Ecological Discourse Modeling in a Low-Resource Setting: A Longitudinal Vietnamese Climate Corpus with Comparative Topic Modeling

Phuong Huyen NGUYEN
Toulouse School of Economics
Toulouse, France
phuonghuyen.ng2710@gmail.com

Abstract

Climate change discourse has expanded substantially in recent decades, yet computational analyses remain concentrated on high-resource languages. In this paper, we construct a longitudinal Vietnamese climate news corpus and examine thematic structure and temporal evolution in a low-resource setting. The corpus comprises 10,401 articles published between 2004 and 2026 and is systematically preprocessed using linguistically informed word segmentation. To ensure domestic relevance, we apply transformer-based Named Entity Recognition and construct a geographically grounded subset of 4,501 Vietnam-focused documents. We analyze this dataset using both Latent Dirichlet Allocation and BERTopic. Results reveal stable thematic dimensions alongside longitudinal shifts from event-driven pollution reporting toward governance- and energy-centered narratives. Embedding-based modeling achieves higher semantic coherence while maintaining comparable topic diversity. The main contribution of this work is thus the compilation of a structured Vietnamese climate corpus and a systematic analysis of discourse evolution in an underrepresented language context.

Keywords: Vietnamese corpus, ecological NLP, topic modeling, environmental discourse

1. Introduction

Climate change is widely recognized as one of the defining challenges of the twenty-first century. While its drivers operate globally, its impacts are unevenly distributed, shaping local vulnerabilities, policy priorities, and public discourse. Countries such as Vietnam face increasing exposure to sea-level rise, coastal erosion, extreme weather events, and air pollution, posing risks to socio-economic development and environmental sustainability. In this context, understanding how climate issues are represented in national media is critical for analyzing thematic prioritization and environmental governance. Media discourse not only reflects environmental conditions but also shapes public risk perception, influences policy agendas, and structures institutional accountability, with implications for broader policy formation and societal response.

Over the past two decades, climate-related reporting in Vietnam has expanded substantially in both volume and scope. However, the scale of this growing body of text renders manual analysis infeasible. Natural Language Processing (NLP) provides scalable approaches for examining large corpora, with topic modeling enabling the extraction of latent thematic structures without manual annotation, making it well suited for analyzing long-term discourse dynamics.

Despite these advances, large-scale computational studies of climate discourse remain concentrated in high-resource languages, particularly English. This limits our understanding of environmental narratives in low-resource contexts, where linguistic, cultural, and institutional factors may differ. Vietnamese presents additional challenges, including multi-syllabic word segmentation, lexical ambiguity, and limited domain-specific resources, complicating preprocessing and downstream evaluation. Standard topic coherence metrics, designed for high-resource languages, may therefore behave unreliably. Consequently, systematic longitudinal analyses of Vietnamese climate media discourse remain scarce.

This study addresses the following research question: *How is climate change thematically structured in Vietnamese national media, and how robust are these structures across different topic modeling paradigms?* In addition to identifying thematic patterns, we examine how methodological choices affect interpretability and evaluation in a low-resource setting, with particular attention to commonly used coherence metrics.

To answer this question, we construct and analyze a longitudinal Vietnamese climate news corpus spanning 2004–2026. We adopt a complementary modeling strategy combining a probabilistic generative approach (LDA) with an embedding-based method (BERTopic) to capture both frequency-based and semantic representations of topics. To ensure domestic relevance, we apply transformer-based Named Entity Recognition to retain articles referring to Vietnam and its provinces. We further evaluate topic coherence metrics, highlighting their limitations and potential misalignment with semantic structure in Vietnamese text, and examine the

Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology) @ LREC 2026, pages 69–78
May 12, 2026 ©European Language Resources Association (ELRA), 2026

temporal evolution of climate discourse.

This study makes three main contributions. First, it introduces a longitudinal Vietnamese climate news corpus spanning more than two decades. Second, it provides a systematic comparison of probabilistic and embedding-based topic modeling approaches in a low-resource language. Third, it critically evaluates topic coherence metrics, showing that standard measures may underestimate topic quality in Vietnamese due to their reliance on surface-level co-occurrence.

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 3 describes the corpus construction and modeling framework. Section 4 presents the empirical results and analyzes thematic structure and temporal dynamics. Finally, Section 5 concludes and outlines directions for future research.

2. Related Work

Computational analyses of climate discourse have developed along several complementary strands. One line of research applies NLP to policy documents and institutional texts to identify adaptation strategies, mitigation priorities, and sectoral trade-offs (Tashakori et al., 2025; Badekale and Akinfaderin, 2025). Another examines environmental communication in social media and public platforms, combining topic modeling and sentiment analysis to study polarization, risk perception, and narrative framing (Gokcimen and Das, 2024; Pruss et al., 2019). These studies demonstrate the value of large-scale text analysis for understanding how climate change is framed across institutional and public arenas.

