BT4103

FINAL REPORT

AY 2021/2022 Semester 1

Exploring ESG Strategies with AI

Team Name	
Student Name	**Student Number**
Clement Harsoyo	A0200756L
Geoffrey Bryan Lion	A0184580Y
Lourdesia Vivyan	A0200692M
Prawira Satya Darma	A0200684L
Putri Darmawan	A0200717R

Table of Contents

Executive Summary

Environmental, Social, and Governance (ESG) criteria are a set of standards that socially conscious investors use to screen potential investment. Nowadays, ESG has become an integral part of sustainable investing, an investing framework which most investors have grown interest in.

However, difficulties arise in gathering and exploring ESG strategies of a company. This is due to the fact that the majority of ESG data is still unstructured and cluttered in different websites and sustainability reports, with limited platforms to summarize them all together.

This project aims to create a platform to assist investors in gathering companies' ESG strategies more effectively. The information visualized are the company's ESG ratings, alliance memberships, summarization of web articles, and distribution of ESG topics in their sustainability reports.

In this report, the four mentioned features will be elaborated in detail, including the rationale, development phase, and the result, which will all be incorporated in an integrated dashboard. Suggested improvements will also be provided to increase the usefulness of the dashboard in the future.

1. Problem Statement

Investors are becoming more interested to invest in companies that have good ESG (Environmental, Social, Governance) performances. As shown in Figure 1, in 2015 there were only less than 10 billion US dollars of ESG assets under management. In 5 years, that number has increased into around 80 billion dollars (Bloomberg, 2020).

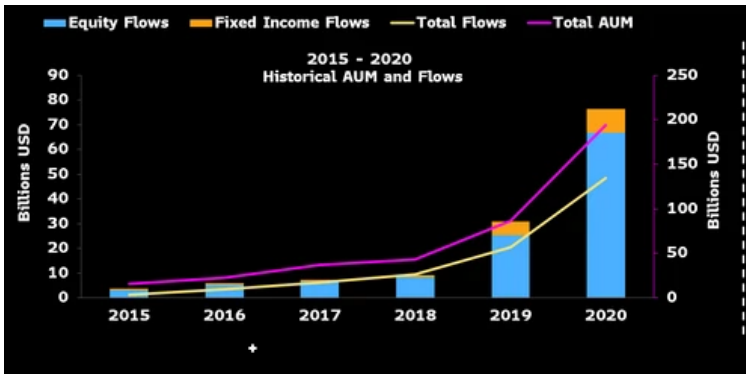

Figure 1. ESG Assets Under Management (2015 - 2020)

However, there are only limited platforms that can provide a comprehensive summary of a company's ESG information, thus making it inefficient for relevant stakeholders to understand and compare between companies' the ESG performances. Specifically, ESG metrics are still uncommon to be included in mandatory financial reporting. Furthermore, typical companies report their ESG performances in the form of sustainability reports, which commonly consist of hundreds of pages, adding on the pain points mentioned previously. Therefore, there is a need to develop a solution which will optimize the process of understanding a company's ESG strategies.

1.1 Analytic Requirements

With the problem mentioned above, business users are in need of a platform that has the capabilities to:

- Speed up the process of gathering ESG information
- Optimize the process of extracting ESG information
- Incorporates different ESG information from different sources

1.2 Functional Requirements

The functional requirements for this project include:

- Gathering relevant ESG information regarding the company that the users have input
- Summarizing the ESG information accurately
- Visualizing the information in a structured and user-friendly manner

2. Objective

To solve the mentioned problem, this project aims to create a platform which can summarize and visualize the ESG strategies and performances of a company in a structured manner and facilitate comparison with other companies.

3. Solution

This project will build a dashboard consisting of 4 main features, which are ESG ratings, alliances membership, topic modelling of sustainability reports, and summarization of ESG articles from the internet. The details for each feature will be shown below.

3.1 ESG Rating

ESG Rating is a measure of a company's ESG performance based on 3 categories: Environment, Social, and Governance. In this project, we gather data from the rating agency

Refinitiv. The information displayed in the dashboard will include ESG overall rating (out of 100), environment rating (out of 100), social rating (out of 100), governance rating (out of 100), and rank in industry.

3.1.1 Rationale

ESG rating is a useful insight to measure the credibility of a company towards Environmental, Social, and Governance issues. The metric rank in industry shows whether a company's ESG performance is better than other competitors in their industry. Moreover, the overall, environment, social, and governance rank can be used as comparison between several companies.

3.1.2 Refinitiv

We chose Refinitiv instead of other rating agencies because it provides us with complete rating information of various companies. Refinitiv's rating methodology itself is divided into 3 sections: environmental, social, and governance. To measure environmental score, Refinitiv takes into account 3 categories: resources use, emissions, and innovation. Moreover, Refinitv considers workforce, human rights, community, and product responsibility for social aspect. Lastly, Refinitiv considers management, stakeholders, and corporate social responsibility (CSR) strategy for the governance category.

