scientific data

Article in Press

Check for updates

OPEN

DATA DESCRIPTOR

CCF Database: A Machine-Learning-Annotated Corpus of 283,964 Canadian Climate Articles (1978–2026)

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

News media shape how societies perceive and respond to climate change, yet fine-grained analysis remains hindered by fragmented corpora and ad hoc coding schemes. The Canadian Climate Framing (CCF) database answers both with a large-scale bilingual annotated corpus and a generalizable annotation framework. It covers 283,964 climate-related articles from 22 Canadian news outlets (1978–2026), processed into 9,908,776 two-sentence units (83.4% English, 16.6% French). Each unit carries 65 hierarchical binary annotations spanning eight main frames, actors, events, solutions, tone, geography, and urgency, produced by 128 BERT and CamemBERT classifiers trained on 4,000+ expert-coded sentences. Against a 1,000-sentence gold standard double-coded by an independent annotator (Cohen's $\kappa = 0.596$, Krippendorff's $\alpha = 0.698$, Gwet's AC1 $= 0.894$), the classifiers reach a pooled macro F_1 of 0.866 (mean per-category macro $F_1 = 0.773$), and every category carries an A/B/C reliability tier. The database—sentence-level annotations, per-article aggregations, and 10,192,740 BAAI/bge-m3 embeddings in six relational tables—is openly released on Zenodo (CC-BY 4.0) as a PostgreSQL edition and an Apache Parquet mirror, refreshed roughly every six months by the CCF Observatory.

Background & Summary

On February 18, 1978, *The Globe and Mail* ran a story headlined "Some think it's warming up" that reported on a survey of leading climate experts. Interestingly, almost half a century ago, the article already laid out for its readers what have since become the well-established root causes of climate change:

> "The concern is that possible warming, the so-called 'greenhouse effect,' is generally associated with the increase of carbon dioxide in the atmosphere by the burning of fossil fuels."

This excerpt clearly embodies the role the media play in informing society and citizens, who often lack direct access to academic research. At the same time, the media also reflect the social and political realities of the country they cover. Indeed, while some newspapers choose to relay the voices of scientists, others choose to report the controversies, so that the media both inform and reflect political and social realities. Research shows that media shape not only what people think about but also how they think about it, with effects on both public opinion and political agendas[1,2]. The media are accordingly not a neutral arena for public debate around climate change. Instead, they actively participate in defining and constructing the problems under discussion[3,4]. Moreover, while digital and social media have grown in prominence, newspapers remain a foundational source of information that continues to shape public discourse and to influence other media formats[5,6]. Studying print media therefore provides a vantage point on a traditional and still highly influential channel of climate communication. Most importantly, it allows for long-term historical comparison, including periods that predate the rise of social media.

The excerpt above was drawn from the oldest article of the *Canadian Climate Framing (CCF)* database, which we introduce in this Data Descriptor. The CCF database is a machine-learning-preprocessed corpus of 283,964

[1]Université de Sherbrooke, Sherbrooke, Québec, Canada. [2]Université de Montréal, Montréal, Québec, Canada. e-mail: antoine.lemor@umontreal.ca; alizee.pillod@umontreal.ca; matthew.taylor@umontreal.ca; richard.nadeau@umontreal.ca

climate-related newspaper articles published between 1978 and 2026 across 22 major Canadian news outlets, accompanied by a rigorous, theoretically grounded annotation framework that we release alongside the data. The corpus contains 9,908,776 two-sentence analytical units, all annotated within a 65-category hierarchical coding scheme. Because Canadian newspaper articles are protected by copyright, the deposit does not redistribute the raw sentence text of the corpus at large (the sole, clearly bounded exception—the expert-coded training and gold-standard sentences—is documented in Data Records). What is shared instead is (i) the 65-category hierarchical annotation framework—a theoretically grounded coding scheme that integrates main frames, actors, events, solutions, emotional tone, geographic focus, and urgency into a single sentence-level taxonomy—and (ii) the full output of an intensive annotation pipeline built around that framework: 65 binary sentence-level labels per analytical unit, the named-entity extractions (persons, organizations, locations), the article-level aggregates and full metadata (newspaper, date, title, author, page number), the per-category reliability tiers, and the BAAI/bge-m3 sentence embeddings (10.19 million 1024-dimensional vectors). The breadth of the annotation scheme and the embedding layer together make the deposit fully analysable at the sentence level without requiring access to the raw text. By focusing on the Canadian press, this database constitutes a historically influential source of climate information.

Canada offers a particularly compelling case for studying climate discourse. Its federal structure, bilingual culture, and vast territory create a microcosm of the diversity—and contradictions—found globally. While Canada is highly vulnerable to the impacts of climate change[7], it remains significantly dependent on fossil-fuel extraction and other resource-based industries[8]. This creates structural tensions between environmental protection and economic interests. Some provinces, notably Quebec and British Columbia, tend to adopt more progressive climate policies, whereas others with resource-intensive economies are more cautious or resistant[9]. This combination of political decentralization, linguistic duality, and regional economic variation produces diverse policy approaches and, in turn, shapes how climate change is discussed in the media. Yet systematic analysis of climate communication in Canada and elsewhere has long been constrained by methodological obstacles: until recently, large-scale corpus collection and precise, reproducible content analysis were virtually impossible. Supplementary Table S1 provides an overview of prior studies in this area, highlighting their scope, focus, and coverage. The dominant pattern is one of corpora restricted to a few outlets, short observation windows, and English-language sources.

The CCF Database was designed to address four major limitations in previous research on climate change media coverage in Canada. First, the CCF draws on a larger set of news outlets and articles, comprising 283,964 articles from 22 news outlets with substantial representation for each outlet. Previous studies generally relied on smaller corpora and a limited number of newspapers, thereby capturing a narrower media landscape. Second, the CCF spans nearly five decades, providing a broader temporal perspective than most existing studies, which often focus on short periods or extend only from the late 1980s to the 2010s. Third, while climate attitudes vary significantly across regions[10], regional climate coverage remains understudied. Existing studies generally focus on nationally circulating newspapers—most commonly *The Globe and Mail*, *National Post*, and *Toronto Star*—and rarely compare a large set of provinces and territories. The CCF therefore includes a wide range of regional newspapers. Finally, the CCF reflects Canada's linguistic diversity through a bilingual corpus that includes a substantial share of French-language content (83.4% English, 16.6% French), with francophone outlets both inside and outside Quebec. While Canada is officially bilingual, previous research has largely excluded French-language media from analysis. Two studies, in particular, have examined francophone coverage: Young and Dugas[11], comparing climate coverage between French-language outlets from Quebec and English-language outlets from the rest of Canada, and King *et al.*[12], analysing climate change's public-health framing with partial inclusion of Quebec.

Methodologically, the CCF database is the first climate media corpus to offer sentence-level annotations derived from supervised machine learning and natural language processing (NLP). Unlike unsupervised approaches such as topic modelling, which remain dominant in the field, the CCF methodology relies on transformer-based classifiers (BERT for English, CamemBERT for French) trained on over 4,000 manually expert-coded sentences. NLP enables scholars to examine much larger corpora in far less time, opening new avenues for systematic inquiry[13–15]. However, as summarised in Supplementary Table S1, when climate researchers do employ NLP it is most often to identify key actors through named-entity recognition, to detect stance, or to conduct sentiment analysis, which tend only to map actors, tone, or evaluative direction rather than interpretive structures. The dominant computational approach in the field is still unsupervised topic modelling, where topics are treated as proxies for frames—an assumption that Nicholls and Culpepper[16] and Khanehzar, and Frermann[17] flag as problematic, as topics are too narrow to capture the defining features of frames, which, following Entman[18], require identifying a problem, specifying causal interpretations, moral evaluations, and possible solutions. By contrast, the supervised approach adopted in the CCF Database enables the identification of frames, actor types, events, solution strategies, and emotional tone with a level of semantic precision that unsupervised methods cannot achieve. Only a few studies have applied comparable transformer-based models in this domain: Luo, Card, and Jurafsky[19] trained a single BERT classifier to analyse the evolution of climate partisanship in U.S. media; Meddeb *et al.*[20] used a CamemBERT model to detect climate misinformation in French news; Frermann *et al.*[21] fine-tuned transformer models to detect narrative roles (conflict, villain, resolution) in media articles on three issues including climate change; Badullovich *et al.*[22] benchmarked a RoBERTa classifier against zero-shot LLMs on issue and rhetorical frames in climate change editorials published in *Nature* and *Science*. None of these works, however, have trained and deployed a comparable suite of transformer models to systematically annotate climate-related media coverage at the scale, breadth, and bilingual coverage of the CCF database.

Validated against an independent gold standard of 1,000 expert-coded sentences whose reliability was confirmed through double-coding (Cohen's $\kappa = 0.596$, Krippendorff's $\alpha = 0.698$, Gwet's AC1 $= 0.894$, on the

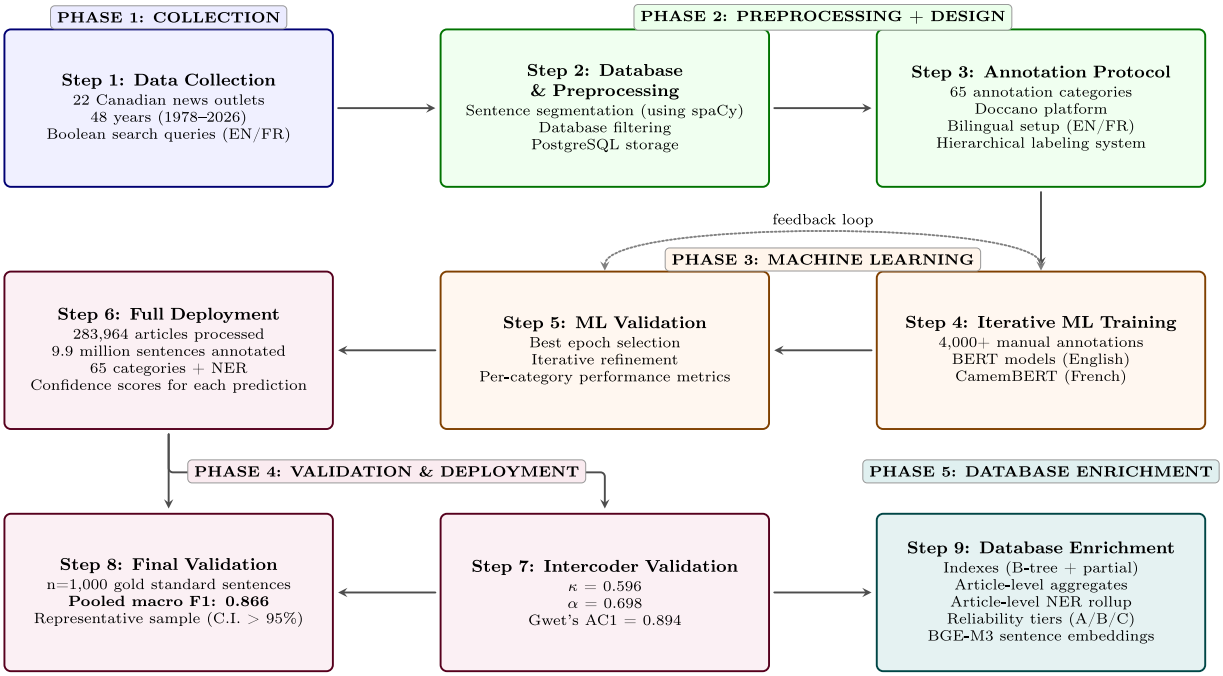

Fig. 1 Methodological pipeline of the Canadian Climate Framing (CCF) project.

