# ESG Reports Using Natural Language Processing

**Yuhe Jiang[#], Jiali Qiu[#]**

School of Mathematics and Physics, Xi'an-Jiaotong Liverpool University, Suzhou, China

[#]These authors contributed equally.

**Keywords:** ESG, Citi Bank, LDA, NLP

**Abstract:** In this study, we examined a lot of information about ESG and analyzed ESG factors using the Latent Dirichlet Allocation (LDA) method. This article proposes the use of natural language processing (NLP) techniques to automatically analyze ESG reports. NLP is a branch of artificial intelligence that enables computers to understand and process human language. By applying NLP to ESG reports, we can extract relevant information, identify key themes, and analyze sentiment to gain a comprehensive understanding of a company's ESG performance. The effectiveness of using NLP techniques to analyze ESG reports is demonstrated by automating the analysis process. We can gain a deeper understanding of a company's sustainability practices and decisions, and have the potential to revolutionize the analysis of ESG reports, enabling investors to more efficiently and accurately assess a company's sustainability practices.

## 1. Introduction

Environmental, social, and governance (ESG) factors are becoming increasingly important in the business world, with investors and stakeholders placing greater emphasis on sustainable and responsible business practices[1] . Among the trends in the international news, the themes of ESG and asset markets were given great importance in the hierarchical thematic composition in the time series analysis[2] . Using a dataset containing sentiment scores for ESG news, the researchers examined investor reactions to various disclosures related to ESG companies, concluding that investors reacted more positively to ESG news [3]. By the 1920s, research on ESG and Natural Language Processing (NLP) was active [4].

The results of this study demonstrate the effectiveness of NLP in analyzing ESG reports. [5] Through the automated analysis process, we can save time and resources while gaining valuable insights into a company's sustainability practices. This approach can help investors, analysts, and other stakeholders make informed decisions based on objective and quantitative ESG performance indicators.

ESG factors have become increasingly important in the business world, with companies striving to achieve sustainable development goals and address social and environmental issues. ESG reports provide a comprehensive overview of a company's performance in these areas, covering topics such as sustainability practices, gender equality, carbon emissions, employee diversity, community engagement, and governance practices [6]. However, manually analyzing these reports is time-consuming and susceptible to human biases.

Natural Language Processing (NLP) offers a solution for automated analysis of ESG reports. NLP technology enables computers to understand and process human language, extracting meaningful information from large volumes of text data [7]. By applying NLP to ESG reports, we can gain deeper insights into a company's sustainability practices, identify areas for improvement, and compare performance between companies and industries.

## 2. The basic funamental NLP

The basic fundamentals of Natural Language Processing (NLP) involve understanding and processing human language using computational techniques. Here are some key concepts in NLP:

1) Tokenization: Tokenization is the process of breaking a text into individual words or tokens. It

  DOI: 10.25236/icceme.2023.019

helps in further analysis by providing a basic unit of meaning.

2) Stop Words: Stop words are common words (e.g., "the," "is," "and") that are often removed from the text during preprocessing as they do not carry much semantic meaning.

3) Stemming and Lemmatization: These techniques are used to reduce words to their base or root form. Stemming involves removing suffixes, while lemmatization considers the word's context and converts it to its base form [8].

4) Part-of-Speech (POS) Tagging: POS tagging assigns grammatical tags to each word in a sentence, such as noun, verb, adjective, etc. It helps in understanding the syntactic structure of a sentence.

5) Named Entity Recognition (NER): NER identifies and classifies named entities in text, such as names of people, organizations, locations, etc [9].

6) Sentiment Analysis: Sentiment analysis aims to determine the sentiment or opinion expressed in a given text. It can be classified as positive, negative, or neutral.

7) Machine Translation: Machine translation involves automatically translating text from one language to another. It utilizes various techniques, such as statistical models or neural networks.

8) Text Classification: Text classification is the process of categorizing text into predefined categories or classes. It is widely used in spam detection, sentiment analysis, topic classification, etc.

