Quantitative Analysis of Citi's ESG Reporting: LDA and TF-IDF Approaches

Zhitong Zhou[1,a,*], Mingkai Liu[1,b], Zheng Tao[1,c]

[1]Faculty of Business and Public Administration, Wenzhou Kean University, Liao Street, Wenzhou, China
[a]1234983@wku.edu.cn, [b]1235791@wku.edu.cn, [c]taozhe@kean.edu
[*]Corresponding author: Zheng Tao

Abstract: The rapid development of society is accompanied by significant changes in the environment. ESG is still in its nascent stages, and ESG policies have yet to be perfected. In this study, we examined Citibank's 2019 annual report and analyzed ESG factors using the Latent Dirichlet Allocation (LDA) method. We collected data through word clouds and identified words with the highest frequency through frequency analysis. Subsequently, we determined the optimal topic number from the perplexity curve obtained via the LDA method. Visualization charts further helped us distinguish these four ESG topics. Our findings indicate that both environmental and social factors are crucial in ESG, with corporate governance playing the most significant role.

1. Introduction

The three key components of environmental, social, and corporate governance (ESG) form the basis of responsible investment. These aspects are deeply interconnected in both economic and non-economic realms. Socially responsible investing, corporate social responsibility, and sustainability have all evolved from the ESG framework. The concept of ESG was first introduced by the UN Principles for Responsible Investment [1], which aims to encourage global investors to take ESG into account in their decision-making processes, thereby promoting sustainable development and CSR [1]. Since 2014, ESG asset management has experienced an annual growth rate of 25%. Today, an increasing number of companies, investors, and regulators are prioritizing ESG, with major stock exchanges around the world expanding their ESG presence. Eccles and Serafeim argue that companies that successfully incorporate ESG factors into their operations can achieve a competitive advantage and long-term value creation [2]. Friede, Busch, and Bassen's findings reveal a positive relationship between ESG performance and financial returns, emphasizing the importance of ESG integration in investment decision-making. [3]

Rapid changes in social, environmental, and economic spheres have intensified complex challenges such as global warming, environmental pollution, and economic and financial crises. Organizations, whether non-profit or not, are increasingly focused on corporate social responsibility and social welfare, often incorporating these factors into their decision-making processes.

ESG represents a set of non-financial performance indicators. Environmental criteria encompass

the effects of business operations on the environment, including energy consumption and renewal, climate change, waste production and disposal, and corporate environmental protection policies. Social factors involve assessing a business's impact on society, such as community relations, employee health, and labor standards. Governance pertains to a company's corporate governance, including effective management oversight and balancing internal power dynamics. In the study of the impact of corporate governance on stocks, it is found that well-governed companies tend to have higher equity valuations, which indicates that strong corporate governance is a key factor to create shareholder value [4].

2. The purpose and significance of studying ESG at Citibank

Citi has consistently recognized ESG as a fundamental component of its strategy, acknowledging the importance of ESG factors as a source of long-term value. Examining Citi's ESG strategy and practices offers valuable insights into the bank's ESG performance and outcomes, as well as the integration of ESG considerations in its investment decisions and business operations. Moreover, the influence of ESG factors on companies is gaining increasing public recognition. Analyzing Citi's ESG performance and outcomes enables us to evaluate the impact of ESG factors on the bank's investment choices and business activities, and how these factors influence aspects such as brand reputation, shareholder returns, and employee satisfaction.

As a leader in ESG, Citibank's practices and performance can offer valuable insights and lessons for other organizations. Exploring Citi's ESG strategies and practices can serve as a benchmark for other organizations, thereby fostering the global advancement of ESG practices and development.

The rest of the study is organized as follows. Section 2 provides a brief overview of the literature. Section 3 provides a data graph and an explanation of the methodology. Section 4 describes the results of the analysis. Section 5 explains the results and summarizes the paper.

3. Literature Review

3.1 ESG and Related Content

ESG, an acronym for Environment, Social, and Governance, is a concept that was first introduced in 2006 by the United Nations Principles for Responsible Investment (PRI). Though not a brand-new concept, ESG is often discussed alongside SRI (Socially Responsible Investing), CSR (Corporate Social Responsibility), and sustainable development.