blind-coding phase of sentences 601–1000), the CCF trained models attain a pooled macro F1 score of 0.866 across the 65 annotation categories (0.773 when the 65 per-category macro F1 scores are averaged without pooling; both quantities are defined and discussed in the Technical Validation section). In addition, named-entity recognition (NER) was applied to extract persons, organisations, and locations mentioned in the articles, and the entire corpus was embedded with the BAAI/bge-m3 multilingual encoder so that semantic-similarity queries can be issued directly against the deposited database. With 128 BERT and CamemBERT classifiers covering frames, actors, events, solutions, and emotional tone, the CCF database fills a clear methodological gap by introducing a supervised NLP approach capable of detecting multiple analytical dimensions at scale, and the absence of raw text in the deposit is offset by the depth of the annotation: every sentence carries 65 binary labels, a deduplicated NER signature, and a 1024-dimensional embedding, so researchers can run frame-based, actor-based, and semantic queries directly on the deposited tables. This technical paper presents the methodology underlying the CCF database and its structure. The remainder of the document outlines the full workflow (data acquisition and preprocessing, model training, validation, and database enrichment) and offers the conceptual rationale and practical steps necessary for researchers wishing to work with the CCF database or to adapt its methodology elsewhere.

Methods

The CCF Database was built through a mixed-methods design that coupled large-scale data collection with state-of-the-art machine learning (ML). The resulting pipeline provides a systematic and reproducible framework for large-scale analysis of climate discourse that allows researchers to track how climate change is discussed, how narratives evolve over time, and who participates in the conversation.

Figure 1 presents the complete methodological pipeline and illustrates how the CCF Database was assembled from raw newspaper articles into a fully annotated database suitable for robust and reproducible climate communication research, both quantitative and qualitative. The pipeline combined human expertise with computational methods through an iterative human-in-the-loop machine learning procedure. This approach ensured both the scalability needed to process 283,964 climate-related articles and the accuracy required for rigorous academic research.

The methodology comprises five distinct phases: (1) the *Collection Phase* involved systematic gathering of climate-related articles from 22 major Canadian news outlets using search queries resulting in 283,964 articles spanning 1978–2026; (2) the *Preprocessing and Design* phase transformed raw text into structured data, including sentence segmentation, language verification, and development of a comprehensive 65-category annotation framework; (3) the *Machine Learning* phase created and validated transformer-based machine learning models (BERT for English, CamemBERT for French) through iterative training with over 4,000 expert manual annotations; (4) the *Validation and Deployment* phase ensured rigorous quality assurance by combining standard performance metrics with a human evaluation protocol based on an expert-coded gold standard that was independently double-coded by a second annotator to assess intercoder reliability, before deploying the trained models to annotate the complete corpus of 9.9 million sentences; and (5) the *Database Enrichment* phase consolidated the sentence-level output into a publication-ready relational system by adding four derived tables (article-level annotation aggregates, article-level named-entity rollups, per-category reliability tiers, and 10.19

Language	Boolean Query
English	"global warming" OR "climate change" OR "climate disruption" OR "climate disturbance" OR "climate disturbances" OR "climate crisis" OR "greenhouse gas"
French	"réchauffement climatique" OR "réchauffement planétaire" OR "changement climatique" OR "changements climatiques" OR "déréglement climatique" OR "crise climatique" OR "gaz á effet de serre"

Table 1. Boolean queries used for article retrieval.

million BAAI/bge-m3 sentence-and-title embeddings indexed for approximate nearest-neighbour search), together with a complete index structure, so that the deposit can be queried directly without rewriting the same aggregation logic on every project.

Phase 1 & 2: Data collection, preprocessing, database creation and annotation protocol.

Step 1: Data collection. The CCF database was built through a systematic protocol to trace the evolution of climate discourse in Canada's bilingual print media, reaching back to the earliest available archives.

Articles were retrieved through university library access to three press databases: Factiva, Eureka.cc, and ProQuest Canadian Major Dailies. These platforms are the standard providers of full-text archives for Canadian newspapers and are routinely used in peer-reviewed media research[11,12,20]. Each platform was queried independently with the same bilingual Boolean keywords (see Table 1 below) restricted to the selected outlets and to the full available time range (the 20 newspapers of the original collection; the CCF Observatory subsequently added the two public-broadcaster news sites through their own feeds), and duplicate articles appearing in more than one platform were collapsed at preprocessing time on the basis of newspaper, date, headline, and a 95% fuzzy similarity of the article body (see Step 2). Full bibliographic metadata are included in the deposit so that the corpus remains transparent and reproducible while staying within the bounds of copyrights. The design relies on the fair-dealing research exception in Canadian copyright law and the analogous fair-use research exception in U.S. law, both of which authorise non-consumptive, transformative use of copyrighted text for academic research; researchers with institutional access to Factiva, Eureka.cc, or ProQuest Canadian Major Dailies can recover the original text from the bibliographic coordinates.

A total of 22 news outlets—20 newspapers and, since the corpus came under the stewardship of the CCF Observatory, the two news websites of the Canadian public broadcaster (CBC.ca and Radio-Canada.ca, covered from 2023 onward)—are included in the corpus (see Supplementary Table S2). The objective was to include both national and regional outlets in order to ensure broad geographic representation. Within this pool, newspapers were primarily chosen based on circulation levels[23], with the highest-circulating outlets retained. Additional criteria also informed the selection, including the desire to capture a range of political orientations and to build a bilingual corpus that reflects Canada's linguistic landscape. The corpus accordingly includes French-language outlets both inside and outside Quebec, together with one English-language outlet located in Quebec; this asymmetry mirrors the country's linguistic geography, in which Quebec is the only province with French as its sole official language while New Brunswick is officially bilingual and several other provinces—most notably northeastern Ontario along the Quebec border—host sizeable francophone communities. The final selection was also partly shaped by availability constraints. When the highest-circulating newspaper of a given province was unavailable, the next highest-circulating outlet was retained instead; for Prince Edward Island, Nunavut, and the Northwest Territories, no eligible newspaper was available for inclusion at all. Coverage windows are bounded by the access periods of the source aggregators: Le Droit is covered until December 2023, The Whitehorse Daily Star until May 2024, and Le Devoir and L'Acadie Nouvelle until March 2025, while all other outlets extend to the corpus cut-off of July 2026; Supplementary Table S2 details the composition and per-outlet counts.

The corpus spans 48 years, from 1978 to 2026, with 1978 marking the earliest article identified, quoted in the introduction of this paper. To identify climate-related articles, the authors used a transparent keyword-based filter with seven climate-domain terms per language (Table 1). To be included in the data set, an article had to contain at least one of these keywords in its title or body. These queries yielded an initial corpus of over 300,000 articles. As with any keyword-based retrieval, the procedure imposes a recall ceiling: it cannot capture paraphrastic, metonymic, or strictly context-only mentions of climate change, and the deposit therefore covers articles that explicitly use one of the query terms rather than the broader set of articles that discuss climate change in some form.

Step 2: Preprocessing. The preprocessing phase transformed the raw data into a structured dataset suitable for machine learning. Each article was segmented into two-sentence contexts using language-specific spaCy models—`en_core_web_lg` for English text and `fr_dep_news_trf` for French text, both optimised for news content. Sliding windows with single-sentence overlaps were implemented to ensure complete coverage while maintaining semantic continuity throughout each article. This level of granularity supported machine-learning applications by increasing the size of the training dataset. More importantly, working at the (near-)sentence level was the only approach that allowed ex-post analysis to remain highly granular compared to paragraphs, where multiple frames, actors, or solutions can co-occur and become indistinguishable. Article-level variables could then be reconstructed with high granularity by aggregating sentence-level annotations (e.g., percentage of sentences in an article that mention a specific frame or actor), whereas paragraph-level annotations would have been less precise and more difficult to interpret.

Quality control measures were applied systematically throughout the data preparation pipeline. Duplicate articles were identified and removed with fuzzy string-similarity algorithms set to a 95% threshold. Any article

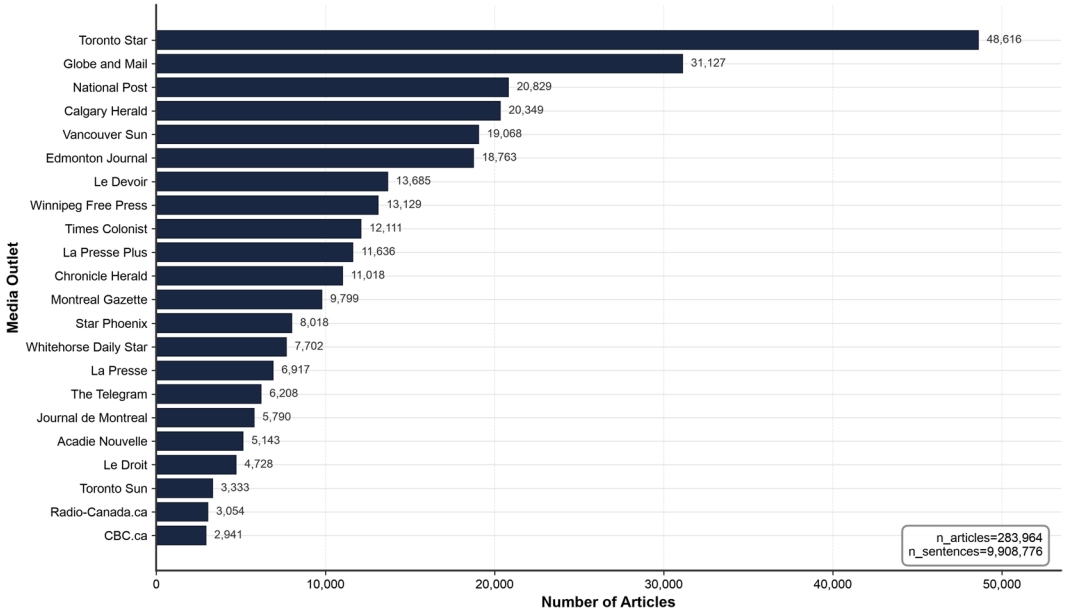

Fig. 2 Distribution of articles by media.

containing fewer than 100 words was also removed to eliminate reader letters, which often did not provide enough material to identify the presence of a frame and did not qualify as a press article per se. However, all other types of text (opinion columns, letters to the editor, news articles, editorials, etc.) were retained. Regarding this matter, Lawrence[24] explains that it is precisely in editorials, or more generally in opinion texts, where journalistic standards are applied with a greater degree of flexibility, that the battle of framing is best observed. Articles published by external news agencies (for example, *Agence France-Presse*, *CanWest News Service*, *Reuters*, or *Bloomberg*) or by other newspapers not included in the analysis but which often belong to the same media conglomerate as the selected outlets (for example, *the Financial Post*) were also retained. These republications were treated as aligned with the editorial direction of the host news outlet, and a republication did not contradict the objectives of the study since it indicated editorial interest in a topic concerning climate change. This process reduced the corpus to 283,964 unique articles. Duplicates of the same article appearing in different media outlets were deliberately kept, since their co-occurrence is itself an object of analysis when measuring media bubbles.

