The University of Hong Kong
Department of Computer Science

FITE4801 Final Year Project
Project Plan

Data-driven ESG Scoring with NLP

Date : 1st October 2022
Supervisor : Prof. YIU, Siu Ming
Students : DING, Ai Wei (3035719741)
          LEONG, Winky (3035664374)
          TANOTO, Nicholas (3035667041)
## Contents

<table>
<thead>
<tr>
<th>No.</th>
<th>Item</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Background</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Objectives</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Scope</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>Methodology</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>Timeline</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>References</td>
<td>9</td>
</tr>
</tbody>
</table>
1 Background

1.1 ESG and Sustainable Investing

Environmental, Social, and Governance (ESG) is a broad taxonomy used in capital markets to evaluate non-financial performance of a particular company, fund, or security [1]. In recent years, sustainable investing has experienced a growth spurt and investors are increasingly integrating ESG metrics into their investment analysis processes. For instance, in 2021, assets in ESG funds have doubled in just one year and flows into ESG-dedicated funds are forecasted to grow from US$8 trillion to US$30 trillion by 2030 [2,3]. Additionally, as of March 2020, the UN Principles for Responsible Investment (UNPRI), an international organization which aims to promote incorporation of ESG issues into investment practices, has approximately US $100 trillion worth of collective assets under management by its PRI signatories [4].

1.2 ESG Disclosures

Following the rise in sustainable investing, companies are increasingly expected to make ESG-related disclosures in their annual reports. In the 1990s, fewer than 20 companies reported ESG-related data; in 2016, the number of ESG reports has increased to nearly 8,000 [5]. There has also been an emerging market for ESG rating agencies, which specialize in providing company ratings upon conducting research, evaluation, and analysis on their ESG practices. Some of the most prominent rating agencies include Sustainalytics, Moody’s ESG (Vigeo-Eiris), S&P Global (RobecoSAM), Refinitiv (Asset4), MSCI, and Bloomberg [6].

1.3 Problem Statement

While such a phenomenon illustrates how ESG factors have become more important in investment decisions alongside the improving availability of ESG data, the lack of high-quality ESG data and scoring indices remains as a main challenge for investors. Firstly, there is a lack of standardization in ESG reporting practices, hindering an effective and intuitive understanding of the information presented in different ESG reports [7]. Secondly, there is a lack of transparency around the methodologies used by rating agencies, such as the sources and credibility of the underlying data, the applied imputation methods and weighting schemes, as well as the sustainability themes considered for assessment. Therefore, it is not uncommon for these agencies to provide different ESG ratings for the same company [8]. As a result, this divergence creates confusion among investors and thereby deters the sustainable investment sector from achieving its full potential in societal and environmental progress.
2 Objectives

To address the problem identified above, this final year project aims to develop a web application which provides more comprehensive and transparent ESG performance data for investors with the use of various Natural Language Processing (NLP) technologies. In particular, the project consists of four main components:

(1) ESG score generation

Fully data-driven ESG scores are generated based on the ESG initiatives disclosed by institutions, allowing investors to obtain a better understanding of an institutions’ ESG performance.

(2) Industry-specific ESG league table

Within a chosen industry, relevant companies will be ranked according to their ESG score, thereby facilitating investors to perform best-in-class investment approaches.

(3) Overview of key ESG initiatives

A summary of ESG initiatives extracted from the annual reports of different companies using NLP will be displayed on the web application dashboard, providing investors concrete insights into how exactly different companies are driving their ESG agenda.

(4) Display of financial metrics

In addition to its ESG-related information, the provision of fundamental financial data of a company will offer investors a more holistic view in their investment choices by displaying both the sustainability as well as the financial performance of a company.

Upon completion, this project aims to facilitate investors in their investment decision-making, as well as to improve stakeholder engagement and thought leadership through a better understanding of a company’s key ESG initiatives.

