School of Computing

Using Machine Learning to Explore Decarbonization Strategies

Group 07 – NatWest Markets

Lam Pei Shi (A0204420B)

Lee Chen Xi (A0220687B)

LIN DA (A0201588A)

Milton Sia (A0217597N)

Su Zifeng (A0222369A)

1. SUMMARY

1.1 Problem Statement

Environmental, Social, Governance (ESG) investing is a rising trend with increased awareness and concerns about sustainability efforts. In addition, there are propositions that better ESG performance tends to yield better financial performance and improved risk-returns. Sustainable funds attracted a record $69.2 billion in net flows in 2021 and ESG assets are expected to hit $53 trillion by 2025 (Bloomberg Intelligence 2021). In particular, carbon emissions have been denoted as the most important measure for measuring ESG by The Economist (2022).

However, the lack of a standardized reporting framework and conflicting ESG ratings from different ESG data providers makes it challenging for investors to obtain curated data for analysis. Manual data retrieval of scattered ESG information across long and unstructured annual and sustainability reports is not only labour intensive, but also prone to human errors. Potential consolidated ESG information and ratings on online sources are likely to be incomplete and unreliable. As such, this project aims to develop an end-to-end data driven solution for our project sponsor, NatWest Markets, to provide insights into decarbonization efforts of companies from publicly available unstructured and structured data.

1.2 Proposed Solution

Based on a list of decarbonization questions provided by NatWest, our final objective is to answer these questions and present them in an interactive and comprehensive manner via a web application dashboard. To do so, we start from a list of companies with attributes such as company name, sector, subsector and sustainability report URLs. In our data collection and pre-processing step, we would retrieve the PDF reports, extract textual and tabular data, and prepare the textual data relevant to each question to be pipelined into machine learning models. In the machine learning step, we use a mix of supervised and unsupervised machine learning methods for binary, scale and relevant sentence prediction. Finally, we present the answers to the questions via a full stack web application hosted on the Dataiku platform, which is the designated data science platform used by NatWest Markets.

1.3 Project Sponsor Use Cases

NatWest Markets is the investment banking arm of NatWest Group, the largest business and commercial bank in the UK. NatWest Markets help corporate and institutional customers manage their financial risks and achieve their short and long-term financial goals, all while navigating changing markets and regulation. NatWest is committed to acting sustainably and responsibly, and actively supports customers in their transition to achieving broader environmental and societal goals, working with issuers and investors to develop holistic sustainability strategies. It intends to at least halve the climate impact of its financing activities by 2030 (NatWest Group 2022). We hope our dashboard would be able to help NatWest achieve their goal and aid the different departments of the bank with their daily tasks more efficiently.

On the client servicing front, we hope to better support clients on their climate journey with easier access to consolidated ESG information, reducing the need of laborious processes prone to human error. For example, the traditional investment banking arm could utilize data from our dashboard to facilitate the creation of decks and memorandums while providing advisory for M&A deals and during fund-raising.

With green financing as a rising trend and an increasing number of investors with a sustainable mandate, the incorporation of ESG information would certainly appeal to NatWest's clients.

Moreover, succinct information on companies' ESG performance could shed some colour on potential financial impacts of climate risk. Such information could support NatWest in making better investment decisions and recommendations according to relevant risk appetite requirements.

In line with NatWest's target to achieve Net Zero emissions by 2050 through its financing activities, detailed aspects on companies' current carbon emissions, future targets and their plans for transition could aid NatWest in structuring better incentives for potential financing targets to reduce emissions and better facilitate business planning for these companies.

2. ANALYTICAL REQUIREMENTS

2.1 Types of Questions to be Answered in the Dashboard

NatWest has provided our team with 21 initial questions to address using publicly available sustainability reports. Each of these questions fall into one of three categories, namely, the current state of emissions, targets for future emissions reduction, and transition plans for the company's decarbonization strategy.

Of these 21 questions, 7 were removed in the process of data exploration, as they were deemed to be either infeasible to obtain or reported inconsistently across different companies that reduced its informativeness. For the remaining 14 questions, they were categorized into 4 different types, namely Binary, Scale, Tabular and Open Ended, depending on the question's requirements and how the data is presented in the reports.

For instance, absolute emissions data is usually stored in the tables, which comes in a variety of sizes and formats. As a result, direct extraction of table values was deemed to be infeasible, and we opted to extract tables in image form instead. On the other hand, "medium-term targets for scope 1-2 emissions" is an open-ended question, as companies use varying target and benchmarking years. The questions addressed and classification are shown in the table below.

Current Emissions Historical Performance	What are your absolute Scope 1 and 2 emissions?	Tabular (JPEG)
	What are your absolute Scope 3 emissions?	Tabular (JPEG)
	Have your Scope 1, 2 and Scope 3 emissions been verified by a third party?	Binary
	Do you have an active program to support increasing green space or promote biodiversity?	Binary
Emission Reduction Target	What does your medium term (5 10 years) Scope 1 - 2 target equate to in % reduction?	Open Ended
	What is your medium term (5 10 years) target for reduction in Scope 3 Downstream emissions (e.g. tenant/purchaser activity) and Upstream emissions (e.g. embodied emissions in purchased construction materials)?	Open Ended
	Do you have a long term net zero target/commitment?	Binary
Transition Plans	Do you have a low carbon transition plan?	Scale
	Do you provide incentives to your senior leadership team for the management of climate related issues?	Binary
	Do you engage in activities that influence public policy on climate related issues to support a net zero aligned transition through any of the following? Direct engagement with (1) policy makers; (2) funding research organizations; (3) trade associations	Binary

	Do you set targets for the production or consumption of renewable energy?	Binary
	Does your transition plan include direct engagement with suppliers to drive them to reduce their emissions, or even switching to suppliers producing low carbon materials?	Binary
	Does you set targets for production of more energy efficient or environmentally friendly products (e.g. Carbon Capture, Hydrogen generation, battery storage)?	Binary
	Do you engage with your value chain on climate related issues?	Binary

2.2 Industry Benchmarks

When evaluating the emissions performance of a company, it is insufficient to merely look at the statistics of the company itself. Companies in different industries face different ESG risks and challenges, while the ESG profile of an industry also changes over time, making it necessary to contextualize emissions data with respect to industry benchmarks in the same time period (Alva Group 2020). For each metric we evaluate a company upon, our dashboard would also provide the percentage of industry peers who meet the same benchmark in that year to provide greater insight into the company's emissions performance.

2.3 Success Metrics

Our data pipeline and dashboard would be evaluated by the sector coverage ratio, question coverage ratio and model accuracy. For the sector coverage ratio, we would be limiting the scope to the Energy sector and aim to cover at least 80% of the companies provided. With respect to the 21 decarbonization evaluation questions, our aim is to answer at least 2/3 of the questions, while ensuring that the fundamental questions about absolute emissions are addressed. In terms of model accuracy, we target to achieve average ROC AUC scores of greater than 0.80 for the binary and scale questions.

