ESGenius

ESG Data-Driven Decision Support System

Project Plan

Supervisor: Dr. Cheng Reynold

Group Members:

Ko Man Sing 3035778216

Ng Tsz Wai Andrew 3035778383

Chan Cheuk Hei 3035786328

Chau Chun Yiu 3035787657

Submission Date: 1st October 2024

Department of Computer Science, The University of Hong Kong

Table of Contents

1. Introduction

1.1. Project Background

In 2004, the term Environmental, Social and Governance (ESG) was first introduced in the paper “Who Cares Wins” [1]. The paper illustrated how ESG factors can be integrated into a company’s operations. As of today, there are more than 10 ESG frameworks, standards and guidelines proposed and used in the ESG arena [2]. Companies across various sectors, including financial services, construction, healthcare, insurance, to even sport, publish the ESG reports to disclose and evaluate their ESG performance. Besides, investors nowadays are increasingly applying the ESG factors as part of their investment strategies to identify material risks and growth opportunities of a company. According to the research of Bloomberg Intelligence, global ESG assets surpassed $30 trillion in 2022 and are predicted to surpass $40 trillion by 2030 [3]. This trend underscores the rise of ESG investment as well as the importance of utilizing effective ESG analytics tools in decision-making.

While rating agencies, such as S&P Global, and Sustainalytics, are present in the market, their rating methodologies and criteria vary significantly due to the use of different ESG frameworks and country-specific policy. Thus, the ESG performance and rating assigned to a company provided by one rating agency may not be reliable and applicable when assessed under a different framework or in other regions. The lack of standardization in reporting frameworks and methodologies hinder the effectiveness of ESG analysis and decision-making processes. This inconsistency lead to confusion among the investors, making it challenging to make informed decisions regarding ESG investments. Investors may face difficulties in comparing companies across different rating systems, potentially compromising the reliability and utility of ESG data in their investment strategies.

1.2. Project Motivations

With the growing emphasis on sustainable business practices, more and more companies report their ESG performance to the stakeholders and the general public. Current solutions often fail to provide real-time insights or struggle with handling data across different ESG frameworks. On the other hand, there are no globally agreed rating methodology and criteria, the ESG ratings and information provided by current solutions may not be accurate and useful for assessment and analysis for companies in different business sectors and regions.

Our motivation stems from the need to improve the accuracy and accessibility of ESG data, enabling organizations to meet different reporting frameworks and market expectations while fostering responsible business practices. This project provides an opportunity to design a cutting-edge solution that supports both business objectives and sustainability goals.

Therefore, we introduce a functional ESG Data-Driven Decision Support System, ESGenius. The system aims to streamline the collection, management, and analysis of ESG data for organizations. The system will offer real-time analytics and risk modelling features that align with different ESG reporting frameworks. By leveraging Natural Language Processing techniques, the system will provide stakeholders with valuable insights to facilitate data-driven decision-making and promote sustainable practices in ESG analytics.

In more particular, the proposed project will include the development of a full-stack web application. Features include a predictor score model, ranking system, risk assessment tools, and a user-friendly dashboard for data visualization. Users will be able to choose one, combine multiple existing rating frameworks, or even customize their own weighing factors for evaluation and analysis accordingly to their needs.

2. Project Objectives

- Implement Machine Learning and Natural Language Processing techniques to generate ratings and qualitative insights for effective ESG performance assessment.

- Provide a user-friendly dashboard to easily compare ESG performance across companies and industries.

- Employ risk modelling techniques in managing exposure to ESG-related risks.

- Develop a full-stack scalable modern web application.

3. Methodology

Our project aims to provide a comprehensive analysis to our audience, together with a user-friendly user interface (UI) and customizable options to maximize the benefits of users from our project. The section explains the methodology used to develop this system and the technical considerations and decisions. This section consists of five subsections: The first subsection covers data collection and preprocessing, the second subsection discusses the Natural Language Processing (NLP) techniques to be used, the third subsection outlines the risk models adopted, and the fourth subsection explains the web application infrastructure of the system, such as the user interface and page structures. The fifth section describes the backend implementation, which includes the database and authentication.

3.1. Data Collection and Preprocessing

This project will be scoped to the S&P 500 companies and the ESG reports will be collected from the companies' official websites. There are two reasons to select the S&P 500 companies as the training data. Firstly, we can compare our test results with the S&P 500 ESG Index when assessing the ESG portfolio performance. Next, the large-cap companies under S&P 500 enforce a consistent reporting format in their published ESG report, which is beneficial to our analysis.