While ESG emphasizes the environmental, social, and governance performance of a company, sustainability predominantly focuses on the environment. Sustainability is primarily concerned with environmental issues and is defined as development that meets the needs of the present without compromising the ability of future generations to meet their own needs [5]. Sustainable development has three objectives: economic sustainability, social sustainability, and environmental sustainability. The main goal is to ensure balanced development for both present and future generations, promoting economic, social, and environmental harmony. Many empirical studies on the relationship between ESG factors and corporate performance have found that companies that perform well in ESG tend to achieve higher long-term financial returns [6]. Overall, CSR and SRI have strong relevance to this area. Revelli and Viviani's article, examining the financial performance of socially responsible investing (SRI) Adopting SRI strategies does not hurt investor returns. SRI strategies that emphasize ESG factors tend to perform better in terms of financial performance [7].

CSR and SRI are closely related to ESG. CSR is defined as a company's voluntary integration of environmental and social considerations into its business operations and interactions with stakeholders [8]. SRI, on the other hand, aims to incorporate specific ethical standards into a

financially driven investment process [9]. Investors' moral motivation and concern for corporate social responsibility promote investors to hold SRI investment funds. The research also finds that holders of socially responsible investment funds are more inclined to participate in corporate governance activities, thus promoting the implementation of ESG factors in companies to a certain extent [10].

3.2 Citibank's Perspective on ESG

In recent years, there has been a substantial increase in the number of organizations integrating ESG factors into their business operations [11]. Citigroup's website clearly states that the company has made ESG a core component of its business strategy and that ESG factors have been an essential source of long-term value for Citi [12].

First, Citibank has made it clear that it will actively support the Sustainable Development Goals (SDGs) and integrate them into its business strategy. Citibank has committed to increasing its sustainable finance investment target to $100 billion by 2025 [12]. In addition, Citibank will strive to reduce its own carbon footprint and become carbon neutral in order to meet the challenge of climate change [12]. At the social level, Citibank focuses on employee well-being and diversity and is committed to creating a more inclusive corporate culture. Citibank has implemented a series of initiatives aimed at improving the diversity and inclusion of its employees, such as the establishment of a "Diversity and Inclusive Development Committee" and the implementation of a diversity recruitment strategy[13]. In terms of governance, Citibank emphasizes transparency and ethical practices as key factors in achieving sustainable development. Citibank regularly publishes ESG reports, showing its latest progress and efforts in environmental, social and governance [14].

Based on the information provided by Citibank, the bank's exceptional performance in environmental, social, and governance (ESG) factors is recognized and backed by investors. Its leadership in ESG allows it to draw in more investors, consequently increasing its share price and market value. Furthermore, Citibank's high ESG ratings have had a positive influence on analyst recommendations. This could result in analysts having a preference for the bank's stock, eventually leading to enhanced market performance for Citi's shares.

Additionally, we collected ESG ratings for different aspects from Yahoo, as shown in Figure 1.

2022-8

totalEsg	26.68
socialScore	13.11
governanceScore	11.82
environmentScore	1.75
esgPerformance	AVG_PERF
percentile	50.69
peerGroup	Banks
highestControversy	4

Figure 1: ESG Score from Yahoo Finance.

4. Research technique

4.1 Data Pre-processing

Prior to the actual analysis, we first pre-processed the dataset. This included text cleaning, word separation and deactivation removal. Text cleaning includes removing punctuation, numbers and special characters, converting text to lower case, etc. Word splitting splits the text into words or phrases for subsequent processing. Deactivated words are those words that appear more frequently in the text but are less informative for topic analysis, such as "and", "the", etc.

4.2 Feature Extraction: TF-IDF

TF-IDF (t, d) = TF(t, d) * IDF(t)
where:
- t represents the term or word
- d represents the document in which the term appears
- TF(t, d) refers to the term frequency, which is the number of times term t appears in document d
- IDF(t) refers to the inverse document frequency, which is calculated as:
IDF(t) = log$(N / DF(t))$
where:
- N is the total number of documents in the corpus
- DF(t) is the number of documents containing the term t

By calculating the TF-IDF score for each term in the document, we can identify the most important words for each ESG topic, as higher TF-IDF scores signify greater relevance to the topic.