3.1.3 Data Gathering

In order to gather the data from Refinitiv, our approach is to use the BeautifulSoup library to get the corresponding company ticker based on the searched company name. Afterwards, we will use the company ticker code to scrape the data from Refinitiv and gather the required ratings and rank.

3.1.4 Result

Figure 2 illustrates an example result of OCBC ESG rating in the dashboard. We can see an overall rating of 53, environment rating of 58, social rating of 47, governance rating of 61, and an industry rank of 254 out of 969.

Figure 2. ESG Rating Result

3.2 Alliance Membership

Alliance memberships are organizations in which companies can collaborate with others to address ESG-related issues together

3.2.1 Rationale

Alliance membership is useful to provide information on how credible financial institutions are in terms of environmental, social, and governance issues. A company that joins more ESG-related alliances can be considered to show more focus on ESG issues and thus, have higher credibility in terms of ESG.

3.2.2 Alliances

There are five alliances that we chose to incorporate in our solution. These five organizations are picked because of its credibility represented by the number of members and its focus on either environment, social, or governance.

3.2.2.1 UNPRI

United Nations of Principles Responsible Investing (UNPRI) is a United Nations-supported international network of investors. The goal of UNPRI is to understand the implications of sustainability for investors and support signatories to facilitate incorporating these issues into their investment decision-making and ownership practices.

3.2.2.2 UNEPFI

United Nations of Environment Programme Finance (UNEPFI) is a global partnership established between the United Nations Environment Program (UNEP) and the financial sector. The focus of UNEPFI is to inspire their members to take Environmental, Social, and Governance issues into their relationships and trade with their customers.

3.2.2.3 ICMA

International Capital Market Association (ICMA) is a self-regulatory organization and trade association for participants in the capital markets. ICMA market conventions and standards have been the pillars of the international debt market, providing the self-regulatory framework of rules governing market practice which have facilitated the orderly functioning and impressive growth of the market. Thus, this alliance focuses more on governance issues

3.2.2.4 UNGC

United Nations Global Impact (UNGC) is a United Nations pact that encourages firms to apply sustainable and socially responsible policies. UNGC itself is a principle-based framework for businesses, stating ten principles in the areas of human rights, labor, the environment and anti-corruption. Thus, this alliance focuses more on social issues.

3.2.2.5 IIGCC

Institutional Investors Group on Climate Change (IIGCC) is a global investor alliance that is focusing on climate change. IIGCC works with business, policy makers and fellow investors to help define the investment practices, policies and corporate behaviours required to address climate change. Thus, this alliance focuses more on environmental issues.

3.2.3 Data Gathering and Methodology

Data gathering and methodology are divided into 3 parts: list of members gathering, member detection with Fuzzy search, and model evaluation.

3.2.3.1 List of Members Gathering

Before we can detect if a company is a member of an alliance, we need to get the list of members of each alliance. In order to extract the data, we use BeautifulSoup library to scrape the list of members from alliance's website page.

3.2.3.2 Member Detection with Fuzzy Search

Using the list of members, we will detect whether a company is a member of the alliance. In order to consider different user inputs, we decided to use Fuzzy search. Fuzzy search uses Levenshtein distance to measure the distances between 2 words using the formula from figure 3.

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Figure 3. Levenshtein Distance Formula

The formula from figure 3 measures the similarity between 2 words. Moreover, we decided to use FuzzyWuzzy library to implement the fuzzy search. In order to optimize the usage of Fuzzy search, we tried 3 different cutoff ratios: 90%, 95%, and 97% as the parameter. We decided to use 95% since it identifies many different cases of input and has higher accuracy. If the ratio is higher or equal to 95%, it will be identified as a member. Otherwise, it will be identified as non-member.

3.2.3.3 Model Evaluation

Model evaluation is crucial since we want to have accurate data for the final dashboard. We did an evaluation of our fuzzy search and received an accuracy of 90%, 91%, 95%, 92%, and 90% for UNPRI, UNEP FI, ICMA, UNGC, IIGCC respectively based on 100 observations for each alliance.

3.2.4 Result

Figure 4 shows the membership table of OCBC in the dashboard. We can see that OCBC only joined UNGC out of the 5 alliances.