Topic modeling remains a central approach in this literature. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) has been widely used to uncover thematic structures in environmental corpora (Gokcimen and Das, 2024; Tashakori et al., 2025). More recently, contextualized language representations have enabled embedding-based approaches such as BERTopic, which leverage transformer embeddings and density-based clustering to capture semantic similarity. Extensions such as dynamic and embedded topic modeling further support temporal analysis (Badekale and Akinfaderin, 2025), reflecting a shift toward embedding-driven frameworks.

Research on low-resource languages has also begun to expand. Wasi et al. (2024) introduce a Bengali climate dataset, while Haque et al. (2025) propose a graph-based hybrid topic model. These studies demonstrate feasibility but typically focus on either dataset construction or a single method, with limited comparative evaluation.

For Vietnamese, climate communication research has largely relied on qualitative approaches (Dang, 2025; Le and Vo, 2026), with limited large-scale computational analysis. To date, there has been no systematic longitudinal study comparing probabilistic topic models with embedding-based approaches in a unified framework.

More broadly, while prior work shows the effectiveness of topic modeling, less attention has been paid to the reliability of evaluation metrics in low-resource settings. Measures such as NPMI and c_v may behave differently across languages due to segmentation and lexical variability, affecting co-occurrence statistics and interpretation.

The present study addresses these gaps by constructing a longitudinal Vietnamese climate news corpus and comparing LDA and BERTopic. In addition to analyzing thematic structure and temporal evolution, it provides insights into the behavior of topic modeling methods and evaluation metrics in a low-resource setting.

3. Experimental Framework

This section outlines the computational pipeline used to build and analyze a large-scale corpus of climate-related news articles in Vietnam. The framework integrates automated data collection, linguistically informed preprocessing, geographic filtering via named entity recognition, and unsupervised topic modeling.

The design emphasizes methodological transparency and reproducibility, while enabling a critical assessment of how standard NLP pipelines behave in a low-resource language context. The overall workflow is illustrated in Figure 1, which summarizes the sequential stages of data collection, preprocessing, geographic filtering, topic modeling, evaluation, and subsequent temporal analysis of discourse dynamics. By combining probabilistic and embedding-based topic modeling approaches, the framework supports a comparative assessment of how different NLP paradigms capture thematic structure, as well as how these themes evolve over time in ecological media reporting.

3.1. Corpus Construction

The target corpus focuses on climate change and related environmental and energy issues in Vietnam. Articles are retrieved using keyword-based searches, including terms such as *"biến đổi khí hậu"* (climate change), *"ô nhiễm môi trường"* (environmental pollution), and *"năng lượng tái tạo"* (renewable energy). Data are collected from VTV, VnExpress, and Nhân Dân, three major Vietnamese news outlets, to ensure wide coverage of mainstream media discourse. The selection of sources introduces an inherent bias toward mainstream and institutional perspectives. As national-level outlets,

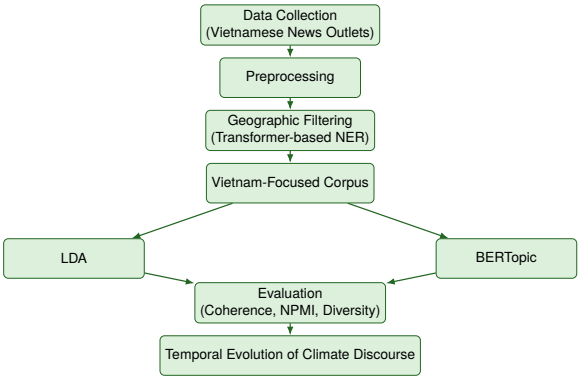

Figure 1: Computational pipeline for geographically grounded Vietnamese climate discourse modeling.

these sources are more likely to emphasize policy, official narratives, and urban issues, potentially underrepresenting local or community-level environmental concerns.

Data Retrieval All articles were automatically scraped from HTML documents using Selenium[1]. For each article, key information, including the title, URL, and publication date, was extracted and stored in a structured JSON format. Due to copyright restrictions, the complete raw text of the articles cannot be redistributed. All data were collected and used exclusively for research purposes, in compliance with the terms of service of the respective websites. To support reproducibility, we provide a curated dataset of article URLs with associated metadata, including title and publication date, publicly accessible at `https://github.com/hnnphuongg/vietnam-climate-news`.

Data Preprocessing Accurate word segmentation is a crucial preprocessing step when working with Vietnamese text. Unlike English, where whitespace reliably separates words, Vietnamese lexical units often comprise multiple syllables that are orthographically separated by spaces. As a result, naive whitespace tokenization incorrectly splits semantically unified expressions into separate tokens, introducing substantial noise in downstream NLP tasks.

