Thematic Categorization on Pineapple Production in Costa Rica: An Exploratory Analysis through Topic Modeling.

Valentina Tretti-Beckles
Potsdam University
Potsdam, Germany
tretti@uni-potsdam.de

Adrián Vergara-Heidke
University of Costa Rica
San Jose, Costa Rica
adrian.vergara@ucr.ac.cr

Abstract

Costa Rica is one of the largest producers and exporters of pineapple in the world. This status has encouraged multinational companies to use plantations in this Central American country for experimentation and the cultivation of new varieties, such as the Pinkglow pineapple. However, pineapple monoculture has significant socio-environmental impacts on the regions where it is cultivated. In this exploratory study, we aimed to analyze how pineapple production is portrayed on the Internet. To achieve this, we collected a corpus of texts in Spanish and English from online sources in two phases: using the BootCaT [1] toolkit and manual search on newspaper websites. The Hierarchical Dirichlet Process (HDP) topic model was then applied to identify dominant topics within the corpus. These topics were subsequently classified into thematic categories, and the texts were categorized accordingly. The findings indicate that environmental issues related to pineapple cultivation are underrepresented on the Internet, particularly in comparison to the extensive focus on topics related to pineapple production and marketing.

1 Introduction

Pineapples are widely available in supermarkets across Europe, Japan, and the United States. This tropical fruit is cultivated in several countries, with Costa Rica ranking as one of the leading producers and exporters (FAO, 2024). While pineapple production has generated significant revenue for companies in Costa Rica, it has also raised concerns regarding its impact on labor rights (Rodríguez and Prunier, 2020; Salgado and Acuña, 2021; León and Montoya, 2021) and the environment.

Numerous academic and NGO studies have documented the environmental and health damage caused by pineapple monocultures in Costa Rica. The adverse environmental effects stem primarily from the use of pesticides, which contaminate water and soil, thereby affecting humans, animals, and plants (Valverde and Chaves, 2020; Carazo and Aravena, 2016). Additionally, these effects result from the excessive use of water for irrigation, which depletes wetlands and underground water reserves (Carazo and Aravena, 2016). Finally, monoculture practices leave the land over-exploited and depleted of essential nutrients (Carazo and Aravena, 2016; Obando, 2020). These issues are emblematic of the current geological epoch: the Anthropocene (de Cózar, 2019).

A notable example of the Anthropocene is the creation of the pink pineapple (known as Pinkglow) in Costa Rica. This new variety was developed in the laboratories of the Del Monte company, which holds the patent for it (Del Monte, 2020). The distinctive pink coloration serves purely aesthetic purposes, catering to a market heavily driven by visual appeal. Due to its high market value, the pink pineapple is exclusively marketed outside of Costa Rica (Riviera, 2024), where it is promoted as an exotic product from a tropical country.

In this context, an interdisciplinary group initiated a study on the relationship between pineapple plantations and the territory, as well as the representation of the pineapple—particularly the Pinkglow variety—as a cultural object. Within this framework, we sought to explore the topics circulating on the Internet about pineapple production

[1] *Bootstrapping Corpora and Terms*: a suite of perl programs implementing an iterative procedure to bootstrap specialized corpora and terms from the web (Baroni and Bernardini, 2004). It is a free toolkit available at: https://bootcat.dipintra.it/

Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025), pages 44–55

and Costa Rica and to examine whether there are differences between texts in Spanish and English. To address these goals, we aim to apply advances in Natural Language Processing (NLP), specifically Topic Modeling, to identify the thematic categories present in the corpus of digital texts. Accordingly, the objective of this exploratory research is to identify and categorize, through Topic Modeling, the thematic categories associated with pineapple production in Costa Rica within a corpus comprising diverse textual genres.

Numerous studies have applied Natural Language Processing (NLP) techniques to address environmental issues. Some focus on identifying and geographically mapping the impacts of climate change by analyzing academic papers and scholarly articles (Mallick et al., 2024) or through sentiment analysis (Stede and Patz, 2021; Sham and Mohamed, 2022; Mi and Zhan, 2023; Krishnan and Anoop, 2023), and stance classification in tweets (Mohammad et al., 2016) or news articles (Luo et al., 2020). Additionally, efforts have been made to create datasets related to environmental topics, such as EcoVerse, an English Twitter dataset for eco-relevance classification, stance detection, and environmental impact analysis (Grasso et al., 2024); GERCCT, a German Twitter dataset for argument mining (Schaefer and Stede, 2022); and ClimaText, a dataset for detecting topics related to climate change (Varini et al., 2021).

In addition, several studies have applied topic modeling techniques within the environmental domain. These include research on climate change discussions in social media (Dahal et al., 2019; Al-Rawi et al., 2021; Uthirapathy and Dominic, 2023; Kim and Kim, 2024), correlations between topics and sentiment in news articles (Jiang et al., 2017; Rabitz et al., 2021; Ejaz et al., 2022; McAllister et al., 2024), analyses of Nature and Science editorials (Stede et al., 2023), and targeted journal publications (Kim et al., 2021), as well as research and policy papers (Werneck and Gomes, 2023). Topic modeling has also been applied to election manifestos and parliamentary debates (Navarretta and Hansen, 2023).

