.

The current practice for evaluating ESG performance is through the use of annual or biennial sustainability reports. These reports are useful sources of information, however, by the time they are reviewed/analysed by investors and analysts they are frequently outdated. In addition, because most of the data produced through these mechanisms is generated on a voluntary basis, the manner in which companies report their data may be inconsistent and/or incomplete. [9].

There are significant limitations in how carbon emissions are reported. Scope 1 emissions (direct emissions from a company's operations) have been tracked in a more advanced manner than Scope 2 (indirect emissions from purchased energy) and Scope 3 (indirect emissions from upstream and downstream activities in a company's value chain), as there is no standardised process to track these types of emissions and they can account for over 70% of a company's total environmental impact at times [7].

The investing community faces many of the same issues as the corporate world in terms of ESG (environmental, social and governance) ratings. ESG rating systems are currently up-dated infrequently, often following very unclear methodologies and, as such, it can be difficult to determine the basis on which an ESG rating was assigned [9]. Furthermore, ESG ratings do not provide investors with real-time risk information on emerging issues such as environmental issues, labour disputes and governance failures. Additionally, because standardised ESG ratings are not based on the unique ethical frameworks under which individual investors make investment decisions, investors are left without any transparency regarding the ap-propriateness of ESG ratings for their respective investments.

EcoLens presents a novel way to provide accountability for carbon emissions as well as up-to-date ESG data on a single platform to solve many of the historical problems associated with these two concepts. EcoLens leverages its ability to read, process, and generate up-to-date insights about corporations based on both unstructured and structured datasets from corporate filings and media, as well as through the use of structured data such as sustainability metrics (i.e., the carbon footprint), and unstructured data such as real-time news articles [5].

EcoLens' three primary contributions to the ESG and Carbon Accounting worlds are: (1) the ability to automatically extract sustainability-related metrics from varying formats of documentation for a holistic view of a corporation's carbon impact by scope; (2) the ability to provide a sentiment-based scoring mechanism that accounts for time-based weighting of real world news signals; and (3) a dual interface system that provides custom analytics tailored to both the corporate sustainability officer and institutional investors.

II. PROBLEM STATEMENT

There are numerous challenges of the current ESG evaluation frameworks: [9] :

- The amount of self-reported and static reporting is too great
- There is no ability to detect ESG risk in real-time
- The explanation for ESG scoring is limited
- The treatment of carbon emissions or other ESG indicators as independent of each other, and insufficient adaptation to regional or sectoral differences in ESG criteria.

These issues have created an overall ineffectiveness of ESG analysis and delaying sustainable decision-making.

III. OBJECTIVES

The goals of EcoLens are:

- Develop an AI analysis tool to assist with real-time ESG analytics.
- Implement a method for assessing Carbon Footprint as part of the score [7].
- Utilize Natural Language Processing (NLP) to gather signals on an organization's ESG reputation derived from unstructured text [5].
- Provide clear and concise data regarding the organizations ESG-based performance.
- Facilitate investments aligned with sustainability and also regulatory compliance [3].

IV. Related Work

A. ESG Performance and Corporate Value

Recent scholarly studies are utilizing more advanced and higher quality methods in analysing the effects of providing evidence to ascertain the link between ERG performance and corporate valuation [8], [10]. As a consequence, stud-ies have been published that utilise two-way fixed effects analysis and/or panel regression analysis as a method of analysis. One of the results from studies that make use of two-way fixed effects modelling techniques was that they found "green technological innovation" acts as a moderating variable connecting ESG scores with corporate market value [10]. These studies conclude that companies with high ESG performance have historically allocated greater amounts of investment to sustainable research and development and have improved operational efficiencies (performance) in addition to increasing their positive market image and value.

Using propensity score matching (PSM) methods to create comparative studies of like firms before and after interven-tions, and using difference-in-differences (diff-in-diff) analysis allows for improvements in the identification of endogeneity-related issues, thereby improving the rigor of causal assess-ments by comparing like firms across multiple time periods. Although those improvements have been made, there remain significant gaps in the literature concerning SME contributions to ESG performance (i.e., SMEs comprise the majority of companies in supply chains), and SMEs remain an underserved population among ESG scholars.

B. Global Financial Modeling and ESG Integration

Researchers gathering data for ESG indices can be very difficult because some of the underlying constructs cannot be directly observed. Therefore, the recommended methodology for analysing these multiple differing and complex relation-ships across different industries has evolved to be Partial Least Squares Structural Equation Modelling. This can be even more helpful to researchers when the amount of differences in the regulatory environments varies widely from one country to another.

Another limitation of this existing body of literature is temporal scope, in that many researchers conduct studies using cross-sectional data (to analyse one-time performance of the International Corporate Sustainability Development) without any consideration of longitudinal data and/or previous trends or time intervals where ESG controversies have occurred. Additionally, variations amongst regulatory systems (e.g., the EU Sustainable Finance Disclosure Regulation and the Indian Business Responsibility and Sustainable

Development Re-porting Framework) also create difficulties when researchers attempt to create an assessment system that can be used universally.