3 Scope

Due to the broad nature of the ESG topic, companies of different industries may emphasize on specific aspects of ESG issues in their disclosure reports. As such, this project will be scoped to the S&P 500 Communication Services and S&P500 Financials constituting institutions. A total of 92 companies – 26 from the communications industry and 66 from the financial sector – will be examined [9,10].

Additionally, among the multiple data sources, such as annual reports, corporate social responsibility (CSR) reports, company websites, and non-governmental organization (NGO) websites, this project will only involve NLP data on annual reports published by the companies in scope.
4 Methodology

4.1 ESG Initiative Retrieval from ESG Reports Using NLP

4.1.1 Gathering and Pre-processing of ESG Reports

To begin with the project, we will first gather the 92 in-scope companies’ latest annual ESG reports from their respective public websites as the primary data source. Since the ESG reports are in usually published in PDF format, they will be pre-processed using the Python library PyPDF2 for sentence extraction. The extracted sentences will then be lemmatized using the Python library Natural Language Toolkit (NLTK) to form well-defined ESG statements.

4.1.2 Topic Modelling with Latent Dirichlet Allocation (LDA)

As an initial analysis of the ESG report, we would like to understand the key topics communicated in the ESG reports. In particular, we will implement Term Frequency-Inverse Document Frequency (TFIDF) from the Python library Scikit-Learn and Latent Dirichlet Allocation (LDA) model from the Python library Gensim to perform topic modelling by extracting the key topics across all of the gathered ESG reports.

4.1.3 ESG Initiative Classification with ESG-BERT Language Model

Bidirectional Encoder Representations from Transformers (BERT) is a language model designed to pre-train deep bidirectional representations from unlabelled texts. BERT can be fine-tuned to perform various downstream NLP tasks without having to exercise significant architecture modification [11]. It has been shown that BERT outperforms the classical NLP approach in text classification tasks [12], and that pre-training BERT on domain-specific corpus yields better classification results [13].

Recent works have resulted in variations of the BERT language model pre-trained on ESG-specific corpus. Examples of such ESG-specific pre-trained BERT models include the ESGBERT [13] and the publicly available ESG-BERT [15]. Tentatively, we aim to leverage the ESG-BERT model as it is accessible online and utilize it to perform classification on the previously derived ESG statements into the key topics obtained from topic modelling step. The main objective of this step is to obtain the key ESG initiatives of different companies based on their ESG reports.

However, if the performance of the pre-trained ESG-BERT available online is unsatisfactory, we will develop an alternative language model ourselves by further pre-training Google’s BERT with other forms of ESG-specific corpus.

4.2 ESG Score Generation

4.2.1 Review of Existing ESG Scoring Methodologies

As previously mentioned, there are currently multiple rating agencies which evaluate a firm’s ESG performance by employing their own proprietary scoring methodology. Given the
variation of such ESG rating methodologies, we first aim to review the scoring systems utilized by few of the most popular rating agencies, namely Refinitiv, Thomson Reuters, and MSCI. Through this process, we aim to decide on which scoring system(s) to replicate or modify upon gaining a deeper understanding of the most relevant factors to consider for this project, given the key initiatives we have extracted from the ESG reports.

4.2.2 Development of ESG Scoring Model

Next, we will generate an ESG score for each company based on its key ESG initiatives using the scoring system(s) identified in subsection 4.2.1. As we progress, we may obtain additional ESG metrics or information from the rating agencies to further enhance the performance of our scoring model. For this purpose, we have currently identified S&P Capital IQ Pro portal as an ideal data sourcing platform.

4.3 Web Application

Finally, we will build a web application to present the insights obtained from the previous steps to end users. Through the web application, end users can access the ESG scoring information of their desired firms or industries in a user-friendly manner.

4.3.1 Frontend Development

The frontend of the web application will be developed using web development technologies such as HTML, CSS, and JavaScript, and some JavaScript frameworks such as ReactJS. The primary user interface will come in the form of a simple user dashboard, which displays the information in either firm-specific or industry-level view based on user input.