Although the aforementioned studies provided a foundation, our exploratory research adopted a distinct thematic and methodological approach. Specifically, we examined pineapple production across various textual genres, compared topics and thematic categories between English and Spanish, and implemented the topic model that yielded the highest coherence score (HDP).

2 Method

The method used consists of five phases (see figure 1). First, the corpus was extracted and retrieved through two stages:

1. Extraction using BootCaT involved nine tuples, each consisting of three keywords, with at least one keyword being "pineapple" for English and either "piña" or "piñera" for Spanish.
   (a) English keywords: "pineapple", "pineapple plantation", "pinkglow", "costa rica" and "pink pineapple".
   (b) Spanish keywords: "piñera", "plantación de piñas", "piña rosada", "pinkglow", "costa rica", and "piña".
2. Manual search on newspaper websites using the keywords related to pineapple plantation.
   (a) Costa Rican newspapers: Delfino, La Nación, El Observador, La República, Semanario Universidad, and Diario Extra.
   (b) International newspapers: CNN, The Guardian, The New York Times, and The Times.

Second, a manual corpus analysis was conducted to remove irrelevant texts (e.g., incomplete texts, texts without mentions of pineapples or the Pinkglow variety, and empty texts) and to classify textual genres (see table 3 in *Appendix A*). Next, topic modeling experiments were carried out on both subsets of the data using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Hierarchical Dirichlet Processes (HDP) (Whye Teh et al., 2006), and Latent Semantic Analysis (LSA) (Landauer and Dumais, 2008). This was followed by the classification of the documents based on their dominant topics. Then, thematic categories were determined according to the dominant topics. Finally, the data documents were classified according to their respective thematic category.

2.1 Data

The dataset was collected between November and December 2024 and consists of 221 texts, which

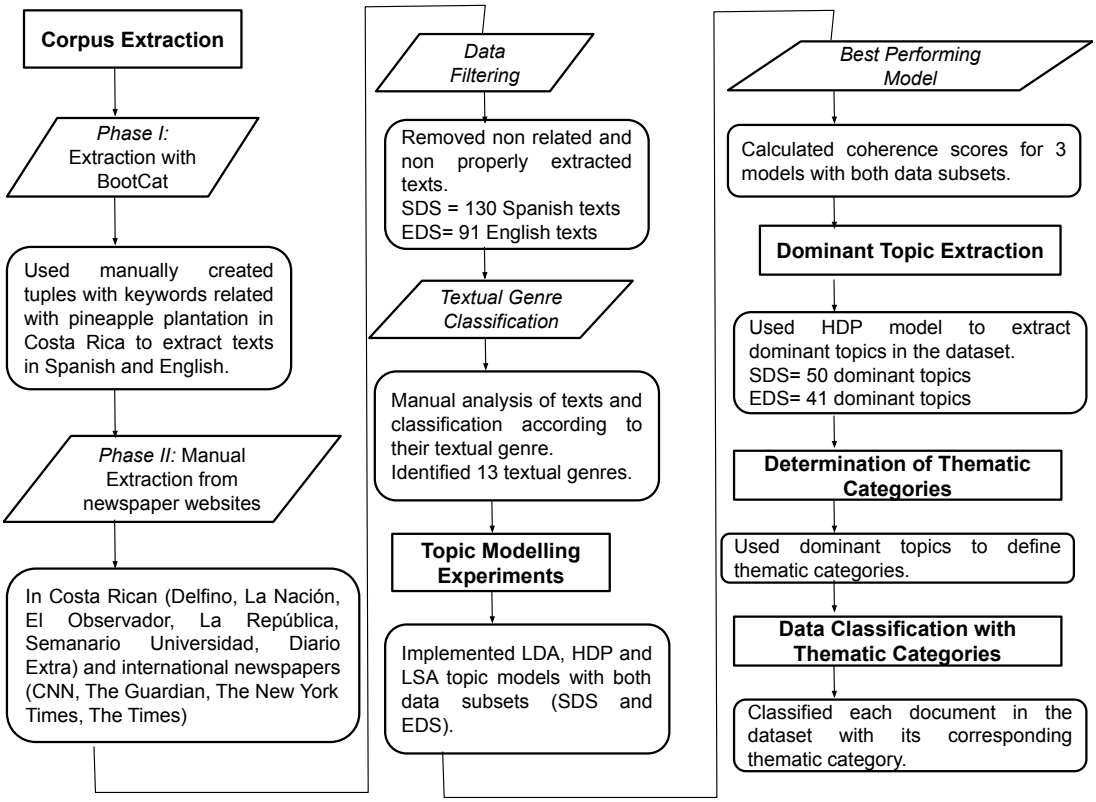

Figure 1: Flowchart with phases followed in the method.

include: blog sites, business websites, news websites, product pages, institutional websites, and documents (e.g., papers, reports, manuals). It comprises a total of 215,241 words. The texts are written in two languages: Spanish (Spanish Data Subset, SDS) and English (English Data Subset, EDS) (see table 1 for more details) and were published between 2001 and 2024. The dataset documents were classified according to their textual genre, with 13 genres identified in total (see table 3 in *Appendix A*). Among these, news articles, divulgative notes, and narrative notes are the most frequent (see figure 2).

Data	**Number of documents**	**Number of words**	**Number of characters**
Spanish	130	161,765	1,032,436
English	91	53,476	333,786
Total	221	215,241	1,366,222

Table 1: Data Distribution considering Language, Words and Characters.