3. IMPLEMENTATION

3.1 Overall Implementation

The overall implementation pipeline is as follows:

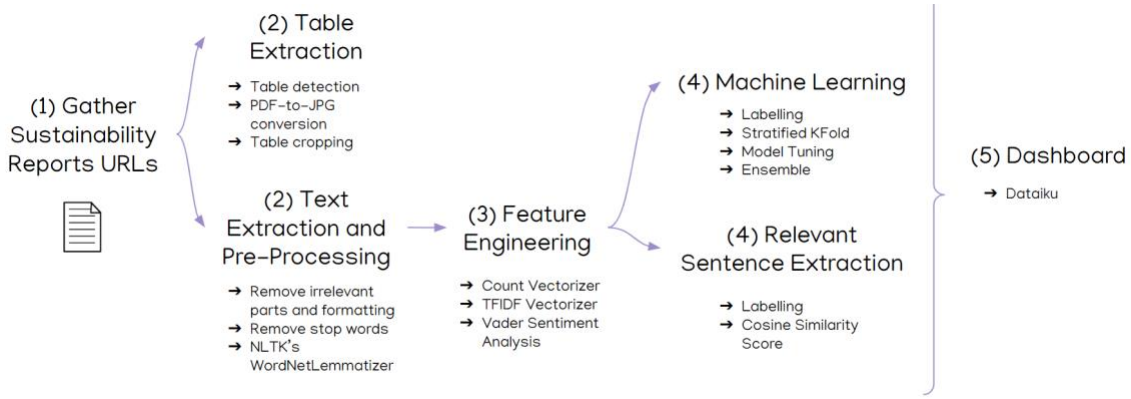


Figure 1: Implementation Pipeline Overview

3.2 Data Gathering

This project utilizes the sustainability reports of companies in the energy sector as its raw data. Given a list of company names and industry sub sectors, we manually retrieved the respective URLs for the sustainability reports. These URLs would be later used to obtain the PDF reports for the table extraction and text extraction steps.

3.3 Tabular Data Processing

3.3.1 Tabular Data Extraction

To process tabular data used to address questions 1 and 2 on absolute Scope 1, 2 and 3 emissions, our table extraction pipeline (1) identifies the page of the emissions table, (2) crops the target table as an image, and (3) uploads the image into Firebase Storage to be displayed in the dashboard.

3.3.2 Table Detection and Page Number Extraction

We first determine the page number that our target table resides in for each report. To do so, we converted the PDF reports into textual format using the report URLs. Subsequently, we filtered for pages that contain the following to locate target tables.

- Keywords regarding the various forms of emissions – Scope 1, Scope 2, and Scope 3
- Permutations of the unit of absolute emissions, such as but not limited to, tonnes co2e, million tonnes co2e and tco2e
- At least 10 numbers
- At least 2 years, such as 2019 and 2020

From the filtered list of page numbers, we then chose pages containing the most numbers. Our study found that pages that contain tables with emission data tend to have more numbers than others.

3.3.3 PDF-to-JPEG Conversion for Target Tables

After identifying the page number containing the required table, the relevant page is converted from PDF to JPEG format and stored locally. The reason for performing image conversion after page number extraction is to minimize memory usage by only storing the necessary image files. Figure 2 illustrates an example of our PDF-to-JPEG conversion algorithm on Suncor Energy Inc's sustainability report.

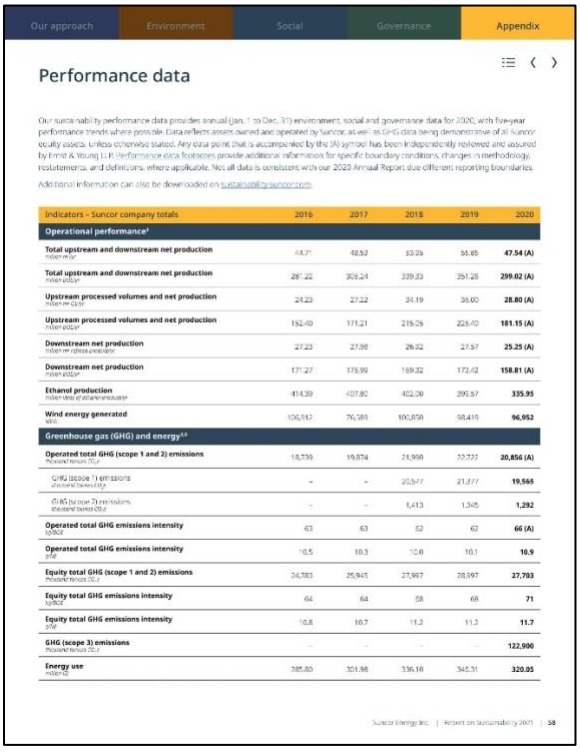


Figure 2: Image of Full Page Containing Targeted Table

3.3.4 Cropping of JPEG Table Images

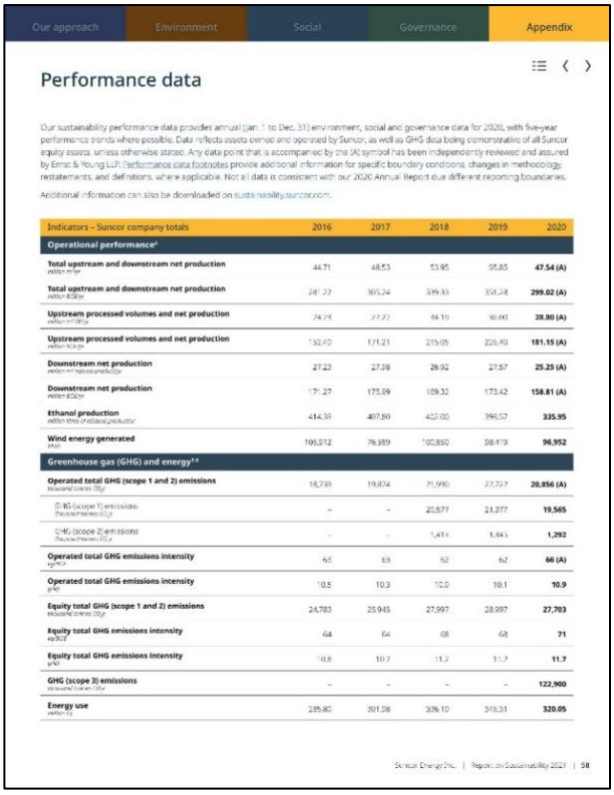
From Figure 2, we observe that the page contains textual data irrelevant to the table we require, which necessitates the table cropping process.