To represent text data, we use the TF-IDF (Term Frequency-Inverse Document Frequency) method, a statistical method for converting text into numerical features that measure the importance of a word in a document. TF indicates the word frequency, i.e., the number of times a word appears in a document. IDF indicates the inverse document frequency, which measures whether a word is common or not. By calculating the TF-IDF value for each word, we can obtain a vector of values that can be used to represent the features of a document. This representation helps to highlight those words that are more important to the subject of the document, thus improving the effectiveness of topic modelling.

4.3 Data Visualization: Word Clouds

Figure 2: Words visualized by word clouds.

Prior to the topic modelling, the data was visualized using a Word Cloud. The Word Cloud is a way of visualizing words according to their frequency in the text, with more frequent words being displayed in a larger font and vice versa. This helps us to visualize the most prominent words and themes in the dataset.

Figure 2 shows the word frequency clearly through visualization, which is easy for readers to read.

4.4 Topic Modelling

4.4.1 Construct the LDA model

The Latent Dirichlet Allocation (LDA) model is an unsupervised machine learning technique used for topic modeling. It is a generative probabilistic model that assumes documents are generated from a mixture of topics, and 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 $\theta_d \sim$ Dirichlet(α) for each document d. b. Choose $\varphi_k \sim$ Dirichlet(β) for each topic k. c. For each word w_{dn} in document d: i. Choose a topic $z_{dn} \sim$ Multinomial(θ_d). ii. Choose a word w_{dn} from the topic z_{dn}, $w_{dn} \sim$ Multinomial(φ_{z_dn}).

3) Key formulas:
- $P(z_{dn}=k \mid w_{dn}) \propto P(w_{dn} \mid z_{dn}=k) * P(z_{dn}=k \mid \theta_d)$
- $P(w_{dn} \mid z_{dn}=k) = \varphi_k(w_{dn})$
- $P(z_{dn}=k \mid \theta_d) = \theta_d(k)$

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: $P(z_{dn}=k \mid w_{dn}) \propto (N_{dk} + \alpha) * (N_{kw} + \beta) / (N_k + M*\beta)$ 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) $\theta_d(k) = (N_{dk} + \alpha) / (N_d + K*\alpha)$
2) $\varphi_k(w) = (N_{kw} + \beta) / (N_k + M*\beta)$

After constructing the LDA model, the output consists of document-topic distributions (θ_d) and topic-word distributions (φ_k). These can be used to analyze the most prominent topics in the document collection and the most relevant words associated with each topic.

4.4.2 Use the LDA model

After feature extraction and visualisation, we carried out topic modelling on the data. Topic modelling is an unsupervised machine learning method for extracting hidden topics from a large number of documents. In this study, we use the Latent Dirichlet Allocation (LDA) model, which assumes that documents are generated from multiple topics, each consisting of a set of related words. By using the LDA model, we can estimate the distribution of topics in each document and the distribution of words in each topic.

In implementing the LDA model, we first determined the number of topics (K), which is a hyperparameter that needs to be adapted to the actual situation. Next, we used the TF-IDF vector as the input data to train the LDA model. Once the training was complete, we obtained a distribution of topics for each document and a distribution of words for each topic. By analyzing these distributions, we were able to identify the hidden topics in the dataset.

5. Evaluation and Presentation of Results

To evaluate the effectiveness of the topic modelling, we used the metrics of Perplexity and Topic Coherence. Perplexity measures the model's ability to predict unknown documents, with lower values indicating better performance. Topic Coherence assesses the semantic relevance of words within a topic, with higher values indicating a more explanatory topic.

Finally, the results of the topic modelling were visualised, including a display of the keywords for each topic and the topic distribution for each document. The visualisation of the results allows us to more intuitively understand the relationships between individual themes and their distribution across the dataset. In addition, we can further analyze the specific meanings represented by each theme to derive insights about the dataset.

5.1 TF-IDF Analysis of Citibank's ESG Data

Figure 3 displays the 30 most frequent words in the report, obtained using TF-IDF technology. The top words in Citibank's data include risk, employees, human rights, client, program, climate, community, financial, business, and others. Among them, "risk" appears most frequently, followed by "employee" and "human rights." The frequency of the subsequent words decreases gradually.