Language was then verified with the pre-trained fastText language-identification model[25,26], in order to ensure consistent categorisation of bilingual content, and publication dates were standardised at the same step. This data preparation produced a clean, standardised corpus ready for machine learning and annotation, stored in two PostgreSQL tables: one with full metadata (titles, main text, authors, dates, and page numbers) and another with the processed sentence segments and their corresponding identifiers.

The resulting database is illustrated in Fig. 2, which displays the distribution of article counts by media (22 outlets), while Fig. 3 shows the total article counts per year across the full time span; the final months of 2026 are incomplete by construction, the corpus being cut off at 31 July 2026. The corpus is now maintained by the CCF Observatory, the continuous extraction–annotation pipeline described above, and is refreshed on Zenodo approximately every six months. Finally, Fig. 4 presents the distribution of articles by province, based on the primary location of each outlet. This geographic breakdown shows the representativeness of the database across Canada, with 83.4% of sentence units in English and 16.6% in French. The national outlets—*The Globe and Mail*, the *National Post*, the *Toronto Star*, and the two public-broadcaster news sites—are omitted from the provincial breakdown because of their pan-Canadian scope; nevertheless, they account for 37.5% of all articles.

Step 3: Annotation protocol. The annotation framework constitutes the conceptual core of the CCF database. It comprises 65 categories organised into a hierarchical taxonomy designed to capture multiple dimensions of climate discourse at near-sentence granularity. The framework was developed through a combined theory-driven and data-driven procedure: it drew on established scholarship in framing, agenda-setting, climate communication, and social problems studies, and was refined inductively through iterative engagement with the corpus and extensive pretesting. The full taxonomy and operational definitions are reported in Supplementary Table S3, while Table 2 provides a compact overview of the framework's structure. The framework was organised around *Primary Categories* corresponding to broad analytical dimensions, and *Sub-categories* capturing internal distinctions within each *Primary Category* (Table 2).

Primary categories. The framework was organised around four primary analytic categories that structure the annotation protocol: *Main Frames*, *Actors/Messengers*, *Events*, and *Solutions* (Table 2). Drawing on climate communication research, these categories acknowledge that media coverage does more than describe the issue, since

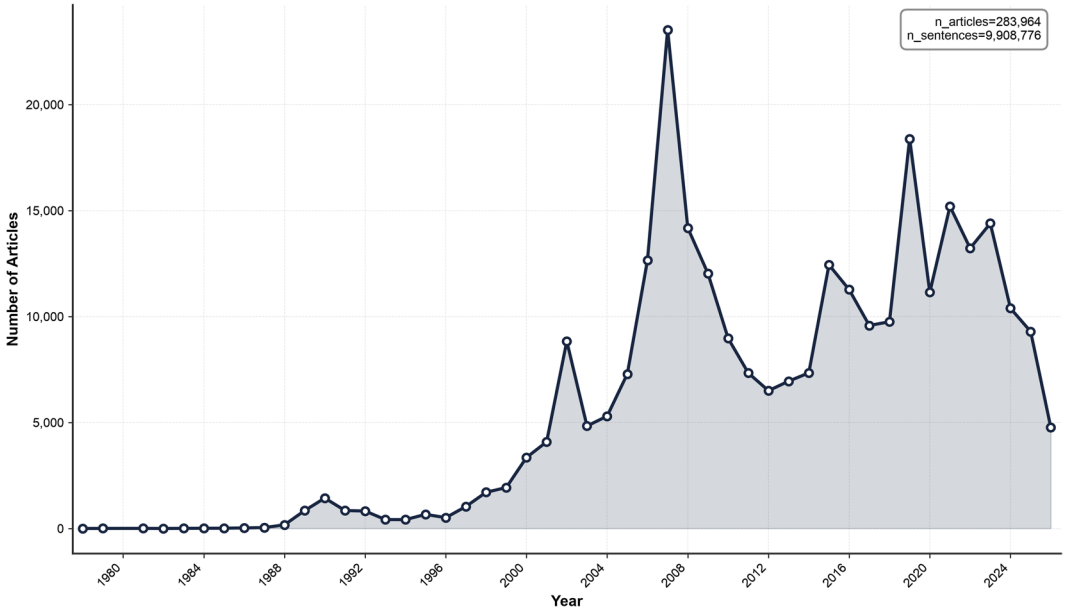

Fig. 3 Total number of articles per year.

it also selects the interpretive lens through which it is read and understood, the actors who are granted voice, the events that anchor attention, and the courses of action that are put forward as plausible.

First, for *Main Frames*, the authors built on the general framing tradition often associated with Entman[18], which encompasses four dimensions: problem definition, causal interpretation, moral evaluation, and treatment recommendations. In practice, however, applying this framework to climate discourse presents significant challenges. Individual statements frequently span multiple dimensions, and "treatments" (i.e., proposed policy responses) proved particularly difficult to identify consistently[27]. This difficulty was amplified by the nature of climate change as a "wicked problem," where proposed solutions remain subject to scientific uncertainty and political contestation[28]. For these reasons, many studies do not apply Entman's four dimensions strictly[29]. The CCF framework addressed these challenges by operationalising main frames around major societal domains in which climate change is repeatedly interpreted in media discourse: economic, security, public health, cultural, environmental, political, scientific, and justice spheres (Table 2 and Supplementary Table S3). This domain-based approach retained the core elements of problem definition, causal interpretation, moral evaluation, and treatment recommendations where they naturally emerged in the data.

Second, the framework treated *Solutions* as a distinct primary family rather than embedding them inside each main frame. This choice reflected both the empirical ambiguity of policy prescriptions in climate coverage and the substantive interest in distinguishing mitigation from adaptation, the latter having historically received less media attention[30]. Third, the *Actors/Messengers* family was grounded in the long-standing finding that *who* speaks about climate change matters for credibility, salience, and persuasion[31,32]. Finally, the *Events* family reflected agenda-setting scholarship showing that media attention is cyclical and often driven by external shocks and institutional moments (e.g., disasters, major conferences, reports)[33,34]. Together, these four primary categories provided a theory-informed structure for capturing how climate change is framed, who is authorised to speak, what triggers attention, and which actions are articulated (Table 2 and Supplementary Table S3).

Operationalisation: hierarchical taxonomy and sub-categories. Operationally, the framework implemented a hierarchical labelling logic that mirrors its conceptual structure (Table 2 and Supplementary Table S3). Each *Primary Category* contains (i) a detection-level category that captures whether the broader dimension is present in a two-sentence unit, and (ii) fine-grained sub-categories that specify which internal distinctions apply when the primary category is detected. This design enabled multi-label annotation: multiple frames, actors, events, and solutions could co-occur within the same unit, and the same applied to sub-categories. Within *Main Frames*, each domain frame (e.g., Economic, Health, Justice) was paired with sub-categories capturing recurring interpretive distinctions observed in the literature and in the corpus (Table 2 and Supplementary Table S3). Importantly, these sub-categories were not treated as mutually exclusive: a unit could simultaneously describe, for instance, the negative economic impacts of climate change and the economic benefits of climate action. The same principle applied to *Actors/Messengers*, where a single person could legitimately be coded as multiple expert types (e.g., a physician would be classified as having both medical and natural science expertise). For full category definitions, see Supplementary Table S3.