These preprocessing decisions, while necessary for model stability, may also influence topic composition by filtering out infrequent but potentially meaningful terms. This trade-off is common in topic modeling pipelines, particularly in low-resource settings where vocabulary sparsity is more pronounced.

In practice, we use `vncorenlp`[2], a state-of-the-art toolkit for Vietnamese natural language processing. It provides linguistically informed word segmentation and tokenization, enabling accurate identification of multi-syllable lexical units. For instance, the expression *"khí hậu"* (climate) is correctly segmented as *"khí_hậu"*, preserving it as a single semantic token.

Additional preprocessing steps include removing non-textual elements such as URLs, redundant whitespace, and Vietnamese stopwords[3]. To improve the stability of topic estimation, we also exclude extremely rare and overly frequent terms prior to dictionary construction. Specifically, words appearing in fewer than 10 documents or in more than 70% of the corpus are removed. These thresholds are selected based on exploratory sensitivity checks to ensure that substantively meaningful terms are retained while enhancing vocabulary stability and topic distinctiveness.

Table 1 summarizes the key properties of the corpus resulting from the preprocessing step. In total, 10,401 articles are collected, spanning 23 years, from April 2004 to February 2026.

#docs	#tokens	#words
10,401	6,708,526	2,588,127

Table 1: Corpus statistics

As illustrated in Figure 2, climate-related reporting remained comparatively limited during the period from 2004 to 2014. Beginning in 2015, coverage increased more visibly, with a marked acceleration after 2017 and several peaks in the early 2020s, culminating in the highest observed level in 2025. This upward trajectory coincides with intensified international climate negotiations and the expansion of domestic energy transition initiatives.

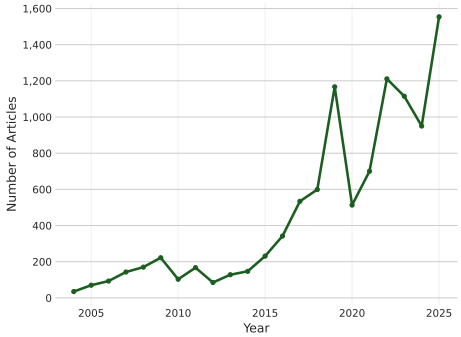

Figure 2: Annual number of climate-related articles published in VTV, VnExpress, and Nhân Dân (2004–2025).

The increase in article counts over time may partly reflect the broader expansion of digital news

[1] `https://www.selenium.dev/`

[2] https://github.com/vncorenlp/VnCoreNLP

[3] https://github.com/stopwords/vietnamese-stopwords

production, rather than a purely substantive rise in attention to the topic itself. Without normalization by the total volume of published articles, these trends should be interpreted as measures of absolute attention rather than relative salience within the overall media landscape. This distinction is important, as growth in raw counts may be driven not only by increased interest but also by the general rise in content production.

Figure 3 presents the overall lexical composition of the corpus through a word cloud representation. Prominent terms cluster around themes of climate change, energy systems, environmental management, and sustainable development. The visibility of development- and governance-related vocabulary indicates that climate issues are frequently framed in connection with infrastructure expansion, investment dynamics, and policy implementation processes.

Figure 3: Word cloud of the climate corpus

These lexical patterns suggest that climate reporting is commonly embedded within an environmental–energy–development nexus, rather than treated as an isolated ecological concern. This integrated framing aligns with the broader policy context in which environmental objectives are closely linked to economic transformation and governance agendas.

3.2. Geographic Filtering

To ensure geographic relevance with an exclusive focus on Vietnam, we employ `NlpHUST/ner-vietnamese-electra-base`[4], a pre-trained Vietnamese Named Entity Recognition (NER) system built upon the ELECTRA transformer architecture. The model identifies four entity categories: `LOCATION`, `ORGANIZATION`, `PERSON`, and `MISCELLANEOUS`.

We apply this model to extract named entities from article texts and retain only those containing `LOCATION` entities referring to Vietnam or any of its 63 former provincial-level administrative units. This filtering procedure is implemented using a predefined lexicon of Vietnamese provinces and municipalities to match the extracted location entities. Articles without domestic geographic references are excluded. While this approach ensures that the resulting corpus is strongly grounded in the Vietnamese context, it may also exclude articles discussing Vietnam implicitly without explicit location mentions, introducing a potential source of selection bias.

Following geographic filtering, the corpus comprises 4,501 articles referring to Vietnam or its provinces. As shown in Figure 4, Vietnam and Hanoi dominate the dataset, appearing in 2,338 and 1,366 articles, respectively. Ho Chi Minh City ranks third with 324 articles, while the remaining provinces are mentioned in substantially fewer documents.