These results indicate that Citibank places significant emphasis on ESG, focusing on risk management, employee rights, and customer service. This aligns with the growing importance that modern enterprises place on ESG issues. Companies increasingly incorporate ESG into their strategic planning and business decisions to establish long-term stability, reputation, enhance their social responsibility, and increase corporate value, ultimately reducing enterprise management risks.

Risk management is a vital aspect of Citibank's ESG strategy. The term "risk" in the ESG context generally refers to identifying and managing environmental, social, or corporate governance risks. Citibank actively manages these risks to ensure the sustainability and soundness of its operations. The experimental results help to further understand Citibank's ESG performance and influence.

The high frequency of "employee" and "human rights" suggests that Citibank faces challenges concerning employee rights, such as employee satisfaction, benefits, diversity, and inclusion. Citibank should pay more attention to and strengthen the management of these issues to avoid a negative impact on its ESG performance.

The words "risk," "employee," and "human rights" are often associated with adverse events and challenges in the ESG space. Their high frequency may indicate issues in Citibank's ESG performance, which could negatively impact its reputation and profits. Citibank should strengthen its ESG strategy to improve its performance and corporate image.
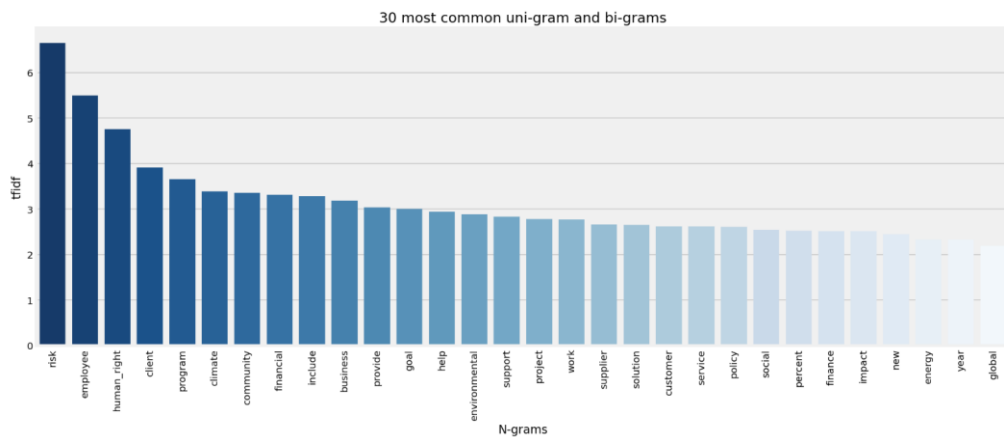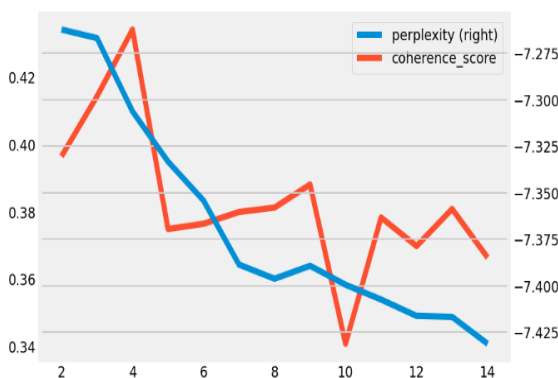
Figure 3: Unigram and Bigram Analysis of words with different frequencies.

2	-7.262310	0.396425
3	-7.266900	0.414277
4	-7.306288	0.434137
5	-7.333262	0.374806
6	-7.354216	0.376360
7	-7.388671	0.379870
8	-7.396271	0.381175
9	-7.389304	0.388052
10	-7.399501	0.340663
11	-7.407429	0.378231
12	-7.416261	0.369714
13	-7.416749	0.380772
14	-7.431229	0.366291

Figure 4&5: Confusion curve and optimal topic number.

5.2 LDA Model Analysis for Citibank's ESG Topics

Combine Figure 4 and Figure 5, observe the best combination of confusion degree and coherence, and get the best combination number.

The perplexity curve was used to visualize the influence of the number of topics on the perplexity and coherence of the language assessment model. When the number of topics reached four, the coherence of the language assessment model was at its highest, and the perplexity decreased to a reasonable value. The LDA model then identified potential topics frequently discussed in Citibank's ESG data.