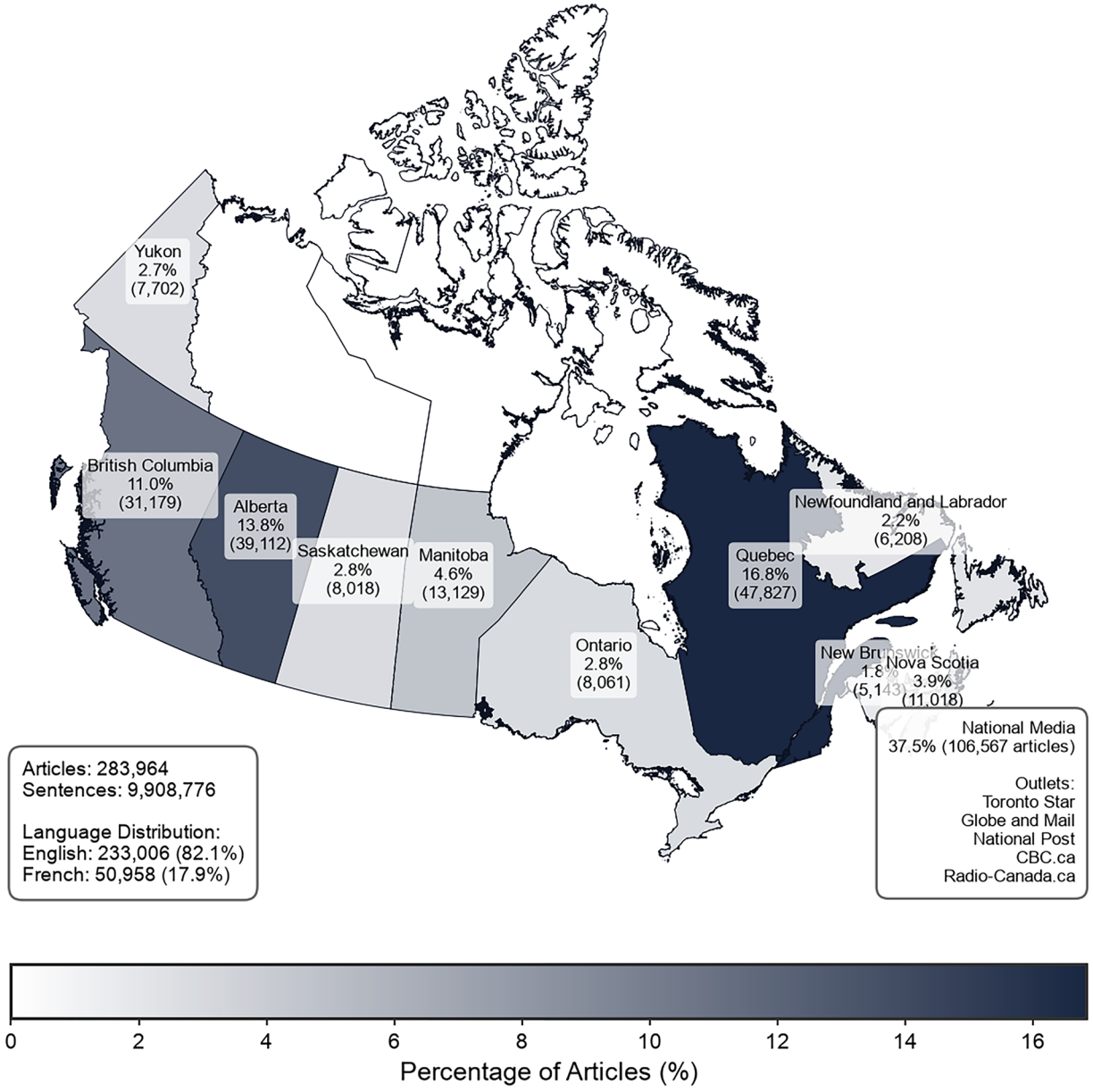

Fig. 4 Distribution of articles by province.

Other dimensions: context, tone, urgency, and entity extraction. In addition to the four *Primary Categories*, the framework included theoretically motivated dimensions that capture context and meaning-making beyond framing in the strict sense (Table 2 and Supplementary Table S3 and S11). A *Geographic Focus* category recorded whether coverage situated climate change in a Canadian context, reflecting concerns that climate change is often portrayed as geographically distant[35]. An *Emotional Tone* category captured positive, negative, or neutral valence, consistent with research showing that emotions shape how audiences interpret and respond to climate discourse[36,37]. An *Urgency to act* category captured alarmist or "code red" signals that communicate immediacy and crisis framing. Finally, Named Entity Recognition (NER) extracted and classified *Persons*, *Organizations*, and *Locations* into a JSON field, enabling downstream analyses such as cocitation networks and the mapping of epistemic authorities (Table 2; Supplementary Table S3).

Pretesting and iterative refinement. Finally, alongside these theoretical design choices, the framework was iteratively refined through inductive coding rounds. As sentences were annotated, emergent themes sometimes prompted targeted additions and clarifications to category definitions, followed by pretesting to ensure internal consistency and coder interpretability. This procedure strengthened the framework's construct validity and ensured that the final taxonomy remained both theoretically anchored and empirically adequate. The following real example from the database—an excerpt from a 2022 article—illustrates the corresponding manual annotation protocol:

Example (annotated sentence). *"In Scientific American, professor of Environmental Studies at Humboldt State University Sarah Jaquette Ray writes, 'Climate anxiety can operate like white fragility, sucking up all the oxygen in the room and devoting resources toward appeasing the dominant group.' But it would be a mistake to conclude that this term isn't applicable in the Global South, period."*

Tagged Dimensions	Tagged categories
Actors/Messengers	Scientists
Main frames	Justice (primary category) + North–South responsibility; Health (primary category)
Emotional tone	Negative

Phase 3: Machine learning. The development of the training data followed an iterative manual annotation protocol aligned with the annotation framework previously described. A single expert annotator (climate policy and communication) labelled sentences in sequential batches (first 1,000 sentences, then three additional batches of 1,000 each) for a total of 4,000 annotated samples. The choice of a single highly-trained expert during training, with an independent second coder reserved for the gold-standard validation, was made to maximise definitional consistency across the 4,000 training sentences. In framing analysis, where 65 conceptually overlapping categories must be applied to fine-grained sentence units, multi-coder training can introduce drift in operational definitions that exceeds the noise it removes; the more reliable protocol is therefore to stabilise the codebook with a single expert and to validate it against an independent coder on a held-out gold standard[38]. The batching served two explicit objectives that were only attainable with a single coder: (1) to iteratively improve and stabilise the annotation guidelines after each round; and (2) to progressively monitor and optimise model training to reach the highest attainable F1 score. This process culminated in a training-phase macro F1 score of 0.826—the unweighted mean of the per-category macro F1 scores obtained on each category's held-out validation split (see Table 3, and Supplementary Tables S4 and S11 for detailed metrics).

The sampling process drew equally from English and French language groups, with final annotation counts reaching 1,927 English and 2,073 French sentences. The training dataset was subsequently partitioned using stratified random sampling to create training (80%) and validation (20%) databases for each category and language. Supplementary Table S5 provides the complete distribution of training and validation samples for all annotation categories. The random sampling approach ensured broad coverage across the temporal span and diverse media sources. The annotation protocol was developed with detailed operational definitions for each of the 65 categories to ensure consistency.

The machine learning pipeline leveraged transformer architectures optimised for each language through a custom fork of the AugmentedSocialScientist library[39,40]. The fork, built on the original library of Do, Ollion, and Shen[39], added several functionalities that were central to ensuring robust machine-learning training (metric logging at every training epoch, best-model selection using a weighted F1 score ($0.98 \times F_1^{(1)} + 0.02 \times$ macro-F_1), and an automated reinforced training protocol for underperforming models, described below)[40]. English text processing employed BERT-base-cased, while French text used CamemBERT-base, both containing 110 million parameters. The English models being case-sensitive, their tokeniser must be instantiated with `do_lower_case=False`; a `tokenizer_config.json` enforcing this ships with every released English model directory. The training strategy implemented hierarchical classification, where detection models were trained first as binary classifiers on the full annotated dataset, followed by sub-classification models trained exclusively on manually annotated sentences that were positive for the corresponding primary category. This approach minimised false-positive propagation and allowed for specialised optimisation for each classification task.

All 128 models shared a single training configuration, reported here in full for reproducibility. Optimisation used AdamW ($\epsilon = 10^{-8}$) with a fixed random seed of 42 (Python, NumPy, and PyTorch generators) for shuffling and weight initialisation; the reinforced-phase sampler draws from PyTorch's default global generator, a known limitation recorded in the deposited configuration file. The training loss was a standard, unweighted two-class cross-entropy in the normal phase. The normal training phase ran for up to 20 epochs with a batch size of 32 and a learning rate of 5×10^{-5}; the reinforced phase (described below) ran for up to 20 additional epochs with a batch size of 64 and a learning rate of 5×10^{-6}. Input sequences were tokenised to a maximum length of 512 tokens. At every epoch the pipeline logged precision, recall, and F1 on the held-out validation split, and the best epoch was selected by the weighted criterion $0.98.F_1^{(1)} + 0.02.\text{macro}-F_1$ introduced above. All training ran on a single Apple Mac Studio (M2 Ultra, 128 GB unified memory); the exact static configuration and the per-model hyperparameters are archived in the deposit (`Database/Training_data/training_static_configuration.csv` and `training_hyperparameters_normalized.csv`) and summarised in Supplementary Table S11.

Models failing to achieve a positive-class F1 score of at least 0.70 during the training phase underwent an automated reinforced training protocol. This protocol was triggered for 45 of the 128 trained models (35%), concentrated in low-prevalence sub-frames—cultural, justice, and security—and in the fine-grained political and scientific sub-categories. The reinforced phase implemented weighted random sampling to oversample minority classes, doubled the batch size to 64, reduced the learning rate to 5×10^{-6}, and applied a class-weighted cross-entropy loss with a positive-class weight of 3.5. Training was extended for an additional 20 epochs, after which the best epochs of the normal and reinforced phases were compared on a combined criterion that gave 98% weight to the positive-class F1 and 2% to the macro F1, i.e. $0.98 \cdot F_1^{(1)} + 0.02 \cdot \text{macro}-F_1$. The reinforced epoch was retained as the final model for 23 of these 45 cases, while the remaining 22 reverted to their best normal-training epoch; averaged over the 23 reinforced models eventually retained, the positive-class F1

Main Category	N[a]	Main category detection *(in italics)* and sub-categories
Primary Categories: Main Frames (8 frames and their sub-categories)		
Economic	6	*Economic frame*, negative/positive impacts on the economy, economic disadvantages/benefits of climate action, carbon footprint of the economic sector
Health	3	*Health frame*, negative impacts on health, health co-benefits of climate action
Security	5	*Security frame*, climate refugees, conflict, post-disaster military assistance, disruption of military operations
Justice	6	*Justice frame*, winners/losers of climate action, differentiated responsibility, unequal vulnerability, unequal access to climate action, intergenerational justice
Political	5	*Political frame*, policy action, political debate, political positioning, public opinion data
Scientific	5	*Scientific frame*, scientific debate, popularisation or discovery, questioning of climate science, defense of climate science
Environmental	3	*Environmental frame*, loss of natural environments, loss of fauna and flora
Cultural	5	*Cultural frame*, artistic representation, disruption of cultural or sports events, loss of indigenous practices, carbon footprint of the cultural and sports sectors
Primary Categories: Actors, Events and Solutions (and their sub-categories)		
Actors/Messengers	9	*Any messenger*, scientists, politicians, activists, medical/economic/security/legal/cultural experts
Events	6	*Any event*, natural disasters, conferences, reports, elections, policy announcements
Solutions	3	*Any solution*, mitigation, adaptation
Emotional Tone	3	Positive (hope, optimism), negative (fear, anxiety), neutral/none
Geographic Focus	1	Canadian places, actors, data, and policies
Urgency to act	1	Conveys immediate danger, crisis, or "code red" urgency
Named Entities	3	Persons (PER), organizations (ORG), locations (LOC)
Total: 65 categories		

Table 2. The CCF annotation framework: 65 hierarchical categories for comprehensive climate discourse analysis. [a]*N* counts the main category detection (shown in italics) together with its sub-categories; e.g., the Economic frame: 1 main category detection + 5 sub-categories = 6. Flat categories (Emotional Tone, Named Entities) and single-detection categories (Geographic Focus, Urgency to act) have no separate main category detection.