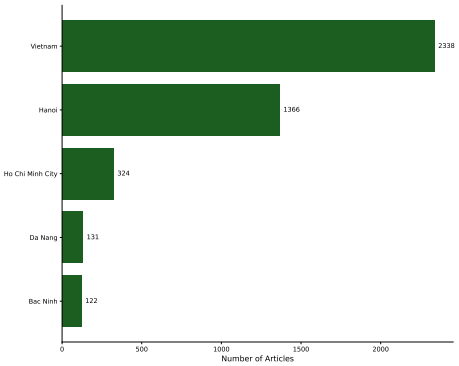

Figure 4: Top 5 most frequently mentioned Vietnamese locations in the corpus

This distribution reflects the structural characteristics of national-level media, where reporting is naturally concentrated in major political and administrative centers such as Hanoi and Ho Chi Minh City. At the same time, the relatively limited representation of smaller provinces highlights a potential imbalance in the visibility of localized climate impacts, suggesting that subnational environmental issues may receive less consistent media attention.

3.3. Topic Modeling

To investigate the thematic structure of the Vietnam-related subset of the corpus, we conduct a series of topic modeling experiments. Topic modeling is viewed as an unsupervised approach to uncovering latent thematic patterns in large text collections. By exploiting word co-occurrence regularities, such models group semantically related terms into clusters that reflect broader themes, without relying on predefined labels or annotated training data (Churchill and Singh, 2022; Gagliardi and Artese, 2020). In order to ensure robustness, we compare a probabilistic generative model with an embedding-based transformer approach.

[4] https://huggingface.co/NlpHUST/ner-vietnamese-electra-base

Latent Dirichlet Allocation (LDA) We first apply Latent Dirichlet Allocation (LDA), a probabilistic generative model introduced by Blei et al. (2003). LDA assumes that each document is represented as a mixture of latent topics, where each topic corresponds to a probability distribution over words. Under this framework, document-specific topic proportions govern the generation of words, which are drawn from the associated topic–word distributions (Uthirapathy and Sandanam, 2023). Estimating these distributions enables the model to infer latent thematic structure directly from the corpus, without requiring labeled data (Gokcimen and Das, 2024).

The model is trained for 15 passes with 1,000 iterations to promote stable convergence. Asymmetric priors are automatically optimized for both the document–topic (α) and topic–word (η) distributions. We evaluate models with six topics and retain this configuration based on semantic coherence, topic distinctiveness, and overall interpretability.

BERTopic To complement the probabilistic framework, we implement BERTopic (Grootendorst, 2022), an embedding-based topic modeling method that leverages contextual representations to capture semantic similarity between documents.

Documents are encoded using the `bkai-foundation-models/vietnamese-bi-encoder`[5], a sentence-transformer model optimized for Vietnamese semantic similarity tasks. The resulting embeddings are reduced using UMAP (McInnes and Healy, 2018) with parameters `n_neighbors=10`, `n_components=5`, `min_dist=0.0`, and `metric='cosine'` to preserve global semantic structure while improving clustering efficiency. We then apply HDBSCAN (McInnes et al., 2017) with `min_cluster_size=25`, `min_samples=10`, and `metric='euclidean'` to group semantically similar documents while allowing noise points to remain unassigned. Each cluster is interpreted as a topic.

For topic representation, we use a `CountVectorizer` configured to extract both unigrams and bigrams (`ngram_range=(1,2)`) and to retain only terms that appear in at least ten documents (`min_df=10`). The inclusion of bigrams enables the identification of multiword expressions, such as policy-related or institutional phrases, which may convey more specific information than isolated tokens. The minimum frequency threshold reduces the influence of rare terms, thereby improving the stability and interpretability of the resulting topics.

Topic representations are derived using class-based TF–IDF (c-TF–IDF), which estimates term importance at the cluster level rather than at the individual document level. This weighting scheme highlights words that are particularly distinctive within each cluster and facilitates thematic interpretation (Grootendorst, 2022).

Evaluation Metrics Model performance is evaluated using Topic Coherence, Normalized Pointwise Mutual Information (NPMI), C_v, and Topic Diversity (Cao et al., 2015; Sawant et al., 2022). Coherence-based metrics assess whether the most representative words of a topic tend to co-occur within the corpus (Röder et al., 2015). NPMI quantifies normalized word association strength (Bouma, 2009), whereas C_v combines sliding-window co-occurrence statistics, cosine similarity, and confirmation measures, and aligns well with human interpretability judgments (Röder et al., 2015). Topic Diversity is computed using pairwise Jaccard diversity, measuring the average dissimilarity between the top-word sets of distinct topics. Higher values indicate greater lexical differentiation across topics and, consequently, clearer thematic separation (Dieng et al., 2020). Unless otherwise specified, coherence and diversity scores are calculated using the top 10 words per topic.

