Utilizing NLP for Comprehensive Scope 3 Emissions Insights: Beyond Traditional Methods

Vittorio Balestrieri

March 5, 2024

1 Introduction

Climate change poses one of the most significant challenges of our time, urging the global community to take decisive actions towards sustainability. Managing greenhouse gas (GHG) emissions is central to this effort, requiring a comprehensive approach that encompasses both direct and indirect emissions sources. Among these, Scope 3 emissions represent a particularly complex category, including all indirect emissions that do occur in the value chain of the reporting company, including therefore both upstream and downstream emissions.

Despite their significance, these emissions are often the most challenging to quantify and manage due to their indirect nature. This complexity not only hinders accurate emissions reporting but also affects the organization's ability to implement effective reduction strategies. As corporations strive towards sustainability, there is an increasing responsibility to adopt innovative approaches for managing their entire emissions portfolio.

My experimental thesis delves into a novel approach leveraging Natural Language Processing (NLP) techniques to estimate Scope 3 emissions byy analyzing financial transactions and procurement data. The proposed model offers a new pathway to understanding and managing these elusive emissions. Furthermore, this paper evaluates the model's methodology in the context of the Greenhouse Gas (GHG) Protocol standards, providing insights into its compliance and potential areas for alignment.

2 Understanding Scope 3 Emissions

Scope 3 emissions represent a significant portion of an organization's greenhouse gas (GHG) inventory. Unlike Scope 1 and Scope 2 emissions, which are directly emitted by the organization or result from the energy it consumes, Scope 3 emissions are indirect and occur in the organization's value chain.

2.1 Definition and Importance

Scope 3 emissions include all indirect emissions that occur in the value chain of the reporting company, including both upstream and downstream emissions. They can often account for the largest portion of an organization's carbon footprint, making their accurate assessment and management crucial for achieving sustainability goals and regulatory compliance.

2.2 Categories of Scope 3 Emissions

The Greenhouse Gas Protocol categorizes Scope 3 emissions into 15 distinct categories to help organizations systematically identify and manage their indirect GHG impacts. These categories facilitate a comprehensive understanding and strategic reduction of emissions across diverse corporate activities:

1. Purchased goods and services
2. Capital goods
3. Fuel- and energy-related activities not included in Scope 1 or Scope 2
4. Upstream transportation and distribution
5. Waste generated in operations
6. Business travel
7. Employee commuting
8. Upstream leased assets
9. Downstream transportation and distribution
10. Processing of sold products
11. Use of sold products
12. End-of-life treatment of sold products
13. Downstream leased assets
14. Franchises
15. Investments

This detailed categorization helps organizations pinpoint specific areas for GHG emission reduction within their value chain.

3 GHG Protocol and Scope 3

The Greenhouse Gas Protocol serves as the most widely used international accounting tool for government and business leaders to understand, quantify, and manage greenhouse gas emissions. The GHG Protocol not only provides the framework for Scope 1 and Scope 2 emissions but also offers comprehensive guidance on how to measure and report Scope 3 emissions.

3.1 GHG Protocol Standards

The GHG Protocol has developed standardized frameworks and tools that enable consistent, comparable, and reliable GHG accounting and reporting across different organizations and sectors. It includes Corporate Standard, Scope 2 Guidance, and the Corporate Value Chain (Scope 3) Standard, among others, to cover the full spectrum of GHG accounting.

3.2 Scope 3 in GHG Protocol

The Corporate Value Chain (Scope 3) Standard under the GHG Protocol guides organizations in assessing their entire value chain emissions impact, encouraging a complete view of greenhouse gases emissions. This standard is crucial for organizations aiming to:

- Identify and engage with the most significant GHG emissions sources within their value chain.
- Develop comprehensive and effective strategies for emissions reduction.
- Improve their reputation and stakeholder confidence by taking accountability for their indirect emissions.
- Comply with regulatory requirements and participate in voluntary initiatives with credible emissions reporting.

Understanding and implementing the GHG Protocol's Scope 3 Standard is essential for organizations committed to reducing their environmental impact and achieving sustainability targets.

4 Data Description

The foundation of our model's ability to accurately estimate Scope 3 emissions lies in the comprehensiveness and quality of the data used for training, validation, and testing. This section describes the dataset curated for the development of our NLP-based emission estimation model, highlighting the methods used for data generation and the sourcing of emission factors.

4.1 Dataset Generation

Given the complexity of accurately capturing the vast array of transactions across the 389 different industry sectors identified in the U.S. Environmentally-Extended Input-Output (USEEIO) model, a novel approach was employed for dataset creation:

- **AI-Generated Examples:** To ensure a broad and representative dataset, approximately 100 examples for each of the 389 industry sectors defined in the USEEIO model were generated using advanced AI tools. This approach facilitated the creation of a rich dataset reflecting the diverse range of goods and services transactions that might impact Scope 3 emissions.
- **Industry Sector Coverage:** The dataset encompasses all industry sectors as delineated by the USEEIO model, ensuring that the model's training and evaluation are grounded in a dataset that mirrors the complexity and diversity of the U.S. economy's industry sectors.