To crop the table out of the page, we first conduct binarization to convert the full colored images into grayscale and threshold the pixel values to reduce image complexity using the cv2 package in Python. Subsequently, we use cv2's morphological operations to apply vertical and horizontal kernels on the image and conduct a connected component analysis to get labels and statistics of the bounding boxes of the table. These statistics returned from the connected component analysis contains the coordinates and area of the detected bounding boxes (Sreekiran 2020).

Using these coordinates and applying our set of filters and heuristics, we locate the coordinates containing the target table and extract the area as an image. Examples of our filters include a minimum and maximum

size of area, as well as thresholds which the coordinates cannot extend beyond. If the coordinates do not satisfy the filtering criteria, we extract a default area guaranteed to contain the target table. Figure 3 illustrates our table cropping algorithm on Suncor Energy Inc's 2021 sustainability report that converts the full page to the target table.

Overall, our table extraction pipeline has produced an 89.7% accuracy rate for both Questions 1 and 2. Figure 4 illustrates the results of our table extraction pipeline in detail.



Before cropping

Indicators – Suncor company totals	2016	2017	2018	2019	2020
Operational performance[a]					
Total upstream and downstream net production million m³/yr	44.71	48.53	53.95	55.85	**47.54 (A)**
Total upstream and downstream net production million BOE/yr	281.22	305.24	339.33	351.28	**299.02 (A)**
Upstream processed volumes and net production million m³ OE/yr	24.23	27.22	34.19	36.00	**28.80 (A)**
Upstream processed volumes and net production million BOE/yr	152.40	171.21	215.05	226.40	**181.15 (A)**
Downstream net production million m³ refined product/yr	27.23	27.98	26.92	27.57	**25.25 (A)**
Downstream net production million BOE/yr	171.27	175.99	169.32	173.42	**158.81 (A)**
Ethanol production million litres of ethanol product/yr	414.39	407.80	402.00	399.57	**335.95**
Wind energy generated MWh	106,912	76,589	100,850	98,419	**96,952**
Greenhouse gas (GHG) and energy[3,4]					
Operated total GHG (scope 1 and 2) emissions thousand tonnes CO_2e	18,739	19,874	21,990	22,722	**20,856 (A)**
GHG (scope 1) emissions thousand tonnes CO_2e	–	–	20,577	21,377	**19,565**
GHG (scope 2) emissions thousand tonnes CO_2e	–	–	1,413	1,345	**1,292**
Operated total GHG emissions intensity kg/BOE	63	63	62	62	**66 (A)**
Operated total GHG emissions intensity g/MJ	10.5	10.3	10.0	10.1	**10.9**
Equity total GHG (scope 1 and 2) emissions thousand tonnes CO_2e	24,783	25,945	27,997	28,997	**27,703**
Equity total GHG emissions intensity kg/BOE	64	64	68	68	**71**
Equity total GHG emissions intensity g/MJ	10.8	10.7	11.2	11.2	**11.7**
GHG (scope 3) emissions thousand tonnes CO_2e	–	–	–	–	**122,900**
Energy use million GJ	285.80	301.98	336.10	346.31	**320.05**

After cropping

Figure 3: Before-and-After Cropping

	Correct	Wrong	Accuracy
Question 1 (Absolute Scope 1 + 2 emissions)	52	11	82.5%
Question 2 (Absolute Scope 3 emissions)	52	11	82.5%

Figure 4: Accuracy for Table Extraction

3.3.5 Cloud Hosting of Image Files

Given that our dashboard takes the form of a web application hosted online, we uploaded the extracted table image files to Firebase Storage, which provides free and easy access via JavaScript APIs to the application backend. As part of the project handover, the Firebase account storing the images would be passed to NatWest and integrated into their existing data infrastructure.

3.4 Textual Data Processing

3.4.1 Text Extraction and Cleaning

pdfminer3 is then utilized to extract the unstructured text from the various PDF files into a String. Subsequently, to convert the unstructured text into cleaned sentences, the following steps were taken.

- Remove header number
- Remove spaces prior to punctuation
- Remove figures not irrelevant to the grammatical structure
- Remove emails and URLs
- Remove multiple spaces
- Remove symbols utilized for formatting
- Join next line with spaces

We also removed sentences with more than 400 words to filter out tables that were unintentionally extracted as sentences.

In addition, abundance of stop words, such as “a”, “the”, “for”, in human language reduces the quality of textual data. Such low-level information is generally deemed as noise to features and contribute little to no unique information in machine learning. (Khanna 2021) As such, stop words were also filtered out from the list of pre-processed sentences.

Finally, to further minimize noise and model complexity, we reduced all words to their root form using *NLTK’s WordNetLemmatizer*. Lemmatization is preferred over Stemming as the former accounts for the context of the word as compared to Stemming which uses a hard-coded rule. (Lang 2022)

3.4.2 Feature Engineering: Vectorization

To consolidate text strings into numerical and vectorial representations that ML models can understand, we performed vectorization. We tried out 2 approaches, using Count Vectorizer and TFIDF Vectorizer. TFIDF Vectorizer focuses on the importance of the word for statistical analysis as compared to Count Vectorizer which simply converts strings into frequency representations using a bag-of-words approach. Generally, TFIDF Vectorizer performs better in ML models as it focuses on words deemed more important. However, there are times where Count Vectorizer may have better performance when than TFIDF. This is especially the case when working with texts with different lengths since Count Vectorizer does not penalize longer texts as much as TFIDF does.

Understanding the benefits and drawbacks of both approaches, we took a data-driven approach to find the most suitable feature engineering process which will be elaborated later in the report.

3.4.3 Feature Engineering: Sentiment Analysis

Research by Alva Group has noted that sentiment analysis is becoming an increasingly significant element of ESG intelligence to understand how companies are performing. As such, another feature we experimented with was to conduct sentiment analysis on the vectorized data. TextBlob and Vader are the most widely used libraries for conducting sentiment analysis. In this case, we selected Vader as research has shown that Vader provides a more granular sentiment than Textblob. Vader is also noted to be more

sensitive to suppositions communicated in media writing. A data-driven approach is also utilized to determine whether sentiment scores are to be included in the ML models.

3.5 Machine Learning

After generating the vector representation of relevant sentences for each question, we perform supervised machine learning to train our models for binary and multi-class classification.

3.5.1 Model Selection

In terms of the base models used, we explored Logistic Regression as a basic classifier, along with the 4 major classification techniques, namely Decision Trees, Bayesian Networks, K-Nearest Neighbours and Support Vector Machines (Soofi and Awan 2017).