To ensure interpretability, the 30 most relevant terms from the TF-IDF analysis were extracted, and the appropriate value of λ was used to define each topic. Based on the LDA model, the word frequency of each related term was obtained, and the one with the highest word frequency was selected as the dominant topic.

Figure 6 to 9 above shows the proportions of the four different topics obtained and the proportions of words linked to their topics.
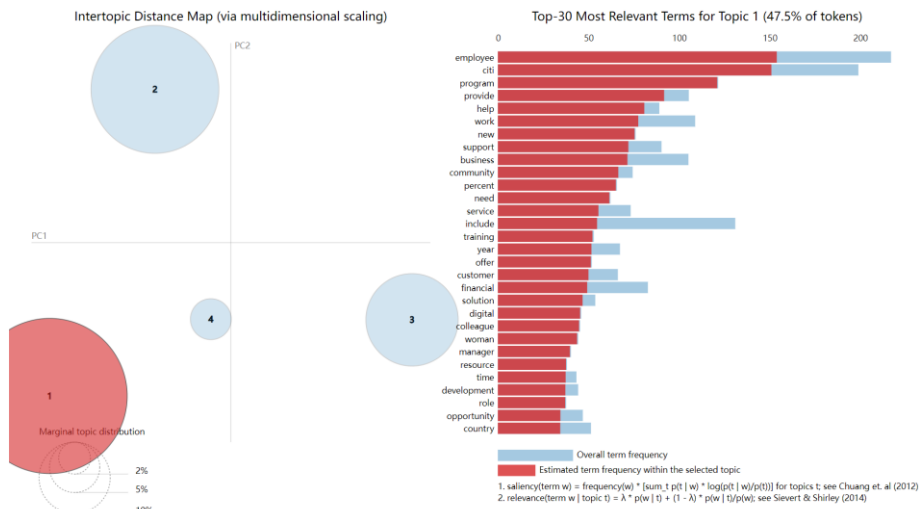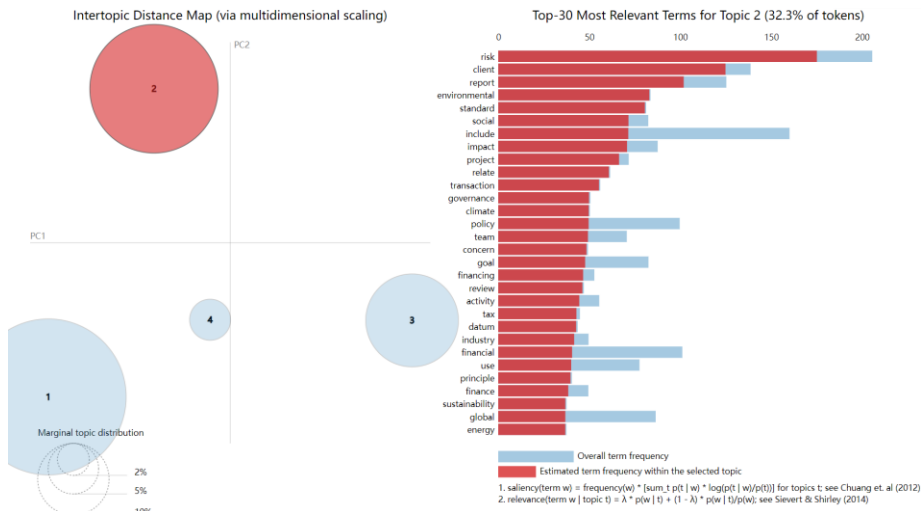
Figure 6: Viewable for Topic 1.
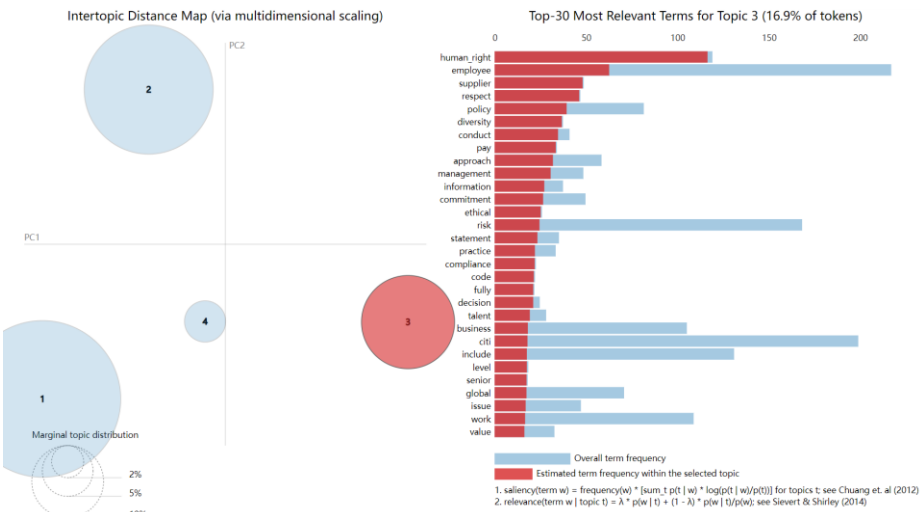