2.2 Topic Modeling Experiments

As mentioned above, the topic modeling was conducted using three models: Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Processes (HDP) and Latent Semantic Analysis (LSA) with both SDS and EDS data subsets. The models were evaluated using the topic coherence (TC) score from *Gensim* (Röder et al., 2015). Among these, HDP demonstrated the highest coherence scores for both subsets (see figures 3 and 4) and was consequently selected for further analysis and the development of thematic categories.

Drawing on reference literature (Del Monte, 2020; Riviera, 2024; Carazo and Aravena, 2016; Obando, 2020; Rodríguez and Prunier, 2020; FAO, 2024), we established 13 thematic categories related to pineapple cultivation in Costa Rica, along with an 'other' category for cases that did not fit within a specific category. Subsequently, we classified the dominant topics within these categories. To do this, we considered that at least four out of the ten words in each topic needed to be associated with a thematic category. When a topic contained words corresponding to multiple thematic categories, it was classified under all relevant categories, with a maximum of two assign-

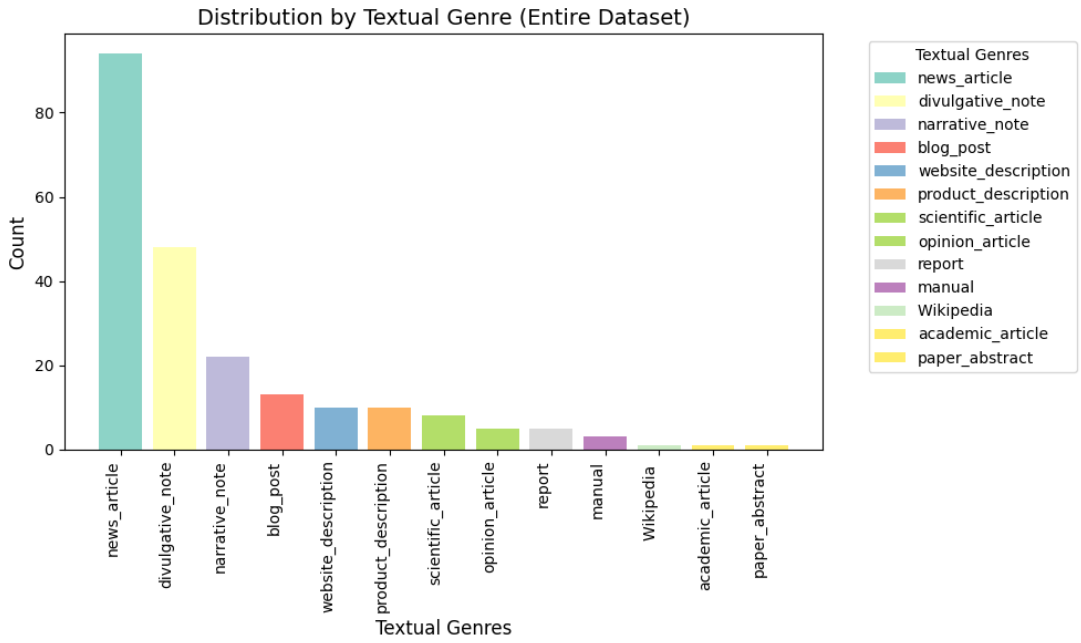

Figure 2: Distribution by Textual Genre (Entire Dataset)

ments per topic. Finally, based on the classification of dominant topics, we analyzed the distribution of texts across thematic categories.

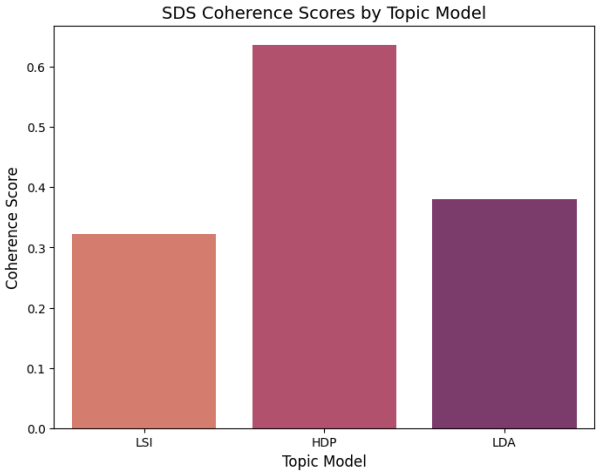

Figure 3: Model's Coherence Score for SDS.

3 Analysis

The HDP model dynamically determines the number of topics. For our data, it identified 73 topics for the SDS and 137 topics for the EDS. Of the 73 topics in the SDS, 50 were dominant, while of the 137 topics in the EDS, only 41 were dominant. The dominant topics in both subsets were classified according to pre-established and emerging categories (see figure 2). It is important to note that a single topic could correspond to multiple categories.