C. Machine Learning Applications in Sustainable Finance

Natural language processing will allow for identification and verification of greenwashing using machine learning al-gorithms to automate analysis of news articles, social media communications, and other (such as external) documents that describe greenwashing as evidenced by nonconformity be-tween disclosed intentions and actual behaviour [12], [13]. Automating the process of verifying corporate sustainability claims (ESG data) addresses one of the key weaknesses of relying on self-reported data.

A significant limitation to this process is the widespread variation across ESG rating systems, sometimes referred to as aggregate confusion. The aggregate correlation among ESG rating systems across data providers is often less than 54 percent (compared to 99+ percent correlation among credit rating agencies) [9]. The reason for this divergence is attributed to discrepancies in attribute count, differences in how each attribute is measured, and differences in the manner of how attributes were calculated.

Multiple approaches exist for improving the accuracy of ESG rating validity (and thereby expanding their use). Some of these approaches may include explainable AI frameworks [6] to enhance the validity of ESG ratings, as well as ensem-ble learning algorithms (e.g., Random Forests and Gradient Boosting) that aggregate multiple non-valid signals to produce a single valid ESG rating and provide clear explanations of how a rating was generated. There are very few examples of how the approaches discussed above have been implemented into operational systems to date.

D. Carbon Footprinting Methodologies

There has been a significant change in the way that EIAs are conducted. In addition to estimating direct emissions, we must estimate the emissions from all value chains for the product/service, including Scope 3 emissions that occur between the time raw materials are sourced until after the product has been disposed of. One of the biggest challenges of estimating Scope 3 emissions is obtaining accurate data on the entire life cycle of a product, in order to include those emissions in the total estimates of a company's greenhouse gas emissions. One possible methodology to estimate Scope 3 emissions is to use an input/output model, combined with machine learning, to estimate Scope 3 emissions when primary data cannot be obtained.

The Greenhouse Gas Protocol is the basis for the methods used to account for carbon emissions as well as many of the current Environmental, Social, and Corporate Governance (ESG) standards; therefore, there is a significant difference between what is required to be verified or reported versus what is available for reporting or verification of carbon emissions in the supply chain.

The literature reviewed indicates that there is a clear trend towards using dynamic, AI driven methods and/or techniques for predicting "real-time" emissions as opposed to static subjectively driven methods and/or techniques for predicting "real-time" emissions; the research objective is to develop a platform that provides both "real-time" sentiment analysis and "real-time" fully loaded carbon accounting systems.

V. SYSTEM ARCHITECTURE

EcoLens uses a modular and scalable system design. There are four layers of data collection, processing, storage, and visualization within its architecture, which separate concerns and provide flexibility, scalability, and maintainability, allow-ing each layer of computation to scale independently based on work load demand, without affecting the operational integrity of other layers in the system. The modular nature of ESG analytics is critical to handling a variety of data types coming from various sources in real time [12], [13].

The architecture contains two main operating environments with different workflows but with a shared data processing structure (back bone for making the two environments work together). This provides continuity in the way the ESG metrics and analytical methodologies are evaluated across the plat-form.

The EcoLens framework has four key layers:

A. Data Ingestion Layer

This layer collects a variety of diverse and different types of data needed to calculate ESG and carbon footprint; this includes:

- Corporate sustainability and annual reports (both of these are ESG disclosures that follow industry standard guide-lines like MSCI and Sustainalytics [1], [2]).