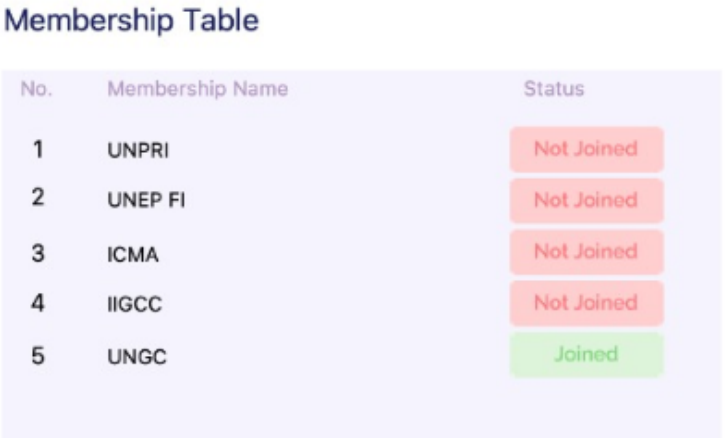

Figure 4. Membership Table Result

3.3 Topic Modelling

Topic modelling is a statistical modelling that is useful for discovering and extracting “abstract” topics that occur in a collection of documents. In this project, we use Latent Dirichlet Allocation (LDA), a three-level hierarchical Bayesian probabilistic model in which each document is modeled as a multinomial distribution of topics, and each topic is modeled as a multinomial distribution of words. In the case of text modelling, the topic distribution generated by the model also provides the explicit representation of the whole document. The machine learning pipeline of topic modelling is as follows.

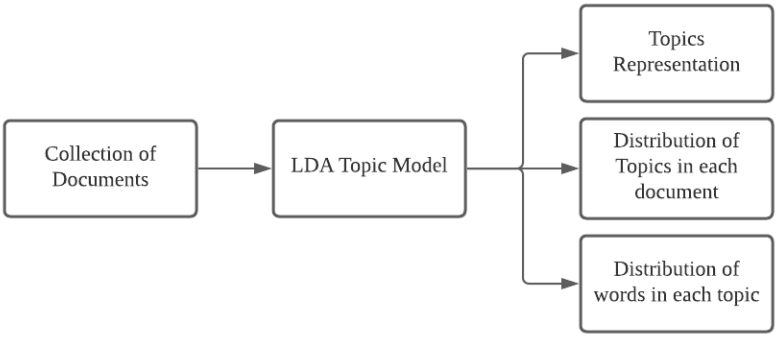

Figure 5. Topic Modelling Pipeline

3.3.1 Rationale

Topic modelling is useful to provide a representation of a company’s sustainability report. As mentioned before, one of the pain-points for stakeholders to review a company’s sustainability report is its length. On average, typical financial institutions disclose their ESG strategies and performance in a 270 pages long report. Therefore, the LDA topic model can speed up the process by providing the explicit representation of ESG topics in the whole document. In other words, topic modelling provides the function of organizing and summarizing the content of a sustainability report with its topic representation. Furthermore, the use of topic modelling is also important for article summarization as it helps to filter articles that are not ESG-related. The details on article summarization will be elaborated on Section 3.4 later in this report.

3.3.2 Development Phase

To build the relevant topic model algorithm for ESG strategies, our development phase consists of three steps. It begins with data preparation, then training and tuning the topic model, and ends with the testing stage.

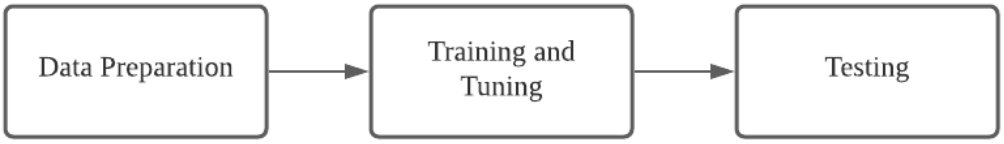

Figure 6. Development Phase Pipeline

3.3.2.1 Data Preparation

The model development starts with data preparation. To have a topic modelling that covers ESG topics, indeed the data fed into the model are ESG-related documents. Here, we use ESG-related articles from the internet, in the form of HTML or PDF, as our training data. It is worth noting that there is not any specific company's sustainability report used as the training data as it may introduce bias.

Several data pre-processing methods are conducted before the data is ready to be fed into the model. Firstly, as we are working with text data, stopwords are removed. Stopwords itself is a set of words which are common and not useful for Natural Language Processing (NLP), such as the word "I", "you", "as", "can", and many more. Secondly, stemming and lemmatization is applied to reduce a word into its root form. For example, the word "fixed" or "fixing" is reduced into "fix". In particular, we use the pre-built Gensim model for our data pre-processing.

After all the pre-processing is finished, each word in the dataset will be collected and fed into a bag of words or TF-IDF model. From this TF-IDF model, we will generate a corpus which will be the input for the training model.

3.3.2.2 Training

Having generated the relevant corpus from data preparation, we build the LDA model using the Gensim library. During building the model, one of the important parameters expected is the number of topics. Therefore, we run an iterative experiment with coherence score as our evaluation metrics to best decide the number of topics.

