



**BT4103**

**Final Report**

**AY 2021/2022 Semester 1**

**Group 7**

---

Team Members
Ho Jheng-Yuan
Ho Lok Sang Kelvin
Lu Xinyi
Soh Jun Heng
Toh Hui Shan Alicia

## Table of Contents

<b>1. Executive Summary</b>	<b>4</b>
<b>2. Analytic Requirements</b>	<b>5</b>
2.1 Text Preprocessing	5
2.2 Sentiment Analysis	6
2.2.1 Overall Sentiment Score Calculation	6
2.3 Topic Modelling	9
2.3.1 Optimal LDA Model	9
2.3.2 Deriving Distribution of Decarbonization Related Topics	10
2.4 Bigram Analysis	11
2.5 Word2Vec Model	12
<b>3. Functional &amp; Non-Functional Requirements</b>	<b>15</b>
3.1 Functional Requirements	15
3.1.1 For Overall Dashboard	15
3.1.2 For First Tab of Dashboard (Individual Company)	15
3.1.3 For Second Tab of Dashboard (Company Comparison)	15
3.2 Non-Functional Requirements	16
3.2.1 Usability	16
3.2.2 Scalability	16
<b>4. User Interface of Dashboard</b>	<b>17</b>
4.1 First Tab (Individual Company)	17
4.1.1 Dropdown Menus	17
4.1.2 Overall Sentiment Analysis	18
4.1.3 Percentage of Decarbonization Disclosure	19
4.1.4 Number of Global Initiatives & Standards	20
4.1.5 Table of Global Initiatives & Standards	21
4.1.6 Top Ten Occuring Bigrams	22
4.2 Second Tab (Company Comparison)	23
4.2.1 Dropdown Menus	23
4.2.2 Comparison - Overall Sentiment Level	23
4.2.3 Comparison - Percentage of Decarbonization Disclosure	24
4.2.4 Top Ten Occurring Bigrams	25
<b>5. Use Case</b>	<b>26</b>
5.1 Deployed Dashboard	26
<b>6. Future Work</b>	<b>27</b>
6.1 Drill Down on Bigram Analysis	27
6.2 Expedite LDA Training Process	27
6.3 Improvements to Dashboard	27
6.3.1 Increase Number of Filter Options for Dashboard	27
6.3.2 Add a ranking interface	27

<b>7. Conclusion</b>	<b>28</b>
<b>8. Appendices</b>	<b>29</b>
8.1 Sentiment Analysis	29
8.1.1 Bag of Words model	29
8.1.2 Predicted sentiment scores	29
8.2 Topic Modelling Topics & Assigned Category	30
8.3 Top 10 Word Count	31

## **Acknowledgement**

We wish to express our sincerest appreciation and gratitude to Professor Lee Boon Kee and Mr Puneet Gupta who have made all this possible. Thank you for being ever so patient with us and our problems and for guiding us through accomplishing the final report.

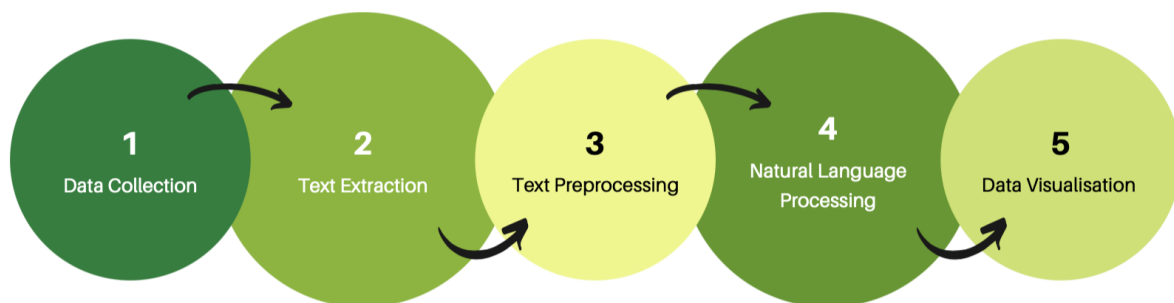
Our group has gained many invaluable insights and experience throughout the course of the project and we hope that the project will benefit Natwest Markets in the long run.

## **1. Executive Summary**

The project titled 'Exploring Portfolio Decarbonization using AI' aims to support NatWest Markets in unstructured data analysis through developing a machine learning model that extracts information related to decarbonization from various sources. This project uses publicly available data sources which comprises mainly of companies' ESG reports, annual reports and corporate social responsibility reports. Reports were collected from over 115 companies across 4 types of financial institutions, namely Asian Banks, Asset Managers, Insurance and Pension Funds.

The goal of our machine learning algorithms is to accurately summarize the decarbonization targets and approaches of a company through understanding the context around a particular word/phrase. Various popular natural language processing techniques including Sentiment Analysis and Topic Modelling were adopted to enhance the performance of our models.

Upon several Artificial Intelligence implementations, unstructured text data was converted into meaningful insights. Finally, an informative dashboard was built to aid users in making business decisions. Figure 1 outlines our approach.



*Figure 1. Overall approach*

The following sections detail the analytic requirements, functional and non-functional requirements, dashboard features, use cases and recommendations of this project.

## **2. Analytic Requirements**

The project has adopted the following analytical approaches. Each analytical approach had been carefully selected with the aim of delivering accurate, reliable and relevant information to users. By implementing a range of Artificial Intelligence (AI) techniques, the goal is to aid NatWest Markets in overcoming the issues of low data quality and data comparability for financial institutions and investors seeking to monitor the environmental impact and flows of climate finance.

### **2.1 Text Preprocessing**

Our data source for the project is various companies' reports which come in the pdf format. Our main tool for text extraction is PyMuPDF, a python package that provides a simple user interface for extracting text from any document. The extracted text however is usually messy. It contains unusual text and symbols that need to be cleaned to improve the reliability of the machine learning models mentioned below.

To prepare the text for further analysis, it undergoes four main processes, using the NLTK and Gensim python libraries:

- **Text Cleaning** - Removes noise and unhelpful parts of data by converting all characters to lowercase, getting rid of punctuation marks, special characters, symbols and non-english sentences.
- **Tokenization** - Breaking the raw text into small chunks, known as tokens to help in understanding the context or developing the model for Natural Language Processing.
- **Stopwords Removal** - Removes words such as "a," "our," "for," "in," to help subsequent machine learning models to consider only key features
- **Lemmatization** - Brings words to their dictionary form with the goal of boosting text mining models

Text preprocessing is a crucial step in our project. It effectively transforms the extracted text into a more analyzable form so that our subsequent machine learning algorithms can perform better.

## 2.2 Sentiment Analysis

Sentiment analysis is a text mining technique used to detect positive or negative sentiments in a text.