4. Results and Discussions

4.1. Topic Modeling by LDA

The six topics identified by the LDA model reflect distinct yet interrelated dimensions of climate and environmental discourse in Vietnam. Their relative prevalence is presented in Figure 5, which indicates noticeable variation in thematic prominence across the corpus.

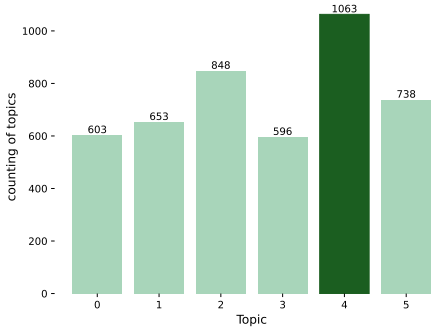

Figure 5: Distribution of articles across LDA topics

Topic 0 – Climate Change and Adaptation. This topic encompasses discussions of climate variability, extreme weather events, and adaptation measures, particularly in coastal and agriculturally exposed regions. It exhibits a consistent but moder-

[5]https://huggingface.co/bkai-foundation-models/vietnamese-bi-encoder

ate presence throughout the dataset, suggesting sustained attention without clear dominance.

Topic 1 – Energy Transition and Green Development. This cluster frames energy transformation within broader development objectives. Renewable energy expansion, technological upgrading, and efficiency improvements are frequently discussed in connection with national green growth strategies, indicating the integration of environmental and economic narratives.

Topic 2 – Environmental Pollution and Local Governance. Among the more prominent themes, this topic highlights pollution-related incidents in urban and industrial contexts. Articles frequently address wastewater discharge, regulatory enforcement, and the role of local authorities, pointing to the governance dimension of environmental management.

Topic 3 – Renewable Energy Projects and Investment. This theme focuses on electricity generation projects, particularly solar and wind power, alongside investment flows and infrastructure expansion. Its distribution appears relatively even, suggesting stable coverage of project-level developments within the broader transition process.

Topic 4 – Air Quality and Public Exposure. Air pollution emerges as the most dominant topic in the corpus. Reporting commonly links deteriorating air quality to everyday environmental conditions and public health concerns in major urban centers, indicating a strong connection between environmental risk and lived experience.

Topic 5 – National Strategy and International Engagement. Climate and energy issues are also embedded within high-level policy narratives, including international cooperation and long-term strategic positioning. The visibility of this topic suggests that environmental discourse is not limited to local or sectoral issues but extends to national and global policy arenas.

Overall, the LDA results point to a multi-layered discourse structure in which different levels of environmental concern coexist. Immediate and observable environmental risks, particularly air pollution and local environmental degradation, receive the greatest attention, while longer-term structural themes such as energy transition and national strategy appear as secondary but stable components.

This distribution suggests that Vietnamese climate discourse is anchored in tangible and locally experienced environmental issues, which are more salient for public communication, while abstract or long-term policy narratives remain comparatively less dominant. At the same time, the coexistence of governance, infrastructure, and international policy topics indicates that climate change is not framed as a purely environmental issue, but rather as a cross-sectoral policy domain embedded in economic development and institutional planning.

However, the relatively broad and partially overlapping nature of several topics, particularly those related to energy transition, investment, and policy, also reflects a limitation of the LDA framework. Specifically, LDA may struggle to disentangle closely related themes when it relies solely on word co-occurrence patterns.

4.2. Topic Modeling by BERTopic

The BERTopic model also yields six topics. Figure 6 presents their semantic composition using word clouds, providing a qualitative illustration of the most representative terms within each cluster.

Figure 6: Word clouds of the six topics identified by BERTopic

Topic 0 – Environmental Pollution and Air Quality. This topic accounts for the largest proportion of documents in the corpus. It centers on environmental pollution, with air quality constituting its core component. Frequent references to pollution indices, environmental conditions, and public health exposure indicate the salience of urban air concerns. The prominence of this theme is consistent with the LDA results, reinforcing its structural importance within the dataset.

Topic 1 – Renewable Energy and Power Sector Development. This cluster captures structural transformation within the electricity system. Generation capacity, grid expansion, and investment dynamics are recurrent elements, reflecting technical and infrastructural change. Compared to the corresponding LDA theme, the focus here appears more concentrated on system-level developments rather than on broader developmental narratives.

Topic 2 – Climate Change Impacts and Adaptation. This topic addresses climate-related impacts in vulnerable sectors, including agriculture, alongside adaptation measures. Its close correspondence with the adaptation theme identified by LDA suggests stability of this strand across modeling approaches.

Topic 3 – Climate Policy and International Commitments. Policy discourse emerges as a distinct cluster. Articles emphasize climate agreements, diplomatic engagement, and national commitments within international governance frameworks. In contrast to LDA, strategic positioning appears more clearly separated from implementation-related themes.