4.2 Emission Factors

Critical to the model's functionality is the accurate mapping of transaction descriptions to their corresponding emission impacts. This was achieved by leveraging emission factors associated with Bureau of Economic Analysis (BEA) codes:

- **Source of Emission Factors:** The emission factors utilized in the model were sourced from the US Environmentally-Extended Input-Output (USEEIO) Technical Content, which provides detailed emission factors for each BEA code. Specifically, the model references "kg CO2 equivalent for each $" for each sector, ensuring that the emission estimations are grounded in scientifically rigorous data.
- **Application of Emission Factors:** For each transaction example, the corresponding BEA code was identified, and the relevant emission factor was applied to estimate the CO_2 equivalent emissions. This process allows for the detailed and accurate calculation of Scope 3 emissions based on transactional data.

4.3 Data Preparation for Model Training

The dataset was meticulously prepared and divided into training, validation, and test sets, with the following considerations:

- **Data Split:** The dataset was split into training (70%), validation (20%), and testing (10%) sets to ensure the model is effectively trained, its performance accurately evaluated, and its generalizability to unseen data assessed. In doing so, we carefully stratified the data in order to ensure that each of the 389 different clusters were equally represented in the training, validation and testing datasets.
- **Preprocessing:** Extensive preprocessing was conducted on the AI-generated examples to normalize the text and ensure compatibility with the RobertaTokenizer specifications, facilitating optimal model performance.

The strategic curation and preparation of the dataset were instrumental in developing a model capable of providing reliable and actionable insights into Scope 3 emissions. This dataset not only supports the model's current needs but also lays a foundation for further refinement and expansion of its capabilities.

4.4 Data limitations

The proposed model would, of course, benefited by having real world data but unfortunately these data are not publicly available and the academic purpose of this thesis is not to have a ready-to-use tool but to test whether the architecture and the idea of using this model could pose the foundation for building a tool able to help in better handling scope 3 emissions.

In respect to the choice of using US-produced datasets, it simply seemed more coherent with the AI-generated texts and it offers an interesting granularity, being therefore more complacient with the GHG protocol.

It is important to note that for real world use of these model it's necessary to promptly coordinate the emission factors with the financial data used in the training phase, from both a granularity and geographical perspective.

5 Model Overview

This section introduces the innovative approach our model adopts to estimate Scope 3 emissions by leveraging natural language processing (NLP) techniques on financial transaction data. The focus is on creating a scalable, accurate method for categorizing transactions into relevant emission factors, ultimately providing a comprehensive view of an organization's indirect emissions.

5.1 Purpose of the Model

The primary goal of our model is to automate the estimation of Scope 3 emissions, which are traditionally difficult to quantify due to their indirect nature and the complexity of value chains. By analyzing descriptions of purchased goods and services within financial transactions, the model aims to classify these transactions into predefined categories that correspond to specific emission factors.

5.2 Problem Addressed

Scope 3 emissions account for a significant portion of an organization's total greenhouse gas emissions. However, the lack of transparency and difficulty in tracking emissions across a company's value chain have posed substantial challenges to accurate reporting and reduction efforts. The problem is compounded by the diversity of goods and services, making manual classification and emission estimation labor-intensive and prone to errors. Our model seeks to address these challenges by:

- Automating the recognition and classification of purchased goods and services from transaction descriptions.
- Mapping classified transactions to their corresponding emission factors based on established standards.
- Providing a scalable solution that can be adapted and refined as more data becomes available.

5.3 Using NLP for Emission Estimation

The model utilizes state-of-the-art NLP techniques to interpret and classify the textual data found in financial transaction records. The process involves several key steps:

1. **Data Preprocessing:** Transaction descriptions are cleaned and normalized to remove noise and standardize the text, making it more amenable to analysis.

2. **Tokenization and Encoding:** The cleaned text is tokenized, and each token is converted into numerical representations using the Roberta tokenizer. This process facilitates the understanding and processing of text by the model.

3. **Classification:** The tokenized text is fed into a Roberta-based sequence classification model, which has been fine-tuned to categorize transactions into specific commodity classes associated with known emission factors.

4. **Emission Estimation:** Once transactions are classified, their associated emission factors are applied to estimate the Scope 3 emissions. This step involves aggregating emissions across all transactions to provide a comprehensive view of an organization's indirect emissions.

By leveraging advanced NLP techniques, the model offers an innovative approach to the complex problem of Scope 3 emission estimation, providing organizations with the tools needed to more accurately report and reduce their environmental impact.