9) Named Entity Disambiguation (NED): NED resolves ambiguous named entities by linking them to specific entities in a knowledge base or ontology.

10) Language Modeling: Language modeling is the task of predicting the next word or sequence of words given a context. It is used in various applications, including speech recognition and machine translation.

## 3. A common topic modeling technique in NLP -LDA

The Latent Dirichlet Allocation (LDA) model is a machine learning technique used for topic modeling without supervision. It assumes that documents are generated from a mixture of topics, where each topic is a distribution over words. Here are the essential formulas, notations, and steps involved in building an LDA model:

1) Notations:

- K: number of topics
- D: number of documents in the corpus
- N: number of words in a document
- M: number of unique words in the vocabulary
- α: Dirichlet prior parameter for document-topic distribution
- β: Dirichlet prior parameter for topic-word distribution
- θ_d: document-topic distribution for document d
- φ_k: topic-word distribution for topic k

2) Generative process: The LDA model assumes that documents are generated through the following process:

a. Choose θ_d from a Dirichlet distribution with parameter α for each document d.

b. Choose φ_k from a Dirichlet distribution with parameter β for each topic k.

c. For each word w_dn in document d:

i. Choose a topic z_dn from a multinomial distribution with probabilities given by θ_d.

ii. Choose a word w_dn from a multinomial distribution with probabilities given by φ_z_dn.

3) Key formulas:

LDA is a typical representative of topic modeling, which has been widely used in the field of text mining, such as text topic recognition, text categorization, and text similarity computation.

Here we define that the corpus D consists of M documents, $D=\{W_1, W_2, \ldots, W_M\}$, where a document W contains n words, $W=\{\omega_1, \omega_2, \ldots, \omega_M\}$. The process of generating documents from corpus D can be expressed as follows:

Step1. Randomly sample a topic distribution from the Dirichlet distribution $\theta$, $\theta \sim Dir(\alpha)$ ;

Step2. Assign a topic to the nth word in document W, and $z_n \sim \text{Multionmial}(\theta)$ ;

Step3. Generate words based on the word frequency distribution $\beta$ corresponding to the assigned topic $z_n$ randomly sampled according to the probability distribution of $P(\omega_n | z_n, \beta)$. ($\beta \in R^{k*V}$, k is the number of topics, V is the size of the word list, and $\beta_{ij} = P(\omega^j | z^i = 1)$.

Step4. Finally we can deduce that, P(z_dn=k | w_dn) $\propto$ P(w_dn | z_dn=k) * P(z_dn=k | θ_d)

- P(w_dn | z_dn=k) = φ_k(w_dn)
- P(z_dn=k | θ_d) = θ_d(k)

4) Detailed steps for constructing an LDA model using the collapsed Gibbs sampling algorithm:

a. Initialize topic assignments for each word in every document randomly.

b. For each document d, compute document-topic counts N_dk (excluding the current word).

c. For each topic k, compute topic-word counts N_kw (excluding the current word) and total words assigned to topic N_k.

d. Iterate through each word w_dn in document d:

i. Calculate the conditional distribution for the current word's topic assignment: P(z_dn=k | w_dn) $\propto$ (N_dk + α) * (N_kw + β) / (N_k + M*β)

ii. Sample a new topic assignment for the current word from the conditional distribution.

iii. Update the document-topic and topic-word counts accordingly.

e. Repeat steps b-d for a predefined number of iterations or until convergence.

f. After the Gibbs sampling process, estimate the document-topic distribution θ_d and topic-word distribution φ_k:

1) θ_d(k) = (N_dk + α) / (N_d + K*α)

2) φ_k(w) = (N_kw + β) / (N_k + M*β)

Once the LDA model is built, it generates document-topic distributions (θ_d) and topic-word distributions (φ_k) as outputs. These distributions can be utilized to analyze the dominant topics in the collection of documents and the most significant words associated with each topic.