After collecting the data, we will use the Python library PdfPlumber to extract the text data from the PDF. To improve our text analysis in terms of speed and accuracy, lemmatization would be adopted by the libraries such as spaCy and/or Natural Language Toolkit (NLTK) to standardize the words to their base form.

3.2. Natural Language Processing (NLP)

Our system aims to perform two NLP tasks, which are Sentiment Analysis and Topic Modelling. These tasks enable our system to analyse and provide feedback for textual reports more easily. The following sections will describe the techniques in more detail.

3.2.1 Sentiment Analysis

Sentiment Analysis aims at capturing the sentiments, opinions and attitude from texts [4]. Since there are paragraphs of texts in different ESG reports, social media sites and forums, it is important to take those words into account as well. Not only the public's opinion affects the image of a company, but it also affects how likely they are going to invest in. Hence, sentiment analysis is essential and should be included in our system for the benefit of both company and investors.

3.2.1.1 Model Selection

There are various models that are suitable for sentiment analysis, such as Support Vector Machine (SVM), Neural Networks and Naive Bayes. To determine which model to be used, many factors should put into consideration, such as the size of the texts and level of accuracy required. It is found that deep learning models, such as Convolutional Neural Networks (CNN), can handle categorization issues more precisely [5].

This project will first pre-process the texts extracted from the documents extracted, such as text cleaning and tokenization, then different machine learning models will be applied to generate the attitude of the texts.

3.2.1.2 Goal

Sentiment analysis can help potential investors to know other's opinions on the company. It can also analyse how different research and articles feel about the ESG performance. By doing that, they can see the overall attitude and understand more whether the company is doing well.

3.2.2 Topic Modelling

Topic modelling aims to find out the theme and category of texts. Unlike other techniques such as semantic tagging, topic modelling works on single corpus and does not consider what are the most differences between corpora [6]. Topic modelling helps company managers to acquire the most trending and most discussed ESG topics, and to develop business strategies based on the most concerned from the public.

3.2.2.1 Model Selection

The most common model is the Latent Dirichlet Allocation (LDA) model, while some ESG reports analysis are also using this model. However, this model has its limitations: fixed number of topics and high computational complexity [7]. Other techniques such as Words2Vec, and BERTopic have been developed for different text lengths. Hence, our project aims to evaluate the performance among these models and select the best model to be applied.

3.2.2.2 Goal

Applying topic modelling can benefit company managers for internal analysis. It can uncover the most critical topics of ESG among forums, news articles and papers. It can also evaluate how the discussion topics changed over time, as to monitor the current trend. Companies can base on these results to formulate their future plans, ultimately aligning strategies with ESG principles to enhance sustainability.

3.3. Risk Modelling

Our project aims to assess risks related to sustainability and ethical practices. Two models will be used, named Time-Series and Monte Carlo. The later subsections will discuss the models in detail.

3.3.1 Time-series

Time-series risk modelling aims to predict risks over time. The process includes collecting past-time data or specific time intervals, then forecast the results in the future, while different models use different algorithms for forecasting [8].

Our project will adopt various models, such as Exponential Time Series (ETS), Pickup Forecasting and Generalized Autoregressive Conditional Heteroskedasticity (GARCH). Since different models produces different results, our project will include multiple models to provide a more comprehensive analysis.

3.3.2 Monte Carlo

Monte Carlo modelling assesses the probability of different outcomes, with the present of random variables [9]. The model can be used for modelling social issues and predict its social impact, for example the labour disputes. The output of Monte Carlo modelling will form a normal distribution, which the middle value is the most likely return [9].

Our project will try to apply this model to estimate the risk that the pattern from previous datasets will be disrupted in the future. If the chances were low, investors will have more confidence on that company.

3.4. Web Application Implementation

This section will illustrate the implementation of our web application, including frontend, backend development, features and page structures.

3.4.1 Technology and Framework

All the features of our project will be embedded into a webpage and dashboards. The frontend will be developed using React, a popular framework with community packages. The backend framework will be developed using Django, which is particularly useful in handling APIs and database management systems and supports multiple databases.

3.4.2 Features and Visualizations

The web application can be categorized into two modes, one being the public mode and the other one being user mode, which can customize the weighting factors for evaluation. Our project aims to serve all kinds of users with different needs, it is crucial to include customizable options, such as different rating frameworks and models, to find out the most suitable piece of data for them.

For the public mode, users will be able select different companies to view their performance from dashboards. For the user mode, users will be able to upload documents for analysis, customize factors for their own model, and do different risk analysis after log-in.

3.4.3 Design Approach

The web application will adopt responsive web design to fit different screen sizes and orientations, such as phones, tablets and computers. By adjusting elements according to the viewport, the website will be adjusted to any size and prevent certain elements to be out of the screen.