Apart from the individual models, we also used ensemble learning methods such as bagging, boosting, and stacking to improve overall model performance and reduce variance. Models used include Random Forest, Extra Trees, Gradient Tree Boosting, as well as the Stacking classifier which combines the top 3 models for each question. Deep learning techniques were not used due to the small amount of data available and limited computational power (Koleva 2020).

3.5.2 Model Evaluation

To evaluate model performance, we have selected ROC AUC for binary classification and macro averaged ROC AUC scores for multi-class classification. There are 3 main reasons for this choice, considering the characteristics of our dataset.

First, ROC AUC provides a probabilistic evaluation of the model's ability to distinguish between a random positive and negative sample, which is easily interpretable. Second, one of the weaknesses of ROC AUC is that in unbalanced datasets, it does not reflect the minority class well (Dataman 2021). However, this does not apply to our dataset with relatively low class imbalance, making ROC AUC an appropriate evaluation metric. Finally, in the context of our project, we value the model's ability to detect false positives and false negatives equally, which is one of the characteristics of the ROC AUC score, as opposed to other metrics such as F1.

3.5.3 Model Validation

Instead of performing a train test split on our dataset, we opted to use Stratified K-fold cross validation for model validation. K-fold cross validation refers to the process of splitting the dataset in k folds, training the model on k-1 subsets and evaluating the model on the last subset. There are 3 main reasons for this choice.

First, cross validation allows for a better utilization of the labelled dataset, which is particularly of value given its small size. Second, taking the average of k splits reduces the effect that the randomness of a particular train-test split has on the evaluation metric. Lastly, using stratification preserves the class distribution in each split to match the complete dataset, which makes each run more representative of a random sample, increasing the reliability of the outcome.

3.5.4 Model Tuning

Model tuning is performed via grid search by iterating through all possible combinations of feature engineering techniques and model hyperparameters for every question. For each model, the best combination is selected based on the ROC AUC score of cross validation. After all other models are tuned, the Stacking classifier is constructed by utilizing the top 3 optimized models as estimators and tuning it across all feature engineering options. Finally, the best model is identified for each question, which is used for the final data pipeline to evaluate new unlabeled datasets.

3.5.5 Model Results

Question	Decision Trees	Random Forest	Extra Trees	Logistic Regression	Gradient Boosting	SVM	K Neighbours	Naive Bayes	Stacking
Q3	0.718	0.692	0.676	0.717	**0.781**	0.754	0.756	0.674	0.635
Q4	0.761	0.900	0.900	0.890	**0.925**	0.780	0.807	0.860	0.820
Q7	0.855	0.905	0.911	0.905	0.885	0.905	0.888	0.899	**0.917**
Q8	0.689	0.709	0.707	0.651	**0.727**	0.680	0.661	0.675	0.629
Q9	0.658	0.747	0.747	**0.770**	0.710	0.477	0.701	0.765	0.753
Q10	0.700	0.663	0.669	0.653	**0.731**	0.675	0.676	0.631	0.593
Q11	0.754	0.906	0.897	0.913	0.891	0.899	0.900	0.891	**0.919**
Q12	0.734	**0.905**	0.894	0.866	0.838	0.854	0.844	0.874	0.834
Q13	0.924	0.971	0.969	0.969	**0.981**	0.969	0.922	0.969	0.925
Q14	0.818	0.812	**0.840**	0.790	0.832	0.807	0.798	0.784	0.782
Counts	0	1	1	1	**5**	0	0	0	2
Mean	0.761	0.821	0.821	0.812	**0.830**	0.780	0.795	0.802	0.781

Fig 5. ROC AUC scores for optimally tuned models.

Figure X shows the complete set of ROC AUC scores for the optimally tuned models for every question. There are 3 major observations with respect to the model results.

First, Gradient Boosted Trees is clearly the best performing model, being optimal for 5 out of 10 questions. However, we recognize that Gradient Boosted Trees are more susceptible to overfitting than Random Forests, which may have been the case. Second, ensemble learning models significantly outperform individual models, accounting for 9 out of 10 questions with only 4 ensemble learning models relative to 5 individual models. This is in line with expectations given that ensemble learning models combines the strengths of multiple models, which tends to improve performance and reduce variance. Lastly, Logistic Regression performed surprisingly well, being the only individual model that was optimal for at least a question, and even surpassing the Stacking classifier in terms of average ROC AUC, despite minimal hyperparameter tuning and the simplicity of the model.

3.6 Relevant Sentence Extraction

The final type of questions are open-ended questions which are answered by filtering up to three most relevant sentences from the report. Empirically, we observe that 3 is the optimal number of sentences to balance accuracy and precision. Figure X illustrates the relevant sentence extraction pipeline.

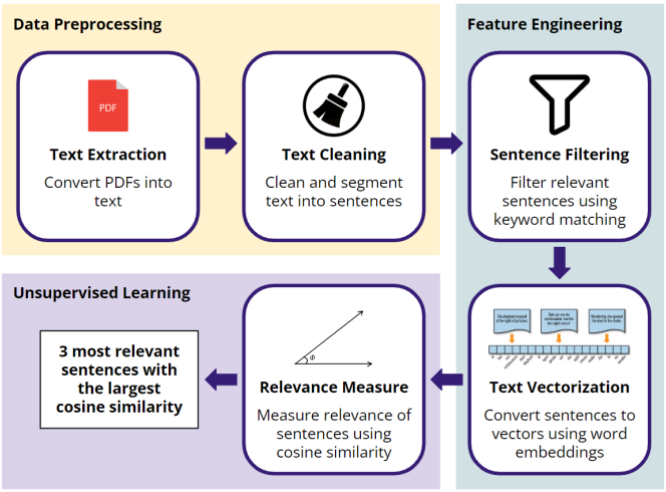


Fig 6. Relevant Sentences Extraction Pipeline.

3.6.1 Cosine Similarity Calculations

The two most common similarity metrics for text classification tasks are Cosine Similarity and Euclidean Distance (Turgay 2022). Euclidean Distance measures similarity using the simple distance between two vectors in a multi-dimensional space, while Cosine Similarity does so by measuring the cosine of the angle between 2 vectors (Nagella 2019).

We decided to use Cosine Similarity as sentence length is not indicative of the relevance of the sentence in our findings. Cosine Similarity measures the cosine of the angle between two vectorized sentences, which is unaffected by the size of the sentence, as opposed to Euclidean Distance (Prabhakaran 2018).

To calculate the relevance score of each sentence, we first vectorize the question to serve as a comparison benchmark. Like the filtered sentences, the question is also vectorized into numerical form using frequency word embeddings. Subsequently, each of the vectors from the previous step are compared with the question vector to generate a cosine similarity score via the Scikit Learn package. When the angle between two vectors is zero, the vectors are overlapping and identical. The smaller the angle, the higher the cosine similarity, and the higher the relevance of the target sentence to the question. However, simply choosing the most relevant sentence is insufficient to achieve a high degree of accuracy. Hence, we selected the top 3 most relevant sentences, which strikes a balance between sentence count and accuracy. Figure 7 shows the full results of relevant sentence extraction.