	F1 (Class 1)		F1 (Class 0)		Macro F1	
Category	**EN**	**FR**	**EN**	**FR**	**EN**	**FR**
Main Frames						
Economic Frame	0.745	0.814	0.944	0.957	0.845	0.885
Health Frame	0.800	0.667	0.989	0.994	0.894	0.830
Security Frame	0.870	0.800	0.996	0.997	0.933	0.898
Justice Frame	0.719	0.717	0.975	0.981	0.847	0.849
Political Frame	0.808	0.774	0.897	0.888	0.853	0.831
Scientific Frame	0.784	0.702	0.953	0.962	0.869	0.832
Environmental Frame	0.842	0.625	0.989	0.980	0.915	0.802
Cultural Frame	0.773	0.833	0.986	0.993	0.879	0.913
Other Primary Categories: Actors, Events and Solutions						
Presence of Messengers	0.912	0.929	0.904	0.915	0.908	0.922
Presence of Events	0.794	0.819	0.935	0.932	0.865	0.876
Presence of Solutions	0.737	0.878	0.914	0.944	0.825	0.911
Mention of Canada	0.942	0.964	0.968	0.980	0.955	0.972
Urgency to act	0.591	0.649	0.975	0.984	0.783	0.816
Overall Average*	**0.794**	**0.782**	**0.956**	**0.962**	**0.875**	**0.872**
Combined (ALL)	**0.753**		**0.900**		**0.826**	

Table 3. Model training performance metrics for primary annotation categories. *Overall average across all active annotation categories (see Supplementary Table S4 for complete metrics).

improved by +0.25, lifting most of the affected categories above the deployment threshold. Two categories—*positive impacts of climate change on health* and *carbon footprint of the health sector*—could not be trained in either language because the manually annotated sentences contained no positive examples for them; they were consequently excluded from the deployed framework, leaving 65 deployed categories out of the 67 candidate categories listed in Supplementary Table S3. Two further categories—*post-disaster military assistance* and *disruption of military operations due to climate change*—could not be trained in English for the same reason but were successfully trained in French. These two remain part of the 65 deployed categories: they are annotated on French

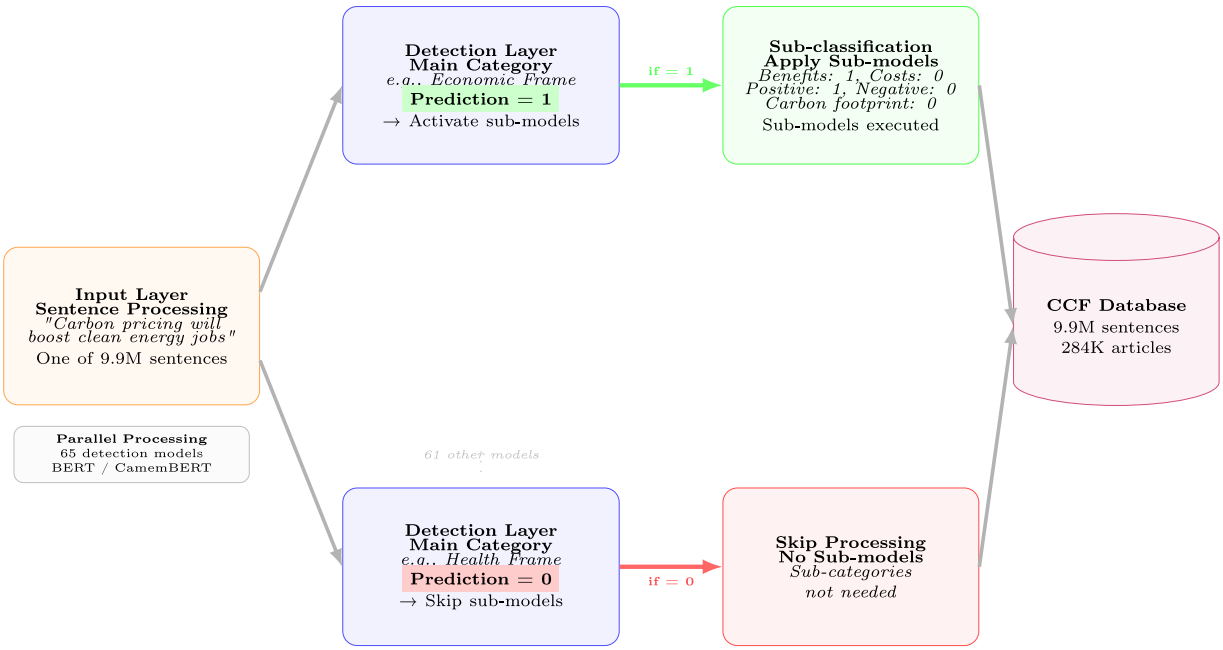

Fig. 5 Hierarchical annotation pipeline. The system processes each sentence through detection models that determine whether corresponding sub-category models should be applied, enabling efficient annotation of the entire corpus.

sentences only, their English annotations are left as `NULL` by construction (and coalesced to zero in the article-level aggregates), and both are flagged Tier C in the reliability lookup (Supplementary Table S10). The deployed suite therefore counts $65 \times 2 - 2 = 128$ models. The complete per-model training provenance, including which models triggered reinforcement and which retained the reinforced epoch, is provided in Supplementary Table S11.

Phase 4: Validation and deployment. Phase 4 chained the full deployment of the trained models (Step 6) with the two validation steps (Steps 7 and 8) that closed the methodological pipeline. Step 6 is described immediately below; the design of the validation steps (sampling protocol, gold-standard construction, model performance metrics, and inter-coder reliability) is described separately and in full in the Technical Validation section below.

Step 6: Full deployment. Once the per-category models had been finalised, all trained models were applied to annotate the complete CCF Database of 9.9 million sentences across 283,964 articles. A hierarchical classification strategy was implemented that directly mirrored the annotation protocol's structure of primary categories and their associated sub-categories, as illustrated in Fig. 5.

The hierarchical approach reflected the conceptual organisation of the annotation framework with eleven primary detection categories, each with specific sub-categories. The eight main frames (*Economic*, *Health*, *Security*, *Justice*, *Political*, *Scientific*, *Environmental*, and *Cultural Frame*) functioned as primary categories with their internal distinctions—for example, *Economic Frame*, when positive (=1), triggered five economic sub-models (*negative/positive economic impacts of climate change*, *economic disadvantages/benefits of climate action*, *carbon footprint of the economic sector*), while a negative detection (=0) bypassed these entirely.

Similarly, the three other primary categories followed this conditional logic: *Actors/Messengers Detection* activated nine actor-type sub-models (*Health Expert*, *Economic Expert*, *Security Expert*, *Legal Expert*, *Cultural Expert*, *Natural Scientist*, *Social Scientist*, *Activist*, *Public Official*), *Event Detection* triggered eight event-type classifications (*Extreme Weather Event*, *Meeting/Conference*, *Publication*, *Election*, *Policy Announcement*, *Judiciary Decision*, *Cultural Event*, *Protest*), and *Solutions Detection* applied two solution sub-categories (*Mitigation Strategy*, *Adaptation Strategy*). The deployment also processed standalone primary categories without hierarchical structure—*Emotional Tone* (*Positive Emotion*/*Negative Emotion*/*Neutral Emotion*), *Geographic Focus* (*Canadian Context*), *Urgency/Alarmism*.

In addition to the custom-trained classification models, three pre-trained Named Entity Recognition (NER) models were employed, selected on the basis of empirical evaluation on the CCF data. A hybrid approach tailored to each language was implemented. For English texts, the publicly available `dslim/bert-base-NER` model (hosted on the Hugging Face Hub) was used for all three entity types: persons, organisations, and locations. For French texts, person entities were extracted with the spaCy `fr_core_news_lg` pipeline, while organisations and locations were extracted with the `Jean-Baptiste/camembert-ner` model (also hosted on the Hugging Face Hub). This hybrid strategy was chosen to leverage the respective strengths of each model: spaCy's superior performance on French person names, and CamemBERT-NER's robust identification

of French organisational and geographical entities. Per-language performance metrics for the three NER models are reported in Supplementary Table S6.

Phase 5: Database enrichment. The deployment described above produced the two original tables of the CCF Database (`CCF_full_data` and `CCF_processed_data`). The database was then further enriched with four additional tables and a complete index structure so that researchers could interrogate the corpus without rewriting the same article-level aggregations on every project. The enrichment procedure was deterministic and idempotent: it ran end-to-end from a freshly restored database dump, and each step (index provisioning, article-level aggregates, article-level entities, sentence embeddings, and reliability tiers) is described in detail below.

Indexes. The first enrichment step provisioned twenty indexes on the two original tables: unique primary keys on `(doc_id)` for `CCF_full_data` and `(doc_id, sentence_id)` for `CCF_processed_data`; routine B-trees on the columns researchers query most frequently (`date`, `media`, `language`); and partial B-trees on `doc_id` restricted to the rows where one of the eleven primary detection flags is positive (*economic_frame*, *health_frame*, *security_frame*, *justice_frame*, *political_frame*, *scientific_frame*, *environmental_frame*, *cultural_frame*, *messenger*, *event*, *solution*). The partial indexes accelerated the most common analytical workflow—retrieving the subset of articles or sentences in which a given frame is positive—without paying the storage cost of a full B-tree on each binary column.

Article-level aggregates. The `CCF_article_aggregates` table was built with one row per `doc_id`; for each of the 65 annotation categories, it stored the article-level proportion (`prop_X`, in [0, 1]) of sentences positive for category *X*, computed as the sum of the binary flag divided by the article's sentence count. The table also recorded the article's sentence count overall and by language (`n_sentences`, `n_sentences_en`, `n_sentences_fr`), the article's top frame (`top_frame`, the main frame with the highest proportion in the article), the proportion of that frame (`top_frame_prop`), the Shannon entropy of the eight main-frame proportions (`entropy_frames`, measuring how concentrated or diffuse the article's framing is), and the number of distinct messenger, event, and solution sub-categories present in the article (`n_unique_messenger_subs`, `n_unique_event_subs`, `n_unique_solution_subs`). Sub-categories whose parent primary detection was negative were coalesced to zero, so a `prop_X` reading of 0 unambiguously means "no sentence in this article is positive for category *X*", regardless of whether the sub-category was hierarchically evaluated.