## 4. Results

The analysis results provide valuable insights into a company's ESG performance. By automating the analysis process, we save time and resources while ensuring consistent and objective evaluations of sustainability practices. The main information from the reports can be understood by adjusting the values of ESG-related parameters. The analysis can reveal industry trends and patterns, identify best practices, and highlight areas for improvement. The discussion section interprets the research findings in the context of existing literature and industry standards. It explores the significance of the results and provides recommendations for improving a company's ESG performance.


Figure 1 ESG-related keywords

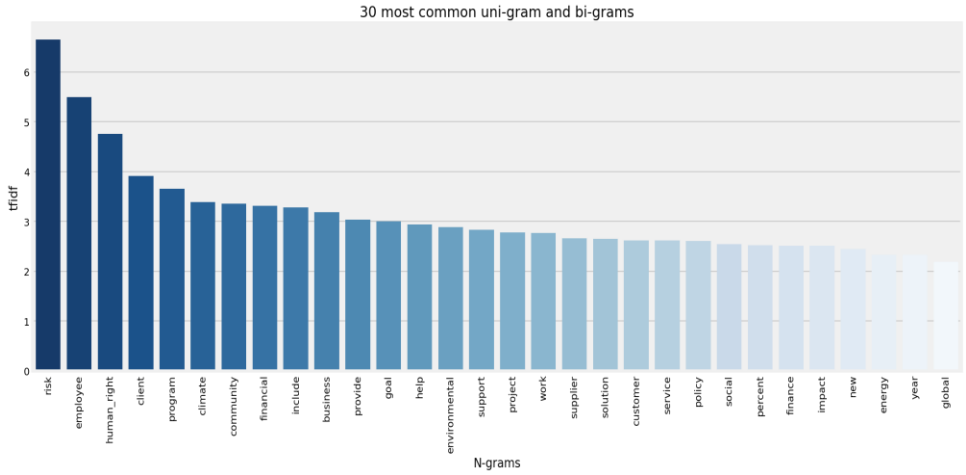


Figure 2. The frequency of ESG-related keywords

The first technique used in this code is natural language processing (NLP), which is a technique to extract and analyze the words in the correct sentences. As shown in Figure 1, it can realize the statistics of the use frequency of all words in the whole article, and export it into the form of tables or word clouds for analysis. As shown in Figure 2, the above table is a statistic of the frequency of all words in an article analyzing ESG projects. The top five words are risk, employee, human rights, customer and project. It was only in the sixth keyword that climate was mentioned. From the keyword analysis, ESG has the strongest correlation with risk.

