CCF DATABASE: A MACHINE-LEARNING-ANNOTATED CORPUS OF 266,271 CANADIAN CLIMATE ARTICLES (1978–2024)

Antoine Lemor[1], Alizée Pillod[2] & Matthew Taylor[3]

The Canadian Climate Framing (CCF) Database is a comprehensive, machine-learning-annotated corpus designed to enable large-scale analysis of climate discourse in Canadian print media. It comprises 266,271 articles from 20 major Canadian newspapers spanning nearly five decades (1978–2024), processed into 9,198,158 bilingual sentences (82.9% English, 17.1% French). Each sentence is annotated with 65 hierarchical categories using transformer-based classifiers (BERT for English, CamemBERT for French), trained on over 4,000 expert-coded sentences. The annotation framework captures multiple dimensions of climate discourse: thematic frames (economic, health, security, justice, political, scientific, environmental, cultural), actor types, event categories, solution strategies, emotional tone, geographic focus, and named entities. The models achieve a macro F1 score of 0.866 against an independent gold standard with confirmed intercoder reliability.

Submission Details

Submitted to *Scientific Data*.

---

[1] Postdoctoral researcher, Université de Sherbrooke, Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Réseau francophone international en conseil scientifique (RFICS). antoine.lemor@umontreal.ca

[2] PhD Candidate, Université de Montréal

[3] PhD Candidate, Université de Montréal