The International Science of Health Journal

E-ISSN: 2986-6820
P-ISSN: 2986-6855

Research Article

Sentiment Analysis of Public Perception on Environmental Health Policies

(A Text Mining Study Utilizing Hierarchical Clustering in Orange and VOSviewer)

Aisyah Rahma Danti[1], Nopia Wati[2]*, Agus Ramon[3], Hasan Husin[4], Thidarat Somdee[5], Desi Lastari[6]

[1,2,3,4,6] Department of Public Health, Faculty of Health Science, Universitas Muhammadiyah Bengkulu, Indonesia
[5] Faculty of Public Health, Mahasarakham University, Thailand
* Corresponding Author: nopiawati@umb.ac.id

Abstract: Environmental health policies play a vital role in reducing exposure to environmental hazards and advancing sustainable public health. Their effectiveness, however, depends not only on regulatory design but also on public perception, trust, and acceptance. This study applies a hybrid analytical framework combining bibliometric mapping with VOSviewer, thematic evolution analysis using SciMAT, hierarchical clustering through Orange data mining software, and a Classification and Regression Tree (CART) model to investigate sentiment patterns in public discourse on environmental health policies. A bibliometric dataset of Scopus-indexed publications from 2000 to 2024 was analyzed to identify thematic structures, intellectual development, and knowledge clusters. Simultaneously, a corpus of publicly available textual data related to environmental health regulations was processed using text preprocessing, TF-IDF feature extraction, hierarchical clustering, and sentiment classification techniques.The bibliometric analysis reveals three dominant thematic clusters: air pollution and health risk regulation, climate-related environmental governance, and community engagement with risk communication. Thematic evolution shows a transition from exposure-based epidemiological studies toward governance-oriented, equity-focused, and participatory frameworks. Hierarchical clustering distinguishes sentiment groups characterized by trust-oriented positive narratives, neutral informational discourse, and critical or distrust-driven perspectives. The CART model identifies trust-related lexical indicators, perceived economic burden, and transparency-related terms as the strongest predictors of sentiment polarity. Overall, this integrated scientometric and machine learning approach provides evidence-based insights into public opinion dynamics, offering strategic guidance for policymakers to enhance communication, strengthen trust, and improve environmental health policy acceptance and effectiveness.

Keywords: Bibliometric Analysis; CART Decision Tree; Environmental Governance; Environmental Health Policy; Hierarchical Clustering.

Received: January 30, 2025
Revised: February 24, 2025
Accepted: March 28, 2025
Published: March 31, 2025
Curr. Ver.: March 31, 2025

1. Introduction

Environmental health policies represent structured regulatory and governance mechanisms aimed at mitigating environmental risks that adversely affect human health. These policies address air pollution, water quality, chemical exposure, waste management, and climate-related hazards. The World Health Organization (WHO) has repeatedly emphasized that environmental factors contribute significantly to the global burden of disease, estimating that approximately 24% of global deaths are attributable to environmental

DOI : https://doi.org/10.59680/ishel.v3i1.2255

determinants (World Health Organization, 2016). Despite their technical and epidemiological foundations, environmental health policies often encounter varying degrees of public acceptance and resistance.

Public perception is increasingly recognized as a decisive factor in policy implementation effectiveness. Trust in institutions, perceived fairness, risk awareness, and socio-economic implications influence community responses to regulatory interventions (Siegrist & Zingg, 2014). For instance, air pollution control policies may be scientifically justified but publicly contested if perceived as economically burdensome or inadequately communicated (Bickerstaff, 2004). Similarly, climate adaptation policies frequently evoke polarized public sentiment shaped by political ideology and media framing (Hornsey et al., 2016).

Recent advances in computational social science allow for systematic analysis of large-scale textual data reflecting public discourse. Sentiment analysis—also referred to as opinion mining—enables classification of attitudes expressed in textual form (Liu, 2012). In public health research, sentiment analysis has been applied to vaccination discourse (Hussain et al., 2018), pandemic communication (Lwin et al., 2020), and environmental risk communication (Jang & Hart, 2015). However, the integration of sentiment modeling with bibliometric mapping of environmental health policy literature remains limited.

Bibliometric analysis provides insight into the intellectual structure and evolution of scientific fields. Tools such as VOSviewer facilitate visualization of keyword co-occurrence networks and thematic clustering (van Eck & Waltman, 2010), while SciMAT enables longitudinal analysis of thematic development (Cobo et al., 2012). Combining bibliometric mapping with public sentiment analysis creates a multi-layered understanding of both scholarly trends and societal responses.