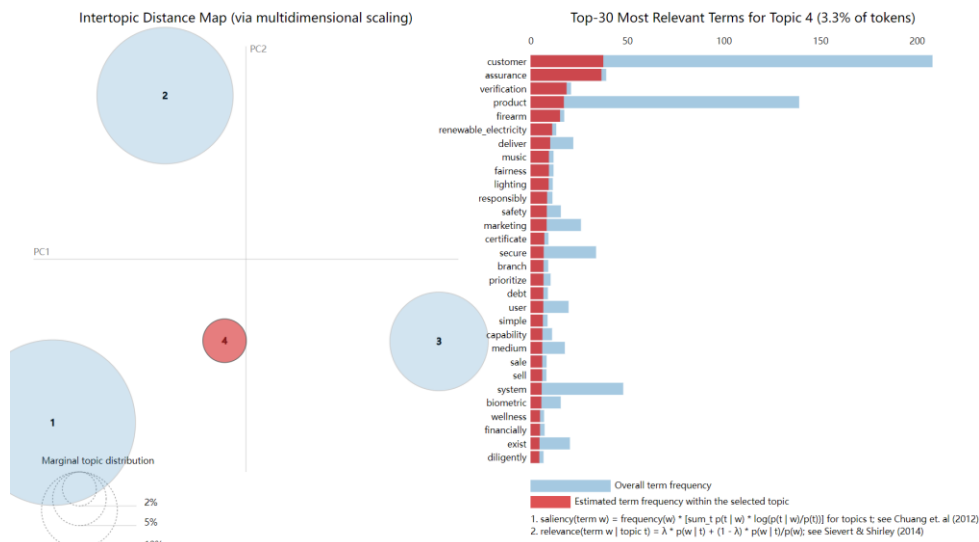Figure 7: Viewable for Topic 2.

Figure 8: Viewable for Topic 3.

Figure 9: Viewaqble for Topic 4.

The four leading issues related to Citibank's ESG, derived from the LDA model, are:

Citibank's ESG Support Community

Citibank's ESG Value Employee

Citibank's ESG Code of Product

Citibank's ESG Ethical Investments

Figure 9 and Figure 10 show the different topics based on the LDA model and the dominant topics derived from the high word frequency.

Figure 10: Citibank's 4 ESG subject maps.

Citibank's ESG Support Community strategies focus on improving social responsibility by

optimizing sustainable development. These strategies involve collaboration with stakeholders to create positive social impact, enhance Citibank's reputation, and maintain long-term financial stability.

Citibank's ESG Value Employee emphasizes the bank's commitment to promoting environmental, social, and governance principles among its employees. By promoting ESG values among its employees, Citibank demonstrates its commitment to sustainability and social responsibility, ensuring that its employees are aligned with its values and goals.

Citibank's ESG Code of Product is a set of guidelines developed to ensure that the bank's products comply with Environmental, Social, and Governance principles. By developing ESG-compliant products, Citibank offers customers more choices and promotes sustainable and socially responsible business practices.