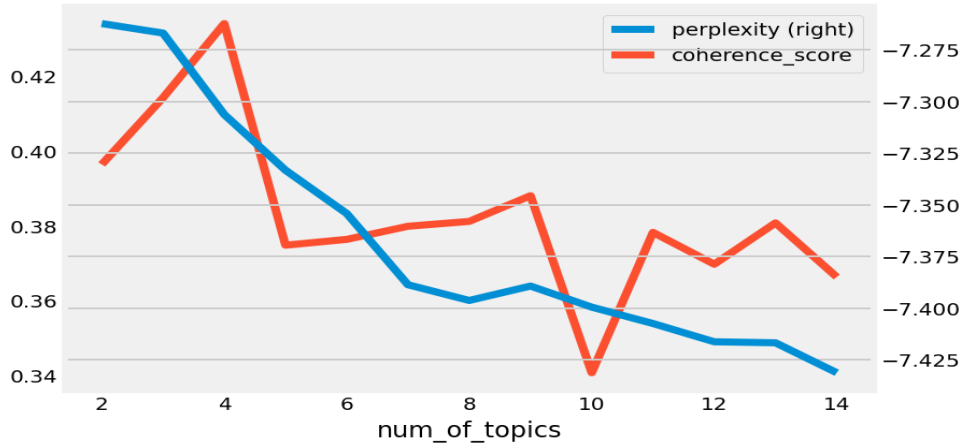


Figure 3. Perplexity and coherence_score line plots for different topics

The second technique used is the potential Dirichlet distribution (LDA). This is an unsupervised learning algorithm, mainly used in the direction of text mining. As shown in Figure 3, there are two metrics in this code, one is to test how surprised a model is to analyze information that has not been seen in the text, and the other is to detect how similar high-frequency words are in terms of semantics. The line chart above is the data analyzed by LDA technology. Through observation, the confusion index has been on a downward trend from the very high level at the beginning, indicating that the information used in this article is very new to the machine, and the information data and data used in this article are very consistent. The graph of correlation degree fluctuates greatly, indicating that the center of gravity of this article is more than one point. From the number of peaks and valleys, this article should be divided into four major directions. Then, using the perplexity and coherence score to determine the group of keywords into several topics. The picture shows that the coherence score is highest at 4 topics and the level of confusion is relatively low. So we've grouped these words into four themes.


Figure 4. ESG-related keywords for 4 topics

The third technique used is visualizing the subject. This technique can automatically form multiple topics from an article being analyzed through code, and can sort the frequency of keyword occurrence in each topic. As shown in Figure 4, we can conclude that the four themes are 1: 'support community', 2: 'value employees', 3: 'code of conduct', 4: 'ethical investments'

These themes can be used to guide company decisions, strategic planning, and business practices.

The company should take community interests into consideration and support the development and improvement of the local community through donations, volunteer activities, or collaborative projects. Meanwhile, focusing on employee welfare and development by providing a good working environment, training opportunities, and competitive benefits to motivate active participation and innovation among employees. In addition, companies should establish a clear code of conduct and ethical standards to ensure that all members of the company adhere to the principles of ethics, integrity and compliance. Second, we need to establish transparent communication and accountability mechanisms. In addition, the company should consider investing in ethical and sustainable development, such as supporting environmental projects, social responsibility investments, or renewable energy projects.

## 5. Conclusions

This paper demonstrates the effectiveness of using NLP techniques to analyze ESG reports. By automating the analysis process, we gain deeper insights into a company's sustainability practices and make informed decisions based on objective and quantitative ESG performance indicators. This approach has the potential to revolutionize the analysis of ESG reports, enabling investors, analysts, and stakeholders to assess a company's sustainability practices more efficiently and accurately.

## References

[1] Henisz, W., Koller, T., & Nuttall, R. (2019). Five ways that ESG creates value.

[2] Lee, H., Lee, S. H., Lee, K. R., & Kim, J. H. (2023). ESG Discourse Analysis through BERTopic: Comparing News Articles and Academic Papers. Computers, Materials & Continua, 75(3).

[3] Sokolov, A., Mostovoy, J., Ding, J., & Seco, L. (2021). Building machine learning systems for automated ESG scoring. The Journal of Impact and ESG Investing, 1(3), 39-50.

[4] Serafeim, G., & Yoon, A. (2022). Which corporate ESG news does the market react to? Financial Analysts Journal, 78(1), 59-78

[5] Ghildiyal, V. (2023).Developing A Chatbot-Based ESG Scoring System Using NLP And Machine Learning Techniques.

[6] Zhou, Z., Liu, M., & Tao, Z. (2023). Quantitative Analysis of Citi's ESG Reporting: LDA and TF-IDF Approaches. Financial Engineering and Risk Management, 6(3), 53-63.

[7] Fischbach, J., Adam, M., Dzhagatspanyan, V., Mendez, D., Frattini, J., Kosenkov, O., & Elahidoost, P. (2022). Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool. arXiv preprint arXiv: 2212.06540.

[8] Curran, J. R., & Clark, S. (2003). Language independent NER using a maximum entropy tagger. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (pp. 164-167).

[9] Khyani, D., Siddhartha, B. S., Niveditha, N. M., & Divya, B. M. (2021). An interpretation of lemmatization and stemming in natural language processing. Journal of University of Shanghai for Science and Technology, 22(10), 350-357.