Topic 4 – Youth Engagement and Social Initiatives. This topic highlights societal participation, including youth movements and community-based initiatives. Its emergence as an independent cluster indicates that social engagement constitutes a recognizable component of climate discourse within the corpus.

Topic 5 – Provincial Energy Projects and Infrastructure. Renewable energy development at the provincial level forms a separate theme. Wind power projects and associated infrastructure are delineated from national policy discussions, suggesting that BERTopic differentiates implementation dynamics from strategic narratives more explicitly than LDA.

Across both modeling approaches, several patterns remain consistent, most notably the dominance of air pollution and environmental quality, as well as the strong presence of energy transition and climate policy themes. This convergence across models increases confidence that these topics represent structurally stable components of Vietnamese climate discourse rather than artifacts of a specific modeling approach.

4.3. Evaluation of Topic Modeling Performance

The quantitative performance of LDA and BERTopic, evaluated using coherence and diversity metrics, is reported in Table 2.

Models	LDA	BERTopic
Topic Coherence (c_v)	0.55	0.66
Topic Coherence (NPMI)	0.06	0.18
Topic Diversity	0.94	0.93

Table 2: Topic Model Performance Metrics

BERTopic achieves higher coherence scores under both measures. The c_v coherence reaches 0.66 for BERTopic, compared to 0.55 for LDA. A similar pattern is observed for NPMI coherence, where BERTopic attains a score of 0.18, whereas LDA yields 0.06. These differences indicate stronger semantic consistency among the most representative words within each topic when using the embedding-based approach.

However, the relatively low NPMI score for LDA (0.06) should be interpreted with caution, as it likely reflects a mismatch between the metric and the characteristics of the data rather than genuinely poor topic quality. NPMI relies on localized word co-occurrence and is therefore sensitive to how frequently related terms appear within a fixed window. This sensitivity makes NPMI particularly vulnerable to underestimating coherence in settings where semantic relationships are distributed across broader textual contexts rather than localized word spans.

In the present corpus, three factors systematically reduce co-occurrence signals. First, lexical synonymy fragments probability mass across multiple surface forms (e.g., *chính sách*, *quy định*, *pháp luật*), reducing pairwise co-occurrence counts even when words are semantically aligned. Second, climate discourse spans multiple subdomains, such as policy, energy, and environment, leading to domain-stratified vocabulary in which related terms rarely co-occur within the same local context. Third, corpus sparsity and high lexical diversity limit the accumulation of reliable co-occurrence statistics under standard window sizes.

These effects are further amplified by characteristics of Vietnamese text, including multi-syllabic word segmentation and lexical variability, which weaken surface-level co-occurrence signals. As a result, NPMI systematically underestimates topic coherence in this setting, particularly for probabilistic models such as LDA that rely on distributed word frequency patterns.

In contrast, the c_v metric yields substantially higher values for both models, reflecting its ability to capture broader semantic similarity beyond strict co-occurrence. The divergence between c_v and NPMI suggests that while topics may not exhibit strong local co-occurrence patterns, they remain semantically interpretable at a higher level.

Topic diversity remains high for both models, with values of 0.94 for LDA and 0.93 for BERTopic. This indicates that both approaches produce largely distinct topic representations, with minimal lexical overlap among top-ranked terms.

Taken together, the results suggest that BERTopic provides improved semantic coherence, particularly under embedding-aware evaluation, while LDA captures broader thematic structures despite lower co-occurrence-based coherence

scores. These findings highlight the importance of using multiple evaluation metrics when assessing topic models, especially in low-resource and linguistically complex settings.

4.4. Temporal Evolution of Climate Discourse

Finally, to examine the temporal dynamics of ecological discourse, we analyze the annual prevalence of LDA-derived themes over 2004–2026. Topic prevalence is computed as the yearly average of document–topic probabilities, enabling systematic identification of shifts in thematic salience over time.

Figure 7 presents the longitudinal evolution of the six themes. The results reveal distinct temporal phases: early dominance of environmental pollution narratives, followed by a rise in adaptation and governance discourse, and more recently, the expansion of energy transition and international policy framing. Several peaks align with major national and global climate-related events, suggesting that media attention responds to both structural policy developments and event-driven environmental crises.

Environmental pollution and local governance constitute a dominant theme in the mid-2000s, reaching an initial peak around 2006–2007. This period coincides with rapid industrial expansion, suggesting a possible association between increased economic activity and heightened media attention to localized environmental degradation. A renewed surge appears in 2016, aligned with the Formosa Ha Tinh marine pollution incident, followed by another increase in 2019–2020 during intensified reporting on urban air pollution in Hanoi and Ho Chi Minh City. These fluctuations indicate that pollution discourse is strongly event-sensitive, intensifying in response to acute environmental crises rather than evolving along a continuous structural trajectory.