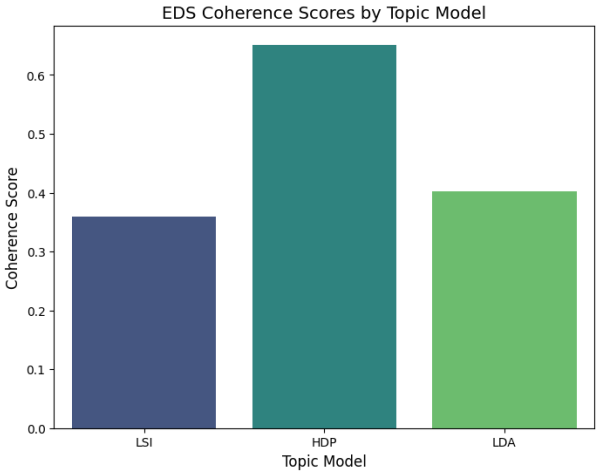

Figure 4: Model's Coherence Score for EDS.

The results (figure 5 and table 2) reveal differences in the distribution of the thematic categories between the two subsets. The most notable distinction is found in the categories 'Pineapple production' and 'Consumption food-aesthetic.' In the SDS, 37% of the topics were classified under 'Pineapple production,' whereas only 11% of topics in the EDS were identified with this category. Conversely, 53% of the topics in the EDS fell within the 'Consumption food-aesthetic' category, compared to only 14% in the SDS. These findings suggest that the primary focus of the SDS is on the production and export of pineapples in Costa

Category	Number of dominant topics SDS	Percentage of dominant topics SDS	Number of dominant topics EDS	Percentage of dominant topics EDS
Contamination	1	2%	5	10%
Erosion	1	2%	1	2%
Lack of Water	2	3%	0	0%
Climate Change	1	2%	0	0%
Health Problems	0	0%	1	2%
Labor Shortages	2	3%	0	0%
Legal Issues	3	5%	0	0%
Consumption Food-Aesthetic	8	14%	25	53%
Pineapple Production	21	37%	5	11%
Health Benefits	1	2%	0	0%
Sales	4	7%	4	8,5%
Cultivation	4	7%	1	2%
Sustainability	0	0%	1	2%
Other	9	16%	4	8,5%
Total Topics Classified	57	100%	47	100%

Table 2: Distribution of dominant topics in SDS and EDS data subsets.

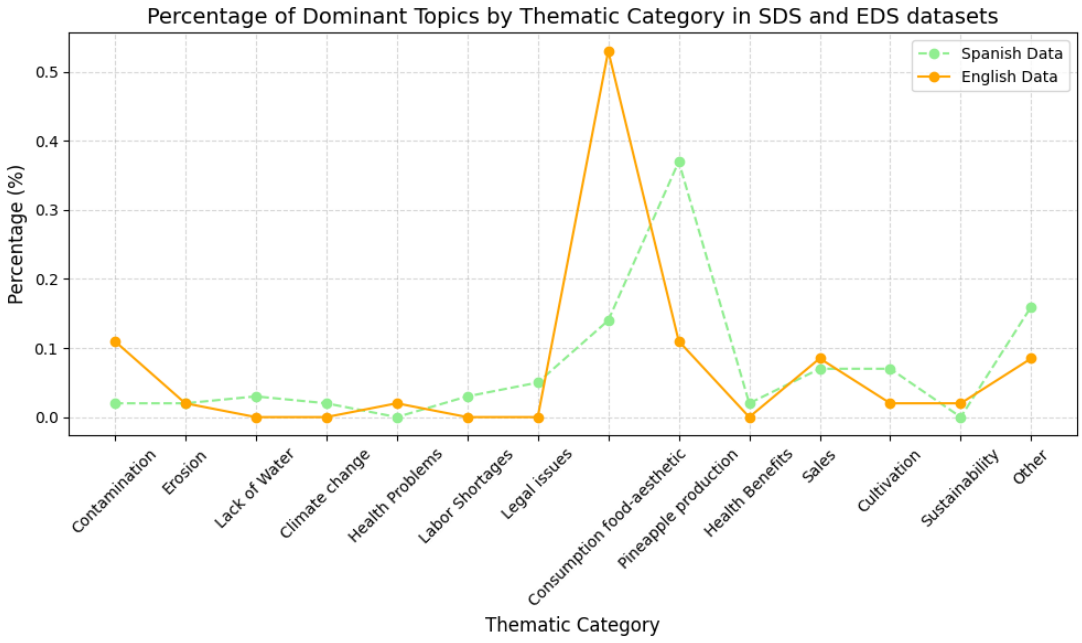

Figure 5: Percentage of Dominant Topics Category by Languages.

Rica, independent of their consumption. In contrast, the results from the EDS indicate a greater emphasis on pineapple consumption, particularly from a food or aesthetic perspective, which may be associated with texts aimed at promoting the international consumption of this fruit, especially the Pinkglow variety.

On the other hand, a greater variety of dominant topics is observed in the SDS compared to the EDS. In the SDS, only two thematic categories lacked a dominant topic, and nine dominant topics could not be classified into any category. In contrast, in the EDS, five thematic categories remained unassigned, and only four dominant topics

were unclassified.

The results (figure 6) indicate that the thematic category 'Pineapple production' appears in the greatest number of SDS texts, with 61 occurrences, representing 41% of the total occurrences across all thematic categories (149 occurrences). This finding suggests that the majority of the Spanish texts predominantly address aspects of pineapple production. This is somewhat expected, as the majority of the dominant topics (21 topics) were classified under this thematic category. In second place, the thematic category 'Consumption food-aesthetic' appeared in 38 texts (26%). Despite being composed of only eight dominant topics, this category's frequency of occurrence is noteworthy.