	Extracted sentence(s) contains expected answer	Extracted sentence(s) does not contain expected answer	Accuracy
Extracting top 1 most relevant sentence	31	32	49.2%
Extracting top 3 most relevant sentences	54	9	85.7%

Fig 7. Accuracy for Relevant Sentences Extraction

4. Dashboard

4.1 Overview

Our end-product is an interactive dashboard that captures all the decarbonization information extracted from the sustainability reports after running them through our complete data pipeline. The dashboard consists of the landing page, three screens each focusing on one of the three topics – Current Emissions, Emissions Reduction Target, and Transition Plans and a comparison function. This allows NatWest Markets to assess ESG information in a structured and systematic manner based on the topic of interest. Additionally, we have included benchmarks against our peers and an Overall Transition Plan Score for users to get a sense of how companies are performing against their peers. Popovers are also included to prompt users on suggested next steps to guide them in navigating the dashboard.

4.2 Dashboard Features

4.2.1 Client Selector

On the left panel of the landing page, we implemented a client selector which consists of dropdown menus for users to view selected company information to be displayed. A search function is also integrated into the dropdown menu to facilitate an easier selection process for the large number of companies.

4.2.2 Company Card

Company details, including the company name, ISIN number, report year, sector, sub-sector and country of incorporation would be displayed on the company card on the right. A hyperlink for easy access to the PDF report is also provided.

4.2.3 Navigation Bar

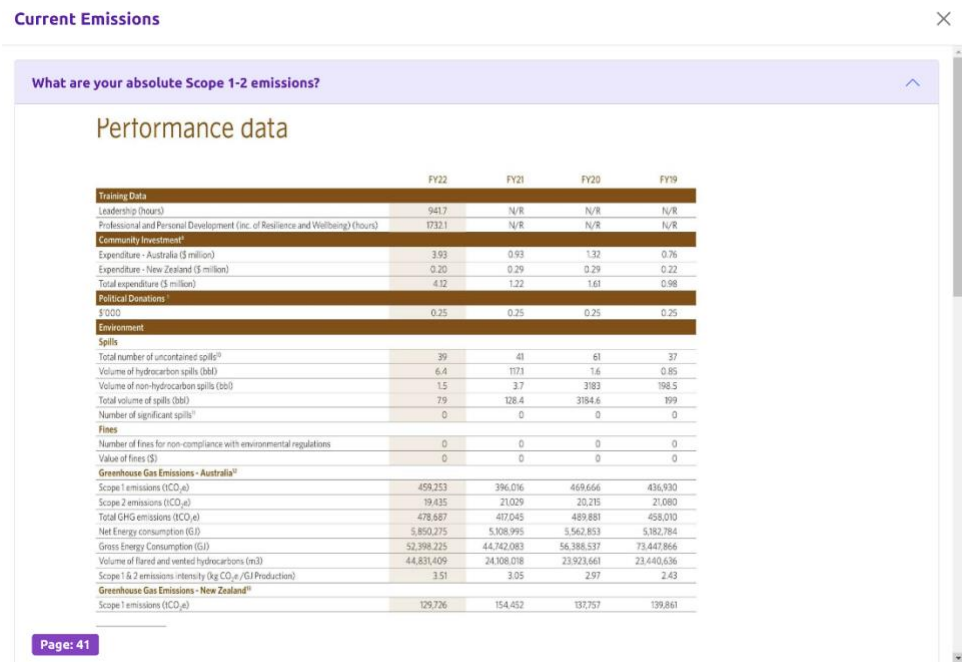
A navigation bar above the company details with four tabs – Current Emissions, Emission Reduction Target, Transition Plans and Comparison is also implemented for easy navigation to the various screens.

Screens are split according to key subtopics to address decarbonization. The Current Emissions tab showcases data to analyze a company's historical emissions performance, the Emissions Reduction Target tab looks at the company's emissions targets moving forward while the Transition Plans tab sheds some color on the company's transition plan to work towards decarbonization. In addition, we have a Comparison tab for users to perform a side-by-side comparison of all 3 aspects of decarbonization.

Fig 8. Main Page of our Dashboard.

4.2.4 Questions with Tabular Answers

To address the first two questions on the absolute Scope 1, 2 and 3 emissions with tabular data, we inserted the previously extracted table images into an accordion pane in the dashboard. An accordion pane is utilized as image files could be large and making image panes collapsible enables better user experience as they navigate through the remaining questions. In the case where no Scope 1, 2 and 3 information is reported in the report, a placeholder image indicating "No absolute emissions is reported" would be displayed instead. In addition, the page number where tables are found would also be indicated below the image for future references.



	FY22	FY21	FY20	FY19
Training Data				
Leadership (hours)	941.7	N/R	N/R	N/R
Professional and Personal Development (inc. of Resilience and Wellbeing) (hours)	1732.1	N/R	N/R	N/R
Community Investment[4]				
Expenditure - Australia ($ million)	3.93	0.93	1.32	0.76
Expenditure - New Zealand ($ million)	0.20	0.29	0.29	0.22
Total expenditure ($ million)	4.12	1.22	1.61	0.98
Political Donations [5]				
$'000	0.25	0.25	0.25	0.25
Environment				
Spills				
Total number of uncontained spills[6]	39	41	61	37
Volume of hydrocarbon spills (bbl)	6.4	117.1	1.6	0.85
Volume of non-hydrocarbon spills (bbl)	1.5	3.7	3183	198.5
Total volume of spills (bbl)	7.9	128.4	3184.6	199
Number of significant spills[7]	0	0	0	0
Fines				
Number of fines for non-compliance with environmental regulations	0	0	0	0
Value of fines ($)	0	0	0	0
Greenhouse Gas Emissions - Australia[8]				
Scope 1 emissions (tCO_2e)	459,253	396,016	469,666	436,930
Scope 2 emissions (tCO_2e)	19,435	21,029	20,215	21,080
Total GHG emissions (tCO_2e)	478,687	417,045	489,881	458,010
Net Energy consumption (GJ)	5,850,275	5,108,995	5,562,853	5,182,784
Gross Energy Consumption (GJ)	52,398,225	44,742,083	56,388,537	73,447,866
Volume of flared and vented hydrocarbons (m^3)	44,831,409	24,108,018	23,923,661	23,440,636
Scope 1 & 2 emissions intensity (kg CO_2e / GJ Production)	3.51	3.05	2.97	2.43
Greenhouse Gas Emissions - New Zealand[9]				
Scope 1 emissions (tCO_2e)	129,726	154,452	137,757	139,861

Fig 9. Tabular Data in Dashboard.