6 Model Architecture

This section explores the architecture of the model, emphasizing its foundation on the RobertaForSequenceClassification transformer, the process of tokenization and encoding, and the strategies adopted for model training. The model's architecture is tailored to understand and classify the natural language data found in financial transactions, thereby estimating Scope 3 emissions more accurately.

6.1 RobertaForSequenceClassification

RobertaForSequenceClassification is a variant of the RoBERTa model, adapted for the task of sequence classification. This model architecture is chosen for its exceptional ability to capture the context and semantics of textual data, making it highly effective for classifying transaction descriptions into emission-related categories.

- **Model Selection:** RoBERTa, a robustly optimized version of BERT, is renowned for its performance on a wide range of natural language processing tasks. Its adaptation for sequence classification enables it to handle the specific challenges posed by the categorization of transactional data.

- **Architecture Details:** The model consists of multiple layers of transformer blocks that process the input text in tokens, capturing the intricate relationships between words in transaction descriptions. It outputs probabilities across the set of predefined categories, indicating the most likely classification for each transaction.

6.2 Tokenizer and Encoding

The preprocessing of text data is a critical step in preparing the input for the model. The RobertaTokenizer plays a pivotal role in this process:

1. **Tokenization:** The tokenizer converts raw text into a sequence of tokens. These tokens represent words or subwords in the vocabulary learned by the model during pretraining.

2. **Encoding:** Each token is then mapped to an integer ID, with special tokens added to mark the beginning, end, and padding of sequences. This encoding step transforms the textual data into a format that the model can process.

6.3 Training Strategies

Training the model effectively requires careful consideration of several strategies, including data split, augmentation, and fine-tuning techniques:

- **Data Split:** The dataset is divided into training, validation, and test sets. This separation allows for the evaluation of the model's performance and generalizability on unseen data.

- **Fine-tuning:** Given the pretrained nature of RobertaForSequenceClassification, fine-tuning involves adjusting the model's weights on the specific task of classifying transaction descriptions. This process is crucial for tailoring the model's predictions to the nuances of emission factor estimation.

- **Hyperparameter Optimization:** Parameters such as learning rate, batch size, and the number of epochs are optimized to balance the model's accuracy and training efficiency. Regular evaluation on the validation set guides the selection of these hyperparameters.

The combination of the RobertaForSequenceClassification model, robust tokenization and encoding practices, and strategic training approaches forms the backbone of our architecture, enabling the precise classification of transactions and the subsequent estimation of Scope 3 emissions.

7 Compliance with GHG Protocol

Assessing the model's compliance with the Greenhouse Gas (GHG) Protocol, specifically regarding Scope 3 emissions, is critical for understanding its applicability and effectiveness in real-world corporate sustainability efforts. This section evaluates the model's methodology against the GHG Protocol's requirements, addressing both its alignment and potential limitations.

7.1 Alignment with GHG Protocol's Requirements

The GHG Protocol provides a comprehensive standard for quantifying and reporting greenhouse gas emissions, including detailed guidance for Scope 3 emissions. Our model's approach to estimating Scope 3 emissions through financial transaction data is aligned with several key aspects of these standards:

- **Comprehensive Coverage:** By classifying a wide range of transaction descriptions, the model facilitates the comprehensive identification and quantification of Scope 3 emissions across all relevant categories, in line with GHG Protocol standards.
- **Data Accuracy and Quality:** Leveraging NLP techniques enhances the accuracy and granularity of emission estimations, contributing to the quality of data reported under the GHG Protocol.
- **Transparency and Consistency:** The model's systematic approach to data analysis promotes transparency and consistency in emission reporting, principles emphasized by the GHG Protocol for credibility and comparability.

7.2 Limitations and Challenges

Despite its strengths, the model may face challenges in fully complying with the GHG Protocol's guidelines, mainly due to the inherent complexities of Scope 3 emission calculations and data availability:

- **Emission Factors Variability:** The model relies on existing emission factors for estimating emissions from classified transactions. Variability and uncertainties in these factors can impact the accuracy of the final emission estimations.
- **Data Gaps and Assumptions:** In cases of incomplete or ambiguous transaction data, the model must make assumptions that could affect compliance with the GHG Protocol's emphasis on specificity and accuracy.
- **Scope of Applicability:** The model's current design may not capture certain indirect emissions that do not directly relate to financial transactions, such as those associated with leased assets or investments, posing a challenge for complete Scope 3 coverage.