### 2.2.1 Overall Sentiment Score Calculation

For each type of financial institution (Asian Banks, Asset Managers, Insurance, Pension Funds), we computed the sentiment score for each of the sentences for each company. The sentiment score is calculated using 2 different components. The first component consists of using the TextBlob library available in python which gives us a general sentiment level of each sentence, with a raw score ranging from -1 to 1. However, this alone does not give an accurate gauge of the company's level of optimism towards decarbonization. Thus, we included a second component which takes into consideration frequency of ESG and decarbonization related keywords for each of the sentences. ESG keywords are an extremely popular way of helping investors evaluate companies in which they are interested to invest in. They also help investors avoid companies that might be deemed as having significantly greater financial risks due to their poor environmental practices, which makes it one of the key metrics to be included in the computation of our overall sentiment score. Some examples of these keywords include "clean technology", "sustainable technology" and "green bonds" that are good indicators of companies overall decarbonization efforts. For each sentence, the calculation of the final sentiment level is summarised using the formula illustrated below.

$$Sentiment(Comp) = TextBlob\ Score + \log(1 + (\Sigma keywords \times \Sigma unique\ keywords))$$

Using the formula, if a company does not have any ESG or decarbonization related keywords in their company reports, their sentiment score will be entirely dependent on the TextBlob score. For companies with a higher number of total keywords and number of unique keywords, their sentiment score will be higher indicating that they are more supportive towards decarbonization. This formula gives a relatively accurate approximation of a company's current overall decarbonization efforts.

With the score generated, we can now move on to conduct sentiment analysis. Using Sentiment(Comp), the label for our machine learning models, optimism, was engineered as shown in Figure 2.

Sentiment(Comp)	Label
Less than 0	Pessimistic
0 to 0.5	Neutral
More than 0.5	Optimistic

Figure 2. Table summarizing the labels for sentiment analysis

Before the company reports sentences are fed into our machine learning algorithms, they are converted into a Bag-of-Words (BOW) model<sup>1</sup> using the CountVectorizer function. The top 1000 most relevant words were selected to predict a sentiment score for the BOW model, which is used as training data for machine learning algorithms. This is to prevent overfitting due to high dimensionality of the data, and to speed up the time taken to train our model.

Document No.	ability	about	above	access	accordance
0	1	0	0	0	1
1	0	1	0	0	0
2	0	0	1	0	0
3	0	1	0	0	1

Figure 3. An Example Of the Bag Of Words Model

As shown in Figure 3, the BOW model helps to convert unstructured text data into fixed length vectors, with 1 indicating the presence of the word and 0 indicating the absence. For example, document 0 contains the words “ability” and “accordance”.

Once created, we splitted the dataset into 70% training data and 30% testing data and proceeded to build two algorithms, Decision Tree and Linear Regression to predict the level of optimism towards decarbonization in unseen sentences. The predicted sentiment score gives the user a gauge of a company's willingness and progress towards decarbonization in the future The list of predicted sentiment scores also allows the user to make comparisons between 2 or more companies of interest. Figure 4 shows a sample of the results from five Asian Bank companies.

<sup>1</sup> The Bag-of-Words (BOW) model extracts the different features and vocabulary from the text data. The columns contain the unique words in the reports, whereas the rows consist of the frequency of each word per sentence.

Company	Type of Financial Institution	Actual Sentiment Score, Sentiment(Comp)	Predicted Sentiment Score from Decision Tree	Predicted Sentiment Score from Regression
AIIB	Asian Banks	0.3979	0.3833	0.3428
BDO	Asian Banks	0.3718	0.3629	0.3453
BEA	Asian Banks	0.3588	0.3587	0.3368
BNI	Asian Banks	0.1806	0.1770	0.2047
BOC	Asian Banks	0.0661	0.0695	0.1047

*Figure 4. Sample sentiment analysis results*

As shown in Figure 4, users can easily view and compare the predicted sentiment score generated by the Decision Tree model and Linear Regression Model. The predicted sentiment scores from the decision tree indicated in Figure 4 will also be displayed in our dashboard.

The relevant metrics such as the root mean squared error (RMSE) and mean absolute error (MAE) were computed to evaluate the performance for each of the models. From Figure 5, the Decision Tree model is superior to the Linear Regression Model due to its lower RMSE and MAE values.

	Linear Regression		Decision Tree	
Type Of Financial Institution	RMSE	MAE	RMSE	MAE
Asian Banks	0.395	0.180	0.273	0.135
Insurance	0.352	0.177	0.262	0.109
Asset Managers	0.341	0.204	0.278	0.129
Pension Funds	0.331	0.197	0.273	0.116

*Figure 5. Evaluation of regression models on test data*

Apart from the prediction of sentiment scores, classification models were also built to predict the sentiment label of companies, whether they are likely to be favourable, against or neutral towards portfolio decarbonisation. Similarly, we also splitted the dataset into the 70% training data and 30% test



data, and computed the average of the accuracy score, recall and F1 score of the model performance on the test data. We built 4 models, which are decision tree, logistic regression, boosting and naive bayes.

Model	Accuracy	Recall (Optimistic)	F1 Score
Decision Tree Classifier	0.848	0.810	0.823
Logistic Regression Classifier	0.862	0.735	0.820
AdaBoost	0.723	0.220	0.363
Naive Bayes Classifier	0.723	0.748	0.675

*Figure 6. Evaluation of classification models on test data*

As shown in Figure 6, the Logistic Regression classifier gave the highest accuracy of 0.862, whereas Decision Tree gave the best recall and F1 score of 0.810 and 0.823 respectively. Recall is the most important metric in determining the best model because investors are more interested in correctly predicting companies which are optimistic towards decarbonisation. Models such as AdaBoost, with a high accuracy but a low recall score implies that the model correctly predicted neutral and pessimistic sentences, but incorrectly predicted the majority of the optimistic sentences. As a result, we concluded that Decision Tree was the best model for sentiment label prediction.

## 2.3 Topic Modelling

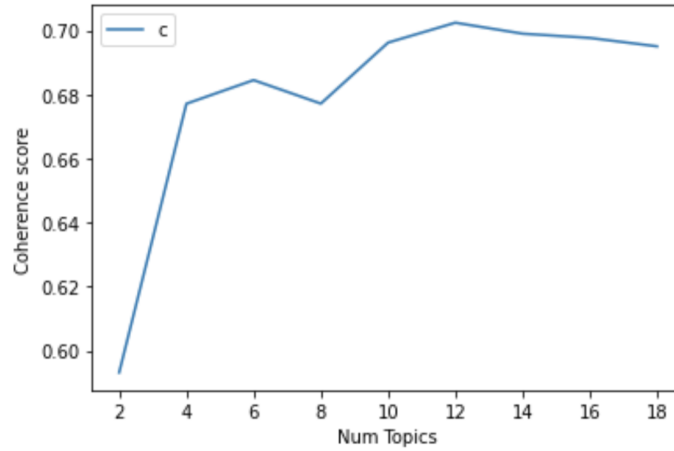
Topic modelling is an unsupervised machine learning technique used to discover a set of topics and words from a collection of texts. This technique is applied to automatically identify the key topics that were discussed in the various reports.