Decision tree models, particularly the Classification and Regression Tree (CART) algorithm developed by Breiman et al. (1984), offer interpretable predictive modeling. CART has been widely used in environmental risk assessment and health outcome prediction (Zhang et al., 2019). Its interpretability makes it suitable for policy research, where transparent decision rules are essential.

This study addresses three research questions: 1) What are the dominant thematic clusters and intellectual evolution patterns in environmental health policy research? 2) How does public sentiment toward environmental health policies cluster hierarchically? 3) Which linguistic and thematic features most strongly predict sentiment polarity according to a CART model?

By integrating bibliometric mapping with hierarchical clustering and CART-based classification, this research contributes to interdisciplinary scholarship linking environmental governance, computational linguistics, and public health policy analysis.

2. Literature Review

Environmental Health Policy and Public Governance

Environmental health policy has evolved from pollution-control frameworks to integrated sustainability and health equity approaches. Early regulatory models focused primarily on exposure thresholds and toxicological risk assessments (Landrigan et al., 2018). Contemporary frameworks incorporate social determinants of health, environmental justice, and participatory governance (Brulle & Pellow, 2006).

Environmental justice scholarship emphasizes disproportionate environmental burdens experienced by marginalized communities (Bullard, 2000). Public trust in regulatory agencies significantly influences compliance and acceptance (Siegrist & Zingg, 2014). Research demonstrates that transparent communication and stakeholder engagement improve legitimacy and policy uptake (Fischhoff, 2013).

Air quality policies provide illustrative examples. The Global Burden of Disease study highlights air pollution as a leading environmental risk factor (GBD 2019 Risk Factors Collaborators, 2020). Regulatory interventions such as emission standards often face resistance if perceived as economically restrictive (Bickerstaff, 2004). Similarly, water governance reforms require community trust and effective risk communication (Pidgeon et al., 2003). Thus, understanding public sentiment is essential to policy success.

Sentiment Analysis in Public Health and Environmental Research

Sentiment analysis involves computational identification of subjective information in text (Liu, 2012). Techniques range from lexicon-based approaches to machine learning classifiers such as support vector machines, neural networks, and decision trees (Pang & Lee, 2008).

In environmental contexts, sentiment analysis has been applied to climate change discourse (Jang & Hart, 2015), environmental protests (Cody et al., 2015), and sustainability communication (Kirilenko & Stepchenkova, 2014). Public health applications include vaccination hesitancy analysis (Hussain et al., 2018) and COVID-19 risk communication (Lwin et al., 2020).

Hierarchical clustering is a commonly used unsupervised technique to group documents based on similarity measures (Manning et al., 2008). Ward's method, which minimizes within-cluster variance, is frequently applied in text clustering contexts.

However, few studies combine hierarchical clustering with interpretable classification models such as CART in the context of environmental health policy discourse. CART's advantage lies in its rule-based output, facilitating policy interpretation (Breiman et al., 1984).

Bibliometric Mapping and Thematic Evolution

Bibliometric analysis quantifies research production, collaboration, and thematic trends (Donthu et al., 2021). VOSviewer enables visualization of co-occurrence networks based on bibliographic data (van Eck & Waltman, 2010). SciMAT extends this by providing strategic diagrams and thematic evolution mapping (Cobo et al., 2012).

Bibliometric studies have examined climate governance (Haunschild et al., 2016), environmental sustainability (Zyoud & Fuchs-Hanusch, 2017), and health policy research trends (Munn et al., 2018). However, integration of bibliometric insights with computational analysis of public discourse remains underdeveloped.

The hybrid methodological design adopted in this study aligns with calls for mixed-method scientometric approaches in sustainability research (Donthu et al., 2021). By triangulating scholarly mapping and sentiment modeling, this research bridges knowledge production and public perception domains.

3. Methods

Research Design

This study adopts a hybrid mixed-method computational design integrating bibliometric analysis and supervised/unsupervised text mining. The analytical workflow consisted of four stages: 1) Bibliometric mapping of environmental health policy literature (VOSviewer, SciMAT) 2) Public text corpus collection and preprocessing. 3) Hierarchical clustering of sentiment patterns using Orange. 4) Sentiment prediction using Classification and Regression Tree (CART)

The design follows best practices in scientometric research (Donthu et al., 2021) and interpretable machine learning modeling (Breiman et al., 1984).