7.3 Advantages in GHG Protocol Compliance

The model presents several advantages for organizations seeking to comply with the GHG Protocol and enhance their sustainability practices:

- **Efficiency in Emission Estimation:** Automating the classification and estimation process significantly reduces the time and resources required for Scope 3 emission calculations.
- **Improved Reporting Accuracy:** The use of advanced NLP techniques can improve the accuracy of categorizing transactions and estimating associated emissions, leading to more reliable GHG reporting.
- **Strategic Emission Reduction:** By providing detailed insights into emission hotspots within the value chain, the model enables organizations to target their reduction efforts more effectively.

In summary, while the model aligns well with many aspects of the GHG Protocol's standards for Scope 3 emissions, addressing its limitations will be crucial for full compliance and maximized effectiveness. Future enhancements should focus on refining emission factors, improving data completeness, and expanding the model's scope to ensure comprehensive coverage of all Scope 3 categories.

8 Conclusion and Future Directions

This document has presented an innovative approach to estimating Scope 3 emissions, leveraging advanced natural language processing techniques to analyze financial transaction data. The model, built on the RobertaForSequenceClassification architecture, demonstrates significant potential for automating the categorization of transactions into emissions-related categories, thereby facilitating more accurate and efficient Scope 3 emissions reporting.

8.1 Key Findings

Our exploration revealed several key findings:

- The use of NLP techniques, particularly tokenization and sequence classification models, can significantly improve the accuracy and granularity of Scope 3 emission estimations.
- Addressing implementation challenges, such as batch size mismatches and data preprocessing, is crucial for optimizing model performance.
- The model's alignment with GHG Protocol standards for Scope 3 emissions underscores its potential utility for organizations seeking to enhance their sustainability reporting and reduction efforts.

8.2 Impact on Scope 3 Emission Reporting

By automating the process of identifying and classifying Scope 3 emissions, the model offers organizations a scalable tool for enhancing their environmental impact assessments. This capability is particularly valuable for companies committed to achieving sustainability targets and complying with regulatory requirements, enabling more strategic decision-making and resource allocation toward emission reduction initiatives.

It's fundamental for actively challenge scope 3 emissions to be precise in individuate and take actions in the sectors, or subsectors more emission-intensive and the granularity of this model in clustering the total scope 3 emissions in more than 350 industrial subsectors can surely help in this challenge.

8.3 Future Directions

While the model represents a significant advancement in the field of Scope 3 emission estimation, there are several avenues for future research and development that could further enhance its utility:

1. **Expansion of Emission Factors:** Incorporating a wider range of emission factors, including those for emerging industries and technologies, would improve the model's comprehensiveness.

2. **Integration with Other Data Sources:** Combining financial transaction data with other relevant datasets, such as supply chain logistics or production metrics, could provide a more holistic view of an organization's emissions profile.

3. **Development of Real-time Monitoring Tools:** Creating tools that leverage the model for real-time emission estimation could support more dynamic and responsive environmental management strategies.

4. **Enhancing Model Accuracy:** Continued refinement of the NLP algorithms and training methodologies, including exploring newer language models, could further improve classification accuracy and reliability.

In conclusion, the development and implementation of this model mark an important step forward in the quest for more accurate and actionable Scope 3 emissions data. As organizations worldwide strive to meet ambitious climate goals, the role of innovative technologies in understanding and managing environmental impacts becomes increasingly critical. We look forward to the continued evolution of this model and its contributions to global sustainability efforts.

References

1. Greenhouse Gas Protocol. (n.d.). *Corporate Value Chain (Scope 3) Accounting and Reporting Standard*. Retrieved from https://ghgprotocol.org/standards/scope-3-standard

2. U.S. Environmental Protection Agency. (n.d.). *USEEIO Models*. Retrieved from https://www.epa.gov/land-research/useeio-models

3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*.

4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv preprint arXiv:1907.11692*.

5. Hugging Face. (n.d.). *Transformers: State-of-the-art Natural Language Processing*. Retrieved from https://huggingface.co/transformers/

6. Yang, Y., Ingwersen, W. W., Hawkins, T. R., Srocka, M., & Meyer, D. E. (2017). USEEIO: A new and transparent United States environmentally-extended input-output model. *Journal of Cleaner Production, 158*, 308–318. https://doi.org/10.1038/S41597-022-01293-7.

7. Klaaßen, L., & Stoll, C. (2021). Harmonizing corporate carbon footprints. *Nature Communications, 12*(1), 1–13.https://doi.org/10.1038/s41467-021-26349-x

8. Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. https://doi.org/10.48550/arXiv.2007.03051

9. Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. https://arxiv.org/abs/2007.03051

10. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

11. Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., et al. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11(1), 233. https://doi.org/10.1038/s41467-019-14108-y

12. Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., et al. (2019). Tackling Climate Change with Machine Learning. *arXiv preprint arXiv:1906.05433*.

13. Luccioni, A., Baylor, E., & Duchene, N. (2020). Analyzing sustainability reports using natural language processing. *arXiv preprint arXiv:2011.08073*.