3.4.4 Page Structure

The web application will consist of six pages with different functions. Detailed explanations as follows:

1. Login page, where users access the user mode by filling in their username and password. Authentication will be performed to allow registered users only.

2. Register page, where users can create a new account.

3. Dashboard page, the first page shows up at visiting the webpage, which shows the most used KPIs by all users.

4. Client page, where users can upload documents (e.g. datasets, textual reports) for analysis

5. Risk analysis page, where uses can select different risk models for forecasting and risk management.

6. Profile page, where users can edit their personal information and page settings.

3.4.5 Hosting

We plan to host the website on cloud hosting platforms, to provide better scalability and reliability. If needed, we will consider upgrading to a paid plan for cloud hosting services.

3.5. Backend Infrastructure Implementation

3.5.1 Features

The backend system will provide users authentication such as checking username and password during login and ensure the newly created credentials matches the system requirement during registration. All data access requests will be made to the backend through HTTP requests.

3.5.2 Database

Our project will use MySQL as the database for high performance. It is also easy to link with web applications. The database will be hosted on Oracle Cloud, which can further enhance the security of the database.

There are three main tables in our database:

1. Ranking: After we predict the score of each company, our project will rank them according to sectors. This table consists of the sector, company name, predicted score and rank by sector and in overall.

2. Report Data: When users upload their documents for analysis, our system will store the data into this table. This table consists of document (i.e. PDF) name, company, and text sentence group by the 3 sectors: Environmental, Social and Governance.

3. Users: This tables stores the credentials by each user and will be used for authentication. Newly registered users will also have their information added into this table. This tables consists of an auto incremental ID field which is unique for each user, a username as well as the password field.

4. Schedule and Milestones

Period	Milestones
Oct 2024	Literature review for ESG scoring methodologies Collect ESG reports and preprocess the data Deliverables of Phase 1 • Project plan • Project website
Nov 2024	Study commonly used ESG scoring models Development of the web application and database • Preliminary GUI design and construction • Design the dashboards for ranking visualization
Dec 2024	Model training and tuning Study risk modelling techniques
Jan 2025	Deliverables of Phase 2 • Demo application • Interim report
Feb – Mar 2025	Integrate the risk modelling techniques Optimization and Testing
Apr 2025	Deliverables of Phase 3 • Finalized implementation • Final report

Table 1. Schedule and Milestones

5. References

[1] Krantz, T. (2024, February 8). The history of ESG: A journey towards sustainable investing. https://www.ibm.com/think/topics/environmental-social-and-governance-history#:~:text=It%20refers%20to%20a%20set,been%20around%20for%20much%20longer

[2] HKEX. (2024). ESG Frameworks. https://www.hkex.com.hk/Listing/Sustainability/ESG-Academy/External-References/ESG-Frameworks?sc_lang=en

[3] Wainwright, S., & Zickus, S. (2024, February 8). Global ESG assets predicted to hit $40 trillion by 2030, despite challenging environment, forecasts Bloomberg Intelligence. https://www.bloomberg.com/company/press/global-esg-assets-predicted-to-hit-40-trillion-by-2030-despite-challenging-environment-forecasts-bloomberg-intelligence/

[4] Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). . Springer Cham. https://link.springer.com/book/10.1007/978-3-319-55394-8

[5] Islam, M. S., Ghani, N. A., & Ahmed, Md. M. (2020, August). A review on recent advances in Deep learning for Sentiment Analysis: Performances, Challenges and Limitations. https://www.researchgate.net/publication/343473374_A_review_on_recent_advances_in_Deep_learning_for_Sentiment_Analysis_Performances_Challenges_and_Limitations

[6] Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017, August). 'What is this corpus about?': using topic modelling to explore a specialised corpus. https://www.euppublishing.com/doi/abs/10.3366/cor.2017.0118?widget=aboutthisjournal#de6dc34d-c714-4ed7-8457-32996500e695-58132d06-cf2f-4e31-a696-f4f2aa0cdd9a

[7] Goel, A. (2024, June 20). Sentiment Analysis in ESG: A Stepping Stone, Not the Destination. https://www.linkedin.com/pulse/sentiment-analysis-esg-stepping-stone-destination-arpit-goel-50wtf/

[8] Banerjee, N., Morton, A., & Akartunal, K. (2020, November). Passenger demand forecasting in scheduled transportation. https://www.sciencedirect.com/science/article/abs/pii/S0377221719308677

[9] Kenton, W. (2024, June 27). Monte Carlo Simulation: What It Is, How It Works, History, 4 Key Steps. https://www.investopedia.com/terms/m/montecarlosimulation.asp