Non-Archival Abstract

Multilingual Topic Modeling of Climate Change and Public Health Discourse in Online News (2020-2025)

Luis Ruiz-Ramirez, Silvan Wehrli, Christopher Irrgang
Centre for Artificial Intelligence in Public Health Research,
Robert Koch Institute, Berlin, Germany
Correspondence: IrrgangC@rki.de

Introduction and Related Work During the last decade, growing attention has been paid to understanding and communicating the effects of climate change on public health. At the same time, online media have become a central space in which climate and health-related issues have been framed, linked, and communicated to broad audiences. However, research on climate change and public health communication has often developed in parallel, examining environmental and health discourses as distinct domains (Schirmag et al., 2025; Ittefaq, 2024). Previous studies have examined climate or health discourse across time, countries, and languages (Pruss et al., 2019; McAllister et al., 2024). However, limited work has explicitly modeled how climate and health narratives are linked within the same multilingual corpus. Unlike prior multilingual topic models that impose cross-lingual correspondence by design (Pruss et al., 2019), we discover the topics independently across languages and then test empirically whether they bridge or not. This allows a cross-lingual convergence to be measured rather than assumed.

Scope Rather than focusing only on their direct intersection, we examine how both issue domains evolve over time in online media, whether they appear together or separately, and how their relationship changes across languages within a global news corpus, and across publisher countries where metadata allows. We distinguish climate-only, health-only, and climate-health linked topics. This allows us to examine how online news media represent links between climate change and public health, instead of estimating public health narratives directly.

Data We constructed a multilingual dataset of online news in English, Spanish, and German from 2020–2025 to compare discourses across languages within a global pool of online publishers. The data were retrieved from the GDELT Project through its Web NGrams service. Publisher country is available for 63% of articles, spanning 94 countries. However, attribution varies sharply by language (German 94%, Spanish 62%, and English 33%): German coverage concentrates in Germany, Austria, and Switzerland. Spanish coverage spans Spain, Mexico, Argentina, and Peru, and English is led by the United Kingdom, Australia, India, and Canada.

Methods Article retrieval was guided by keyword lists derived from the Lancet Countdown reports (Romanello et al., 2025). To address strong language and temporal imbalances, we created a balanced corpus by applying stratified down-sampling by language and year. We then applied BERTopic with the multilingual Qwen3-Embedding-4B model (Zhang et al., 2025), UMAP projection, and HDBSCAN for the topic clustering. Lastly, we grouped related topics into higher-level macro-clusters to better represent global narratives.

Preliminary Results The balanced 2020–2025 baseline contains 608,076 articles, of which 572,030 were assigned to 1,526 fine-grained topics. The relevance rule identified 185 climate-health linked candidate topics covering 42,782 articles (7.5% of articles assigned to a topic). These articles were then summarized into seven higher-level narrative groups. Fine-grained topics are largely siloed (around 86% are dominated by a single language), and cross-lingual narratives emerge when these topics are bridged into macro-clusters. Language composition varied substantially: two policy-oriented groups were 83-84% German. Narratives about environmental pollution and agriculture were more evenly distributed across languages.

Future Work We will build on this analysis by linking topic prevalence to an extreme-events dataset (EM-DAT) to test whether environmental shocks causally affect the prevalence of climate-health narratives in online media.

References

Muhammad Ittefaq. 2024. Climate communication, public health, and social media: Examining the role of health agencies in addressing the impacts of climate change on human health. *J. Health Commun.* DOI: 10.1080/10810730.2024.2360021.

Lucy McAllister, and 3 others. 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):024046.

Dasha Pruss, and 6 others. 2019. Zika discourse in the americas: A multilingual topic analysis of twitter. *PLOS ONE*, 14(5):e0216922.

Marina Romanello, and 127 others. 2025. The 2025 report of the lancet countdown on health and climate change: climate change action offers a lifeline. *The Lancet*. DOI: 10.1016/S0140-6736(25)01919-1.

Tatjana Schirmag, and 3 others. 2025. Neural topic modeling reveals german television's climate change coverage. *Commun. Earth Environ.* DOI: 10.1038/s43247-025-02402-1.

Yanzhao Zhang, and 11 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint*. DOI: 10.48550/arXiv.2506.05176.