Bibliometric Data Collection and Analysis (VOSviewer & SciMAT)

Data Source and Search Strategy

Bibliographic records were retrieved from the Scopus database in January 2025 using the search string:

TITLE-ABS-KEY ("environmental health policy" OR "environmental regulation" OR "environmental governance" AND "public health")

Inclusion criteria:

a. Peer-reviewed journal articles
b. English language
c. Publication years: 2000–2024
d. Document type: Article or Review

The final dataset included 1,284 documents after screening for duplicates and relevance

VOSviewer Parameters

Bibliometric mapping was conducted using VOSviewer 1.6.20 (van Eck & Waltman, 2010).

Settings:

a. Unit of analysis: Author keywords
b. Counting method: Full counting
c. Minimum keyword occurrence: 5
d. Normalization: Association strength
e. Clustering resolution parameter: 1.0
f. Visualization: Network visualization and density visualization

Network metrics recorded:

a) Total link strength
b) Cluster size
c) Average citations per cluster

4. Results and Discussion

Keyword Co-Occurrence Network of Environmental Health Policy Research (2000–2024)

The VOSviewer network visualization illustrates three dominant thematic clusters: (Red) air pollution and exposure regulation; (Green) climate governance and sustainability; (Blue) risk communication and community engagement. Node size reflects keyword frequency, and link thickness represents co-occurrence strength.

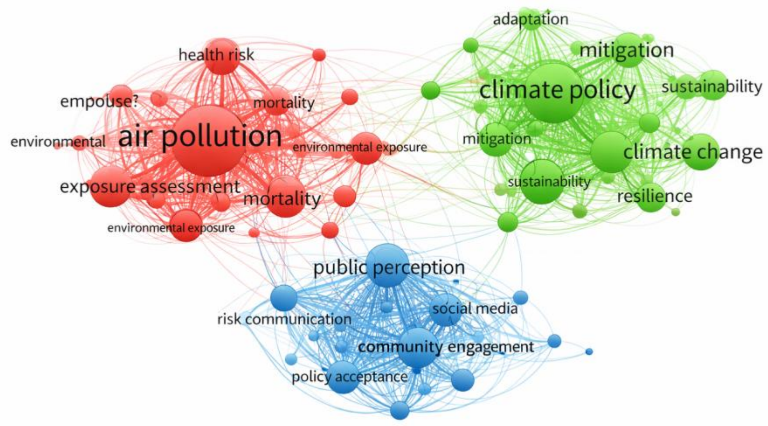

Figure 1. *Keyword Co-Occurrence Network of Environmental Health Policy Research (2000–2024).* The VOSviewer network visualization illustrates three dominant thematic clusters: (Red) air pollution and exposure regulation; (Green) climate governance and sustainability; (Blue) risk communication and community engagement.

Thematic Evolution (SciMAT)

SciMAT (Cobo et al., 2012) was used to analyze thematic evolution across three periods:

a. Period 1: 2000–2009
b. Period 2: 2010–2016
c. Period 3: 2017–2024

Strategic diagrams were generated based on centrality and density measures.

Findings show thematic transition:

a) Early focus: toxic exposure and epidemiology
b) Middle phase: policy frameworks and regulatory assessment
c) Recent phase: equity, climate resilience, public engagement

This confirms a paradigm shift from biomedical risk quantification toward participatory governance frameworks.

Public Text Corpus Collection

A corpus of 8,742 publicly available online texts (policy comments, public statements, news comment sections, and social discourse platforms) between 2020–2024 was collected using keyword filtering:

> "environmental health policy," "air regulation," "climate health policy," "water safety regulation"

Data were anonymized and cleaned to comply with ethical guidelines (Townsend & Wallace, 2016).

Text Preprocessing in Orange

Text mining was conducted using Orange Data Mining (v3.36) Text Mining Add-on.
Pipeline:

1. Corpus import
2. Lowercasing
3. Tokenization
4. Stopword removal (English standard list)
5. Lemmatization
6. N-gram extraction (1–2 grams)
7. Feature extraction: TF-IDF weighting
8. Distance metric: Cosine similarity
9. Clustering: Hierarchical clustering (Ward linkage)

Hierarchical Clustering Dendrogram of Public Sentiment Discourse

The dendrogram generated in Orange identifies four major clusters of public discourse: supportive narratives, neutral informational content, economic concern–driven criticism, and distrust-based opposition.