### 2.3.1 Optimal LDA Model

Specifically, the Latent Dirichlet Allocation Mallet (LDAMallet) algorithm from the Gensim package was used in our implementation. The performance of the model depends heavily on the optimal number of topics and is measured by a coherence score<sup>2</sup>. The base model was initialised with 20 topics, 10 words in each topic and achieved an initial score of 0.695. Subsequently, hyperparameter tuning was conducted to find the optimal model by building many LDA models with different numbers of topics, ranging from 2 to 20.

---

<sup>2</sup> Coherence score is the evaluation metric for LDAMallet models.



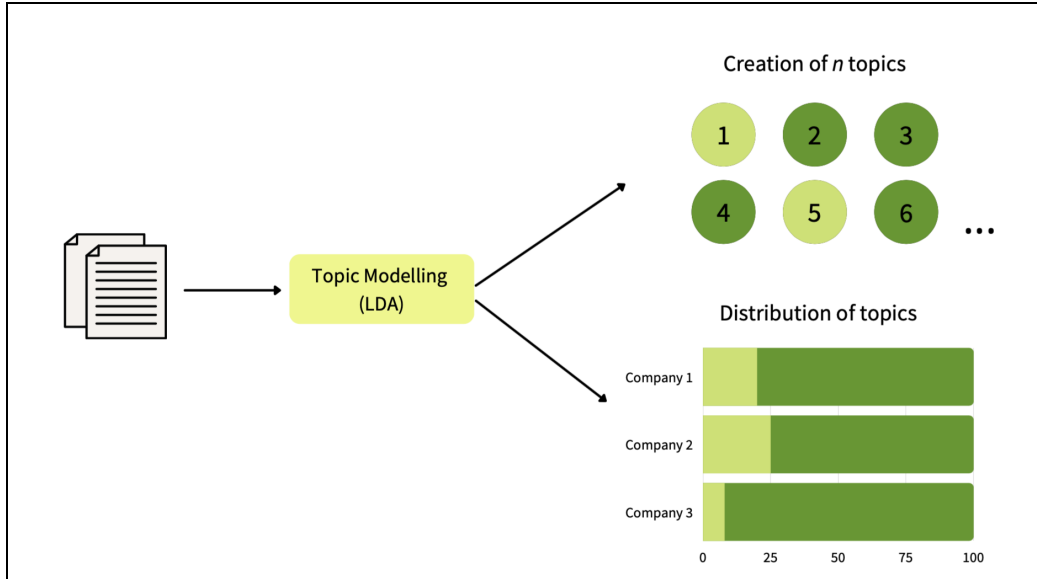
*Figure 7. Optimal LDAMallet Model*

Based on Figure 7, the model with 12 topics is chosen as the optimal model as it attained the highest coherence score of 0.7024.

### **2.3.2 Deriving Distribution of Decarbonization Related Topics**

Following which, our algorithm assigned a dominant topic to each sentence of a report. Topics 2, 4 and 5 contain more keywords related to portfolio decarbonization and were assigned topic 'E', while the remaining topics were assigned topic 'SG' (Appendices 8.2). Finally, for each company, the percent of 'E' sentences was calculated using the formula below. This percentage represents how much decarbonization related information contributed to a company's report and a company's degree of decarbonization commitment can be inferred.

$$Percent(comp) = \frac{\sum "E" sentences}{\sum sentences} * 100\%$$



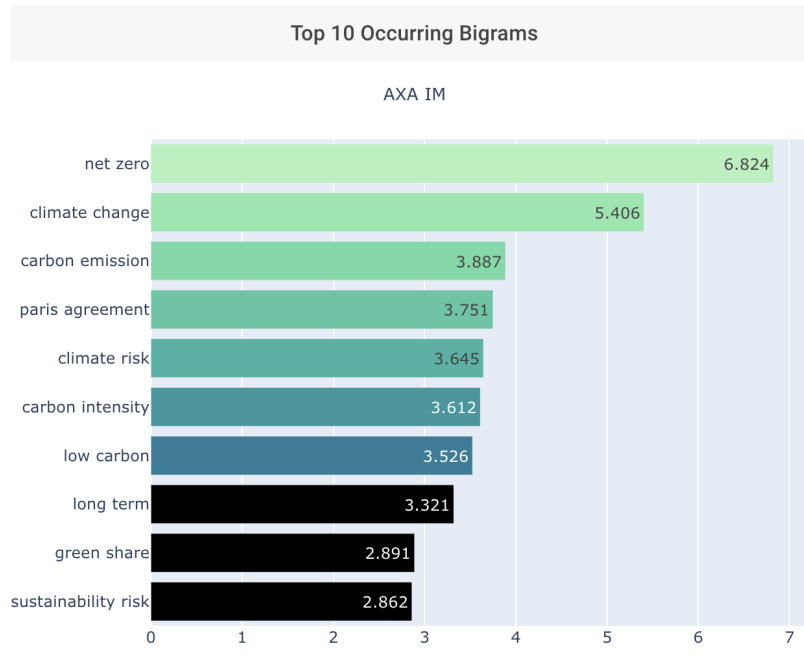
*Figure 8. Topic modelling technique and steps*

Figure 8 shows a summary of how topic modelling is being used to determine the distribution of decarbonization topics.

## 2.4 Bigram Analysis

Besides examining individual words in each company's report, we expanded our analysis to exploring common word pairs, also known as bigrams. After successfully extracting decarbonization related text through topic modelling as mentioned in 2.3, the TfidfVectorizer library was utilised to further extract the top ten occurring bigrams in each company's reports.

Bigrams are especially useful since a very large text dataset is used. Pairs of consecutive words might capture structure that is not present when one is just counting single words, and may provide context that makes words more understandable. For example, "green bond" is more informative than "green" and similarly, "sustainable finance" is more explanatory than "finance".



*Figure 9. An example of a result obtained from bigram analysis*

Figure 9 displays the top ten bigrams extracted from the 2021 TCFD Report of AXA Investment Managers and the value displayed in each bar chart represents the TF-IDF, short for term frequency–inverse document frequency, a numerical statistic that reflects how important a word pair is to a document in a collection. It is evident that bigrams effectively displays the key decarbonization topics that are discussed by a company and informs users of its decarbonization approach and targets. This increases the relevance of the information intended to be shown on the dashboard.