3.3.2.3 Testing

During the testing process, we fed ESG-related documents into the topic model. At the early stage of testing, we were surprised that our topic model can only extract the distribution of environmental topics. Therefore, we decided to examine our training and testing data in depth to tune and further strengthen the model's performance.

Reading our training and testing documents, we realized that our data is actually biased. This is due to the uneven distribution of environmental, social, and governance related topics in most ESG-related articles, whereby disclosure about environmental related information outweighs the proportion of the other two topics significantly. Therefore, we decided to include more training data which covers more on social and governance to ensure that the topic distribution of all the three topics are represented. An example of the topic result is as follows.

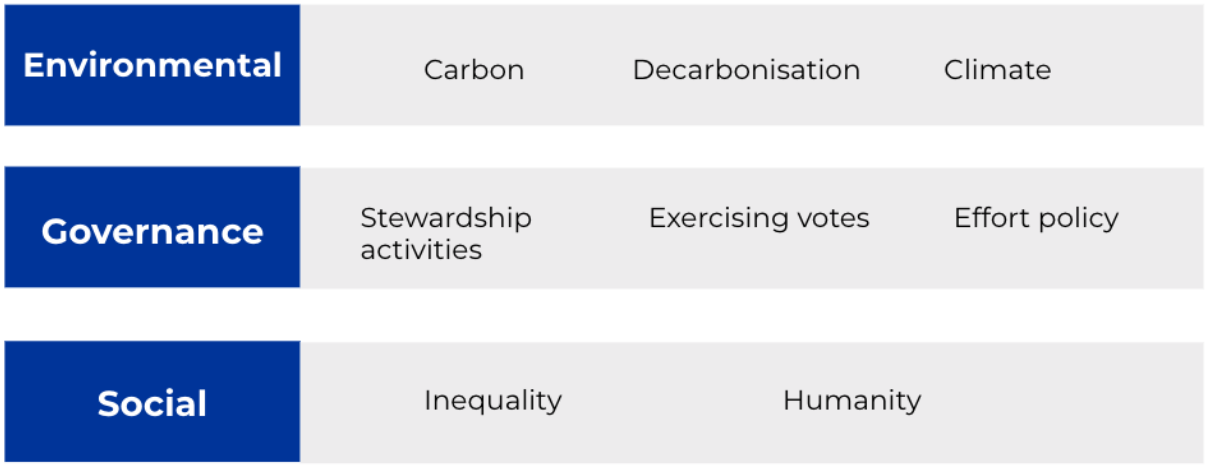

Figure 7. Topic Keywords

In short, during the development phase, we have managed to extract out topics to represent each environmental, social, and governance, although the number of topics for environmental related information outnumbers the topics related to social and governance.

3.3.3 Result

The topic model built will then be used to provide explicit representation of a company's sustainability report. It will first extract sustainability reports from our database rather than searching through the internet as it is computationally heavy to scrape for hundreds pages of reports from the internet. The generated topic distribution with a bar graph, and it is important to note that we fuse several topics that represent each environmental, social, or governance related information. As such, the final result of the topic modelling is visualized with at most 3 bars in a bar graph, each representing the environmental, social, and governance topic proportion in the sustainability report.

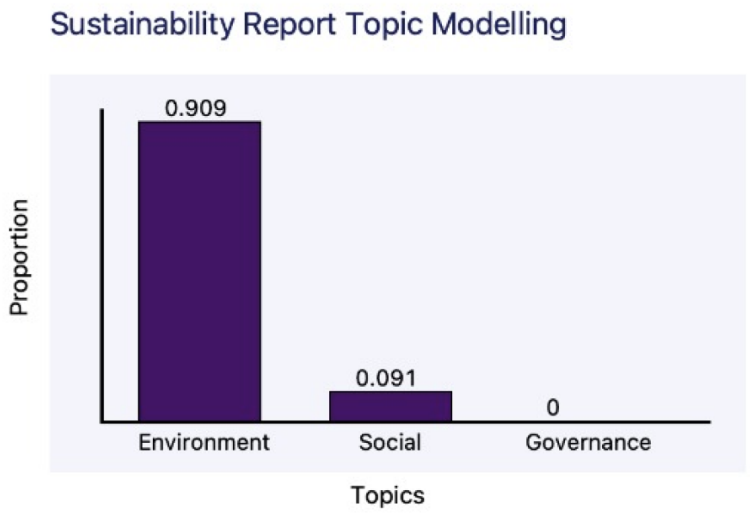

Figure 8. Topic Modelling

3.4 Article Summarization

For this article summarization, we use extractive summarization where the methods rely on extracting several parts, such as phrases and sentences, from a piece of text and stack them together to create a summary. Unlike abstractive summarization, extractive summarization will use phrases from the text and combine them to form a summary, rather than creating new sentences. The text summarization will be done with cosine similarity to measure the similarity between two sentences. Cosine similarity is a measure of similarity between two sentences represented as vectors.