Climate change and adaptation discourse increases markedly between 2008 and 2012, peaking around 2012. This expansion corresponds to major international policy milestones, including the IPCC Fourth Assessment cycle and the Copenhagen and Cancun negotiations, as well as domestic initiatives such as the National Target Program to Respond to Climate Change (2008) and the National Strategy on Climate Change (2011). The subsequent decline after 2013 suggests thematic differentiation, with climate concerns increasingly embedded within sectoral governance and energy-related narratives rather than remaining an isolated strand.

Energy transition follows a U-shaped trajectory. After declining until approximately 2015, coverage increases steadily from 2016 onward, aligning with the Paris Agreement and Vietnam's Nationally Determined Contribution, suggesting that international policy developments may influence domestic media attention. A more pronounced acceleration after 2022 corresponds to COP26-related commitments and participation in the Just Energy Transition Partnership. Unlike pollution-related themes, this pattern reflects cumulative policy alignment and long-term strategic repositioning of the national energy system, rather than episodic crisis response.

Renewable energy projects and investment display sustained growth beginning around 2014, following the introduction of feed-in tariff mechanisms and renewable incentives. Stronger expansion after 2019 parallels rapid solar deployment and intensified clean energy investment. This trajectory suggests that renewable discourse is closely linked to infrastructural scaling and regulatory reinforcement, indicating structurally anchored implementation dynamics.

Air quality and public exposure exhibit a pronounced spike in 2019, coinciding with severe pollution episodes in Hanoi. The decline after 2020 likely reflects reduced mobility during the COVID-19 period, followed by stabilization as air quality becomes integrated into broader environmental governance discussions. This pattern illustrates the transition of specific environmental risks from acute public attention to normalized policy framing.

National strategy and international engagement peak around 2015 and again in 2021-2022, corresponding to the Paris Agreement process and subsequent COP26 commitments. The cyclical nature of this theme indicates temporal synchronization between national discourse and international negotiation cycles, highlighting the external anchoring of domestic climate narratives within multilateral processes.

Jointly, these patterns indicate that different thematic domains follow distinct temporal logics. Pollution-related topics exhibit short-term volatility driven by discrete events, whereas energy and policy-related themes evolve more gradually in response to institutional and regulatory processes. This divergence highlights the coexistence of reactive and structural dynamics within climate discourse.

Across these trajectories, Vietnamese climate media discourse undergoes a gradual transformation. Early coverage is characterized by reactive responses to localized environmental crises, whereas later periods show increasing institutionalization, international alignment, and structural framing of climate and energy issues. Pollution-related themes remain event-driven, while energy transition and renewable investment narratives exhibit cumulative, policy-oriented dynamics. This shift points to a transition from crisis-centered environmen-

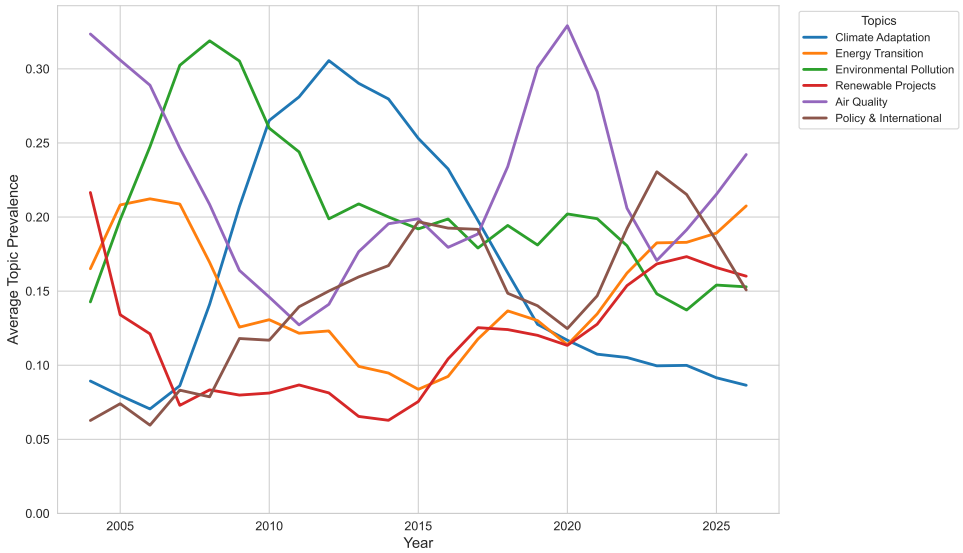

Figure 7: Temporal evolution of LDA-derived climate discourse topics in Vietnamese national media (2004–2026)

tal reporting toward a more institutionalized and policy-driven discourse, in which climate change is increasingly framed as a long-term governance and development challenge rather than a series of isolated events.