4.2.5 Question with Binary Answers

Answers to binary questions in all 3 tabs are presented with 2 columns. The 'Value' column contains the answers to the binary questions, and the 'Performance vs Peers' column indicates the percentage of companies that have a 'Yes' response to the question based on filters in the client selector panel on the landing page. Users can choose to benchmark the company based on other companies in a specific sector and from a specific country of incorporation. This allows users to gain insight on how the company is performing in the various areas compared to its peers.

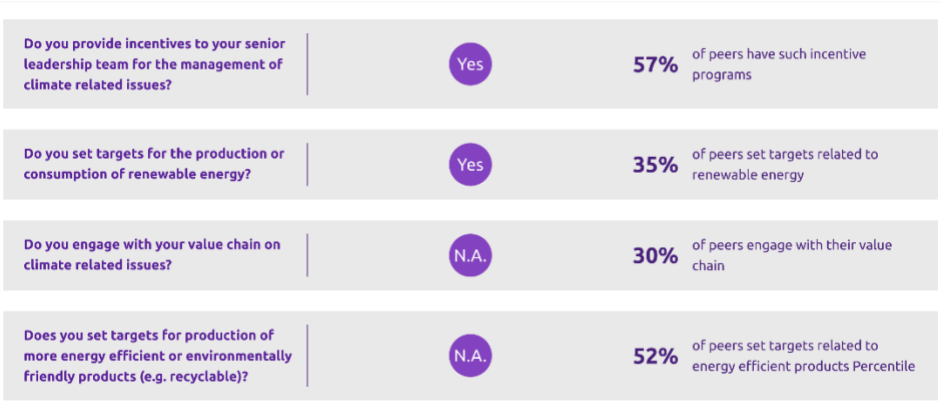


Fig 10. Binary Questions in Dashboard

4.2.6 Questions with Open Ended Answers

The two open-ended questions on the emission reduction target tab would be answered using up to three most relevant sentences extracted from the sustainability reports. In cases where no relevant sentence is captured, the question is answered with "N.A.".

Fig 11. Open Ended Questions in Dashboard

4.2.7 Tabulation of Overall Score

The Transition Plan tab displays the answers to 6 binary and 1 scale questions which are more quantifiable. Hence, we implemented an Overall Transition Plan Score to provide a holistic understanding of the company's transition plans performance. The Overall Transition Plans Score is calculated by taking the sum of the scores for each question. The scoring framework is described as follows:

Binary Questions

Answer	Score
Yes	1
N.A.	0

Scale Questions

Answer	Score
Established Carbon Transition Plan	1
Plans to Transition to Low Carbon Environment	0.5
N.A.	0

Based on the calculated score, a percentile value is also computed to provide insights on the company's overall performance in transition plans compared to its peers.

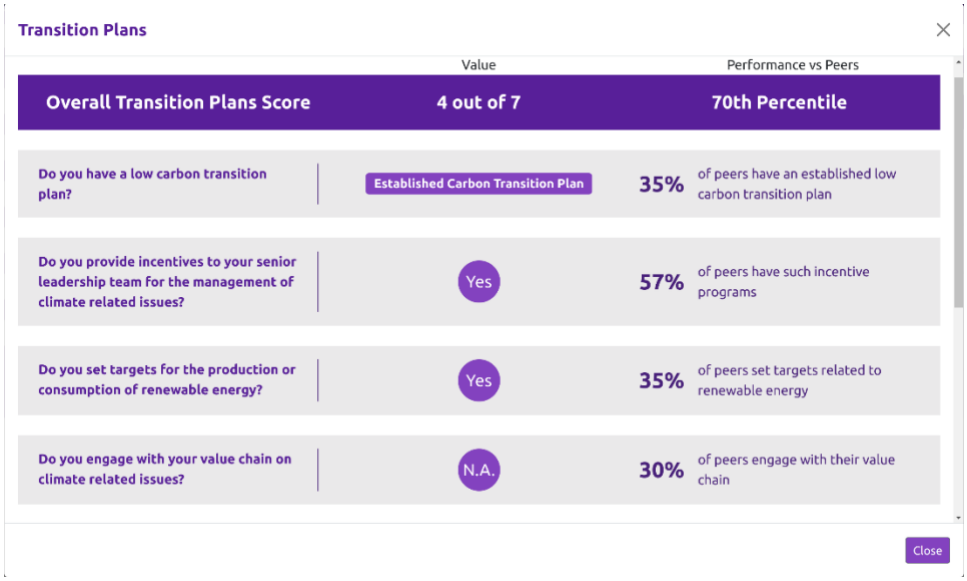


Fig 12. Transition Plans Tab of our Dashboard.

4.2.8 Comparison Function

A comparison tab is also implemented to enable users to view the answers to questions across all three categories for two companies simultaneously as seen in Figure X below. The two companies are selected via the client selector under the fields 'Company A' and 'Comparison Company'. This allows users to carry out direct metric comparisons across companies to make a more informed decision, further easing the ESG-related workflows for stakeholders in NatWest Markets.

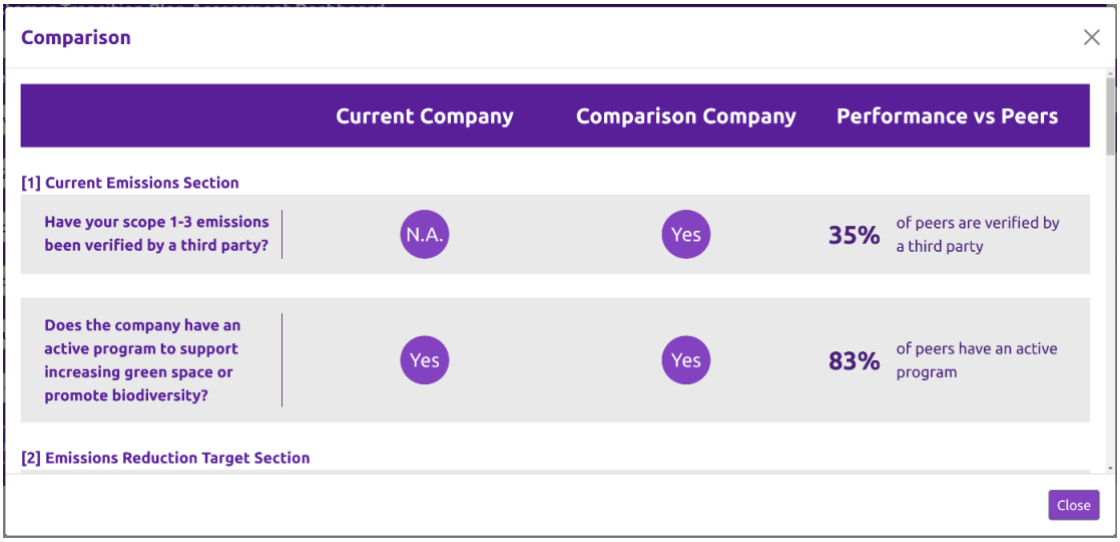


Fig 13. Comparison Tab of our Dashboard.