3.4.1 Rationale

Knowing what ESG strategies that each company has will be useful to assist investors during the decision-making process. Other than sustainability reports, these ESG strategies can be obtained from websites. However, it will be a tedious process for investors to read and understand many articles from the internet. Therefore, we aim to alleviate this problem by using article summarization. By using text summarization, the investor can get the general idea and the important points of a particular company's ESG strategies. Furthermore, it will be very time efficient to read since all of the key points are summarized into a maximum of 120 words.

3.4.2 Development Phase

To make the summary, we use requests, requests_html, urllib libraries from Python to mimic the Google search. From this stage, we will have URL links that cover information about the

company's ESG strategies. Then, we will filter out the links that we do not want, such as youtube, instagram, and linkedin. After we filter the links, we will use BeautifulSoup to scrape the text from the articles. Having the text ready, we will use topic modeling to determine whether the article is actually ESG-related or not according to the decision rules, e.g. if the article has a score larger than 0.5 for the environment topic, we will include the article. We choose to use the top three most ESG-related articles (the score of the topic distribution is the highest). Lastly, we summarize those articles using gensim.summarization with a maximum of 40 words for each article. Therefore, identifying the right sentences for summarization is of utmost importance in an extractive method. By using topic modeling, it is ensured that the summary will be related to ESG.

However, there are some limitations of the text summarization. Using a real time google scraping will give a different result for every time it is run. In addition, some articles also cannot be scraped since some of them are protected with cookies and have privacy settings. Nevertheless, these limitations are resolved by topic modeling. Even though the summary changes every time, it is still related to ESG. In addition, if the article cannot be scraped, it will have a very low score for the topic distribution. Thus, it will be filtered out by the topic modeling.

3.4.3 Result

The result will be in the form of a string, which consists of the summarization of 3 articles that have been filtered with topic modeling. Sample result of text summarization for OCBC:

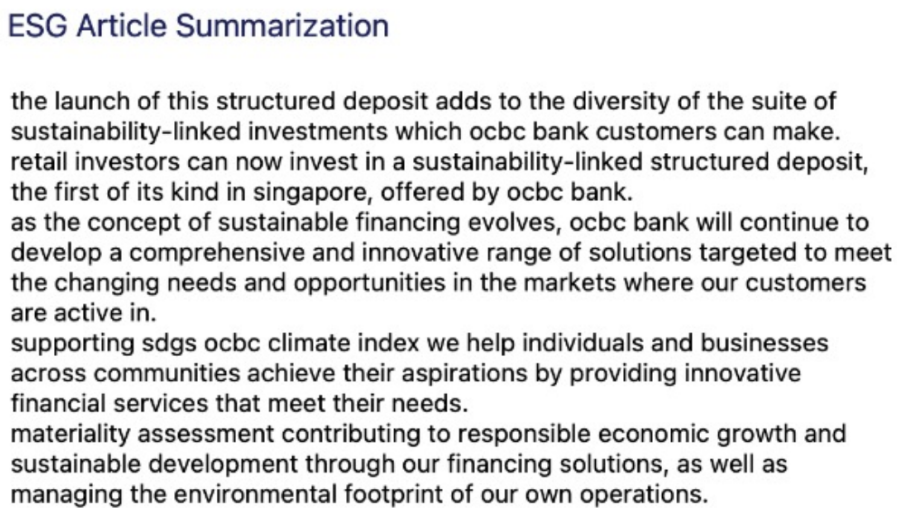

Figure 9. Article Summarization

4. Final Product: Dashboard

The dashboard is the incorporation of all the solutions proposed in Section 3. The features will be described in this section. The dashboard picture is as below:

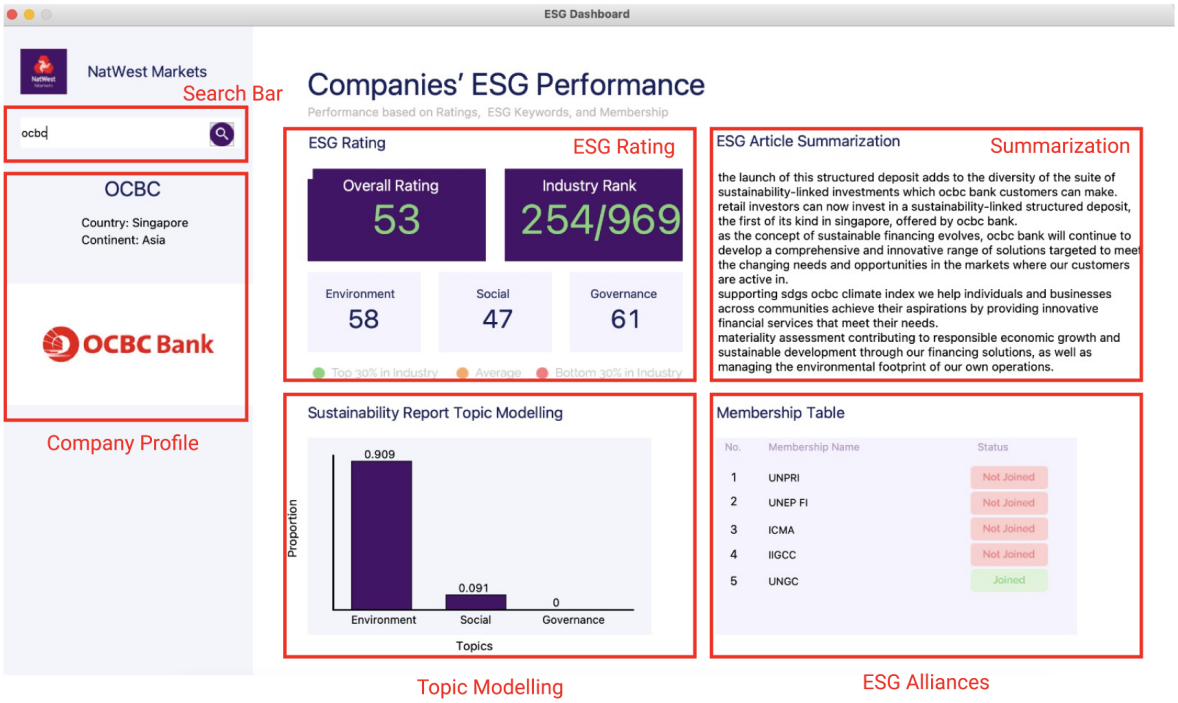

Figure 10. Dashboard

4.1 Search Bar

The search bar is a place to input the company's name. After inputting the name, users will need to press the magnifier button to start the process. The button will embark the process and show a loading screen.

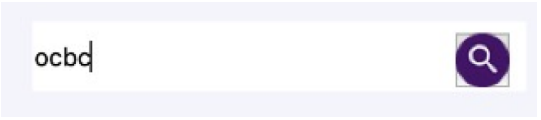

Figure 11. Search bar

4.2 Loading Screen

A loading screen is added after the user key in the company name and before the dashboard can show the results. We added this component to give assurance to the user that the program is loading, not crashing.

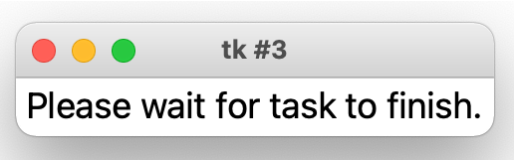

Figure 12. Loading Screen

4.3 Company Profile

For the investors to know the details of the company such as country and continent, it will be useful if we include the company profile in the dashboard. We include 4 things in the dashboard for the company profile, company name, logo, country, and continent.

We use requests, requests_html, urllib to get the links of the websites that contain the company's logo, country, and continent. To get the company logo, we use the scrape function that we made to search for company name + "logo". After that, we will find websites that contain a jpg or png file, copy the image link and use tkinter to open the image. Since it is real time, the logo might change every time it runs.

To get the country and continent, we search for company + "country". Then, we take the information from wikipedia and other websites that provide the country name. We have a dictionary that contains every country name and the continents, and compares it with the text. The country name will then be sorted and we will get the most frequent country name.

Figure 13. Company Profile

4.4 ESG Rating

The ratings shown are overall rating, industry rank, environment rating, social rating, and governance rating. The overall rating and industry rank has a deeper colour and a bigger size because it is relatively more important than the individual ratings of ESG components. Moreover, the industry rank is crucial for the user to be able to compare between other companies in the same industry.

Additionally, the colors: green, yellow, and red represent the company's rank within the industry. Green colored rating and rank means that the company is in the top 30% of industry. Yellow colored means the company is average (between green and red). Red colored means that the company is in the bottom 30% of the industry.

Figure 14. ESG Rating

4.5 Topic Modelling

The bar chart shows three main topics we are exploring, which are Environment, Social, and Governance. Each of them is represented in the topic proportion of the sustainability report.

Figure 15. Topic Modelling

4.6 Article Summarization

This is the summary from articles that are scraped from the internet. Hence, it is a “real time” data, which means that it is able to catch new articles.

Figure 16. Article Summarization

4.7 Alliance Membership

The table shows whether the company is involved in that particular ESG alliance membership with a simple “Joined” and “Not Joined”.