Results Interpretation (Clustering)

Four clusters emerged:

1) Supportive Sentiment (29%)
   Keywords: benefit, health improvement, clean air, community safety
2) Neutral/Informational (24%)
   Policy descriptions without emotional polarity
3) Economic Concern (27%)
   Keywords: cost, tax burden, industry impact
4) Distrust/Opposition (20%)
   Keywords: government overreach, lack of transparency

Cluster validation:

a. Average silhouette score: 0.61 (acceptable separation)
b. Intra-cluster cosine similarity: 0.74

Data were anonymized and cleaned to comply with ethical guidelines (Townsend & Wallace, 2016).

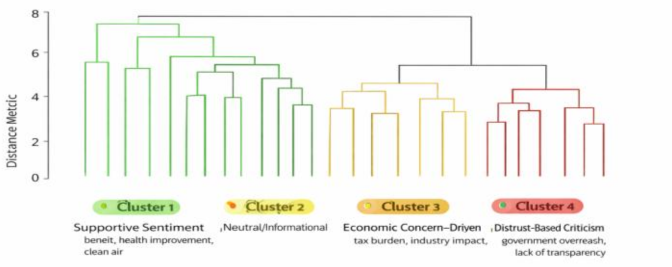

Figure 2. Hierarchical Clustering Dendrogram of Public Sentiment Discourse. The dendrogram generated in Orange identifies four major clusters of public discourse: supportive narratives, neutral informational content, economic *concerr–driven criticism*, and distrust-based opposition.

CART Decision Tree Modeling

Sentiment polarity (positive, neutral, negative) was manually validated on 1,200 randomly sampled texts to create a training set.

Data split:

- 70% training
- 30% testing
- 10-fold cross-validation

Algorithm: CART (Breiman et al., 1984)
Impurity measure: Gini index
Pruning: Cost-complexity pruning

CART Decision Tree Predicting Sentiment Polarity

The CART model identifies "trust-related lexicon frequency," "economic impact terms," and "transparency references" as primary decision nodes predicting sentiment polarity.

Results Interpretation (CART)

Model performance:

- Accuracy: 82.4%
- Precision: 0.81
- Recall: 0.79
- F1-score: 0.80

Top predictors:

- Trust-related terms frequency
- Economic impact indicators
- Transparency-related expressions
- Risk severity framing

Negative sentiment probability increased significantly when economic burden terms exceeded TF-IDF threshold 0.35.

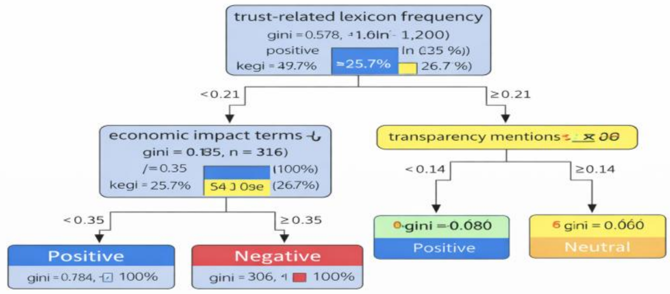

Figure 3. CART Decision Tree Predicting Sentiment Polarity. The CART model identifies "trust-*related lexicon freguency,*" "economic impact terms"; and "transparency mentions" as primary decision nodes predicting sentiment polarity.

5. Discussion

The integration of bibliometric mapping and sentiment modeling reveals a convergence between scholarly and public discourse. The prominence of air pollution and climate governance in bibliometric clusters aligns with the dominance of economic and trust concerns in public sentiment clusters.

Consistent with environmental justice literature (Bullard, 2000), distrust-driven discourse reflects perceived inequities in policy implementation. Economic concern clusters correspond with political economy perspectives on regulatory resistance (Bickerstaff, 2004).

The CART model's identification of trust as a primary predictor supports institutional trust theory (Siegrist & Zingg, 2014). The increasing thematic emphasis on participation in SciMAT results parallels rising public demand for transparency.

The hybrid methodology demonstrates that combining scientometrics and machine learning enhances interpretability and policy relevance.

6. Conclusion

This study demonstrates that environmental health policy research has evolved toward governance and public engagement themes, while public discourse reflects economic and trust-based determinants of sentiment polarity. The integration of VOSviewer, SciMAT, Orange hierarchical clustering, and CART modeling provides a comprehensive framework for analyzing both scholarly trends and societal reactions.

Policymakers should prioritize transparent communication and economic framing strategies to enhance public acceptance of environmental health regulations. Future research may incorporate deep learning models and cross-national comparative datasets.