Article-level named entities. The `CCF_article_entities` table consolidated the sentence-level Named Entity Recognition output (the `ner_entities` JSON column in `CCF_processed_data`) into one row per article. Three JSONB arrays (`entities_per`, `entities_org`, `entities_loc`) stored the deduplicated set of persons, organisations, and locations mentioned anywhere in the article; three counters (`n_unique_per`, `n_unique_org`, `n_unique_loc`) recorded their cardinalities; and three further counters reported how many sentences of the article contained at least one entity of each type. The presence of an entity in this rollup does not by itself disambiguate whether the entity is the speaker, the subject, or a passing reference; users wishing to recover speaker–statement attribution should combine the entity rollup with the sentence-level `messenger` family of annotations or with downstream attribution models that are out of scope for this database.

Sentence embeddings. The `CCF_sentence_embeddings` table was populated, for every sentence in `CCF_processed_data` and for the title of every article (encoded as `sentence_id` = 0), with the dense vector representation produced by the BAAI `bge-m3` multilingual encoder[41]. `bge-m3` was selected for three reasons: (i) it natively supports the 100+ languages relevant to bilingual Canadian coverage and is reported to perform competitively on French–English cross-lingual tasks; (ii) it produces 1024-dimensional dense vectors that combine acceptable retrieval quality with manageable storage overhead; and (iii) it has been stable and widely adopted since its January 2024 release, which is favourable for the long-term reproducibility expected of a published dataset. Vectors were L2-normalised at encoding time, so cosine similarity reduces to a dot product and can be queried directly through the `pgvector<=>` operator. Storage used the `halfvec(1024)` type (2 bytes per dimension, identical to the upstream `float16` representation), and an HNSW index on the cosine operator class was built to deliver sub-second approximate top-*k* retrieval over the 9.9 million-row table.

Reliability tiers. The last enrichment step recorded, for each annotation category, a recommendation on how confidently it can be used in downstream analyses. The recommendation was summarised as a discrete *reliability tier* derived from the two evaluation criteria reported in full in the Technical Validation section below—the validation macro F_1 score against the gold standard, and the blind-phase Cohen's κ between the expert annotator and the independent second coder. The `CCF_reliability_tiers` table was built as a 65-row lookup classifying each category into one of three tiers: Tier A (*research-grade*) required $F_1 \geq 0.80$ *and* $\kappa \geq 0.60$; Tier C (*exploratory*) corresponded to $F_1 < 0.60$ *or* $\kappa < 0.40$; Tier B (*use with caution*) captured everything in between. The table also flagged the categories affected by *prevalence-induced reliability deflation*—that is, categories for which $F_1 \geq 0.70$ but $\kappa < 0.40$, where the chance-corrected agreement is depressed by very low prevalence rather than by systematic disagreement between coders[42,43]—and the two categories without an English model (annotated on French sentences only). Because the annotation tables store one *column* per category while the tier table stores one *row* per category, the two are not combined through a relational join: `CCF_reliability_tiers` is a lookup that tells the analyst *which columns* of `CCF_processed_data` or `CCF_article_aggregates` to keep. Selecting the codes where `tier_overall = 'A'` and restricting the analysis to the corresponding columns limits it to the 27 categories that the authors recommend without reservation;

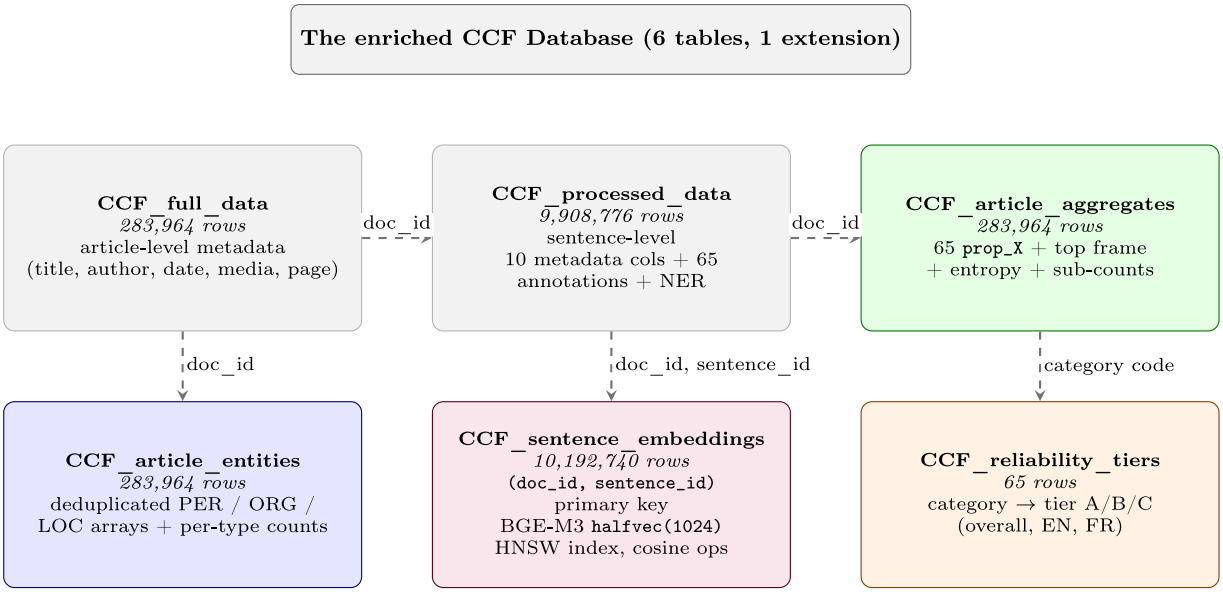

Fig. 6 Schema of the enriched CCF Database. Dashed arrows indicate the join paths between tables; column types and full schemas are documented in Supplementary Table S12. `CCF_full_data` and `CCF_processed_data` are the original sentence-level resource; the three coloured groups are the article-level rollups (green), the reliability lookup (orange), and the embedding table (purple).

a worked example is given in the Usage Notes section, and the full tier assignment by language is provided in Supplementary Table S10.

A diagram of the enriched database, together with the join paths that connect the six tables, appears in the Data Records section (Fig. 6). The complete enrichment procedure ran end-to-end in approximately 2.5 hours on a single Apple Mac Studio with the M2 Ultra chip and 128 GB of unified memory.

Data Records

Before working with any table of the deposit, users are advised to consult the `CCF_reliability_tiers` lookup described in this section: it assigns each of the 65 annotation categories an A/B/C reliability tier, and the authors recommend restricting substantive conclusions to Tier A categories (27 of 65) unless the analysis explicitly accounts for the uncertainty of lower tiers. The database is versioned: new releases are published on Zenodo approximately every six months, as the CCF Observatory—the continuous extraction, annotation, and enrichment pipeline that now maintains the corpus—ingests newly published articles. The concept DOIs (https://doi.org/10.5281/zenodo.20346363 for the PostgreSQL edition, https://doi.org/10.5281/zenodo.20346372 for the Parquet mirror) always resolve to the latest release, while version DOIs remain permanently resolvable so that any analysis can be reproduced against the exact release it used. This paper describes version 2.1.0.

Overview of the deposit. The CCF Database is deposited on Zenodo under a CC-BY 4.0 licence in two complementary editions sharing identical schemas. The canonical PostgreSQL edition (DOI https://doi.org/10.5281/zenodo.22019287 [44]) ships as a single tarball (`CCF_Database.tar`, ≈ 40 GB) wrapping a `pg_dump` *directory* archive of the six relational tables together with all B-tree, partial, and HNSW indexes; restoration requires PostgreSQL 16 or 17 with the `pgvector` extension (≥0.8.2 for `halfvec` support) and the deposit's README provides the exact three-line restore command (`tar -xf`, `createdb`, `pg_restore -j 8`). A sister column-oriented Apache Parquet mirror (DOI 10.5281/zenodo.22019288 [45], one ZSTD-compressed file per table, ≈ 17.5 GB) is provided for users without a PostgreSQL backend; the `halfvec(1024)` embedding column is materialised as a 1024-element `LIST<FLOAT>` and the JSONB entity arrays are serialised as UTF-8 JSON strings, so each Parquet file can be read directly with `pandas`, `polars`, R/`arrow`, DuckDB, or Spark. The companion code and documentation—the full annotation pipeline (`Scripts/Annotation/`), the enrichment and export scripts (`Scripts/Database_creation/`), and the canonical CSVs underlying the supplementary tables (`Database/Training_data/`)—are archived in a dedicated, versioned Zenodo deposit (DOI 10.5281/zenodo.22019289 [46]) and, for convenience, also bundled with both data editions as `ccf_code_and_paper.tar.gz`; a development mirror is maintained on GitHub (`antoinelemor/CCF-canadian-climate-framing`, referenced in the Code Availability statement).

Because Canadian newspaper articles are protected by copyright, the deposit does not redistribute the raw sentence text or full article bodies of the corpus at large. What is shared is everything else: article-level metadata (titles, authors, publication dates, newspapers, page numbers), sentence-level identifiers, the 65 binary annotations, the named-entity extractions, the article-level aggregates and entity rollups, the per-category reliability tiers, and the BGE-M3 sentence embeddings. One deliberate, clearly bounded exception is made so that the single-coder training protocol and the gold standard can be independently audited: the code and documentation deposit releases the expert-coded sentences themselves—the ~ 4,000 training sentences (`Database/`

Table	Rows	Granularity	What it stores
CCF_full_data	283,964	Article	Bibliographic metadata for every article in the corpus.
CCF_processed_data	9,908,776	Sentence	The 65 sentence-level binary annotations and the per-sentence NER output.
CCF_article_aggregates	283,964	Article	Per-article proportions, top frame, framing entropy and sub-category counts.
CCF_article_entities	283,964	Article	Deduplicated JSONB arrays of PER / ORG / LOC mentions, with per-type cardinalities and per-type sentence counts.
CCF_reliability_tiers	65	Category	Per-category reliability tier (A / B / C) with the underlying F_1 and κ values.
CCF_sentence_embeddings	10,192,740	Sentence + title	L2-normalised BGE-M3 1024-dimensional vectors, queryable through `pgvector`.

Table 4. Overview of the six tables in the deposited CCF Database. The `pgvector` extension underpins the `CCF_sentence_embeddings` table.

Column	Type	Definition
`doc_id`	BIGINT (PK)	Unique article identifier; foreign-key target for every other table.
`news_type`	TEXT	Editorial category: news, opinion, editorial, etc.
`title`	TEXT	Article headline.
`author`	TEXT	Author byline when available.
`media`	TEXT	Newspaper name.
`words_count`	BIGINT	Total word count of the article.
`date`	DATE	Publication date.
`language`	TEXT	"EN" or "FR".
`page_number`	TEXT	Print page (free text, e.g. "A1" or "B3").