4.3 Technology Stack Overview

We developed our dashboard using Dataiku's standard webapp to provide a seamless user experience since NatWest uses Dataiku as their primary analytics software. Although Dataiku supports more user-friendly dashboarding tools without the need to implement the backend, such as Dash, we opted to use the standard webapp with our own back end implementation to ensure flexibility and customizability to meet NatWest's functional requirements. With access to the back end, we can also enhance the dashboard's scalability since we can easily develop, append, and maintain REST API endpoints and routing to ensure a smooth end-to-end data pipeline.

4.3.1 Front-End

We used HTML, CSS, and JavaScript to develop our Dataiku standard webapp front-end. To enhance our dashboard's scalability, we integrated the Bootstrap framework into our front-end code base to reduce its complexity, providing developers more time to design the dashboard. Bootstrap also ensures standardized front-end designs by ensuring code consistency across developers. Furthermore, Bootstrap is an industry standard framework that is well-documented and supported by a large developer community.

4.3.2 Back-End

We used Flask, which is written in Python, to develop our Dataiku standard webapp backend as it is well-documented and lightweight. Compared to other common Python-based back-end frameworks, such as Django, Flask requires less code and components to develop a web application. This enables more efficient implementation, maintainable, and scalable, especially given the time constraints of the project timeline. Additionally, we integrated a Firebase storage database for our image files, using a JSON file of the account authentication credentials. This is necessary for the dashboard to access and display the images since Dataiku does not support direct image file uploading or storage.

4.3.3 Data Rendering on Dashboard

The front end draws our machine learning and relevant sentences pipeline outputs from the consolidated_output.csv file, which we manually upload onto Dataiku's server. On the other hand, the extracted table images are drawn from our integrated Firebase storage account. To render the CSV or image outputs onto the dashboard, we make GET requests to the respective REST API endpoints. Each feature on the dashboard is implemented with a function and its respective get request.

5 Challenges Faced

5.1 Time Constraints

NatWest Markets required us to undergo compliance checks before we could proceed with any discussion of the project details and requirements. With the checks only completed in Week 5, we were left with a little over half a semester to complete the entire project. Hence, time constraint was a major challenge. To mitigate this issue, we took the initiative to do our own initial research based on the limited information we had. Examples of this include exploratory analyses of sustainability reports, reviewing the literature of decarbonization research and evaluating the options for dashboarding. By doing so, we were able to fast track our progress with respect to the table extraction and text extraction pipelines, which were vital in allowing us to complete the project on time.

5.2 Lack of Domain Knowledge

Due to the lack of a single universal ESG reporting standard (Bose 2020), domain knowledge in ESG is key in determining the appropriate analysis framework for our project. However, this is something our team was lacking in. To overcome this issue, we raised our concerns to the NatWest team and requested for a specific decarbonization evaluation framework that we would reference to build our solutions around. This came in the form of a list of guiding questions, which we subsequently refined. This process was done through sampling several sustainability reports, attempting to extract the answers to the questions, and eventually modifying or dropping some questions after consulting with NatWest. While the framework provided a list of guiding questions, it failed to clarify the dashboard requirements or the decarbonization performance scoring methodology. To circumvent this, we designed a Figma prototype to align the dashboard expectations with NatWest. We also proposed an aggregated scoring framework to obtain an objective performance rating that grades companies solely on NatWest's ESG framework.

5.3 Complications of Tabular Data

Extracting Scope 1/2/3 emissions values out of report tables proved to be extremely tedious and time-consuming. Our first strategy was to convert the report tables into Pandas dataframes using the tabula.py package. However, the inconsistent table formatting across company reports resulted in high occurrences of incorrect column names and misaligned rows, making the dataframes unusable. Hence, we explored an alternative strategy to convert the report tables into image files instead. While this image extraction algorithm using the CV2 package was largely successful, there were instances of extracted areas being too small to view the table. This was overcome by setting a lower limit threshold for the extracted area and using a default image size that is guaranteed to contain the table when the area is below the threshold. Extracting images instead of values presented another issue of image hosting, because NatWest's primary

analytics software Dataiku does not support image storage. We resolved this issue by hosting the images on Firebase Storage and integrating the Firebase authentication credentials with the Dataiku dashboard.

6 Future Extensions

6.1 Extensions to the Decarbonization Framework

After eliminating 7 questions, our decarbonization framework currently consists of 14 questions from the initial list provided by NatWest. While these questions were deemed infeasible during our exploratory data analysis, it is conceivable that some of the questions could be integrated given sufficient time. For instance, 'What is the current emissions intensity for Scope 1-2 emissions?' is an extension of question 1 and requires company revenue numbers found in annual reports. In the future, annual reports and sustainability reports could be analyzed together to provide a more comprehensive evaluation of clients' decarbonization strategies and performance.

6.2 Automatic Retrieval of Sustainability Reports

Our machine learning pipeline takes as input a CSV file containing a list of company names and their sustainability report URLs. Currently, the URLs are manually retrieved, and this process must be repeated for new reports every year. A possible extension is to automate the URL retrieval process by using the Google Search API to emulate the manual Google search process. The top few search results can be retrieved for further filtering to correctly identify the URL that contains the sustainability report. This will remove the laborious manual retrieval process and allow the reports to be updated automatically each year.

6.3 Expansion of Industry and Year Coverage

Constrained by time, we were only able to conduct manual labelling for Energy sector reports released in 2021, resulting in a dataset that is relatively small. Hence, we propose to train the models on a larger dataset by including data from reports in sectors other than Energy as well as years prior to 2021. A larger dataset will likely improve the performance of machine learning models, allowing the models to generate more accurate predictions (Brownlee 2020). This can be validated by training the models on the expanded dataset and testing on the 2022 reports when they are released. As new reports are released in subsequent years, they can also be labelled and added to the training set. With a large enough dataset, deep learning models, which tend to outperform other techniques with sufficient training data, could also be incorporated (Mahapatra 2018).

6.4 Expansion of Feature Engineering Techniques

In terms of the Natural Language Processing techniques applied to textual data, our pipeline currently only uses frequency-based word embeddings. A possible extension is to utilize prediction-based word embeddings such as word2vec which provide the additional advantage of capturing meanings and semantic relationships between words and sentences (Ramachandran et al. 2021). Another technique is the Bidirectional Encoder Representations from Transformers (BERT) developed by Google. Unlike word2vec, BERT generates contextualized word embeddings, i.e. it generates two different vectors for the

same word appearing in two different contexts (Jeske 2019). The use of these sophisticated embedding techniques could result in better model performance at the expense of using more computational resources (Satvika, Thada, and Singh 2021).