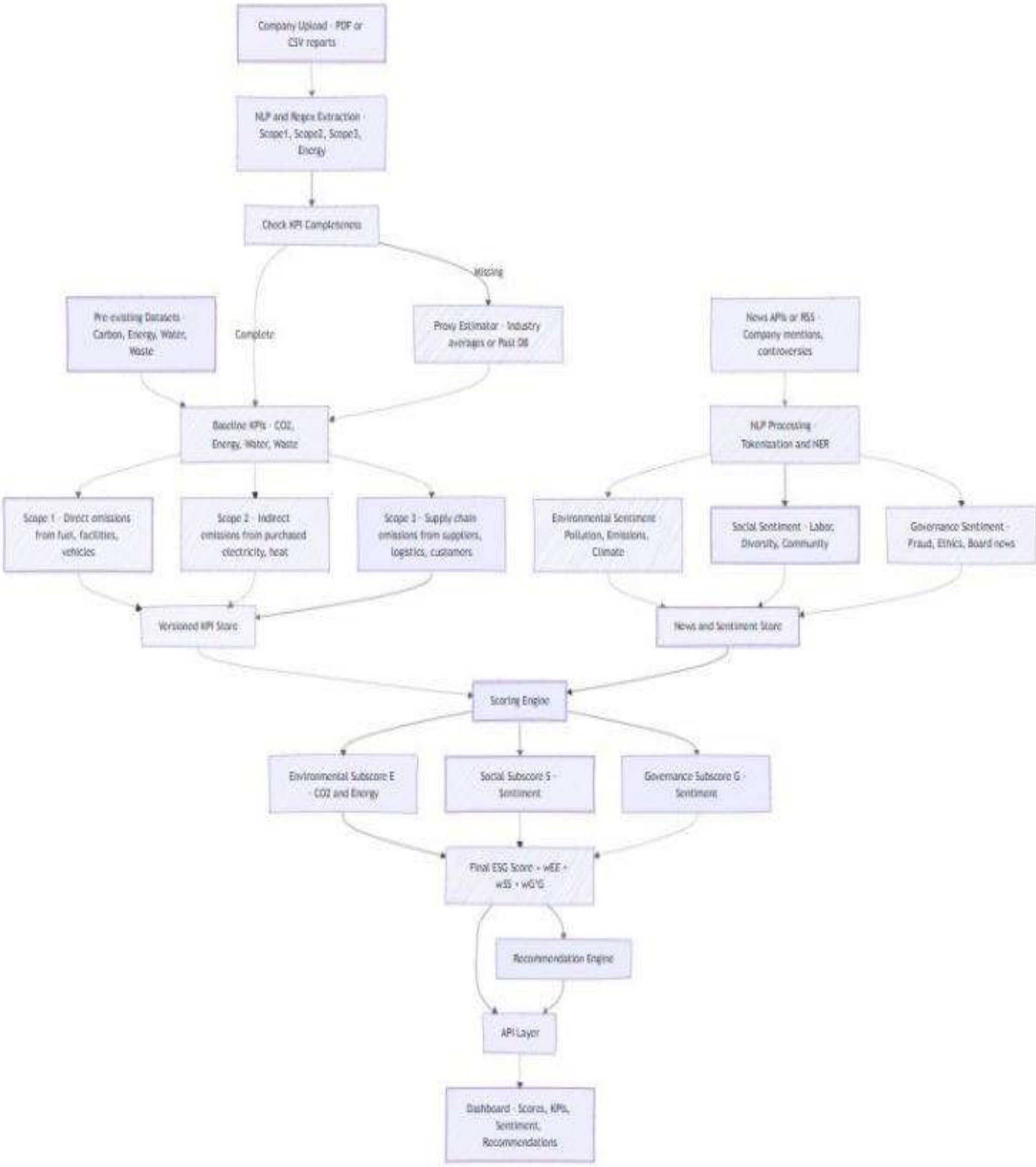

Fig. 1. EcoLens System Architecture

- News articles from around the world, as well as press re-leases, which provide a way to detect real-time sentiment and controversies.
- Government and regulatory databases based on guidelines or frameworks such as the OECD [3]
- Data such as carbon emissions and energy consumption based on standards like the GHG Protocol [7]

The aggregation of the data from many different sources creates a more balanced and holistic view of a company's ESG and environmental performance.

B. Data Processing Layer

The raw data collected in the previous layer will be preprocessed (e.g., cleaned, normalized and transformed) extensively to prepare it for use, with the goal of enhancing the data's quality and consistency. When the data is in a structured format, it will be standardized to allow for interoperability across multiple sources; this will help eliminate known issues related to the differences in ratings from the various ESG rating providers [9].

Sophisticated Natural Language Processing (NLP) tech-niques will be used to preprocess textual unstructured data and will consist of:

- Tokenizing and Lemmatizing
- Performing Named Entity Recognition (NER)
- Performing Sentiment Analysis

For example, transformer models such as BERT will be used to perform contextualization and semanticization of the text related to finance and sustainability [5]. These preprocessing activities are essential in enabling effective future analytical and modeling activities.

C. Analytics and Modeling Layer

The purpose of this layer is to provide actionable insights obtained from data that has been processed through machine learning and deep learning methods. The key functional pieces of this layer include the following:

- Estimating ESG score predictions based on multi-factor model approaches using recognized methodologies [1], [2].
- Analyzing sentiment or unusual patterns in order to identify ESG risk and controversy.
- Providing a detailed estimate of carbon footprint in accordance with applicable climate science metrics [4].
- Classifying ESG themes in order to inform sector specific insights.

Furthermore, XAI (explainable artificial intelligence) meth-ods, such as SHAP, will be used to make models more explainable and accountable [6]. This aspect is particularly important when it comes to making financial and sustainable decisions that require accountability for those who make them.

D. Visualization and Reporting Layer

The main objective of this layer is to provide insights that can be easily visualized and interactively explored through user-friendly and intuitive dashboards. The following is a high-level summary of some of the major features of these dashboards:

- Graphs showing ESG trends over time and comparisons of multiple companies to each other.
- Detailed breakdowns of emissions based on their source and category.
- Alerts for real-time ESG risk and controversy.
- Automated reports for stakeholders.