Table 5. Columns of `CCF_full_data`.

`Training_data/manual_annotations_JSONL/Annotated_sentences.jsonl`, together with the per-category train/validation splits under `annotation_bases/`) and the 1,000-sentence double-coded gold standard (`blind_verification.jsonl` and `consensus_gold_standard.jsonl`). These files contain short excerpts of two sentences each, quoted for the purpose of research and validation with full bibliographic attribution through the corpus metadata; this reproduction of brief extracts falls under the fair-dealing exception for research and private study of the Canadian *Copyright Act* (s. 29) and does not substitute for the original works. Each article in the corpus is identified by its bibliographic coordinates (`media`, `date`, `title`, `author`, `page_number`), which is the standard bibliographic reference for Canadian newspapers and is sufficient for any researcher with institutional access to Factiva, Eureka.cc, or ProQuest Canadian Major Dailies to retrieve the original text. No DOIs exist for the historical articles covered by the corpus.

The database is organised around two original tables produced by the annotation pipeline (`CCF_full_data`, `CCF_processed_data`) and four enrichment tables that pre-compute the queries most frequently issued by analysts (article-level aggregates, article-level entities, reliability tiers, sentence embeddings). The schema is summarised in Table 4 and depicted graphically in Fig. 6; the join paths in that diagram point to the recommended foreign keys whenever a query needs to combine two tables.

`CCF_full_data` – article-level metadata. This table is the canonical inventory of articles in the corpus and the natural entry point for any analysis that operates at the article level. Each of its 283,964 rows describes one unique article and is uniquely keyed by `doc_id`; every other table in the database joins back to `CCF_full_data` through this identifier. The columns are summarised in Table 5.

`CCF_processed_data` – sentence-level annotations. This is the workhorse table of the database. It contains one row per two-sentence analytical unit, so 9,908,776 rows in total, and every annotation produced by the machine-learning pipeline lives here. The same ten metadata columns as `CCF_full_data` are denormalised on every row so that sentence-level queries can run without joining back to the article table whenever the analyst only needs the date, the newspaper, or the language. On top of that, the table stores the 65 binary annotation columns named exactly as in Supplementary Table S3, and one JSON column (`ner_entities`) holding the per-sentence Named Entity Recognition output. Sub-category columns are NULL when their parent primary detection is negative; researchers who want to treat "absent" as equivalent to "not detected" should coalesce NULL to zero, which is exactly the convention adopted by `CCF_article_aggregates`. The structural columns are summarised in Table 6.

`CCF_article_aggregates` – article-level rollup. `CCF_article_aggregates` is the recommended starting point for most article-level analyses—bearing in mind the general recommendation that opens this section: its 65 `prop_*` columns inherit the reliability of the underlying sentence-level classifiers, so analyses should prioritise the columns whose category is Tier A in `CCF_reliability_tiers`. It collapses the 9.9 million sentence-level rows into one row per article (283,964 rows) and pre-computes the quantities that almost every project needs: the proportion of sentences positive for each of the 65 categories, the top frame, a measure

Column	Type	Definition
doc_id	BIGINT	Foreign key to CCF_full_data.
sentence_id	BIGINT	Position of the sentence within its article (starting at 1). sentence_id = 0 is reserved for the article title in the embeddings table.
(metadata)	various	Ten metadata columns duplicated from CCF_full_data for query convenience: news_type, title, author, media, words_count, date, language, page_number.
65 annotation columns	INTEGER (0/1, nullable)	One binary flag per category from Supplementary Table S3. Sub-categories are NULL when the parent primary detection is negative.
ner_entities	TEXT (JSON)	Per-sentence NER output in the format {"PER": [..], "ORG": [..], "LOC": [..]}.

Table 6. Structural columns of CCF_processed_data. The 65 binary annotation columns are defined in Supplementary Table S3.

Column	Type	Definition
doc_id	BIGINT (PK)	Foreign key to CCF_full_data.
n_sentences	INTEGER	Total number of two-sentence analytical units in the article.
n_sentences_en, n_sentences_fr	INTEGER	Same count restricted to English / French sentences.
65 prop_X columns	DOUBLE PRECISION ([0, 1])	Article-level proportion of sentences positive for category *X*. Sub-categories with a negative parent detection are coalesced to zero.
top_frame	TEXT	Main frame with the highest prop_X; NULL when no frame is positive.
top_frame_prop	DOUBLE PRECISION	Proportion of the top frame in the article.
entropy_frames	DOUBLE PRECISION	Shannon entropy (natural base) of the eight main-frame proportions; 0 when no main frame is positive.
n_unique_messenger_subs	SMALLINT	Number of distinct messenger sub-categories present in the article (0–9).
n_unique_event_subs	SMALLINT	Same for event sub-categories (0–8).
n_unique_solution_subs	SMALLINT	Same for solution sub-categories (0–2).

Table 7. Columns of CCF_article_aggregates.

of how concentrated or diffuse the framing of the article is, and the number of distinct sub-categories activated. Because these aggregates are computed once at enrichment time and indexed, a sentence-level GROUP BY on the 9.9-million-row sentence table is rarely necessary once the aggregates are available. The columns are summarised in Table 7.

CCF_article_entities – article-level named-entity rollup. CCF_article_entities consolidates the sentence-level Named Entity Recognition output (the ner_entities JSON column in CCF_processed_data) into one row per article. The table is designed to answer the recurring analytical question *which entities does an article mention?* without forcing the analyst to parse the JSON for every sentence. It contains 283,964 rows and its structure is summarised in Table 8.

CCF_reliability_tiers – per-category quality lookup. CCF_reliability_tiers provides a single-row-per-category recommendation on how confidently each of the 65 annotation categories may be used in downstream analyses. The tier assignment combines validation F_1 and blind-phase Cohen's κ following the rules defined in Section ??; the underlying metrics and flags are kept in the same row so that researchers can refine the tier rule for their own use case if they wish. This table is a *lookup*, not a join target: annotation categories appear as columns (not rows) in CCF_processed_data and CCF_article_aggregates, so the tier table is used to decide *which columns* to retain—select the code values where tier_overall = 'A' and restrict the analysis to those 27 columns (worked example in the Usage Notes). Columns are listed in Table 9.

CCF_sentence_embeddings – dense vector representations. CCF_sentence_embeddings enables semantic retrieval and similarity-based analyses across the whole corpus. It holds one row per sentence (and one extra row per article title, encoded with sentence_id = 0), for a total of 10,192,740 rows. Each row stores a 1024-dimensional BAAI/bge-m3 vector that is L2-normalised at encoding time, so cosine similarity reduces to a dot product and can be queried directly through the pgvector<=> operator. Storage uses the halfvec(1024) type (2 bytes per dimension, identical to the upstream float16 representation), and an HNSW index on the cosine operator class delivers sub-second approximate top-*k* retrieval over the full table. Columns are summarised in Table 10.

The complete data dictionary, including every column type and the full set of join paths, is provided in Supplementary Table S12.

Column	Type	Definition
doc_id	BIGINT (PK)	Foreign key to CCF_full_data.
entities_per	JSONB	Deduplicated array of persons mentioned in the article.
entities_org	JSONB	Deduplicated array of organisations.
entities_loc	JSONB	Deduplicated array of locations.
n_unique_per, n_unique_org, n_unique_loc	INTEGER	Cardinalities of the three arrays above.
n_sentences_with_per, n_sentences_with_org, n_sentences_with_loc	INTEGER	Number of sentences in the article that mention at least one entity of the corresponding type.

Table 8. Columns of CCF_article_entities.

Column	Type	Definition
number	SMALLINT (PK)	Sequential index of the 65 deployed categories, in canonical framework order. Note that Supplementary Table S3 enumerates all 67 *candidate* categories (including the two excluded ones) and therefore numbers differently: categories should always be matched on code, never on number. S3 reports each category's code to make this crosswalk unambiguous.
category	TEXT	Readable category name.
code	TEXT (UNIQUE)	Column code as used in CCF_processed_data.
tier_overall	CHAR(1)	Tier on the combined corpus: "A", "B", or "C".
tier_en	CHAR(1)	Tier on the English subset.
tier_fr	CHAR(1)	Tier on the French subset.
paradox_kappa	BOOLEAN	True if the category exhibits prevalence-induced reliability deflation ($F_1 \geq 0.70$ and $\kappa < 0.40$).
exclusion_reason	TEXT	"insufficient_training_data" for the two categories that could not be trained, NULL otherwise.
f1_macro_all, f1_macro_en, f1_macro_fr	DOUBLE PRECISION	Validation macro F_1 score on the gold standard, overall and per language.
support_all, support_en, support_fr	INTEGER	Number of positive (class 1) validation examples.
cohens_kappa_blind_phase, gwet_ac1_blind_phase, krippendorff_alpha_blind_phase	DOUBLE PRECISION	Inter-coder agreement on the blind phase (sentences 601–1000).
n_blind_phase	INTEGER	Sentence count of the blind phase.
prev_coder1_full, prev_coder2_full	DOUBLE PRECISION	Prevalence (proportion of positives) per coder on the full 1,000-sentence sample.

Table 9. Columns of CCF_reliability_tiers.

Column	Type	Definition
doc_id	BIGINT (PK, with sentence_id)	Foreign key to CCF_full_data.
sentence_id	INTEGER (PK)	Sentence position within the article. sentence_id = 0 is reserved for the article title.
embedding	halfvec(1024)	L2-normalised BGE-M3 vector. Cosine distance is computed via the pgvector<=> operator; an HNSW index on the cosine operator class is provisioned by the enrichment pipeline.

Table 10. Columns of CCF_sentence_embeddings.

Data Overview

This section consolidates three reproducible summaries of the deposited CCF Database. Every figure and table is generated by the script Results/Scripts/3_data_overview.py, which pulls the underlying numbers directly from the deposited PostgreSQL tables: no manual transcription is involved, and re-running the script regenerates the artefacts byte-for-byte.

Article-level descriptive statistics. Table 11 reports the corpus size (article count, two-sentence analytical units, and per-article word statistics) computed on CCF_full_data and CCF_processed_data, and the mean per-article share of sentences positive for each annotation category, computed on the prop_X columns of CCF_article_aggregates. The left block lists the eight frames together with the percentage of articles for which each frame is the top_frame; the right block lists the primary categories (messenger, event, solution, Canadian context, urgency) and the three tones.

Temporal coverage of the main frames. Figure 7 plots, for every calendar year between 1980 and 2026, the mean of each prop_X main-frame column of CCF_article_aggregates, grouped by the year extracted from CCF_full_data.date. No smoothing or regression is applied; each point is a yearly mean.