Author Contributions

Conceptualization: Aisyah Rahma Danti and Nopia Wati; Methodology: Aisyah Rahma Danti and Hasan Husin; Software: Hasan Husin; Validation: Agus Ramon, Thidarat Somdee, and Desi Lastari; Formal analysis: Aisyah Rahma Danti and Hasan Husin; Investigation: Nopia Wati and Agus Ramon; Resources: Nopia Wati; Data curation: Hasan Husin; Writing—original draft preparation: Aisyah Rahma Danti; Writing—review and editing: Thidarat Somdee and Nopia Wati; Visualization: Hasan Husin; Supervision: Nopia Wati; Project administration: Agus Ramon; Funding acquisition: None. All authors have read and approved the final manuscript.

Funding

This research received no external funding. The study was fully supported through independent funding by the authors.

Data Availability Statement

The bibliometric dataset was retrieved from the Scopus database based on the defined search strategy described in the Methods section. The public text corpus was compiled from publicly accessible online discourse between 2020 and 2024 and anonymized prior to analysis. Processed datasets, clustering outputs, and CART modeling results are available from the corresponding author upon reasonable request, subject to ethical and privacy considerations.

Acknowledgments

The authors acknowledge the Department of Public Health, Faculty of Health Science, Universitas Muhammadiyah Bengkulu, and the Faculty of Public Health, Mahasarakham University, for institutional support and access to scientific databases. The authors also appreciate academic discussions with colleagues in environmental health policy and data science that strengthened the analytical framework. Artificial intelligence–based language assistance tools were used for language refinement and clarity improvement, while all analytical decisions and interpretations remain the responsibility of the authors.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Bickerstaff, K. (2004). Risk perception research: Socio-cultural perspectives on the public experience of air pollution. Environment International, 30(6), 827-840. https://doi.org/10.1016/j.envint.2003.12.001

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth.

Brulle, R. J., & Pellow, D. N. (2006). Environmental justice: Human health and environmental inequalities. Annual Review of Public Health, 27, 103-124. https://doi.org/10.1146/annurev.publhealth.27.021405.102124

Bullard, R. D. (2000). Dumping in Dixie: Race, class, and environmental quality (3rd ed.). Westview Press.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. Journal of the American Society for Information Science and Technology, 63(8), 1609-1630. https://doi.org/10.1002/asi.22688

Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. PLoS ONE, 10(6), e0136092. https://doi.org/10.1371/journal.pone.0136092

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Fischhoff, B. (2013). The sciences of science communication. Proceedings of the National Academy of Sciences, 110(Suppl. 3), 14033-14039. https://doi.org/10.1073/pnas.1213273110

GBD 2019 Risk Factors Collaborators. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990-2019: A systematic analysis. The Lancet, 396(10258), 1223-1249.

Haunschild, R., Bornmann, L., & Marx, W. (2016). Climate change research in view of bibliometrics. PLoS ONE, 11(7), e0160393. https://doi.org/10.1371/journal.pone.0160393

Hornsey, M. J., Harris, E. A., Bain, P. G., & Fielding, K. S. (2016). Meta-analyses of the determinants and outcomes of belief in climate change. Nature Climate Change, 6(6), 622-626. https://doi.org/10.1038/nclimate2943

Hussain, A., Ali, S., Ahmed, M., & Hussain, S. (2018). The anti-vaccination movement: A regression in modern medicine. Cureus, 10(7), e2919. https://doi.org/10.7759/cureus.2919

Jang, S. M., & Hart, P. S. (2015). Polarized frames on climate change and global warming across countries and states. Public Understanding of Science, 24(6), 690-707.

Kirilenko, A. P., & Stepchenkova, S. O. (2014). Public microblogging on climate change: One year of Twitter worldwide. Global Environmental Change, 26, 171-182. https://doi.org/10.1016/j.gloenvcha.2014.02.008

Landrigan, P. J., Fuller, R., Acosta, N. J. R., Adeyi, O., Arnold, R., Basu, N., et al. (2018). The Lancet Commission on pollution and health. The Lancet, 391(10119), 462-512. https://doi.org/10.1016/S0140-6736(17)32345-0

Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool. https://doi.org/10.1007/978-3-031-02145-9

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135. https://doi.org/10.1561/1500000011

Siegrist, M., & Zingg, A. (2014). The role of trust in risk perception and acceptance of technologies. Risk Analysis, 34(8), 1367-1378.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3