The use of these visualizations allows stakeholders to use data-driven insights to confidently make decisions that support their goals of sustainable investment strategies and account-ability to their stakeholders [8], [11].

VI. Methodology

The EcoLens methodology consists of:

A. Data Gathering and Preprocessing

This process collects information from many different sources asynchronously. Corporate sustainability reports are collected in the form of either PDF or CSV (comma-separated values). The reports contain a combination of structured and semi-structured text. Global news feeds are also collected in real-time on an ongoing basis via the GDELT knowledge graph, which collects media coverage from all over the world, with updates provided every fifteen (15) minutes.

To prepare the data for analysis, we will begin by nor-malizing the formats of the different documents. The PDF documents will be converted to a more usable digital format (e.g., an MS Word document) which will remove the structure of the PDF document (e.g., tables, headings, etc.) and create a new digital copy of the text. During this process, all text from PDF documents will be cleaned to remove OCR artifacts from the original document but retain its original meaning.

Data in CSV format must first be validated to ensure that it is in the correct schema before proceeding. In addition, we must ensure that any records in the CSV file that do not have data have been accounted for.

Cleaning the news data is more challenging than cleaning the GDELT; however, it still requires a similar set of processes, such as attaching the requisite data fields and getting rid of data encoding problems via standardising all delimiters used and extracting required metadata. In addition, in order to facilitate the publication date provided, sentiment score, and semantic topic classification associated with each news article for the purposes of reporting: GDELT has readied some positive and negative emotion scoring methods, which will further allow us to apply a categorization and scoring system to the news articles.

B. Natural Language Processing Pipeline

By processing through the natural language processing (NLP) pipeline, the company disclosures and news articles will have two separate NLP workflows, optimizing them to suit their specific structures [5]. Sustainability reporting uses regular expressions to determine specific sustainability report-ing metrics, e.g., greenhouse gas emissions, energy used, water used, and waste generated, using the appropriate reporting nomenclature and by listing variations of the reporting and nomenclature across the various companies and levels or frameworks of reporting.

Named Entity Recognition is the process that will be used to identify and label company names, locations, organisational structures, and other entities referenced within the document. This will allow for the correct metrics to be attributed to the identified company and will allow for the disambiguation of entities from multiple company references in news articles. Dependency Parsing will extract the relationship of the iden-tified entities with their numeric values.

Sentiment analysis will be performed on news articles, and sentiment models will be trained in either the financial or the sustainability domain. Following sentiment analysis, an article will have an output score based upon whether the sentiment expressed towards the event or entity referenced is negative or positive. Each article will also be coded to one of the three environmental, social, or governance (ESG) topic codes using a content analysis approach rather than by performing a simple keyword search.

C. Carbon Footprint Analysis

There are numerous ways in which one can measure greenhouse gas emissions as defined by a widely accepted international standard of classification by scope. The three scopes of measurement relate to the different categories of greenhouse gas emissions, as follows: (1) Scope 1 refers to direct emissions that occur from the burning of fossil fuels through owned or controlled facilities and vehicles that an organization operates. (2) Scope 2 refers to indirect emissions associated with the electricity, steam, and heat or cooling purchased from a third party and consumed by a facility (i.e., not consumed through the burning of fossil fuels). (3) Scope 3 refers to all other indirect greenhouse gas emissions associated with the entire value chain (i.e., from the acquisition of purchased goods or services to the disposal or treatment of the organization's products following the sale of those products).

To identify emissions in each scope, a series of extraction algorithms were developed to assist with identifying reported values for each scope of measurement found in sustainability reports and other sustainability documents prepared for each of these scopes. Where no reportable (i.e., published) value was available, proxy values (i.e. an estimation) were calculated for those sectors by using benchmarks specific to the sector and extrapolating from historical data [4]. All estimations of greenhouse gas emissions will have a flag indicating they are estimations for the purpose of being transparent about the source of the data.

The emissions from numerous sources are allocated to the various working stages of an organization's supply chain as well as to the organization itself so that hotspots can be detected throughout the entire organization and supply chain. In addition to the continuous assessment of the various emis-sions associated with each work stage along the supply chain, emissions sources that are similar are grouped together with the use of clustering algorithms based on machine learning and then ranked based on their overall contributions to the organization's total emissions footprint. This kind of analysis will be instrumental in ongoing development of specific rec-ommendations for mitigating emissions from similar sources or work stages within a supply chain that would otherwise go undetected.

D. ESG Sentiment Scoring Framework

The basic concept is a scoring system that accepts raw signals from news articles as inputs and produces normalized ESG (Environmental, Social, and Governance) scores as out-puts via certain processing steps [1], [2], [9].

The first processing step is called entity resolution and involves the use of fuzzy matching techniques to compare and validate news references to organizations against the official list of publicly-traded companies to identify how entities referenced in news articles relate to entities analyzed in the investment universe. This process allows you to remove a significant amount of noise from your ESG analytical data before performing any future investments analyses because only mentions of the organization that can be tied to publicly-traded companies will be included in your ESG analysis dataset.