On the other hand, in the EDS, the results show that the dominant topics in most texts belong to the thematic category 'Consumption food-aesthetic' (50 occurrences, 47%), followed by 'Pineapple production' (26 occurrences, 24%). This suggests that the content of most EDS texts focuses on consuming pineapples in general, and Pinkglow in particular. This is consistent with the fact that the majority of the dominant topics (25 topics) were classified under this thematic category. However, it is notable that 24% of texts (26 occurrences) feature dominant topics related to 'Pineapple production,' despite being represented by only five topics. Additionally, the category 'Contamination' appears in just 10 texts (9%), further emphasizing the thematic focus of the EDS.

Each dominant topic contributes a different percentage within each text. This contribution reflects the extent to which the topic is represented in the text, thereby indicating its importance in terms of coverage or reiteration. For this reason, we chose to examine the percentage contribution of topics within each thematic category across the texts where they were dominant.

Considering the thematic categories with the highest number of texts ('Pineapple production,' 'Other,' 'Consumption food-aesthetic,' and 'Sales'), the results show that in the SDS, the topics within the 'Consumption food-aesthetic' category (80%) contribute the most to the texts when dominant, as their average contribution exceeds that of the other categories (figure 7). This is followed by 'Sales' (77%) and 'Pineapple production' (75%), while 'Other' has the lowest average contribution (65%). These results suggest that when the 'Consumption food-aesthetic' category appears, its topics occupy a greater proportion of the text or are repeated more frequently, indicating a stronger presence compared to other topics within the same text. Additionally, the high average (80%) suggests that the primary topics in these texts are more specific in nature.

Finally, the average contribution of the dominant topics in the 'Other' thematic category (65%) is the lowest among the four categories with the highest number of texts. It is important to refrain from providing detailed explanations for this result, as the 'Other' category encompasses a wide range of topics. As a result, the content of these texts may vary significantly, therefore it would not be advisable to analyze them as a group.

In the EDS (figure 8), two thematic categories clearly group the largest number of texts: 'Pineapple production' and 'Consumption food-aesthetic.' The mean for 'Pineapple production' (75%) suggests that, in the texts where its topics are dominant, there is a greater thematic specialization compared to 'Consumption food-aesthetic' (73%). Nevertheless, the topics in both thematic categories contribute over 70% on average in most of the texts in which they are dominant. This substantial contribution indicates a significant presence of these topics, either through their thematic focus or repeated mention within the texts.

4 Discussion

The results indicate that in most texts from both subsets, the most frequent categories are 'Pineapple Production' and 'Consumption: Food-Aesthetic'. However, the subsets differ in their dominant thematic categories. In the SDS, the most frequent category is 'Pineapple Production', whereas in the EDS, 'Consumption: Food-Aesthetic' is predominant. This distinction suggests that English-language texts circulating on the Internet focus more on promoting pineapple consumption, including products such as Pinkglow. Conversely, Spanish-language texts more frequently address topics related to pineapple production.

This difference can be attributed to the fact that pineapples are produced in several Latin American countries (e.g., Costa Rica, Ecuador, Cuba, Nicaragua) and represent an important export product, particularly in Costa Rica. Consequently, Spanish-language texts prioritize

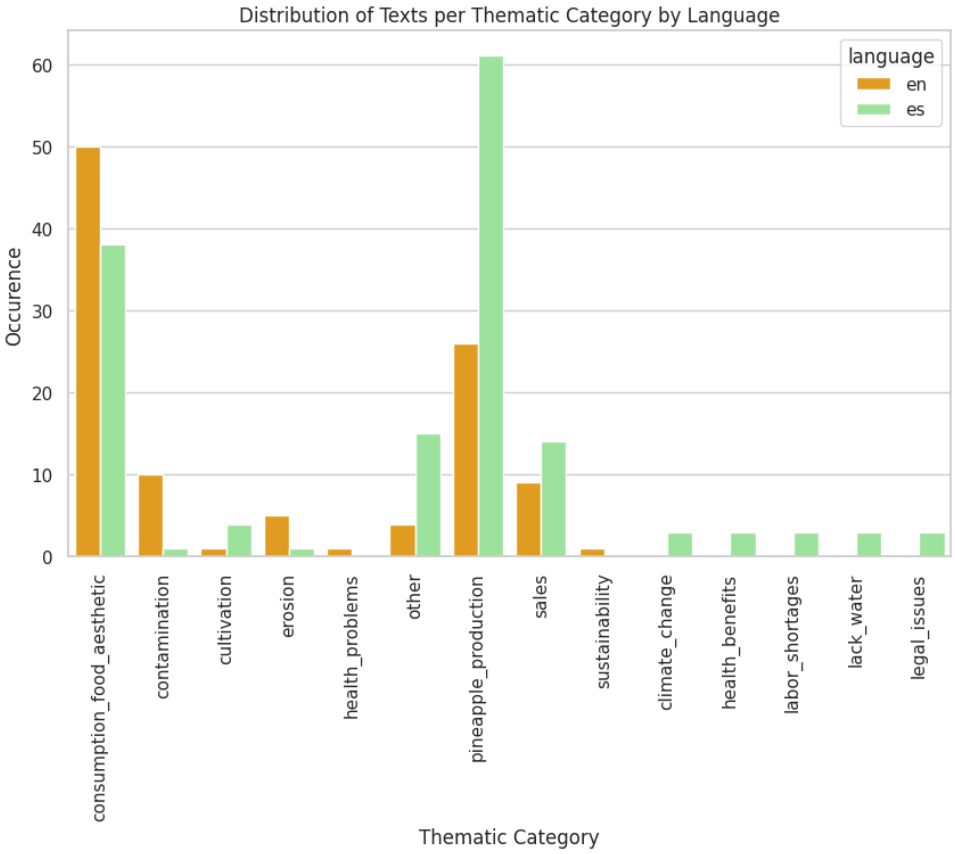

Figure 6: Distribution of Texts per Thematic Category by Language.