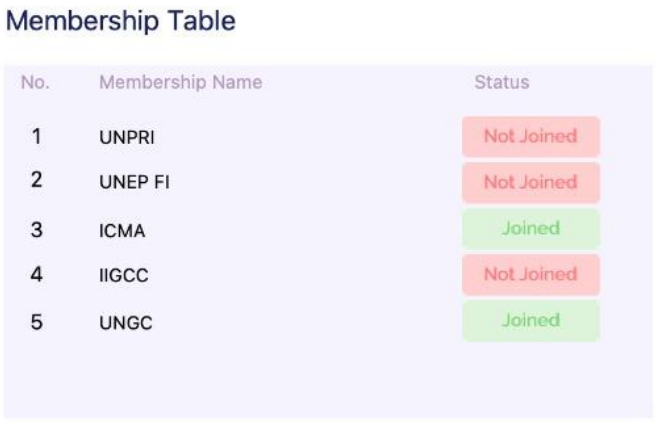

Figure 17. Alliance Membership

5. Limitation and Further Improvement

5.1 Numerical Extraction

Our team attempted to extract numerical figures related with ESG keywords such as green bond and social bond value from the sustainability reports. We attempted several methods to do numerical extraction, with manual coding and Spacy. We realized that manual coding would be very limited as we need to manually adjust the code, possibly rewriting the code, for different keywords and different companies. Thus, we focused more on exploring Spacy as it is a well-known python library for NLP. We managed to construct a function to extract

water and electricity consumption numerical figures from 5 different companies using models from Spacy. However, the models need a massive amount of data to train to have a high accuracy. As we were not able to compile a large amount of training data, we incorporated manual coding in the function. As a result, we had troubles in expanding the function to extract from different companies. Moreover, we were limited by the text extraction results from the pdf since there were problems with the results that we could not solve. We believe that the problems were caused by the PDF format which caused the information in tables, text, columns to be difficult to extract accurately. In addition, there were special characters which came up in the result but were not in the original text or some words were not extracted. Since we haven't completed the functions to produce satisfactory results, we decided not to include them in the final dashboard.

5.2 Dashboard Deployment

Currently, users will need to run the dashboard locally on their computer, which is a limitation. As a further improvement, we hope to be able to deploy this using web hosting, such as netlify, etc. This is to ease access to the dashboard, as users do not need to install various libraries on their local computer. They can just access it through their preferred web browser.

The reason behind this limitation is because currently web hosting deployment does not support Python yet. They mainly support Javascript. Our codes are purely from Python, which makes us unable to deploy yet. In the future, we can try deploying using other web hosting that can support Python or change to Javascript.

5.3 Improve Processing Speed

The average speed for querying one company is around 3 minutes. However, this process can still be a lengthy time. The long processing time can be attributed to topic modelling and summarization which uses real time data. As such, for every query, the dashboard's information needs to be processed from scratch.

An idea to tackle this is by having a database with regular updating, such as daily or weekly. When we implement this, there will be no need to process the data for every company query. The trade off between speed and up to date data can be tackled by giving users a choice. Hence, we can speed up the process.

6. Conclusion

In conclusion, the problem of gathering ESG information quickly and effectively is real. This is due to the non-existent of a platform that supports the mentioned problem, and the fact that the disclosure of a company's ESG information is mostly using hundred pages long pdf reports. We aim to tackle this problem by developing a dashboard consisting of 4 different features, numerical ESG rating, alliance memberships, topic distribution of sustainability report, and article summarization from the internet. Further improvements can be made in the form of numerical extraction from sustainability reports and deployment to increase the usefulness of our current dashboard.

7. Installation Procedure

We have uploaded our codes into github, with the requirements.txt and documentations on each file. To install and run our project, users can easily follow these steps:

1. Type git clone https://github.com/putridar/esg-dashboard.git in your terminal/command prompt
2. Type cd esg-dashboard
3. Run sudo pip install -r requirements.txt to install dependencies
4. Run python -m spacy download en_core_web_sm
5. Run python -m spacy download en_core_web_lg
6. Run python Dashboard.py
7. A Tkinter window will show up and the platform can be used

