The University of Hong Kong
Department of Computer Science
FITE4801 Final Year Project

Interim Report
FYP22026

Data-driven ESG Scoring with NLP

Date of Submission: 22\textsuperscript{nd} January 2023

Supervisor: Prof. YIU, Siu Ming

Team Members:

DING, Ai Wei (3035719741)
LEONG, Winky (3035664374)
TANOTO, Nicholas (3035667041)
TABLE OF CONTENTS

1  Background 3
2  Objectives 4
3  Methodology 5
   3.1 Extraction of Key ESG Initiatives 5
   3.2 Generation of ESG Score 7
   3.3 Display of Financial Metrics 8
   3.4 Development of Web Application 8
4  Progress 10
   4.1 Work Accomplished to Date 10
   4.2 Upcoming Deliverables 12
5  Conclusion 13
6  References 14

LIST OF FIGURES

Figure 1: High-level overview of the project workflow. 5
Figure 2: Topic modelling for a report in the finance industry. 6

LIST OF TABLES

Table 1: Main components with respective functionalities of the proposed web app 8
Table 2: Proposed project timeline 10
Table 3: Sample result of raw text lemmatization on Royal Bank of Canada’s ESG report 11
Table 4: Sample result of topic modelling on Royal Bank of Canada’s annual report 12
1 BACKGROUND

Environmental, Social, and Governance (ESG) is a framework that helps stakeholders understand how an organization manages risks and opportunities related to sustainability issues [1]. It is increasingly prominent for businesses and investors today as study has shown that companies that adopted sustainability policies tend to be less volatile and outperformed their competitors in financial performance [2]. Therefore, sustainable investing has experienced exponential growth, where investors increasingly integrate ESG metrics into their investment analyses. According to Bloomberg Intelligence, ESG investing may surpass US$53 trillion in assets by 2025 [3].

The rising demand for sustainable investing led to more companies making ESG-related disclosures in their annual reports. In 2020, more than 90% of the S&P 500 companies and approximately 70% of the Russell 1000 companies published ESG reports [4]. Additionally, more than 60 stock exchanges globally provide ESG disclosure guidance [5]. Given more ESG reports disclosure, there is an emerging market for ESG rating agencies, which specialise in providing company ratings based on their ESG practices. Some of the most prominent rating agencies include MSCI, S&P Global, and FTSE Russell [6].

While such a phenomenon shows the importance of ESG in investment decisions and the improving availability of ESG data, the lack of global standardization in ESG reporting practices remains a challenge for investors to intuitively understand the information presented in different ESG reports. Furthermore, the lack of transparency around the scoring methodologies across rating agencies result in varying scores for a single company, further hindering investors’ decision-making process. Research by State Street shows a less than 55% correlation between the scores of two major ESG rating agencies, implying that their ratings are only consistent for about half of the coverage universe [7]. As a result, it deters sustainable investing from achieving its full potential in societal and environmental progress.
2 OBJECTIVES

Given the above problems, this project aims to develop a web application that provides more intuitive and transparent ESG performance data to investors while leveraging various Natural Language Processing (NLP) technologies. Upon completion, this project hopes to help investors understand a company’s key ESG initiatives better and facilitate their ESG-based investment decision-making process. In particular, the project consists of four main components:

(1) Extraction of key ESG initiatives

A summary of ESG initiatives extracted from the companies’ ESG reports using NLP will be displayed on the web application dashboard, providing investors with intuitive insights into the ESG focus of different companies.

(2) Generation of ESG score

Data-driven ESG scores are generated based on the ESG initiatives disclosed by institutions, allowing investors to obtain a more transparent understanding of their ESG performance. The scoring methodology used will also be disclosed on the web application.

(3) Display of industry-specific ESG league table

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

(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 of their investment choices. Therefore, the financial performance will be displayed alongside the ESG performance data on the web application.
3 METHODOLOGY

An overview of the project workflow is shown in Figure 1, in which the project is broken down into 3 subsections — ESG initiatives retrieval, generation of ESG score, industry-specific ESG league table, and web app development. In this section, the adopted technical frameworks alongside the justification of choices will be discussed.

![Figure 1: High-level overview of the project workflow.](image)

3.1 Extraction of Key ESG Initiatives

To retrieve key ESG initiatives of a company, a total of four steps needs to be taken, starting with report collection in PDF formats. This is followed by pre-processing work, topic modelling and text classification tasks as elaborated in this section.

3.1.1 ESG Reports Gathering and Pre-processing

This project is scoped to all 500 companies in the S&P 500 index, which includes 11 industries from communication, financial services, to healthcare. The reason for retrieving data from these companies is that most of these large-cap firms do publish relatively comprehensive annual reports, which provides more credible data compared to smaller firms or indirect sources such as social media and news outlets. We have also considered the possibility of further expanding the scope to include HKEX-listed companies if more data are strongly required to improve the performance of our NLP models.