## 2.5 Word2Vec Model

The final model developed is the Word2Vec Model, which is a deep learning algorithm used to determine the similarity of words in the sustainability reports. Word2Vec is a method of computing vector representation of words and text using Continuous Bag Of Words, which is an extension of the bag of words model used in the prediction of our overall sentiment score. It helps to suggest similar words in the sustainability report to the word inputted by the user. The similarity measure between words in the vector space is computed using the cosine similarity formula. The values of cosine similarity ranges from 0 to 1, with 1 indicating that the word is perfectly similar and identical to the inputted word, whereas a value of 0 indicates that there is completely no relevance to the inputted word.

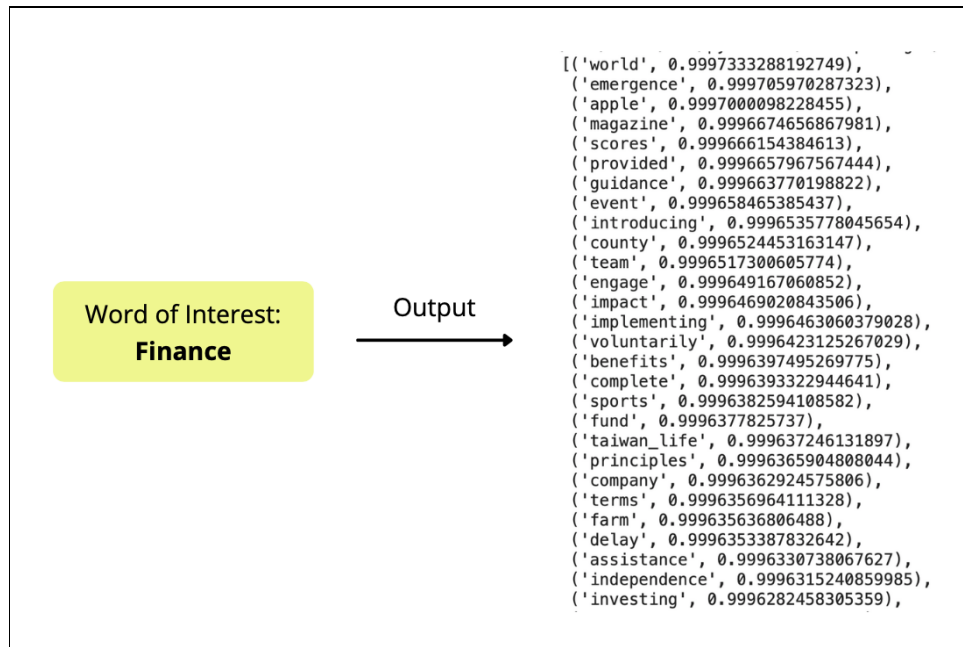


Figure 10. Example of the word2vec similar word function

Figure 10 gives an illustration of how the algorithm works. The function will produce a list of the top 50 most similar words to the word the user inputted, together with its similarity score with a value ranging from 0 to 1.

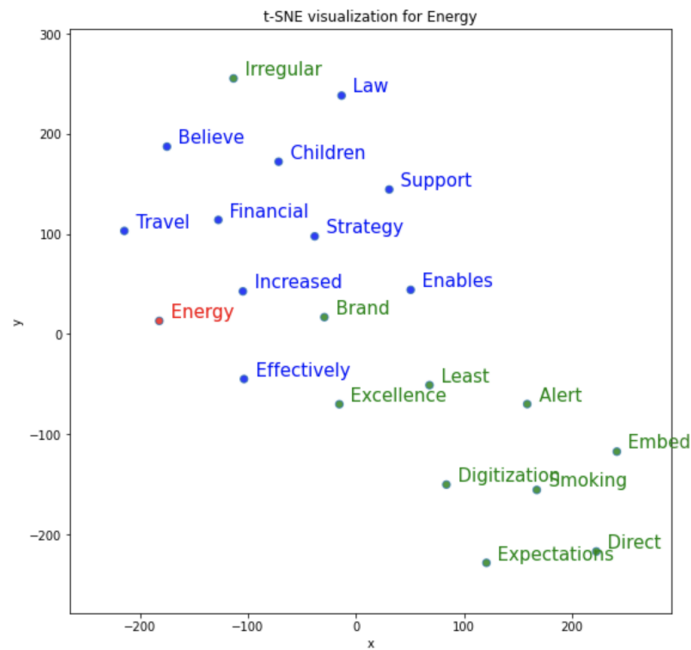


Figure 11. Example of a TSNE visualization

In addition, a TSNE (t-distributed stochastic neighbour embedding) visualisation was also generated (Figure 11). Unlike the Word2Vec model which gives the similarity measure between words in the form of numbers, the TSNE visualisation gives the user a visual representation of similar words. Each word is represented in the 2 dimensional feature space and visualised. Words of high similarity will be closer to each other, whereas unrelated words will be further apart. As shown in figure 11, the word in red represents the word inputted into the function by the user, while the other words in blue and green represent the words of greater similarity to the input word.

### **3. Functional & Non-Functional Requirements**

The following functional and non-functional requirements describe how the dashboard should behave and the necessary system attributes. This is a critical step for the success of the project and to ensure that users' expectations are met.

#### **3.1 Functional Requirements**

##### **3.1.1 For Overall Dashboard**

- a. The system will display a dashboard that contains key performance indicators related to decarbonization that conveys useful information to users.
- b. The dashboard will provide users with two tabs in order to cater to different purposes: (1) To view a specific company; (2) To compare between two companies at once.

##### **3.1.2 For First Tab of Dashboard (Individual Company)**

- a. The dashboard will enable users to select the type of financial institution and the specific company to be displayed through dropdown menus.
- b. The dashboard will enable users to view data that is relevant to the company chosen.
- c. The dashboard will display a company's approach and attitude towards decarbonization.
- d. The dashboard will display how widely a company discusses its decarbonization approaches in their report.
- e. The dashboard will display a company's level of commitment towards global standards and initiatives regarding sustainable finance.
- f. The dashboard will display the most frequently used bigrams and its respective count for users to get a rough idea of the most pertinent issue discussed by each company.
- g. The dashboard will provide appropriate and relevant comparison with the average performance of all companies belonging to the same type of financial institution.

##### **3.1.3 For Second Tab of Dashboard (Company Comparison)**

- a. The dashboard will enable users to select the type of financial institution and two companies for comparison through dropdown menus.
- b. The dashboard will provide comparison of the decarbonization performance of any two selected companies at a glance.

- c. The dashboard will compare the top ten occurring bigrams in each company's report to compare and contrast both companies' decarbonization approaches.