(1) Firm-specific view

The dashboard will allow users to select individual firms, and then display key information, including a brief introduction of the company, its industry, ESG score, key initiatives, as well as a summary of its financial data. Various visualization techniques such as a word cloud or interactive charts may be employed to enhance the overall user experience. Through this feature, we aim to present meaningful insights for end users who are looking into making investment decisions while considering the ESG performance of a firm.

(2) Industry-level view

The dashboard will also offer a feature to select a specific industry out of the available sectors according to S&P 500. According to the scope of this project, there will be tentatively only two sectors to select from (Communication Services and Financial). The dashboard will then display a list of all companies within that industry, ranked according to their ESG scores, allowing end users to easily compare the ESG performance of various firms within a specific industry.
4.3.2 Backend Development

The backend of the web application will comprise a database to store the data obtained from the processing and analyses as described in previous sections. Relevant information will then be pulled to the frontend display when the end user makes an input. The database will be hosted on a cross-platform database systems such as MongoDB, which offers capabilities for scaling, consistency, fault tolerance, and more.

4.3.3 Cloud Hosting

Finally, the web application will be hosted on a cloud application platform such as Heroku, which allows for more seamless deployment, easy management, and on-demand availability.

4.4 Project Management

4.4.1 Product Backlog and Issue Tracking

Adhering to modern software engineering practices, we will create a product backlog to document process and rising issues throughout the development life cycle of the project. For this purpose, we will utilize Jira by Atlassian as the main project management tool. This allows us to keep track of the overall progress of the project and delegate tasks efficiently.

4.4.2 Source Control

Given the collaborative nature of this project, different team members may be working on different aspects of the product development simultaneously. As such, Git will be used as the main source control tool to keep track of working changes. A shared repository for the project codebase will also be hosted on GitHub.

![Figure 1: High-level overview of the project workflow, illustrating ESG initiative retrieval and ESG score generation processes as well as information display through web application.](image-url)
5  Timeline

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Milestones</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 1: Inception</strong></td>
<td></td>
</tr>
</tbody>
</table>
| Sep 2022       |  - Project brainstorming  
  - Literature review  
  - *Deliverables of Phase 1: Detailed project plan and initial setup of project webpage*                                                                 |
| **Phase 2: Elaboration**                                                                                                                   |
| Oct 2022       |  - Gather and pre-process ESG reports  
  - Review and enhance or develop ESG-BERT language model                                                                                                      |
| Nov 2022       |  - Review existing ESG scoring methodologies  
  - Develop own ESG scoring model                                                                                                                                 |
| Dec 2022       |  - Retrieve key ESG initiatives using LDA and ESG-BERT  
  - Develop word cloud visualization for key initiatives                                                                                                    |
| Jan 2023       |  - First presentation  
  - *Deliverables of Phase 2: Preliminary implementation (ESG scoring model) and interim report*                                                                 |
| **Phase 3: Construction**                                                                                                                   |
| Feb 2023       |  - Fine-tune ESG scoring model  
  - Develop web application                                                                                                                                 |
| Mar 2023       |  - Continue development of web application  
  - Perform testing and debugging  
  - Code review and documentation                                                                                                                                 |
| Apr 2023       |  - *Deliverables of Phase 3: Finalized implementation and final report*                                                                                                                                  |
| **Final Presentation**                                                                                                                     |
| May 2023       |  - Final presentation                                                                                                                                                                                   |

6  Conclusion

In recent years, there has been a rise in sustainable investing, and ESG data is increasingly integrated by investors into their asset allocation decisions. However, quality ESG performance data remains one of the biggest obstacles for investors, stemmed from the lack of standardization in ESG reporting and scoring transparency among ESG rating agencies. With the help of NLP, this project aims to streamline the extraction of ESG data from annual reports, followed by the development of a web application to display key insights such as ESG score, industry ranking, and initiatives of different institutions. Ultimately, this project strives to facilitate a tectonic shift towards sustainable investing by enabling access to ESG information.
7 References