The second step in processing articles involves using a Se-mantic Mapping technique to identify which of the three ESG dimensions are referenced in each article. Articles covering Environmental topics will be classified using four different classifications: Climate Change; Pollution; Resource Manage-ment; and Biodiversity. Articles covering Social issues will be classified using five different classifications: Labour; Human Rights; Diversity and Inclusion; Community; and Product Safety. Articles covering Governance issues will be classi-fied using four different classifications: Executive Compensa-tion; Board Structuring; Ethics/Compliance; and Shareholder Rights.

The third component of the methodology is called sentiment weighting, which takes into account the tone of the article and its semantic relationship to the article and provides an article-level ESG signal [12]. An article can also serve as a contributor to multiple dimensions if it contains more than one topic or subject matter; the degree of contribution will be determined by semantic analysis.

The fourth component is temporal aggregation, where an exponential decay method will be used so that more recent events will have a greater impact on their scale than older events. The decay parameter will be set so that an event that occurred three months prior to the current event will contribute approximately 20% to the weight of the current event.

The last step is Normalizing. To facilitate ESG performance comparisons across all organizations (irrespective of size / sector / frequency in reporting on ESG metrics), raw scores will be normalized on a scale of 0-100 (using min-max scaling) using data from all companies in the universe of companies included in our study.

E. Recommendation Engine

The recommendation engine generates recommendations for how best to implement changes based on the integration of carbon analysis and ESG assessment [13]. For the purposes of carbon footprint reduction, the recommendation engine uses optimization algorithms to analyze different intervention options based on expected ROI and the complexity associated with executing an intervention. The recommendations gener-ated will be grouped into five categories: supplier selection, logistics optimization, energy procurement, facility improve-ment, and operational improvement.

To make improvements to ESG components of companies, recommendations will be generated by comparing the orga-nization with peers in the same sector and assessing their performance concerning other variables [8], [11]. Areas where there are deficiencies will be prioritized based on either the material risk associated with the deficiency or the potential for a significant competitive advantage through implementing the improvement. The above text outlines recommendations, the relevant impacts of each recommendation, and how they would be prioritized against a given organization. Thus, after thorough scrutiny of the current options, the estimated total impact, es-timated total resources for implementation, and the supporting evidence will be provided with each recommendation for ease of reference.

Vii. Datasets used

The system utilizes:

- Corporate ESG and sustainability reports
- Global Database of Events, Language, and Tone (GDELT)

Viii. Implementation details

EcoLens is implemented by utilizing Python and is aug-mented by Pandas and NumPy for data manipulation, Scikit-learn and TensorFlow for machine learning algorithms, spaCy and transformer models for natural language processing mod-els (NLP), and MongoDB for persistent data storage and scalable data systems (across multiple nodes). The interface utilized is a Flask-based web application that allows the end-user to interact with EcoLens [12], [13].

A. Technology Stack

The entire system is developed in Python and is capable of processing large quantities of data while handling complex operations. FastAPI is used for the creation of the web service (RESTFUL), providing the capability of responding to multi-ple simultaneous user requests very quickly. To manage the use of data and to retain data, MySQL is used for data storage and relational database management systems (RDBMS) to provide efficient transactions while also providing useful relationship management between all the related tables of data.

Analyzing our provided text by providing options of existing natural language processing (NLP) tools. A tool called spaCy helps us achieve those goals by performing tokenization (split-ting our text into words), determining the intuition for part of speech tagging (understanding what nouns, adjectives, and verbs are), and identifying important features of the words that are found within our text (for example, the names of companies or the location of companies).

Using scikit-learn, we can implement machine learning methods (i.e., Classification, Prediction, Grouping) to classify, make predictions, and group our data. Furthermore, we can utilize regular expressions (RegEx) to help us locate and extract patterns or elements from our structured or semi-structured text.

We have developed the user interface using React.js, which allows us to build user-friendly interfaces that require mini-mal maintenance due to their component-based approach. To present our data to users visually, we have established D3.js and Plotly as tools to create user-friendly graphics (charts and graphs) so that the user will be able to explore the data more thoroughly and better understand it. Finally, we utilized Tailwind CSS to create visually appealing layouts and consistent style for the layout of the user interface.

B. Data Storage Strategy

Corporate sustainability data is stored in a versioning stor-age framework to allow both retaining prior history, and thereby to facilitate any updates to that prior history. The system assigns each report that is uploaded onto it a unique identifier, and also provides metadata containing the time that the report was uploaded. Archived metrics are stored, along with a wide range of provenance information including reference documents and confidence levels in the extraction process.

All news that is stored in the database has a historical record because of the append-only nature of the architecture. All news that is stored has gone through a deduplication process to prevent the storage of duplicate copies of the same article; yet, all unique signals are preserved. Query performance for historic review and real-time updates is maximized through the indexing strategies used to establish the database structure and indexes [9].