## **3.2 Non-Functional Requirements**

### **3.2.1 Usability**

- a. The fonts and font size should be clear and easy to read.
- b. The background color of the page should be simple and not distracting.
- c. The dashboard should be easily navigated.
- d. The dashboard should render well on a variety of devices and windows with different screen sizes. The displayed information should not be cut off or non-visible.
- e. The dropdown menu and search function enables users to look for their desired companies in the shortest time possible.

### **3.2.2 Scalability**

- a. The dashboard will be able to handle more data as time progresses without compromising on the speed that data will be processed.



## **4. User Interface of Dashboard**

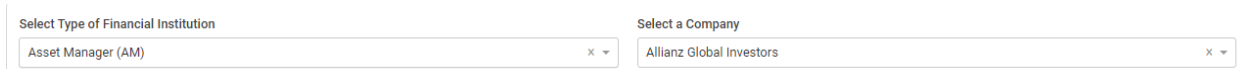
The dashboard is divided into two tabs. The first tab enables users to view the relevant metrics of a selected company while the second tab enables users to compare the performance of any two selected companies of the same type of financial institution.

We have chosen key performance indicators which aim to convey insightful and comprehensible information to users. These performance indicators effectively capture the decarbonization targets and approaches of companies from the various financial institutions.

### **4.1 First Tab (Individual Company)**

The first tab focuses on displaying data of the selected company.

#### **4.1.1 Dropdown Menus**



The image shows two side-by-side dropdown menus. The first menu is titled 'Select Type of Financial Institution' and has 'Asset Manager (AM)' selected. The second menu is titled 'Select a Company' and has 'Allianz Global Investors' selected. Both menus have a small 'x' icon and a downward arrow on the right side.

*Figure 12. Dropdown menus of the first dashboard tab*

Two dropdown menus are provided:

1. Users can choose one out of the four possible types of financial institution (Asset Manager, Asian Banks, Insurance Company, Pension Funds).
2. Users can select or search for a company that belongs to the chosen type of financial institution. The dropdown list is arranged in alphabetical order to facilitate the user's search process.

### 4.1.2 Overall Sentiment Analysis

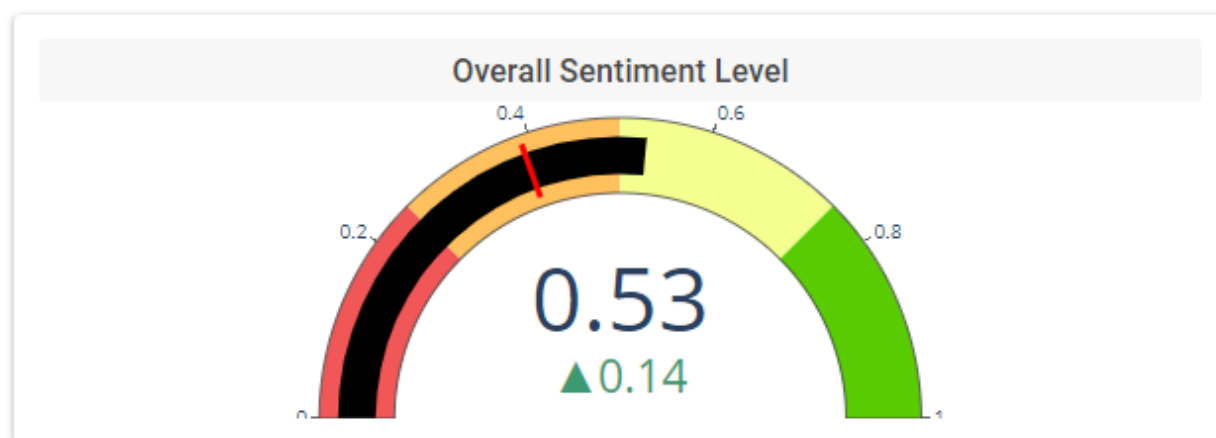


Figure 13. Gauge chart displaying the overall sentiment level

The overall sentiment level of a company is displayed as a gauge chart. The black bar represents the company's score while the red bar represents the average sentiment score of all companies under the same type of financial institution. The score is calculated based on the polarity of the sentence and any mention of ESG related keywords.

The overall sentiment level is calculated to show the company's sentiments/views towards portfolio decarbonization. A greater sentiment score (indicator pointing more towards the green side) suggests that a company is more supportive and more willing to work towards decarbonization. In short, a greater sentiment score represents that a company is more willing to make progress in reducing exposure to carbon risk and align with a low-carbon future.

### 4.1.3 Percentage of Decarbonization Disclosure

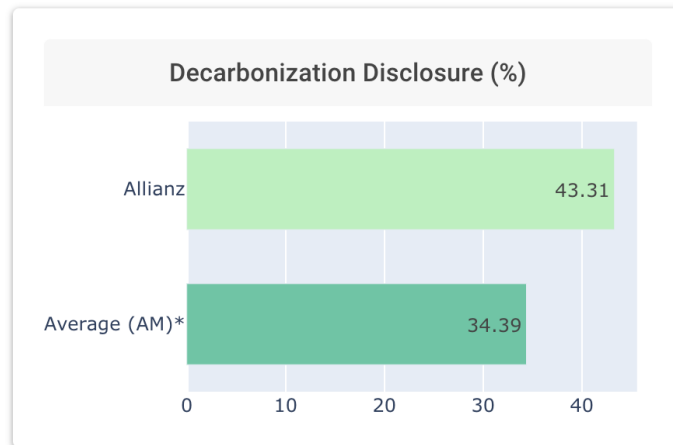
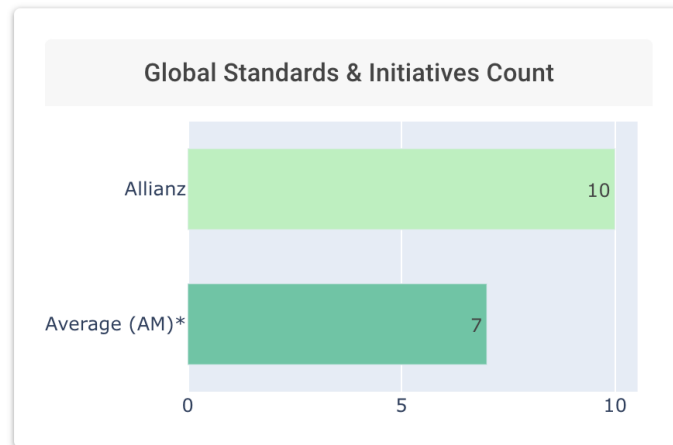


Figure 14. Bar chart displaying the percentage of decarbonization disclosure

The percentage of decarbonization disclosure is displayed in a bar chart. It measures how much of a company's report is dedicated to discussing decarbonization. We conducted topic modelling and obtained 12 key topics that were covered in each company's report. We further categorised these topics into 2 sub-categories: (1) topics related to decarbonization; (2) topics unrelated to decarbonization. Next, our model assigned a dominant topic to each sentence. We then calculated the percentage of sentences related to decarbonization across reports to see how widely each company discusses decarbonization in their reports.