6.5 Extracting Textual Data from Tables

Currently, our table extraction pipeline extracts the entire table containing information on greenhouse gas emissions and converts it to an image file to be displayed on the dashboard. A possible extension is to utilize Optical Character Recognition (OCR) to convert the image file into machine-encoded text so that the specific emission values can be retrieved (Chaudhuri et al. 2016). This will allow for more useful visualizations which will be elaborated in the 'Dashboard' segment below.

6.6 Dashboard Features Extensions

Following up from the modelling extension to use OCR to retrieve the specific emission values, we could utilize these values to create more useful visualizations that provide more insights about the companies' decarbonization strategies and performance. For instance, we could create line graphs to visualize the trend of the emissions over time. A steeper downward sloping line could suggest more effective decarbonization strategies to reduce greenhouse gas emissions.

Another extension we are proposing is to include an input field which takes in the URL of a new report. Currently, our dashboard is static and only displays decarbonization information of clients passed into the pipeline. This additional feature will allow the user to add and analyze a new report and could be useful in cases where NatWest onboards a new client.

6.7 Database Maintenance

Currently, image files are stored on Firebase Storage. As more reports are released each year, more image files will also have to be stored. Hence, Firebase which only provides up to 10GB of storage at no cost may not be a sustainable solution (Google Firebase 2022). To prevent storage shortage issues, we propose for NatWest to host the image files on their own cloud or their on-premises storage server.

References

Bose, Satyajit. “Evolution of ESG Reporting Frameworks.” Values at Work, 2020, 13–33. https://doi.org/10.1007/978-3-030-55613-6_2.

Brownlee, Jason. “Impact of Dataset Size on Deep Learning Model Skill and Performance Estimates.” Machine Learning Mastery, August 25, 2020. https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/.

Chaudhuri, Arindam, Krupa Mandaviya, Pratixa Badelia, and Soumya K. Ghosh. “Optical Character Recognition Systems.” Optical Character Recognition Systems for Different Languages with Soft Computing, 2016, 9–41. https://doi.org/10.1007/978-3-319-50252-6_2.

Dataman, C. K. D. “Why can't I just use the ROC curve?” Medium. Medium, July 12, 2021. https://medium.com/dataman-in-ai/sampling-techniques-for-extremely-imbalanced-data-281cc01da0a8

Economist, The. “ESG Should Be Boiled down to One Simple Measure: Emissions,” July 21, 2022. https://www-economist-com.libproxy1.nus.edu.sg/leaders/2022/07/21/esg-should-be-boiled-down-to-one-simple-measure-emissions.

Firebase, Google. “Usage and Limits.” Google. Google, 2022. https://firebase.google.com/docs/firestore/quotas.

Group, Alva. “How to Measure ESG Performance,” October 6, 2020. https://www.alva-group.com/blog/how-to-measure-esg-performance/.

Group, NatWest. “Climate,” 2022. https://www.natwestgroup.com/who-we-are/at-a-glance/our-purpose/climate.html.

Intelligence, Bloomberg. “ESG Assets May Hit $53 Trillion by 2025, a Third of Global Aum,” February 23, 2021. https://www.bloomberg.com/professional/blog/esg-assets-may-hit-53-trillion-by-2025-a-third-of-global-aum/.

Jeske, Stephen. “Google BERT Update and What You Should Know.” MarketMuse, November 7, 2019. https://blog.marketmuse.com/google-bert-update/#:~:text=%E2%80%9CBERT%20is%20a%20technology%20to,scale%2C%20it%20can%20become%20costly.

Khanna, Chetna. “Text Pre-Processing: Stop Words Removal Using Different Libraries.” Towards Data Science, February 10, 2021. https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a.

Koleva, Nancy. “When and When Not to Use Deep Learning.” Dataiku, May 1, 2020. https://blog.dataiku.com/when-and-when-not-to-use-deep-learning.

Lang, Niklas. “Stemming vs. Lemmatization in NLP.” Towards Data Science, February 19, 2022. https://towardsdatascience.com/stemming-vs-lemmatization-in-nlp-dea008600a0.

Mahapatra, Sambit. “Why Deep Learning over Traditional Machine Learning?” Towards Data Science, March 22, 2018. https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063.

Munková, Daša, Michal Munk, and Martin Vozár. “Influence of Stop-Words Removal on Sequence Patterns Identification within Comparable Corpora.” ICT Innovations 2013, 2014, 67–76. https://doi.org/10.1007/978-3-319-01466-1_6.

Nagella, Vijaya Sasidhar. “Cosine Similarity Vs Euclidean Distance.” Medium. Medium, December 26, 2019. https://medium.com/@sasi24/cosine-similarity-vs-euclidean-distance-e5d9a9375fc8.

Prabhakaran, Selva. “Cosine Similarity – Understanding the Math and How It Works (with Python Codes).” Machine Learning Plus, October 22, 2018. https://www.machinelearningplus.com/nlp/cosine-similarity/#:~:text=The%20cosine%20similarity%20is%20advantageous,angle%2C%20higher%20the%20cosine%20similarity.

Ramachandran, Rahul, M. Ramasubramanian, Iksha Gurung, Carson Davis, Derek Koehl, Manil Maskey, and Tsengdar Lee. “Augmenting Data Systems with Prediction Based Embeddings.” 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2021. https://doi.org/10.1109/igarss47720.2021.9555031.

Satvika, Vikas Thada, and Jaswinder Singh. “A Primer on Word Embedding.” Data Intelligence and Cognitive Informatics, 2021, 525–41. https://doi.org/10.1007/978-981-15-8530-2_42.

Sinha, Arka, Johannes Bayer, and Syed Saqib Bukhari. “Table Localization and Field Value Extraction in Piping and Instrumentation Diagram Images.” 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 2019. https://doi.org/10.1109/icdarw.2019.00010.

Soofi, Aized Amin, and Arshad Awan. “Classification Techniques in Machine Learning: Applications and Issues.” Journal of Basic & Applied Sciences 13 (2017): 459–65. https://doi.org/10.6000/1927-5129.2017.13.76.

Sreekiran, A R. “Checkbox/Table Cell Detection Using OpenCV-Python.” Towards Data Science, November 22, 2020. https://towardsdatascience.com/checkbox-table-cell-detection-using-opencv-python-332c57d25171.

Turgay, Gulden. “Two Most Common Similarity Metrics.” Towards Data Science, March 25, 2022. https://towardsdatascience.com/two-most-common-similarity-metrics-39c37f3fe14d.