The master datasets in the database are validated on a periodic basis to provide consistency and to identify potential anomalies. Safety inspections check distributions of scores for adherence to acceptable limits and provide experience for determining whether there are expected correlations between measures that are related.

C. Security Implementation

OAuth2.0 and JSON Web Tokens provide the authentication basis and control of sessions; and role based access control identifies the type of user that is accessing documents. Additionally, there are multiple levels of permission that allow complete granularity for separating the rights that corporate employees have to access specific functions/data from the rights that investors have to access those same functions/data.

D. Performance Optimization

This system's design is focused on maximum performance. It remains quick and responsive, even with numerous concurrent users [12]. Asynchronous processing is performed on long-running jobs, such as documents being parsed and analysis done on news articles. Long-term jobs do not impede or delay faster user interaction.

All heavy calculations are assigned to available resources in a task queue. These resource assignments are re-configurable based on dynamic workload.

The system uses caching to reduce repeated calculations, thus improving overall efficiency. All computed scores are stored in our datastore and re-used; access to these scores occur only if a change has occurred in the underlying data. The use of indexing and selective denormalization improves the database's ability to perform given its most frequently executed queries.

At the front-end, performance has improved through loading only what is necessary when loading the application for the first time; additional components will be loaded based on whether they are needed and therefore can be "on demand" loaded. Large amounts of data may require that virtualization be used in order to only display data that would appear in the viewport of the application so that the application feels very smooth and responsive to users.

Ix. Eveluation metrics and results

Performance is evaluated using:

- Accuracy, Precision, Recall, and F1-score
- Mean Absolute Error for emission estimation
- ROC-AUC for ESG risk classification

A. ESG Metric Extraction Pipeline

An 11-layer transformer-based pipeline presented in this work extracts automatically a total of 11 ESG metrics from unstructured corporate sustainability reports (available as PDF formats) — Scope 1/2/3 emissions, Energy Consumption, Water Usage, Waste Generated, ESG Score and the corresponding Environmental/Social/Governance Scores and Carbon emissions. At the Data Preparation stage, raw synthetic ESG corpora are transformed into BIO-tagged NER sequences, classification samples, and bases for relation extraction along with document section labels (Environmental, Social and Governance) and synthetic non-ESG negative cases to balance classes. Datasets are stratified into a 70/15/15 train/validation/test splits. Three transformer architectures are trained in succession: a DistilBERT binary ESG relevance filter, a BERT token classifier that performs NER over 23 BIO labels, and finally, a BERT sequence classifier that maps metric spans to 11 target classes — all fine-tuned with weighted F1 optimization and early stopping through the Hugging Face Trainer API. During inference time, PDFs go through bi-phasic extraction of structured tables and free-form text, where the TableReconstructor leverages metric-specific regex patterns, as well as a multi-stage fallback strategy that loosens its constraints successively if strict extraction returns 0 results. The outputs are scored for confidence by the provenance of the source — context on a table with a total-row (0.90), only a table (0.70) and paragraph fallback (0.50) — as well as validated in canonical unit maps and plausibility bounds for each metric specifically.) Robustness is tackled by injecting both five categories of real-world PDF noise such as OCR errors and encoding artifacts, as well as table formatting noise into the training data at a rate of 30%. System evaluation includes layer-wise F1, confidence calibration using Expected Calibration Error (ECE), and end-to-end accuracy over metric identification, value extraction, unit normalization, and validation aiming for ¿82% full-pipeline accuracy.

B. ESG Sentiment Analysis Performance

1) Fuzzy String Matching based Entity Resolution: To reliably map unstructured news articles to proper corporate entities, our approach leverages a solid entity resolution framework based on fuzzy string matching. In actual media, news outlets often mention corporates with different spellings, informal abbreviations or short names and thus exact character matching will not work. To tackle this, we employed a Token Set Ratio algorithm which determined the lexical similarity of the raw company mentions extracted from news sources, against that of the official National Stock Exchange (NSE) registry. To ensure high data integrity, we combined a stringent confidence threshold of 85% with the requirement for two-word lexical overlap. This focused method also was successful in closing the gap between informal news reporting and sys-tematic corporate datasets to eliminate noise, where fewer than 1% of all companies ended up receiving non-false-positive hits only those greatly weighted towards perfect ESG scoring.

2) Temporal Weighting by Exponential Time Decay: Given the dynamic nature of corporate sustainability profiles, recent events contribute to a more accurate representation of a company's current ESG profile than are aging historical occurrences. To accommodate this decay in relevance over time, we added a continuous exponential time-decay function to our scoring pipeline. Instead of equally treating all chronological data in the framework, we apply a computed decay rate = 0.05 according to the day-age of each news article. In this mathematical model, a breaking news event has an impact weight of 100%, while in the same way, the influence of a ten-day-old article is naturally diminished to around 61%. The temporal weighting mechanism is vital, as it aligns our final aggregated ESG scores closely to real-time institution behaviours—encouraging companies ahead of the curve with sustainable practices whilst hedging out the statistical noise that close scandals or redeem nonvalue added milestones.