The percentage of decarbonization disclosure is a representation of how widely a company discusses decarbonization compared to all other companies in our dataset. Our machine learning model works on the belief that greater mentions of decarbonization in a company's report is equivalent to a company putting in more effort to integrate climate considerations into their investment portfolio. Additionally, a greater decarbonization rating also implies that a company has greater transparency in disclosing information related to decarbonization.

#### 4.1.4 Number of Global Initiatives & Standards



*Figure 15. Bar Chart displaying the number of global standards & initiatives*

The number of global standards and initiatives that a company is committed to is displayed in a bar chart. Our algorithm will run through the extracted text to capture the various global initiatives and standards that were adopted by each company and subsequently keep track of the total count. These initiatives are widely adopted and recognised by organisations globally. Some examples include UN Global Compact, UNEP Financial Initiative and the Carbon Disclosure Project.

The number of global initiatives participated by each company is an estimation of how committed a company is towards decarbonization. We may infer that a greater number of global initiatives means that a company is more proactively working out their strategic approaches to decarbonise their investment portfolios.

#### 4.1.5 Table of Global Initiatives & Standards

Global Standards & Initiatives		
Initiatives	Acronym	Details
Carbon Disclosure Project	CDP	A global environmental initiative that helps corporations disclose their climate change response information and mitigate climate change risk
Climate Action 100+	-	An initiative to ensure the world's largest corporate greenhouse gas emitters take necessary action on climate change
Climate Bonds Initiative	CBI	An initiative that works on mobilising the \$100 trillion bond market for climate change solutions
Green Bond Principles	GBP	An initiative that supports companies in financing environmentally sound and sustainable projects that foster a net-zero emissions economy
Institutional Investors Group on Climate Change	IIGCC	A collaboration among European members to help define the investment practices, policies and corporate behaviours required to address climate change
International Corporate Governance Network	ICGN	An initiative that advances the highest standards of corporate governance and investor stewardship worldwide in pursuit of a sustainable economy
Investor Group on Climate Change	IGCC	A collaboration of Australian and New Zealand institutional investors with the aim of building a climate-resilient net zero emissions economy by 2050
Principles for Responsible Investment	PRI	A principle-based framework that offers a list of possible actions for incorporating ESG issues into investment practice
Sustainability Accounting Standards	SASB	An initiative that facilitates the disclosure of material and decision-useful climate-related information to investors

*Figure 16. Table summarizing the list of global standards & initiatives*

The table summarizes the different global initiatives and standards that a company is committed to as well as provides a brief description of each initiative. It shows users how a company is staying ahead of climate-related risks by adopting best practices which are globally recognised. The table serves as an evaluation of how committed a company is in aligning ESG management with sustainable finance.

The table also has a scroll function to cater to companies that have more initiatives to be displayed. This also helps to improve the user's experience when dealing with a huge amount of data.

#### 4.1.6 Top Ten Occuring Bigrams

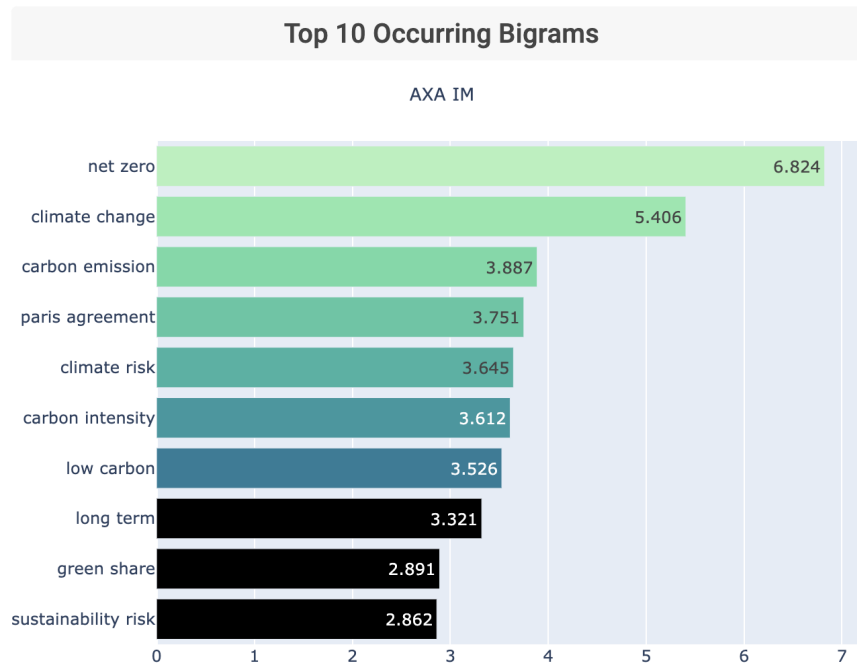


Figure 17. Bar chart displaying top 10 occurring bigrams

The top ten occurring bigrams that appear most frequently in a company's report and their respective count will be displayed in a bar chart. These words highlight the commonly discussed decarbonization related topics by a company. For example, looking at the top ten bigrams of AXA Investment Managers, we can conclude that some of their decarbonization targets include achieving 'net zero' emissions and aligning with the 'paris agreement'.

## 4.2 Second Tab (Company Comparison)

The second tab of the dashboard enables users to compare between any two selected companies faster and more efficiently.

### 4.2.1 Dropdown Menus

Select Type of Financial Institution	Select the First Company	Select the Second Company
Asian Bank (AB) ×	BDO Unibank ×	China Merchants Bank ×

Figure 18. Dropdown menus of the second dashboard tab

Three dropdown menus are provided:

1. Users can choose one out of the four possible types of financial institution (Asset Manager, Asian Banks, Insurance Company, Pension Funds).
2. Users can select or search for two companies that belong to the chosen type of financial institution for comparison. (second and third dropdown menus)