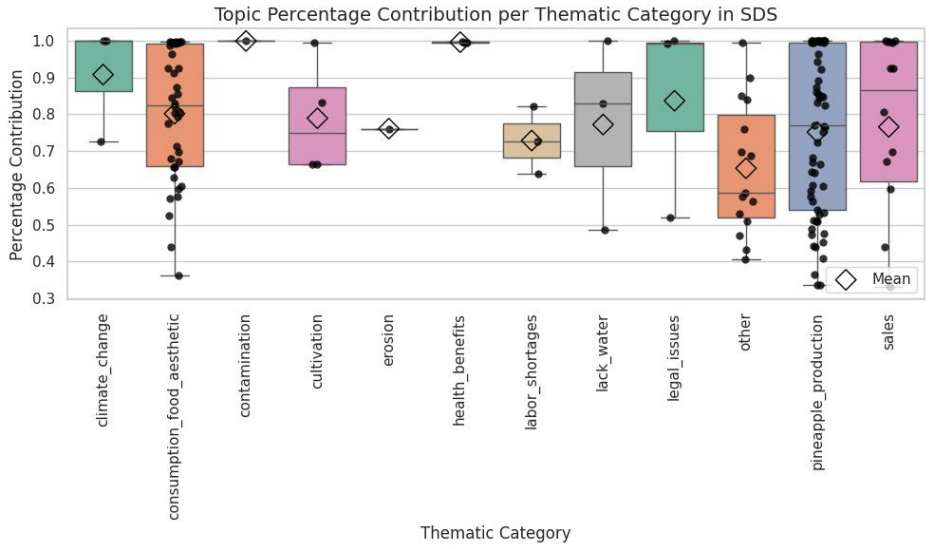

Figure 7: Topic Percentage Contribution per Thematic Category in SDS.

production-related themes over the promotion of consumption. On the other hand, English-language texts have broader international reach, as English is the primary language for global marketing communication. Therefore, it is unsurprising that 'Consumption: Food-Aesthetic' appears more

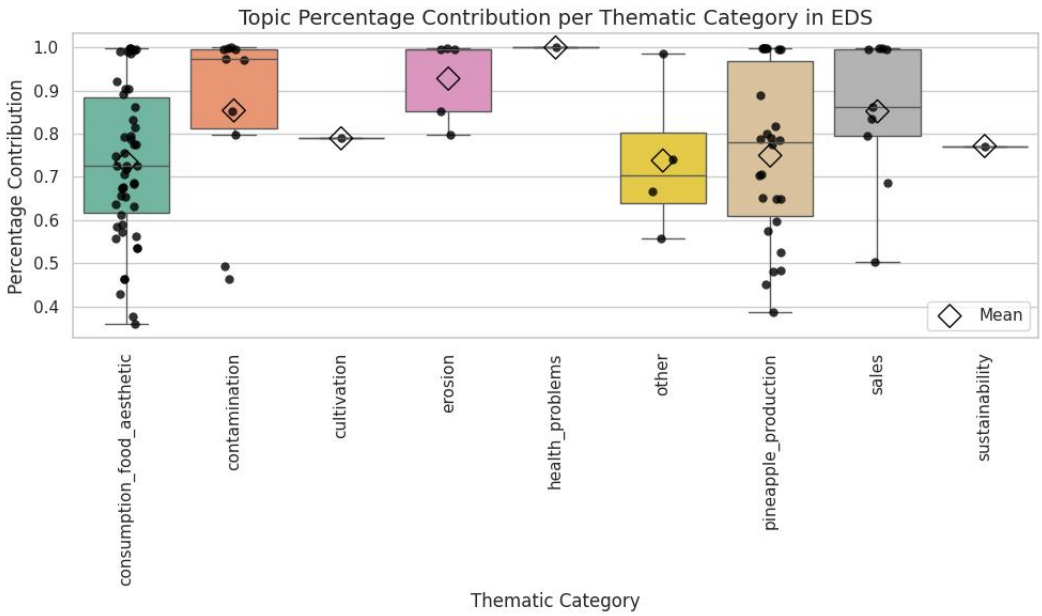

Figure 8: Topic Percentage Contribution per Thematic Category in EDS.

frequently in English-language texts.

In both the SDS and EDS, references to environmental topics are minimal. When combining categories such as 'Contamination', 'Erosion', 'Lack of Water', 'Climate Change', and 'Health Problems', these topics constitute only 9% in the SDS and 13% in the EDS. This finding indicates that most online texts in Spanish and English fail to address the serious environmental issues caused by pineapple monoculture in Latin American countries, particularly in Costa Rica. Instead, the Internet perpetuates an image of pineapples as an appealing fruit for consumption, obscuring the environmental consequences of their production. Such an image may not only bolster a positive perception of pineapples but also introduce biases for researchers working with Internet-based datasets on this topic.

