Analyzing ESG Reports Using Natural Language Processing, Topic Modeling and Text Classification of Sentences Using The ESG-BERT Models

Aswad Hardi Mohammed

East Tennessee State University

CSCI 5037 – NLP and Text Analysis

Fall 2024

Contents

[1. Introduction 3](#_Toc184165110)

[2. Problem Statement 4](#_Toc184165111)

[3. Motivation 4](#_Toc184165112)

[4. Research Questions 4](#_Toc184165113)

[5. Literature Review 5](#_Toc184165114)

[6. Dataset Description 5](#_Toc184165115)

[6.1 Source 5](#_Toc184165116)

[6.2 Data Characteristics 5](#_Toc184165117)

[7. Methodology 5](#_Toc184165118)

[7.1 Data Collection 5](#_Toc184165119)

[7.2 Data Cleaning 6](#_Toc184165120)

[7.3 Topic Model and ESG Classification with ESGBERT 6](#_Toc184165121)

[7.4 ESG Classification 7](#_Toc184165122)

[7.6 Results and Insights 7](#_Toc184165123)

[7.7 ESG Sentence Classification 8](#_Toc184165124)

[8. Research Question Answers 9](#_Toc184165125)

[9. Conclusion 9](#_Toc184165126)

[10. Future Research Directions 9](#_Toc184165127)

[11. Team Member Roles 9](#_Toc184165128)

[12. Reflections 9](#_Toc184165129)

[References 10](#_Toc184165130)

# 1. Introduction

Environmental, Social, and Governance (ESG) have become crucial indicators of a company's sustainable and ethical performance. This project uses Natural Language Processing (NLP) techniques to analyze and extract meaningful insights from corporate ESG reports, focusing specifically on the Pecan Energies 2023 Sustainability Report in the demonstration.

Environment, Social, and Governance (ESG) is a framework used to evaluate a company's sustainability and social impact. By assessing a company's performance in these three areas, investors can better understand potential risks and opportunities. The ESG report gives an outlook to understand the non-financial performance of a company.

Using Natural Language Processing (NLP) techniques, including ESGBERT, a specialized language model for ESG text classification, to extract key information from an ESG report (Pecan Energies 2023) helps to mitigate the manual process of analysis these reports. ESGBERT was used to categorize sentences within the report into one of the three ESG dimensions: Environmental, Social, or Governance. I also then visualized the extracted data using Word Clouds and Topic Modeling (LDA) to identify the company's primary ESG trend and focus. With an efficient esg report analysis to stakeholders can gain insight at almost real-time into ESG practices to assess the risks and opportunities of a company. This also supports making informed decisions and recommendations for improvement towards sustainable growth of a company.

Different companies prioritize different aspects of ESG based on their industry, size, and geographic location. ESG reports are complex and multifaceted that poses challenges and some bias in the results of this project. However, it is a good point to develop a more robust tool for analysis. Text classification of sentences can be improved in the future by identifying sentences with multilabel and irrelevant sentences to the ESG dimension. This also helps to calculate the true esg performance of a company’s report.

This is to demonstrate my understanding and application of the theocratical knowledge from the NLP class in ETSU CSCI 5037.

# 2. Problem Statement

The primary challenge addressed in this project is developing a systematic and data-driven approach to:

- Extract meaningful textual information from ESG reports

- Identify key themes and trends within the ESG narrative

- Classify sentences across environmental, social, and governance dimensions

# 3. Motivation

The motivation for this project stems from several critical factors particularly from my work experience as an ESG/CSR professional, growing investor and stakeholder interest in corporate sustainability, the rising need for automated, scalable methods to analyze ESG reports, limitations of traditional manual report analysis and potential to provide evaluate insights from esg data for true esg performance using advanced NLP techniques and AI.[1]

# 4. Research Questions

The study aims to answer the following research questions:

1. How can advanced NLP techniques be used to systematically analyze ESG reports?

2. What are the dominant themes in Pecan Energies' sustainability narrative?

3. How can topic modeling and sentence classification provide insights into a company's ESG performance?

4. What are the relative proportions of sentences related to environmental, social, and governance dimensions?

# 5. Literature Review

Recent studies highlight the importance of computational techniques in ESG analysis. For instance, research shows a link between improved municipal ESG performance and reduced credit risk. Additionally, strong ESG policies can benefit business growth, migration, and tax revenue, while potentially reducing racial inequities. Machine learning models, such as BERT, have shown promise in analyzing ESG data. However, challenges remain, including diverse reporting formats, subjective nature of ESG reporting, and lack of standardized reporting. Overcoming these challenges is crucial for harnessing the full potential of ESG data to drive sustainable practices and positive social impact.[1][2]

# 6. Dataset Description

The dataset is made up of a Companies report that highlights its sustainability efforts and initiatives.

## 6.1 Source

For this demonstration Pecan Energies Ghana, an oil and gas company’s Sustainability report for 2023 was used. The report was a pdf version and locally available. Other reports attached that are equally suitable are scraped from the company’s website using the URL of the report.

## 6.2 Data Characteristics

The number of pages and textual content varies from report to report with each having a significant proportion of textual data for the analysis.