### 4.2.2 Comparison - Overall Sentiment Level

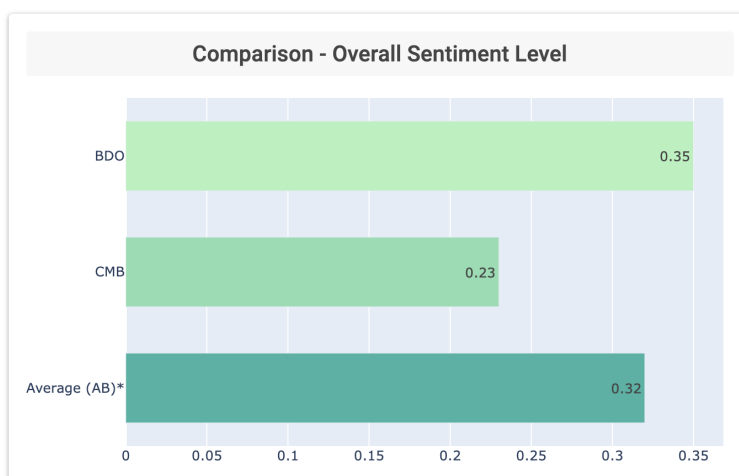
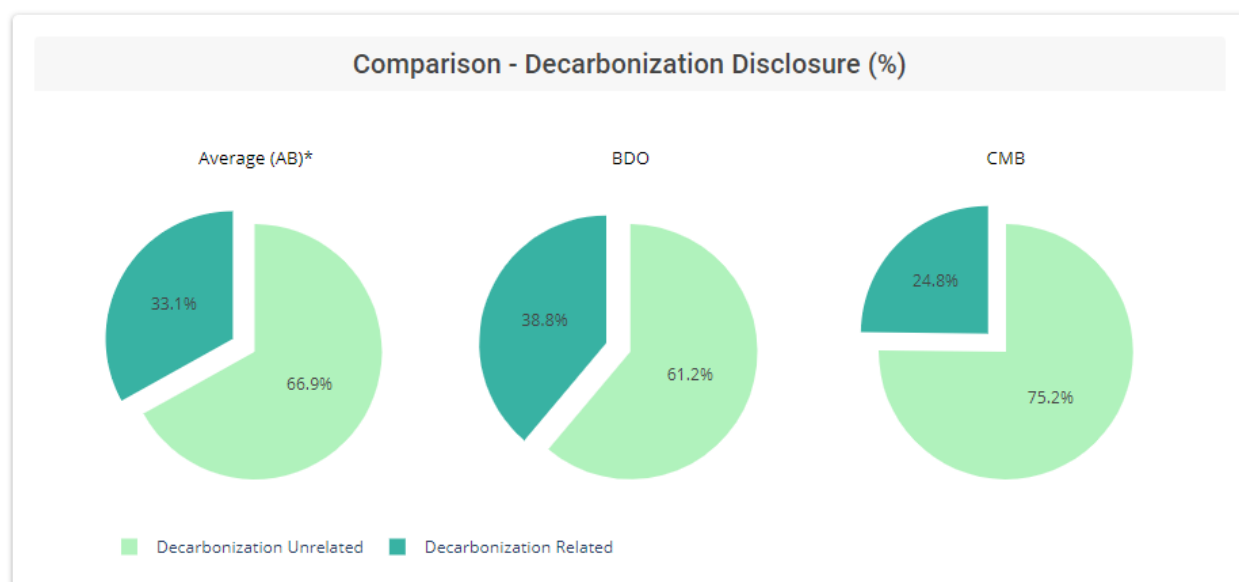


Figure 19. Bar chart comparing the overall sentiment level

The bar chart compares the attitude and engagement of both companies towards decarbonization. A higher sentiment score implies that a company has a more positive outlook of decarbonization and is likely to be more confident and better prepared to adapt to a low-carbon future.

### 4.2.3 Comparison - Percentage of Decarbonization Disclosure



*Figure 20. Pie charts displaying the breakdown of information in a company's report*

The respective pie charts display the breakdown of decarbonization related and decarbonization unrelated text in a company's report. These values are obtained through topic modelling and provides users with a visualization of how much focus a company puts on decarbonizing their portfolio.

A company that has a greater percentage of decarbonization related text in their company is more likely to be incorporating decarbonization and climate related factors in their investment decisions. Besides the comparison between both companies, the average performance of all companies under the same type of financial institution is also included for easier comparison with their relevant counterparts.



4.2.4 Top Ten Occurring Bigrams

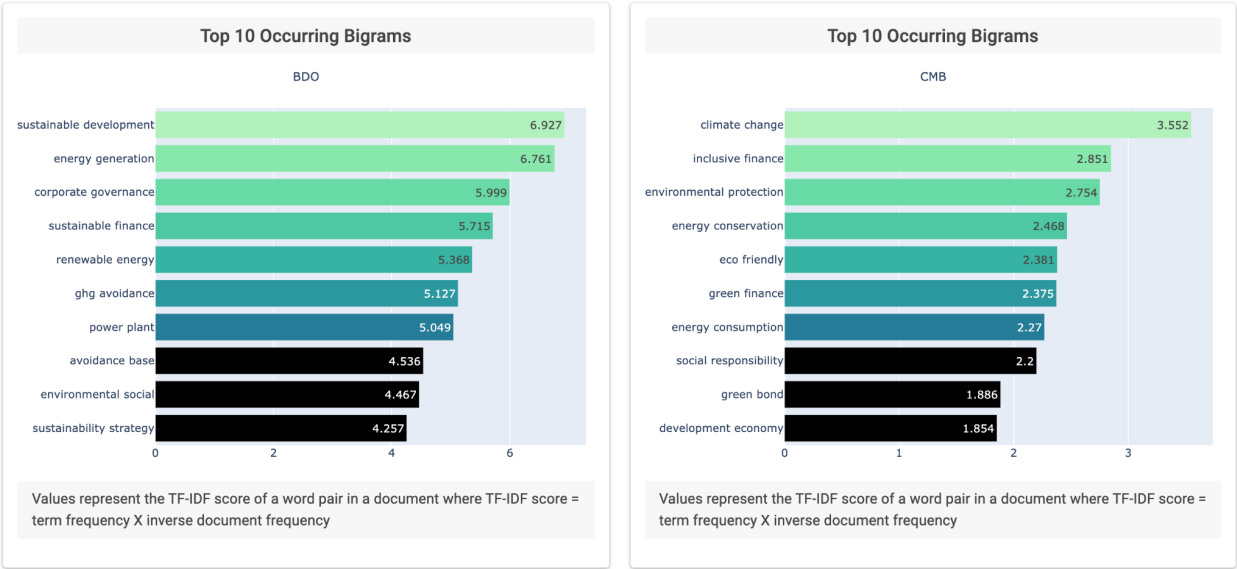
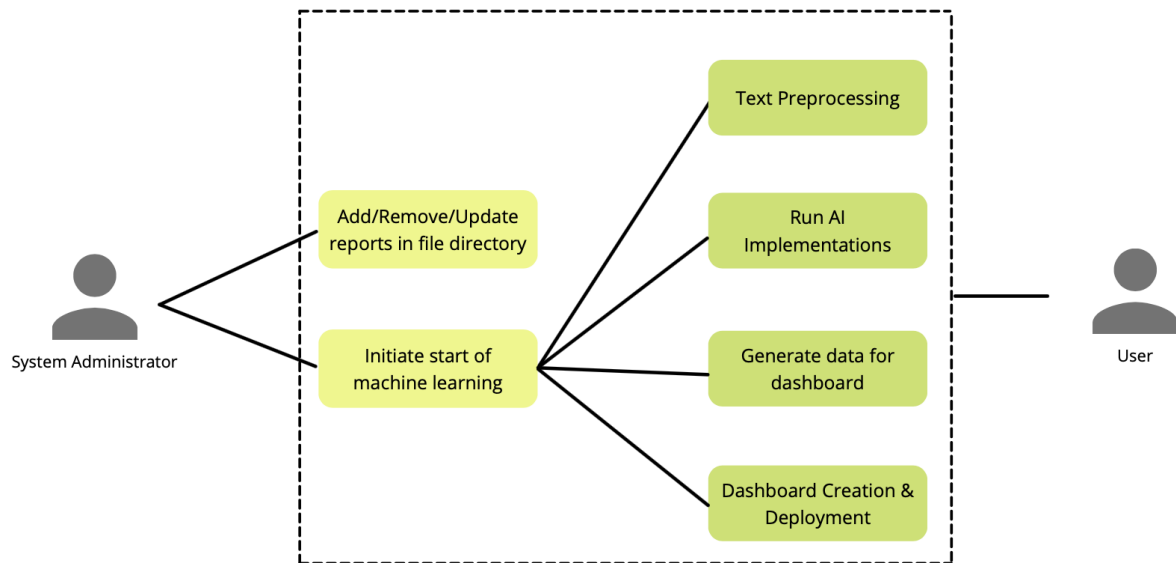