In this step, the two most recent reports in the form of sustainability reports, Corporate Social Responsibility (CSR) reports, ESG reports, or annual integrated reports, were gathered from the companies' public websites as the primary data source and...
downloaded to serve as a training corpus for later usage in Section 3.1.3 and Section 3.1.4.

Since the ESG reports are in PDF format, the reports were pre-processed using the Python library PyPDF2 for text extraction. The extracted texts were segmented into sentences, then lemmatised using the Python library Spacy to form well-defined ESG statements.

### 3.1.2 Topic Modelling with Latent Dirichlet Allocation (LDA)

To understand the key topics communicated in the ESG reports, Term Frequency-Inverse Document Frequency (TFIDF) from the Python library Scikit-Learn, and Latent Dirichlet Allocation (LDA) model from the Python library Gensim was implemented to obtain groups of keywords that describe the top topics discussed. The team reviewed the groups of keywords extracted, each corresponding to a topic, and manually named nine common topics discussed in the ESG reports. This step was vital to determine the word groups relevant to E, S, or G themes.

Since different industries may have different emphases on various aspects of ESG issues, the topic modelling step is conducted separately for each industry. For instance, Figure 2 shows the mapping of the groups of keywords to topics for the finance industry.

![Figure 2: Topic modelling for a report in the finance industry.](image)
3.1.3 ESG Initiatives Classification with ESG-BERT Language Model

Upon conducting topic modelling, a Bidirectional Encoder Representations from Transformers (BERT) model will be pre-trained for tasks classifying ESG statements into the key topics from steps 3.1.1 and 3.1.2. ESG-BERT model - a publicly available pre-trained model on ESG-specific corpus, will be used to categorise the key ESG initiatives [8].

There are 3 reasons behind choosing BERT among other NLP models. Firstly, it is designed to train deep bidirectional representations from unlabelled texts, i.e., it can process text from both left-to-right and right-to-left directions. The training results can be fine-tuned to perform various downstream NLP tasks, such as question answering and sentiment analysis, without significant architecture modification [9]. Besides, it has been shown that BERT outperforms the classical NLP approach in text classification tasks [10], and that pre-training BERT on domain-specific corpus yields better classification results [11].

However, if the performance of the ESG-BERT is unsatisfactory, an alternative language model will be developed by further pre-training Google’s BERT with other forms of ESG-specific corpus.

3.2 Generation of ESG Scores

As previously mentioned, there are currently multiple rating agencies that evaluate a firm’s ESG performance by employing a variation of proprietary scoring methodology. Literature review will be conducted by the team on the ESG scoring models of a few prominent rating firms, e.g., Refinitiv, Thomson Reuters, and MSCI, to decide on which data-driven 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.

Next, an ESG score will be generated for each company based on its key ESG initiatives using the scoring system(s) selected. The team has currently identified S&P Capital IQ Pro portal as our data-sourcing platform to obtain additional ESG metrics or information from the rating agencies to further enhance the performance of the scoring model.
3.3 Display of Financial Metrics

The fundamental financial data of each data will be pulled from Alpha Vantage Application Programming Interface (API) as it provides financial market data through a set of powerful and developer-friendly APIs. The data obtained will be displayed alongside the key ESG initiatives on the firm-specific view of the web application to be built in Section 3.4.

3.4 Development of Web Application

A web application will be developed to deliver insights obtained from the previous steps to investors via a web application, where investors can access and compare the ESG scoring information of their desired firms or industries. The design of the app features 4 main components, as shown in Table 1.

Table 1: Main components with respective functionalities of the proposed web app

<table>
<thead>
<tr>
<th>Component</th>
<th>Functionalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESG Scores Generation</td>
<td>Based on the ESG information disclosed in company reports, an ESG score will be generated for the in-scope companies.</td>
</tr>
<tr>
<td>Industry-specific ESG League table</td>
<td>Users are allowed to select an industry out of the 11 available sectors on the S&amp;P 500 index. Within the chosen industry, companies are ranked according to their ESG scores, thereby facilitating investors in performing best-in-class investment strategies.</td>
</tr>
<tr>
<td>Overview of key ESG initiatives</td>
<td>The ESG initiatives of companies will be extracted from their annual reports using NLP, and displayed on the web app dashboard in the form of a word cloud. This summary of initiatives provide investors insights into the concrete ESG efforts of a company.</td>
</tr>
<tr>
<td>Display of financial metrics</td>
<td>Aside from ESG-related data, fundamental financial analyses of a company are also provided on the dashboard to offer a more holistic view of both financial and non-financial performance of a company.</td>
</tr>
</tbody>
</table>

3.4.1 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.