It is important to note that computational models perform more accurately with English texts than with Spanish texts. This discrepancy may influence the results, which can only be validated through human review. While such a process is slow and labor-intensive, it is essential for improving computational models and the research relying on them.

The thematic categories were developed based on texts addressing pineapple production and marketing in Costa Rica. The classification of dominant topics within each category was derived from the ten most frequent words [2] associated with each topic. However, the analysis did not involve a thorough review of the content of individual texts to confirm the alignment of the dominant topics with the thematic categories. Consequently, there was no human verification to ensure that: (1) the dominant topics were consistent with the main themes or frequent words in the texts, and (2) the primary topics in the texts aligned with the thematic categories used in the research. To address these limitations, future studies should incorporate human verification steps to confirm the relationship between topics, keywords, and the content of the texts.

Future research could also focus on improving the methodology in several ways. First, the corpus should be expanded, as the number of texts collected for this study was relatively small compared to the volume of texts available online. Alternative methods for collecting online texts should be explored. Second, the research scope could be broadened to include other agricultural products from Costa Rica, such as bananas, cocoa, coffee,

[2] These words do not include fillers or stop words.

and palm. This expanded scope would increase the corpus size and help determine whether the low prevalence of environmental topics is consistent across other crops.

Third, new categories could be added for corpus classification, such as main topics, perspectives and source ideologies. Depending on the size of the corpus, a dataset could be created to train computational models to classify texts according to these new categories. This approach would provide greater clarity regarding the relationship between topics, content, perspectives, and sources. Finally, sentiment analysis could be integrated to assess the sentiment associated with different topics and thematic categories. This would allow researchers to identify the probable sentiment of various themes and assign it to the thematic categories more systematically.

Lastly, it should be noted that due to space constraints, this paper does not present the results for the most frequent words or the relationship between textual genres and thematic categories.

5 Conclusion

Following this exploratory study, we conclude that combining the application of Hierarchical Dirichlet Processes (HDP) with the human construction of thematic categories could effectively identify the main content patterns in texts across different languages. The findings suggest that significant environmental and labor rights issues associated with pineapple production are rarely disseminated on the Internet. This lack of coverage hinders the visibility of the environmental and health problems caused by pineapple monocultures and pesticide use in Costa Rica and other producing countries. Furthermore, this scarcity of information contributes to a lack of awareness among global consumers, who continue to purchase pineapples and represent a potential market for new laboratory-developed variants created by corporations.

Despite the insights gained, it is important to acknowledge the limitations of this exploratory study when interpreting the results and considering directions for future research. First, the corpus obtained through web scraping was relatively small, limiting the generalization of the findings. Second, we have not yet conducted a human evaluation of the assignment of dominant topics to individual texts, which would provide a more robust verification of topic classification within the thematic categories.

References

Ahmed Al-Rawi, Oumar Kane, and Aimé-Jules Bizimana. 2021. Topic modelling of public Twitter discourses, part bot, part active human user, on climate change and global warming. *Journal of Environmental Media*, 2(1):31–53.

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Eva Carazo and Javiera Aravena. 2016. *Condiciones de producción, impactos humanos y ambientales en el sector piña en Costa Rica*. Asociación Regional Centroamericana para el Agua y el Ambiente, San Jose, Costa Rica.

José M. de Cózar. 2019. *El Antropoceno*. Catarata, Madrid, Spain.

Biraj Dahal, Sathish Alampalayam Kumar, and Zhenlong Li. 2019. Topic Modeling and Sentiment Analysis of Global Climate Change Tweets. *Social Network Analysis and Mining*, 9(1):1–20.

Del Monte. 2020. Pinkglow. pineapple. `https://www.pinkglowpineapple.com/`. Last accessed on 2024-12-12.

Waqas Ejaz, Muhammad Ittefaq, and Sadia Jamil. 2022. Politics Triumphs: A Topic Modeling Approach for Analyzing News Media Coverage of Climate Change in Pakistan. *Journal of Science Communication*, 22:1–18.

FAO. 2024. *Principales Frutas Tropicales. Análisis del mercado. Resultados preliminares 2023*. FAO, Roma, Italy.

Francesca Grasso, Stefano Locci, Giovanni Siragusa, and Luigi Di Caro. 2024. EcoVerse: An annotated Twitter dataset for eco-relevance classification, environmental impact analysis, and stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5461–5472, Torino, Italia. ELRA and ICCL.

Ye Jiang, Xingyi Song, Jackie Harrison, Shaun Quegan, and Diana Maynard. 2017. Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation. In *Proceedings of the 2017 EMNLP Work-*

shop: Natural Language Processing meets Journalism, pages 25–30, Copenhagen, Denmark. Association for Computational Linguistics.

Joohee Kim and Yoomi Kim. 2024. Using Structural Topic Modeling to Explore the Climate Change Discourse about the Paris Agreement on Social Media. *Telematics and Informatics Reports*, 15:1–13.