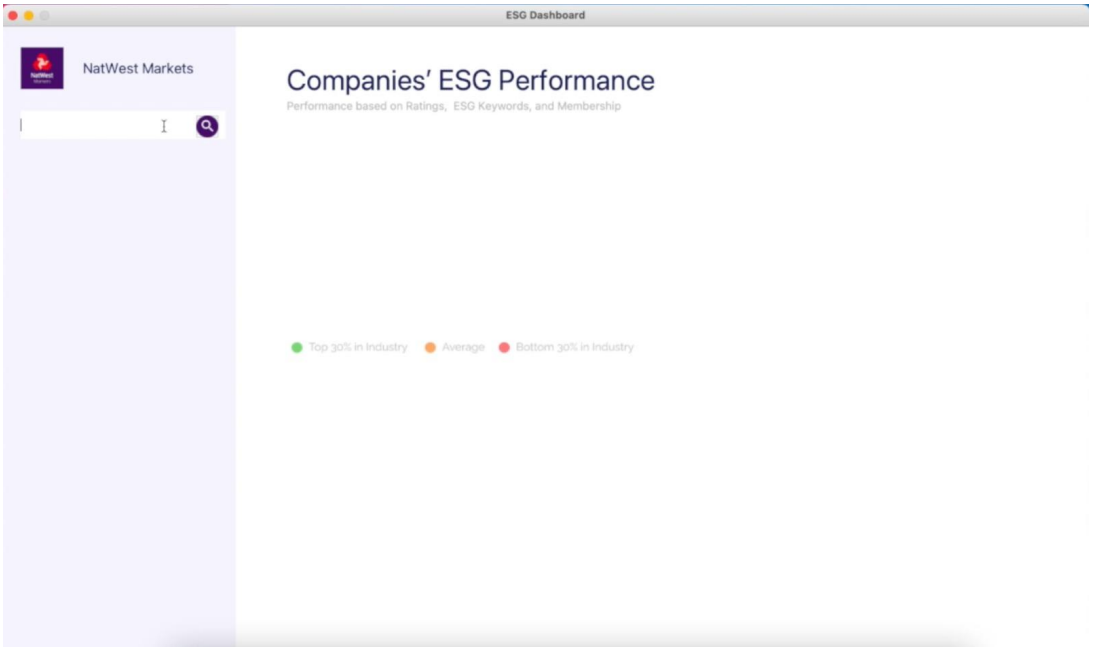

Figure 18. A newly opened window

8. To use the dashboard, users only need to input the company keyword in the search bar and features explained in section 4 will be displayed.

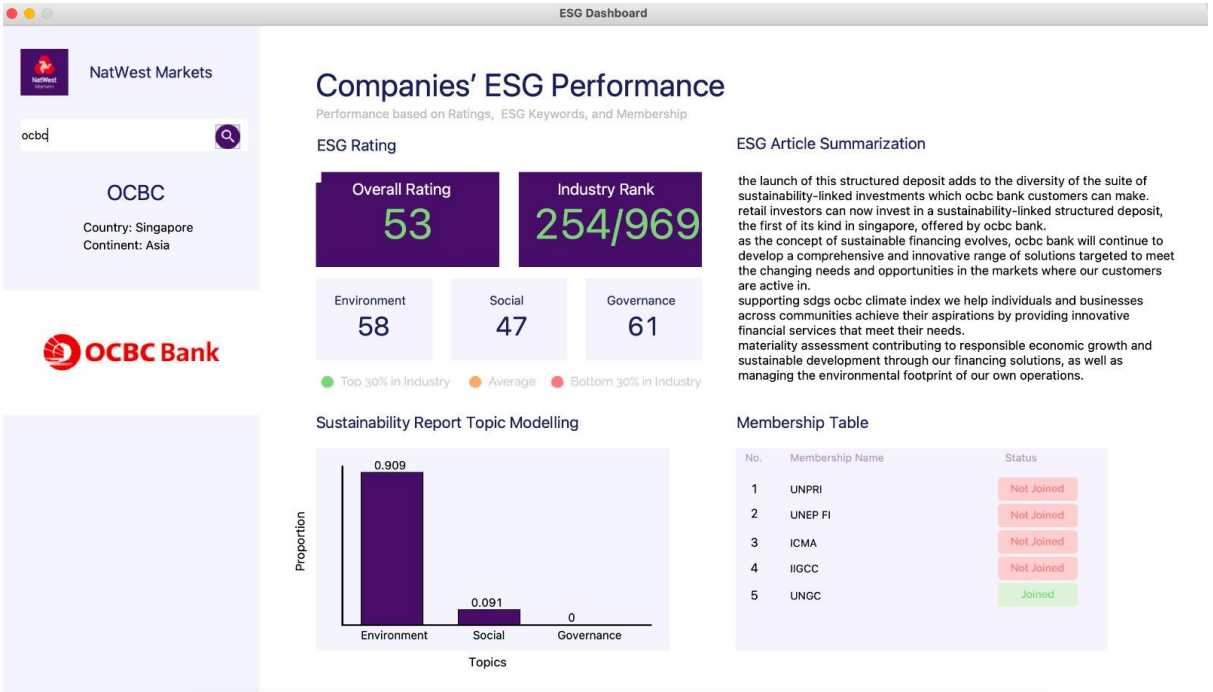

Figure 19. Dashboard result

8. References

Bloomberg. (2020). Bloomberg.com. Retrieved November 12, 2021, from https://www.bloomberg.com/professional/blog/esg-assets-may-hit-53-trillion-by-2025-a-third-of-global-aum/#:~:text=ESG%20assets%20are%20on%20track%20to%20reach%20%2453%20trillion%2C%20based,from%20%2422.8%20trillion%20in%202016.