Figure 21. Bar charts displaying top 10 occurring bigrams

The final bar charts compare the top ten occurring bigrams that are extracted from both selected companies. The bigrams are able to display the most commonly discussed topics by each company. Similar to the top ten words, if most of the bigrams are related to the theme of decarbonization, we may infer that a company is more proactively contributing to decarbonizing the economy.

## **5. Use Case**



*Figure 22. Use case diagram*

Please refer to the user guide document (BT4103\_Group 7\_Project\_Documentation.pdf) for the detailed procedure on how to use our system.

### **5.1 Deployed Dashboard**

URL: <https://bt4103-esg-dashboard.herokuapp.com/>

## **6. Future Work**

Due to time constraints, we did not manage to implement all the features that we initially planned to have. This section summarises the features that can be added for future enhancements.

### **6.1 Drill Down on Bigram Analysis**

In the second interface of the dashboard, we currently only present the top ten occurring bigrams for a company report. It would be more informative if users could view the specific sentences that contained the particular bigram to get the full picture and gain greater insights of the decarbonization approach of a company.

### **6.2 Expedite LDA Training Process**

The current time required to train our LDA model is very long and may be a hindrance if users are seeking a time-efficient implementation. There are other options that may be explored which have the potential of speeding up the current process. Users may wish to explore with trigrams and different lemmatization options.

### **6.3 Improvements to Dashboard**

#### **6.3.1 Increase Number of Filter Options for Dashboard**

We can expand the number of filter options available for users to improve their search experience. Other possible filter options include, filter by country and filter by regions.

#### **6.3.2 Add a ranking interface**

Developing an overall ranking of companies would be helpful for users if they want to find out which company ranks higher than another after taking into consideration all the metrics. In addition, users will be able to take a quick glance at the top few companies pledging to portfolio decarbonization.

## **7. Conclusion**

Overall, we understand that there is an increasing number of financial institutions that are setting decarbonization targets in hopes of transitioning to a sustainable low carbon economy. Yet, there lacks automated and effective methods to consolidate and compare between the large amounts of text data sources for investors, who are seeking to make investment decisions and monitor environmental impact and flows of climate finance. Therefore, this project aims to deliver a reliable and useful machine learning algorithm to turn unstructured text into informative data that translates it into an intuitive dashboard. Subsequently, by delivering a consistent flow of content related to decarbonization, our dashboard hopes to help investors evaluate their options across a comprehensive set of KPIs which would aid them in making informed business decisions.

## 8. Appendices

### 8.1 Sentiment Analysis

#### 8.1.1 Bag of Words model

	ability	able	about	above	access	accordance	according	account	accounting	accounts	achieve	acros
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
49033	0	0	0	0	0	0	0	0	0	0	0	0
49034	0	0	0	0	0	0	0	0	0	0	0	0
49035	0	0	0	0	0	0	0	0	0	0	0	0
49036	0	0	0	0	0	0	0	0	0	0	0	0
49037	0	0	0	0	0	0	0	0	0	0	0	0

49038 rows × 1001 columns

#### 8.1.2 Predicted sentiment scores

	name	sentiment_score	type	predicted_sentiment_tree	predicted_sentiment_regression
0	AIIB	0.397096	ab	0.413725	0.394479
1	BDO	0.350231	ab	0.346698	0.324633
2	BEA	0.338333	ab	0.338799	0.332083
3	BNI	0.198146	ab	0.200520	0.213761
4	BOC	0.092336	ab	0.093589	0.112981
...	...	...	...	...	...
113	Prudential	0.433567	ins	0.421702	0.406408
114	Shin Kong	0.256182	ins	0.253850	0.246324
115	Sun Life	0.311841	ins	0.302115	0.293463
116	Taiwan Life	0.225408	ins	0.215581	0.223386
117	Tokio Marine	0.392865	ins	0.387729	0.369711

118 rows × 5 columns

## 8.2 Topic Modelling Topics & Assigned Category

Topic No.	Keywords for Each Topic (in decreasing importance)	Assigned Topic Category (E / SG)
1	year, increase, number, time, plan, include, rate, period, benefit, pay	SG
2	risk, climate, change, impact, related, manage, include, process, relate, analysis	E
3	work, support, employee, training, community, people, program, promote, programme, local	SG
4	energy, project, emission, green, carbon, reduce, activity, sector, sustainable, low	E
5	sustainability, social, governance, corporate, sustainable, environmental, approach, stakeholder, gri, performance	E
6	business, development, continue, develop, focus, strategy, improve, build, society, growth	SG
7	policy, information, system, datum, conduct, internal, control, ensure, compliance, establish	SG
8	market, term, equity, include, level, fund, bond, base, share, return	SG
9	company, issue, client, financial, responsible, good, engagement, invest, industry, make	SG
10	customer, provide, service, product, health, support, branch, covid, launch, offer	SG
11	asset, total, interest, loan, net, fair, loss, amount, statement, note	SG
12	director, annual, executive, member, committee, review, shareholder, meeting, report, audit	SG

### 8.3 Top 10 Word Count

This was an initial data visualization that was planned to be included in our dashboard. Upon discussion with Natwest Markets, we have decided to replace it with the top ten occurring bigrams bar chart instead.

