CCF DATABASE : UN CORPUS DE 266 271 ARTICLES CANADIENS SUR LE CLIMAT ANNOTÉ PAR APPRENTISSAGE AUTOMATIQUE (1978–2024)

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

La base de données Canadian Climate Framing (CCF) est un corpus exhaustif annoté par apprentissage automatique, conçu pour permettre l'analyse à grande échelle du discours climatique dans la presse écrite canadienne. Elle comprend 266 271 articles issus de 20 grands quotidiens canadiens couvrant près de cinq décennies (1978–2024), traités en 9 198 158 phrases bilingues (82,9 % en anglais, 17,1 % en français). Chaque phrase est annotée selon 65 catégories hiérarchiques à l'aide de classificateurs basés sur des transformers (BERT pour l'anglais, CamemBERT pour le français), entraînés sur plus de 4 000 phrases codées par des experts. Le cadre d'annotation capture de multiples dimensions du discours climatique : cadrages thématiques (économique, sanitaire, sécuritaire, justice, politique, scientifique, environnemental, culturel), types d'acteurs, catégories d'événements, stratégies de solutions, tonalité émotionnelle, focus géographique et entités nommées. Les modèles atteignent un score F1 macro de 0,866 par rapport à un étalon-or indépendant avec une fiabilité intercoder confirmée.

Détails de soumission

Soumis à *Scientific Data*.

---

[1] Chercheur postdoctoral, 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] Doctorante, Université de Montréal

[3] Doctorant, Université de Montréal