5. Conclusion

To summarise, we constructed a geographically grounded longitudinal corpus of climate-related news articles in Vietnam covering 2004–2026. The corpus was automatically collected from major national media outlets and processed using Vietnamese-specific NLP tools to ensure linguistic and geographic consistency. Using Latent Dirichlet Allocation and BERTopic, we modelled the thematic structure of the corpus and traced its evolution over time, enabling a comparison between probabilistic and embedding-based topic modeling approaches in a low-resource setting.

The results reveal stable thematic dimensions across models, including air pollution, renewable energy development, and climate governance. Longitudinal analysis highlights a gradual shift in media attention: early discourse centers on localized environmental pollution, followed by increasing emphasis on adaptation and institutional policy frameworks, and more recently, energy transition and international climate engagement. These domains exhibit distinct temporal dynamics, with pollution-related topics showing event-driven fluctuations, while energy and policy themes evolve more gradually in response to structural developments. The embedding-based approach achieves higher topic coherence while maintaining comparable topic diversity, indicating stronger semantic consistency.

Beyond methodological comparison, this study contributes a new Vietnamese climate corpus and provides empirical evidence on how ecological discourse evolves in a low-resource, policy-driven media environment. These findings underscore the value of combining modeling approaches to capture both high-level thematic structure and fine-grained semantic variation.

Several directions for future research emerge. The corpus could be extended to include regional outlets and social media to capture a broader spectrum of communication. Dynamic topic modeling may enable more fine-grained temporal analysis, while domain-adaptive pretraining could improve representation quality. Integrating sentiment or uncertainty modeling would further support the analysis of framing and discursive dynamics.

This study has several limitations. Vietnamese remains a comparatively low-resource language, and domain-specific pretrained models are limited. Automatic word segmentation may introduce noise affecting downstream modeling, while reliance on mainstream national media may introduce structural bias and underrepresent local perspectives. Finally, the absence of human qualitative validation limits full assessment of topic interpretability.

6. Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback and helpful suggestions. Their comments have contributed to improving the clarity and quality of this work.

7. References

Rafiu Adekoya Badekale and Adewale Akinfaderin. 2025. Temporal analysis of climate policy discourse: Insights from dynamic embedded topic modeling. *ArXiv*, abs/2507.06435.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, pages 2210–2216. AAAI Press.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35.

Thi Kim Phung Dang. 2025. Climate change communication in vietnam's online newspapers and its implications for climate actions. *Sustainability*, 17:1354.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Isabella Gagliardi and Teresa Artese. 2020. Semantic unsupervised automatic keyphrase extraction by integrating word embedding with clustering methods. *Multimodal Technologies and Interaction*, 4(2):30.

Tunahan Gokcimen and Bihter Das. 2024. Exploring climate change discourse on social media and blogs using a topic modeling analysis. *Heliyon*, 10(11):e32464.

Maarten R. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

F. M. Anamul Haque, Md. Abdur Rahman, and Sumon Ahmed. 2025. Ghtm: A graph based hybrid topic modeling approach in low-resource bengali language. *ArXiv*, abs/2508.00605.

Long Le and Lien-Huong Vo. 2026. Vietnam in the climate change narratives: A discursive news values analysis of english-language news. *Studies in Media and Communication*, 14:267–280.

Leland McInnes and John Healy. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2:205.

Dasha Pruss, Yoshinari Fujinuma, Ashlynn R. Daughton, Michael J. Paul, Brad Arnot, Danielle Albers Szafir, and Jordan Boyd-Graber. 2019. Zika discourse in the americas: A multilingual topic analysis of twitter. *PloS one*, 14(5):e0216922.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Sahil Sawant, Jinhong Yu, Kirtikumar Pandya, Chun-Kit Ngan, and Rolf Bardeli. 2022. An enhanced bertopic framework and algorithm for improving topic coherence and diversity. In *Proceedings of the 2022 IEEE 24th International Conference on High Performance Computing & Communications (HPCC/DSS/SmartCity/DependSys)*, pages 2251–2257. IEEE.

Ehsan Tashakori, Yaser Sobhanifard, Adel Aazami, and Rahim Khanizad. 2025. Uncovering semantic patterns in sustainability research: A systematic nlp review. *Sustainable Development*.

Samson Ebenezar Uthirapathy and Domnic Sandanam. 2023. Topic modelling and opinion analysis on climate change twitter data using lda and bert model. *Procedia Computer Science*, 218:908–917.

Azmine Toushik Wasi, Wahid Faisal, Taj Ahmad, Abdur Rahman, and Mst Rafia Islam. 2024. Dhoroni: Exploring bengali climate change and environmental views with a multi-perspective news dataset and natural language processing. *arXiv preprint arXiv:2410.17225*.