C. System Performance Metrics

Simulations of production workload were performed to evaluate the overall performance of the system. During tests, the platform was able to analyze sustainability reports of about 50 pages in under a minute (60-90 seconds). This included extracting data from the PDF, deciding which ESG metrics were important, and putting that information in a database. The data pipeline has been carefully orchestrated to run once per day at a fixed scheduled time with a rolling windows of 24 hours for the ESG scores generated to ensure they are in line with current market realities. As an instance, if the automated framework is activated at 12:30 PM on a Wednesday, the system starts a backward-looking retrieval all the way back to 12:30 PM exactly one week prior (Tuesday). As the ongoing Global Database of Events, Language and Tone (GDELT) project publishes real-time global news updates on rigid fifteen-minute intervals, this 24-hour analytical envelope precisely collates 96 separate, high-density CSV files. After the data has been securely downloaded, this large data set is processed through a series of steps in the computational pipeline—through raw data extraction, fuzzy entity resolution based on NLP, thematic mapping of the data, and finally calculation of ESG scores. Because so much global news is parsed over this time frame, the end-to-end execution of our pipeline takes roughly 2.5 to 3 hours. Choosing to run this framework at a regular hour each day is the right methodological middle ground since it keeps dataset timeliness exceptionally high and still tackles the rather intensive computational operations of the system architecture.

X. APPLICATIONS

EcoLens supports:

- Sustainable investment analysis
- Regulatory compliance monitoring
- Corporate ESG strategy development
- ESG risk management

XI. Discussion

A. Platform Capabilities and Limitations

Carbon accounting and the analysis of ESG sentiment might, as shown in EcoLens, be integrated into one platform used by both corporate sustainability teams and institutional investors [12], [13]. The extraction of relevant ESG information is automated which saves time and labor, while significantly improving structured metric accuracy. Furthermore, traditional ESG ratings can be greatly insightful compared with it through the real-time news analysis by which the platform acts as an indicator of recent changes of ESG actions in addition [9].

The platform is also fairly limited, though. It mainly bases on sustainability and news reports published on the Internet, which is why quality of analysis relies on companies ESG activities reporting quality and their media coverage level. Companies with little or no publicity cannot be appraised as such. Similarly, it is not straightforward to analyze a private company that does not publish extensive sustainability reports. Scope 3 emission is also hard to assess since it covers complex supply chains and lack of data [4], [7].The system estimates these emissions based on proxy values that it would still be useful but not as accurateas data that had been gathered from the companies themselves. In contrast with this, the platform makes clearly established what estimations are done so that reading it is easy to users during decision making process. Despite its general performance, the sentiment analysis model is not a perfectly accurate one [5]. It can confuse more complex language, sarcasm or ambiguous implications in news articles. In addition to this, news coverage can also be biased in itself and it is able to impact the perception of a company with regard to its ESG performance. This problem is alleviated by cross-referencing multiple news sources, but even then there may still be a significant level of bias.

B. Comparison with Existing Approaches

Many traditional ESG rating agencies update their ratings once/twice per year and these updated ratings generally come months after the publication of sustainability reports [1], [2]. Unlike the static data of Google Trends, EcoLens updates its information instantaneously whenever new data becomes available, thus delivering much more timely insights. The second advantage is that EcoLens has a clear and easily understandable explainable scorecard, in contrast to most commercial ESG rating systems whose scorecards do not clearly indicate on what basis they are computed [9].

Besides Carbon accounting systems are specifically focused on emission monitoring not other sustainability indicators. The EcoLens does not end with this though as it provides a hybrid between carbon and environmental, social and governance (ESG) factors usable for quantifying both risk and opportunity which would give users the ability to have a more holistic view of the sustainability activity of an organization [4], [7]. The most prevalent portfolio screening applications are based solely on third-party ESG ratings, with limited customization [8]. Sustainability data is being filtered for examples of ethical filtering by EcoLens which gives the investor the ability to set their own criteria based on their values or investment priorities rather than solely relying on standard ESG filters.

C. Practical Implications

The platform supports organizations in enhancing their sustainability reporting from a compliance measurement to an ongoing performance tracking activity [10]. With the ability to retrieve real-time data, companies now know where they have emissions and can monitor and resolve issues before they are reported. Additionally, through the use of AI-generated recommendations, organizations can identify effective sustainability performance improvement options based on data analysis. [12].

Investors utilizing this platform are able to monitor ESG risk on a real-time basis, allowing them to track instances of corporate scandals, environmental catastrophes, and governance failures in real-time [13]. Furthermore, every investor using this platform can develop their own evaluation methodology for their specific investment strategies based on the consideration of ESG risks and the exigency of financial performance in relation to their risk tolerance.