### 3.4.2 Backend Development

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

### 3.4.3 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 PROGRESS

An overview of the project timeline is shown in Table 2. The following subsections will elaborate on the completed work and an evaluation of the important findings till date, alongside ongoing and scheduled work.

Table 2: Proposed project timeline.

<table>
<thead>
<tr>
<th>Period</th>
<th>Milestones</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep 2022</td>
<td>• Project brainstorming&lt;br&gt;• Literature review on current ESG investing landscape&lt;br&gt;<strong>Deliverables</strong>&lt;br&gt;  o Detailed project plan&lt;br&gt;  o Initial setup of project webpage</td>
<td>Completed</td>
</tr>
<tr>
<td>Oct 2022</td>
<td>• ESG reports gathering and pre-processing&lt;br&gt;• Design functionalities of web application</td>
<td>Completed</td>
</tr>
<tr>
<td>Nov 2022</td>
<td>• Key initiatives word cloud visualization&lt;br&gt;• Review of existing scoring methodologies</td>
<td>Completed</td>
</tr>
<tr>
<td>Dec 2022</td>
<td>• Topic modelling with LDA&lt;br&gt;• Fine-tuning of ESG-BERT for ESG initiatives classification</td>
<td>In progress</td>
</tr>
<tr>
<td>Jan 2023</td>
<td>• Development of ESG scoring model&lt;br&gt;• Preparation for first presentation&lt;br&gt;<strong>Deliverables</strong>&lt;br&gt;  o Preliminary implementation&lt;br&gt;  o Interim report</td>
<td>In progress</td>
</tr>
<tr>
<td>Feb 2023</td>
<td>• ESG scoring model fine-tuning&lt;br&gt;• Development of web application</td>
<td>Under planning</td>
</tr>
<tr>
<td>Mar 2023</td>
<td>• Continue development of web application&lt;br&gt;• Perform testing and debugging&lt;br&gt;• Code review and documentation</td>
<td>Under planning</td>
</tr>
<tr>
<td>Apr 2023</td>
<td><strong>Deliverables</strong>&lt;br&gt;  o Finalized implementation&lt;br&gt;  o Final report</td>
<td>Under planning</td>
</tr>
<tr>
<td>May 2023</td>
<td>• Preparation for final presentation</td>
<td>Under planning</td>
</tr>
</tbody>
</table>

4.1 Work Accomplished to Date

In early October, Phase 1 (Inception) has been concluded. The team conducted literature review on existing industry challenges and submitted the relevant deliverables comprising a detailed project plan and an initial setup of the project web page. With sufficient understanding of the current pain points of the sustainable investing, the team finished drafting the functionalities, user interface design and system architecture of the ESG web app, as elaborated in section 3.4, surrounding the design of the web app.
Next, the team has retrieved around 700 reports in total, from the public websites of the S&P 500 companies to be used as the primary dataset for our project. This is followed by building a pipeline to automate the sentence extraction and lemmatization processes from these ESG reports. For a given website address to an ESG report, the pipeline will automatically retrieve and return the pre-processed sentences from the report for further analysis. The pipeline works satisfactorily for reports that are uploaded in PDF format, but it may require additional modification for those posted in different formats, for instance reports shown on an embedded interactive PDF viewer.

This is followed by the pre-processing of all the PDF reports, in particular the lemmatization of raw texts into de-capitalized text in singular forms and present tenses. Table 3 provides a sample result of PDF pre-processing for Royal Bank of Canada (RBC)’s annual ESG report. After the data collection and pre-processing step, we performed topic modelling with the LDA model on the same RBC report. In this step, word groups relevant to E, S, G themes were grouped under the same topic. As seen in Table 4, the terms “green, bond, tax” are categorised under the same theme, i.e. topic #5 on sustainable finance.

Table 3: Sample result of raw text lemmatization on Royal Bank of Canada’s ESG report

<table>
<thead>
<tr>
<th>company</th>
<th>statement</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>rbc</td>
<td>ESG PERFORMANCE REPORT 1 Royal Bank of Canada Environment, Social and Governance (ESG) Performance Report 2019</td>
<td>make written or oral forward-looking statement within the meaning of certain security law include the safe harbour provisions of the United States Private Securities Litigation Reform Act of 1995 and any applicable Canadian securities legislation.</td>
</tr>
<tr>
<td>rbc</td>
<td>We may make forward-looking statements in this 2018 Environmental, Social and Governance (ESG) Performance Report, in filings with Canadian regulators or the U.S. Securities and Exchange Commission (SEC), in reports to shareholders and in other communications.</td>
<td>may make forward looking statement in this environmental social and governance esg performance report in filings with canadian regulator or the securities and exchange commission sec in report to shareholder and in other communication.</td>
</tr>
<tr>
<td>rbc</td>
<td>We have included forward looking information in this document to assist our stakeholders in understanding our noncash performance objectives, vision and strategic goals, our tax equity investments, as well as our social, economic, environmental and governance-related impacts and objectives.</td>
<td>have included forward looking information in this document to assist shareholder in understanding noncash performance objective vision and strategic goal as well as social economic environmental and governance related impact and objective</td>
</tr>
<tr>
<td>rbc</td>
<td>Forward-looking statements are typically identified by words such as believe, expect, forecast, anticipate, intend, estimate, can, could and similar expressions of future or conditional verbs such as will, may, should, could or</td>
<td>forward look statement are typically identified by word such as believe expect forecast anticipate intend estimate can could and similar expression of future or</td>
</tr>
</tbody>
</table>
Table 4: Sample result of topic modelling on Royal Bank of Canada’s annual report