Taeyong Kim, Hyemin Park, Junyong Heo, and Minjune Yang. 2021. Topic Model Analysis of Research Themes and Trends in the Journal of Economic and Environmental Geology. *Journal of Economic and Environmental Geology*, 54(3):353–364.

Ajay Krishnan and V. S. Anoop. 2023. ClimateNLP: Analyzing Public Sentiment Towards Climate Change Using Natural Language Processing.

Thomas K. Landauer and Susan Dumais. 2008. Latent Semantic Analysis. *Scholarpedia*, 3.

Andrés León and Valeria Montoya. 2021. La función de la frontera en la economía política de las plantaciones piñeras en Costa Rica. *Trace*, 80:116–137.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting Stance in Media On Global Warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Tanwi Mallick, John Murphy, Joshua David Bergerson, Duane R. Verner, John K Hutchison, and Leslie-Anne Levy. 2024. Analyzing Regional Impacts of Climate Change using Natural Language Processing Techniques.

Lucy McAllister, Siddharth Vedula, Wenxi Pu, and Maxwell Boykoff. 2024. Vulnerable Voices: Using Topic Modeling to Analyze Newspaper Coverage of Climate Change in 26 Non-Annex I Countries (2010–2020). *Environmental Research Letters*, 19(2):1–14.

Zhewei Mi and Hongwei Zhan. 2023. Text Mining Attitudes towards Climate Change: Emotion and Sentiment Analysis of the Twitter Corpus. *Weather, Climate, and Society*, 15(2):277–287.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Costanza Navarretta and Dorte H. Hansen. 2023. According to BERTopic, what do Danish Parties Debate on when they Address Energy and Environment? In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 59–68, Ingolstadt, Germany. Association for Computational Linguistics.

Alexa Obando. 2020. Acciones y omisiones del Estado costarricense en la expansión piñera: el caso de la Zona Norte-Norte de Costa Rica. *Anuario del Centro de Investigación y Estudios Políticos,*, 11:22–55.

Florian Rabitz, Audronė Telešienė, and Eimantė Zolubienė. 2021. Topic Modelling the News Media Representation of Climate Change. *Environmental Sociology*, 7(3):214–224.

Riviera. 2024. Pinkglow pineapple – everything you need to know about this new summer sensation. `https://www.rivieraproduce.com/pinkglow-pineapple-everything-you-need-to-know-about-this-new-summer-sensation/`. Last accessed on 2024-12-12.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Tania Rodríguez and Delphine Prunier. 2020. Extractivismo agrícola, frontera y fuerza de trabajo migrante: La expansión del monocultivo de piña en Costa Rica. *Frontera norte*, 32.

Moisés Salgado and Marylaura Acuña. 2021. Trabajo asalariado en el monocultivo de piña en la Región Huetar Norte. *Revista Reflexiones. Dossier especialX Jornadas de Investigación*, pages 1–17.

Robin Schaefer and Manfred Stede. 2022. GerCCT: An Annotated Corpus for Mining Arguments in German Tweets on Climate Change. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6121–6130, Marseille, France. European Language Resources Association.

Nabila M. Sham and Azlinah Mohamed. 2022. Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches. *Sustainability*, 14(8).

Manfred Stede, Yannic Bracke, Luka Borec, Neele Charlotte Kinkel, and Maria Skeppstedt. 2023. Framing Climate Change in Nature and Science Editorials: Applications of Supervised and Unsupervised Text Categorization. *Journal of Computational Social Science*, 6:485–513.

Manfred Stede and Ronny Patz. 2021. The Climate Change Debate and Natural Language Processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online. Association for Computational Linguistics.

Samson E. Uthirapathy and Sandanam Dominic. 2023. Topic Modelling and Opinion Analysis on Climate Change Twitter Data USing LDA and BERT Model. *Procedia Computer Science*, 218:908–917.

Bernal Valverde and Lilliana Chaves. 2020. The Banning of Bromacil in Costa Rica. *Weed Science*, 68(3):240–245.

Francesco S. Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2021. Climatext: A Dataset for Climate Change Topic Detection.

Marcelo Werneck and André Gomes. 2023. The Interface between Research Funding and Environmental Policies in an Emergent Economy using Neural Topic Modeling: Proposals for a Research Agenda. *Review of Policy Research*, pages 1–25.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

A Appendix. Description of Textual Genres.

Gender Categories	**Description**
News article	Texts produced by the media on various current topics.
Divulgative note	Texts that present scientific topics to the general public.
Narrative note	Texts in which the author describes an experience (e.g., visiting a restaurant or place, tasting food or drink, or using an object) without a critical perspective.
Blog post	Texts published on blogs.
Website description	Texts that describe or briefly summarize the content of a webpage.
Product description	Texts that describe or introduce products.
Scientific article	Texts that present research results and are published in academic journals or books.
Opinion article	Texts that express individuals' opinions and perspectives on various topics.
Report	Texts that present the results of a professional analysis or research conducted for companies or organizations.
Manual	Texts that explain how to use or apply an object, methodology, or theory, or how a person should act.
Wikipedia	Texts that explain various topics within the Wikipedia platform.
Academic article	Texts that present scientific topics to an academic audience but are not published on web pages, in academic journals, or in books.
Paper abstract	Texts that summarize the contents of a scientific article.

Table 3: Typology of Textual Genres in Dataset (own creation).