Another key attribute of the platform is its methodology that is clearly defined; it openly shows how data is collected, processed, and converted to resulting ESG scores. This transparency facilitates confidence in the outcome. Many traditional ESG rating systems have been accused of being black-box systems with little to no explanation of how they calculate ESG scores [9].

XII. FUTURE SCOPE

There are several things that could be used to further develop the platform in the future. For example, predictive models that could be used to calculate future ESG performance based on current trends and historical data are one improvement that would assist both organisations and investors in understanding their risks and sustainable future. Sector-based benchmarks could provide companies with an opportunity to compare themselves against their competitors, which would improve accuracy and make comparisons more meaningful.

Another option to improve the platform is to integrate IoT devices into the platform that would allow for real-time emissions tracking. Installing smart meters and environmental sensors would allow for automatic collection of emissions data and help reduce the amount of time and resources required to perform manual reports.

There are a couple of major areas of change in all probability related to the manual reporting process Another significant aspect of the proposed extension will be the regulatory compliance aspect that will have the ability to assist businesses with preparing sustainability disclosures required under many global standards. For example, the system could potentially automatically map internal environmental, social, and governance (ESG) related information to the standard reporting frameworks (such as TCFD, SASB, GRI), therefore fulfilling these requirements in a more efficient manner and accelerating the reporting process. Additionally, enhanced supply chain analysis may be included as part of the extension on the platform in the future, allowing organizations to trace emissions from numerous levels within the supply chain of suppliers and identify who is responsible for contributing to the Scope 3 emissions. Additionally, utilizing supply chain data from chain suppliers would also improve the accuracy and completeness of the analysis.

XIII. Conclusion

Despite certain drawbacks of the system, in particular, regarding the availability of the data and the accuracy of estimations, it has a firm basis supporting the use of sophisticated ESG analytics systems. The tools that can unify various data sets with sophisticated analysis techniques will be even more valuable as the concept of sustainability gains relevance in the corporate strategy and investment decision-making. Additional features that can be implemented on the platform in the upcoming times include predictive ESG analysis, industry-specific benchmarking, and regulatory support. The architecture that is designed during this study gives it an adaptable and scalable foundation on which these future improvements will be carried out.

This system has its own limitations, especially in terms of data availability, and estimates are difficult to calculate using available forms. Having said that, there is an important rationale for developing highly sophisticated analytics systems to assess ESG data. Over time, systems that incorporate different types of data with advanced analytical processes will become even more useful, as sustainability becomes an increasingly important component of the way companies conduct their business and make investment decisions.

New features that could be implemented on the system within a short period of time include predictive ESG analytics, industry-specific benchmarks, and support for regulatory compliance. The system architecture that was developed as part of this research provides an adaptive and scalable platform upon which these new features can be developed in the future.

XIV. ACKNOWLEDGMENT

The authors wish to thank Dr. Meera Narvekar for her support and advice during the course of this research project. We also appreciate the Department of Computer Engineering at Dwarkadas J. Sanghvi College of Engineering for providing the resources and infrastructure necessary for this project.

References

1. MSCI, "MSCI ESG Ratings Methodology," MSCI Inc., 2023.
2. Sustainalytics, "ESG Risk Ratings Methodology," Morningstar, 2022.
3. OECD, "OECD Guidelines for Multinational Enterprises," OECD Publishing, 2021.
4. Intergovernmental Panel on Climate Change, "AR6 Climate Change 2023: Synthesis Report," IPCC, 2023.
5. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
6. S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
7. GHG Protocol, "Corporate Accounting and Reporting Standard," World Resources Institute and World Business Council for Sustainable Development, 2015.
8. G. Friede, T. Busch, and A. Bassen, "ESG and financial performance: Aggregated evidence from more than 2000 empirical studies," Journal of Sustainable Finance & Investment, vol. 5, no. 4, pp. 210–233, 2015.
9. F. Berg, J. F. Ko¨lbel, and R. Rigobon, "Aggregate confusion: The divergence of ESG ratings," Review of Finance, vol. 26, no. 6, pp. 1315–1344, 2022.
10. X. Zhang, Y. Li, and H. Chen, "ESG performance, green technology innovation, and corporate value: Evidence from industrial listed companies," Journal of Cleaner Production, vol. 312, pp. 127–145, 2021.
11. M. K. Sharma and L. Thomas, "ESG impact on corporate financial strategy of energy and utilities companies worldwide," Energy Economics, vol. 108, pp. 105–120, 2022.
12. R. Gupta and S. Mehta, "AI in sustainable finance: Identifying green investments through machine learning," IEEE Access, vol. 10, pp. 98765–98778, 2022.
13. A. Kumar, P. Singh, and D. Rao, "Sustainable artificial intelligence in finance: Impact of ESG factors," Expert Systems with Applications, vol. 198, pp. 116–134, 2022.