<table>
<thead>
<tr>
<th>topic</th>
<th>keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic #1:</td>
<td>employee, people, work, training, support, business, woman, diversity, senior, include, program, day, diverse, talent, provide</td>
</tr>
<tr>
<td>Topic #2:</td>
<td>information, sustainability, page, include, annual, billion, statement, performance, management, business, governance, reporting, strategic, approach, chain</td>
</tr>
<tr>
<td>Topic #3:</td>
<td>climate, business, client, help, nancial, change, risk, customer, global, impact, financial, work, service, community, support</td>
</tr>
<tr>
<td>Topic #4:</td>
<td>investment, company, risk, management, portfolio, governance, environmental, responsible, social, policy, asset, team, process, business, committee</td>
</tr>
<tr>
<td>Topic #5:</td>
<td>sustainable, green, bond, development, finance, sustainability, nance, project, market, billion, company, include, tax, sdg, support</td>
</tr>
<tr>
<td>Topic #6:</td>
<td>new, million, support, launch, provide, people, program, community, income, include, job, initiative, care, development, programme</td>
</tr>
<tr>
<td>Topic #7:</td>
<td>customer, banking, service, management, asset, include, uk, global, product, year, investment, market, complaint, client, provide</td>
</tr>
<tr>
<td>Topic #8:</td>
<td>company, risk, climate, change, investor, portfolio, investment, engagement, include, analysis, sector, shareholder, base, issue, management</td>
</tr>
<tr>
<td>Topic #9:</td>
<td>energy, emission, reduce, renewable, use, carbon, water, waste, gas, scope, company, power, environmental, consumption, source</td>
</tr>
</tbody>
</table>

The current findings and results for the RBC annual report as shown in Table 3 and Table 4 have been acceptable, which serve as a foundation to extend the same process to all other PDF reports within the project scope.

Additionally, the ESG rating methodologies used by some of the most well-known rating agencies were reviewed as a vital step to start developing our ESG scoring model.

4.2 Upcoming Deliverables

Currently, the team is working on evaluating and exploring the use of BERT and ESG-BERT to do more advanced processing of our extracted statements and topics. For instance, we are considering the use of ESG-BERT to perform the classification of different statements and topics into the three broad ESG families: Environmental, Social, and Governance. However, we are facing slow progress as all the team members are new to NLP and have dedicated additional time to self-learning and experimenting with the implementation of various NLP technologies into this project.

After this step, the project focus will be on ESG scoring, in particular the development of the scoring model and generation of ESG scores. Upon preparing the codebase and database, front and back-end development work of the web app will be conducted. The application is scheduled to be developed by March 2023, to allow sufficient time for bug fixing before its official deployment in April.
5 CONCLUSION

In recent years, ESG as a sustainability framework has gained growing attention among investors and is increasingly integrated into asset allocation decisions. Although more companies are disclosing their ESG policies today, the lack of standardization and transparency across the ESG reporting formats and scoring methodologies remain the biggest challenges for investors. Therefore, this project aims to leverage NLP to streamline the key ESG initiatives extraction, then provides ESG performance data to investors via a web application.

Currently, the team has completed drafting functionalities of the web app, data collection, alongside conducting pre-processing and topic modelling using LDA model on one of the reports in-scope. Preliminary results on this sample report are deemed acceptable, therefore the next step is to extend the topic modelling process to all the in-scope reports, as well as the development of BERT model, scoring model, and the web app. However, the project has a few limitations, particularly on the evaluation of ESG performance solely based on the companies’ disclosure is at risk of greenwashing.

Nevertheless, this project has addressed the theoretical and practical improvement on the efficacy and integrity of the ESG data market, serving as a foundation for further development of NLP-driven alternatives to labour-intensive analyses. Therefore, real-time news sentiment analysis can be included in future ESG scoring work to improve the credibility of the score. The scope of companies can also be extended to smaller market-cap or private firms, as these firms are typically not covered by overseas data service providers.

Ultimately, this NLP-driven web app is expected to provide transparent and objective ESG information of companies for investors, thereby facilitating a tectonic shift towards sustainable investing by enabling better access to ESG information.
REFERENCES