6. Conclusion

In conclusion, this study provides valuable insights into Citibank's ESG performance and practices. The analysis of the bank's ESG data using TF-IDF and LDA model reveals four main topics: Citibank's ESG Support Community, Citibank's ESG Value Employee, Citibank's ESG Code of Product, and Citibank's ESG Ethical Investments. These topics highlight Citibank's commitment to social responsibility, sustainable development, employee engagement, and responsible product offerings.

The findings suggest that Citibank has integrated ESG factors into its core strategy, emphasizing risk management, employee rights, and customer service. However, the high frequency of terms like "risk," "employee," and "human rights" may indicate some challenges in Citibank's ESG performance. To mitigate potential negative impacts on its reputation and profits, Citibank should continue to strengthen its ESG strategies and address potential shortcomings.

The results of this study not only contribute to the understanding of Citibank's ESG performance and practices but also provide valuable insights for other organizations seeking to incorporate ESG factors into their decision-making processes. By learning from Citibank's experiences, other organizations can adopt similar ESG practices, ultimately promoting global ESG development and responsible business conduct.

Figure 1 is an ESG Score collected from another Yahoo Finance. Figure 1 depicts ESG scores for social, governance, and environment, as well as average ESG Performance and percentiles. It shows the peer group are banks. Social scores are the highest, followed by governance scores, and environmental scores are the lowest. It can be seen that people's awareness and evaluation of ESG environment is not very high.

References

[1] United Nations Global Compact & United Nations Principles for Responsible Investment. (2006). Principles for responsible investment. Retrieved from https://www.unpri.org/download?ac=1320
[2] Eccles R G., & Serafeim G. (2013). The performance frontier: innovating for a sustainable strategy. Harvard Business Review, 91(5), 50–56, 58, 60, 150.
[3] Friede G., Busch T., & Bassen A. (2015). ESG and financial performance: aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), 210–233. https: // doi. org/ 10. 1080/ 20430795. 2015. 1118917
[4] Gompers P A., Ishii J., & Metrick A. (2003). Corporate Governance and Equity Prices. Quarterly Journal of Economics, 118(1), 107–156. https://doi.org/10.1162/00335530360535162
[5] Brundtland G H. (1987). Our common future. World Commission on Environment and Development. Retrieved from https://sustainabledevelopment.un.org/content/documents/5987our-common-future.pdf
[6] Clark G L., Feiner A., & Viehs M. (2015). From the Stockholder to the Stakeholder: How Sustainability Can Drive Financial Outperformance. Social Science Research Network. https://doi.org/10.2139/ssrn.2508281
[7] Revelli C., & Viviani J L. (2015). Financial performance of socially responsible investing (SRI): what have we learned?

A meta-analysis. Business Ethics: A European Review, 24(2), 158–185. https://doi.org/10.1111/beer.12076

[8] Strand R., Freeman R E., & Hockerts K. (2015). Corporate Social Responsibility and Sustainability in Scandinavia: An Overview. Journal of Business Ethics, 127(1), 1–15. https://doi.org/10.1007/s10551-014-2224-6

[9] Sandberg J., Nilsson J., & Lodh S. (2008). Corporate social responsibility and investors' assessments of company value. Journal of Business Ethics, 82(4), 413-428.

[10] Riedl A M., & Smeets P. (2017). Why do investors hold socially responsible mutual funds? Journal of Finance, 72(6), 2505-2550.

[11] Amel-Zadeh A., & Serafeim G. (2018). Why and how investors use ESG information: Evidence from a global survey. Financial Analysts Journal, 74(3), 87-103. https://doi.org/10.2469/faj.v74.n3.3

[12] Citigroup. (2021). Environmental, social, and governance. Retrieved from Citi 2021 Environmental, Social & Governance Report https://www.citigroup.com/rcs/citigpa/akpublic/storage/public/Global-ESG-Report-2021.pdf

[13] Citigroup. (2020). Environmental, social, and governance. Retrieved from Citi 2021 Environmental, Social & Governance Report https://www.citigroup.com/rcs/citigpa/akpublic/storage/public/Global-ESG-Report-2020.pdf

[14] Citigroup. (2019). Environmental, social, and governance. Retrieved from Citi 2021 Environmental, Social & Governance Report https://www.citigroup.com/rcs/citigpa/akpublic/storage/public/Global-ESG-Report-2019.pdf