# 7. Methodology

## 7.1 Data Collection

To extract relevant information from PDF reports, I initially checked if the file source was a pdf or a URL then used the right method to access the PDF documents using a custom extraction function, leveraging the robust capabilities of pdfminer3. To ensure data quality, page and sentence filtering techniques were applied. Extracted text is preprocessed, including sentence-level extraction, lemmatization, and stopword removal. The resulting clean text is then stored in a JSON format for subsequent analysis. This ensures that data quality was optimal for the analysis.

## 7.2 Data Cleaning

In the cleaning of the textual data, I removed punctuation, lower casing of characters, removed non-ASCII characters, filtered pages that are less than 500 words to remove content such as table of contents and other irrelevant information such as pages with more graphical content, headers and footers. I used lemmatization and stopword removal with a combination of domain specific keywords.

## 7.3 Topic Model and ESG Classification with ESGBERT

I made a word cloud as seen in Figure 1 using word cloud packages to get a visual representation of the most common words. It helped to understand the results of the cleaned data and ensure if any more preprocessing is necessary before training the model.

A close-up of words

Description automatically generated

Figure 1

I start by tokenizing the text and removing stopwords. Then convert the tokenized object into a corpus and dictionary to serve as an input for the training LDA model. I built a model with 12 topics where each topic is a combination of keywords, and each keyword contributes to a certain weightage to the topic. The results of the optimal topics in the Pecan report is shown below.

A black screen with white text

Description automatically generated

Figure 2

TF-IDF vectorization was done to convert data into numerical format and Bigram and unigram was done for term frequency in the document.

Topic modeling using Gensim Latent Dirichlet Allocation (LDA .I used a popular visualization package, [pyLDAvis](https://github.com/bmabey/pyLDAvis) to understand and interpret individual topics, and understanding the relationships between the topics. I also view the topmost frequent or relevant topics and how the topic relates to each other. For an optimal and reliable results I manipulated the parameters of the model while evaluating the outcome.   
Resulting perplexity score of the Optimal Model is -6.746085355851031, Coherence Score of the Optimal Model is 0.41286577438344596.

## 7.4 ESG Classification

Used Hugging Face Transformers pipeline with specialized BERT models for sentence classification. i.e. ESGBERT/EnvironmentalBERT – This model helped to categorize ESG environmental related sentences. ESGBERT/SocialBERT - This model helped to categorize ES social related sentences. ESGBERT/GovernanceBERT - This model helped to categorize ESG governance related sentences. I used Matplotlib to visualize the number of sentences for each ESG dimension in the corpus.[3]

## 7.6 Results and Insights

Topic Modeling Insights

- Optimal number of topics: Determined through coherence score

- Key identified topics:

1. "Business Ethics" – which indicates the prevalence of governance ESG topic

2. "Employée Value" - which indicates the prevalence of social ESG topic

A comparison of words on a screen

Description automatically generated

## 7.7 ESG Sentence Classification

Results of the ESG Sentence Classification of a total of **201** values.

A graph of a number of sentences assigned to esg dimensions

Description automatically generated

# 8. Research Question Answers

1. NLP Techniques Effectiveness: Demonstrated robust and scalable analysis of ESG reports

2. Dominant Themes: Focus on business ethics and employee value was identified as a dominant topic in Pecan Energies report.

3. Topic Modeling Insights: Revealed terms in sustainability narratives indicating the ESG focus of the report.

4. ESG Dimension Distribution: Provided quantitative breakdown of report content and how the company was performing better in social impact noting their strong community impact, engagement, and social impact programs.

# 9. Conclusion

The project successfully demonstrated the potential of advanced NLP techniques in ESG report analysis, offering a data-driven approach to understanding corporate sustainability narratives.

# 10. Future Research Directions

1. Improve analysis using new technologies like GPT to achieve the true ESG performance of a company.
2. Design and develop analysis tools to enable real-time analysis of ESG reports. [4]
3. Expand analysis to more companies for comparative studies
4. Develop a multi-label classification model for ESG topics/dimensions
5. Integrate financial performance data with ESG data (non-financial performance data) to achieve the ultimate performance evaluation of a company

# 11. Team Member Roles

The work was done solely by Aswad Hardi Mohammed.

# 12. Reflections

I learnt that complex text analysis requires robust preprocessing and specialized NLP models provide deeper insights. NLP can be used for a broad range of problem solving within different special domains. Without the right approach and techniques, the results will not be reliable.

# References

[1] KPMG, “Addressing the Strategy Execution Gap in Sustainability Reporting,” *kpmg.com*, Feb. 2024. Available: https://kpmg.com/us/en/articles/2024/esg-organization-survey.html

[2] W. Heather, “Navigating the labyrinth: The complexity of ESG data collection,” *Oracle.com*, Sep. 2024. Available: https://blogs.oracle.com/modernfinance/post/navigating-labyrinth-complexity-esg-data-collection. [Accessed: Nov. 04, 2024]

[3] T. Schimanski, A. Reding, N. Reding, J. Bingler, M. Kraus, and M. Leippold, “Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication,” *Finance Research Letters*, vol. 61, p. 104979, Mar. 2024, doi: https://doi.org/10.1016/j.frl.2024.104979.

‌[4] J. Fischbach *et al.*, “Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool,” *arXiv.org*, 2022. https://arxiv.org/abs/2212.06540 (accessed Nov. 25, 2024).