Main frames			Primary categories & tone	
	Mean share (%)	Top frame (%)		Mean share (%)
Political frame	35.4	53.7	Messenger present	52.9
Economic frame	17.6	21.0	Event present	23.0
Scientific framew	7.9	9.4	Solution present	27.2
Environmental frame	3.6	4.5	Canadian context	46.7
Justice frame	3.6	1.4	Urgency to act	2.6
Cultural frame	3.7	7.2	Positive tone	9.6
Health frame	1.7	1.7	Negative tone	39.6
Security frame	0.5	0.4	Neutral tone	62.7

Table 11. Article-level descriptive statistics of the deposited CCF Database. Mean share is the average per-article proportion of sentences positive for the category; the top-frame column reports the percentage of articles for which the frame has the highest share among the eight frames. Computed from `CCF_full_data`, `CCF_processed_data`, and `CCF_article_aggregates`. *Corpus:* 283,964 articles, 9,908,776 two-sentence units; 736 words/article on average (SD = 536; median = 680).

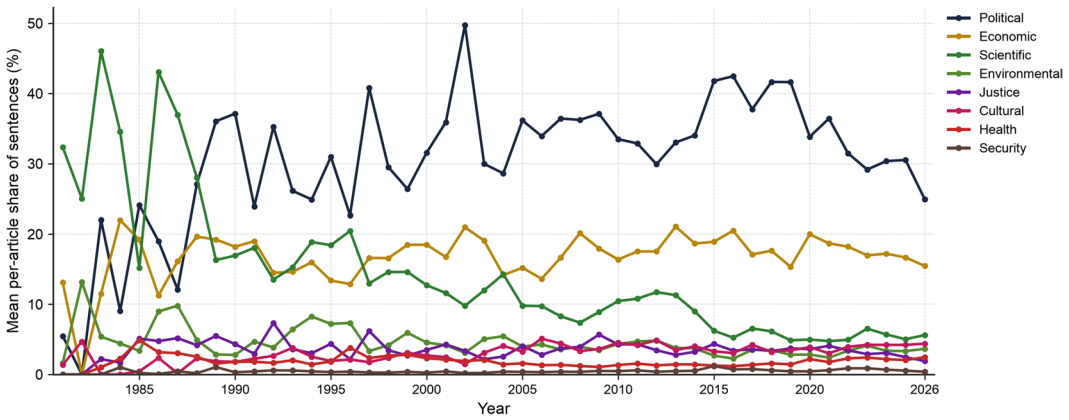

Fig. 7 Mean per-article share of each main frame, year by year, 1980–2026. Each point is the mean of the corresponding `prop_X` column of `CCF_article_aggregates` across the articles published in that year, joined with `CCF_full_data` on `doc_id`. The pre-1988 portion of the curve is computed on fewer than fifty articles per year.

The earliest years of the corpus carry very few articles—fewer than fifty per year before 1988 and fewer than five hundred per year before 1990—so the pre-1988 portion of the curve is computed on small samples and should be read with that caveat.

Named-entity coverage. Table 12 reports the ten most frequently mentioned entities of each NER type. Entities are extracted from the JSONB arrays `entities_per`, `entities_org`, and `entities_loc` of `CCF_article_entities`, expanded with `jsonb_array_elements_text`, filtered to remove pure punctuation and one-character tokens, and counted in articles—an entity is therefore counted once per article regardless of how many sentences mention it. Variants such as *Trudeau* vs. *Justin Trudeau* are kept separate because the underlying NER pipeline emits them as distinct strings; analysts who need canonicalised entities can apply their own resolution step on top of these arrays.

Technical Validation

Technical validation of the CCF database and the trained models proceeded in three steps. First, a stratified validation set was constructed in line with established best practices for multi-label classification. Second, an expert annotator in climate change politics produced the reference annotations that define the gold standard used for all model evaluation. Third, an independent second annotator replicated the coding following a structured protocol comprising a training phase (sentences 1-600) and a blind coding phase (sentences 601-1000), enabling assessment of inter-coder reliability. The 1,000-sentence gold standard is strictly disjoint from the 4,000-sentence training/validation pool used in Phase 3. We explicitly excluded all sentence identifiers present in the training data before drawing the stratified sample, so that no test sentence was ever seen by a trained model at training time.

Construction of the stratified validation set. The validation set comprised 1,000 sentences (500 per language) extracted from the fully annotated database using a root-inverse probability weighting scheme combined with strict structural constraints, summarised in Fig. 8. Each candidate sentence *i* received a sampling weight

	Persons (PER)		Organisations (ORG)		Locations (LOC)	
	Entity	Articles	Entity	Articles	Entity	Articles
1	Justin Trudeau	24,996	NDP	21,366	Canada	136,896
2	Stephen Harper	24,174	United Nations	13,632	U.S.	59,003
3	Trudeau	20,909	UN	13,158	Ottawa	49,617
4	Harper	20,069	OTTAWA	9,940	Alberta	47,836
5	Donald Trump	13,486	Star	9,836	United States	38,626
6	Barack Obama	11,936	European Union	8,011	Ontario	36,709
7	Trump	10,417	Environment Canada	7,967	Toronto	35,531
8	Obama	8,660	House of Commons	7,519	Quebec	23,264
9	George W. Bush	7,282	National Post	6,969	U.S	22,501
10	Bush	6,885	Senate	6,953	Québec	21,761

Table 12. Ten most frequently mentioned entities of each type in the deposited CCF Database. "Articles" is the number of articles in which the entity appears at least once. Computed from the JSONB arrays of `CCF_article_entities` (`entities_per`, `entities_org`, `entities_loc`).

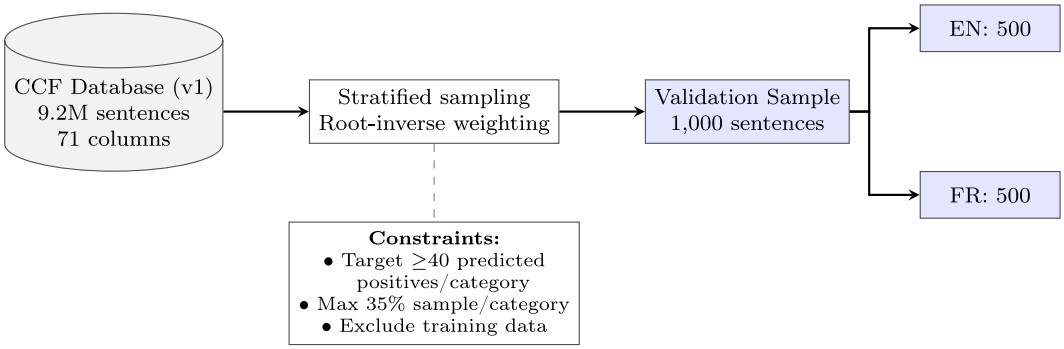

Fig. 8 Stratified sampling procedure for final validation. Root-inverse probability weighting ensures balanced representation across all 65 annotation categories.

$$w_i \propto \sum_{j \in C(i)} \frac{1}{\sqrt{f_j}}, \tag{1}$$

where $C(i)$ denotes the set of categories that are positive for sentence i and f_j denotes the corpus-wide prevalence of category j. The square root in the denominator deliberately moderated the influence of extremely rare categories, while still privileging them over common ones. Concretely, a sentence annotated with both *Economic Frame* (15.4% prevalence in the corpus) and *Loss of Indigenous Practices* (0.28% prevalence) received a weight of

$$\frac{1}{\sqrt{0.154}} + \frac{1}{\sqrt{0.0028}} = 2.55 + 18.90 = 21.45,$$

which was roughly eight times the weight of a sentence containing only *Economic Frame*, and ensured that under-represented categories nevertheless reached the validation set.

Two structural constraints were then applied on top of this weighted draw. First, a floor of 40 positive examples per category was targeted at sampling time—counted on the *model-predicted* labels, since sampling necessarily precedes the expert coding of the gold standard—so that each category would enter the validation set with as much statistical support as its corpus prevalence allowed. This target could not always be met for rare categories, and even where 40 predicted positives were drawn, the number of *expert-confirmed* positives can be substantially lower: the effective support ranges from 2 (*Disruption of military operations*, whose corpus prevalence is on the order of 10^{-5}) to several hundred, and 23 of the 65 categories fall below 40 confirmed positives. The per-category support is reported in Supplementary Table S7 so that readers can weigh each per-category estimate accordingly. Second, a 35% maximum allocation per category was imposed to prevent any single prevalent category from monopolising the validation set; without this ceiling, *Actors/Messengers Detection* alone, present in 48.8% of database sentences, could plausibly have dominated the entire sample and left insufficient space for the rarer categories.

Expert gold-standard annotation and model performance. All 1,000 sentences in the validation set were manually annotated by an expert annotator specialising in climate change politics. These expert annotations defined the gold standard against which all model performance metrics reported in this section were computed.

Language	F1 Macro	F1 Micro	F1 Weighted
English (EN)	0.869	0.909	0.911
French (FR)	0.864	0.902	0.905
Combined (ALL)	**0.866**	**0.905**	**0.908**

Table 13. Overall validation performance metrics.

Model performance against the gold standard is reported in Table 13 and is summarised through two complementary quantities, both reproducible from the deposited per-category metrics (`Database/Training_data/final_annotation_metrics_normalized.csv`, Supplementary Table S7). The *pooled* macro F1 aggregates the predictions of all 65 categories over the 1,000 gold-standard sentences into a single confusion matrix before computing the score: it reaches 0.866 overall (0.869 for English, 0.864 for French) and measures the quality of the annotation layer as a whole, weighting every annotation decision equally. The *mean per-category* macro F1—the unweighted average of the 65 per-category macro F1 scores, i.e. the same aggregation as the training-phase figure of 0.826 reported in Phase 3—is 0.773. The like-for-like comparison is therefore 0.826 in training versus 0.773 on the gold standard: the ordinary out-of-sample decrease expected when models face an independent, disjoint sample, concentrated in the low-support categories whose estimates are noisiest. Two design features of the validation set should be kept in mind when reading the per-category values: the sampling deliberately over-represents sentences positive for low-prevalence categories (the root-inverse weighting above), which are precisely the categories targeted by the reinforced training protocol, so post-reinforcement gains are more visible here than in the training-time average; and the per-category support varies widely (see above), so estimates for low-support categories carry wide uncertainty. As detailed in Supplementary Table S7, primary detection categories achieved the highest performance, with *Canadian Context* (F1=0.982) and *Actors/Messengers* (F1=0.961) showing near-perfect classification. The frames maintained strong performance despite their semantic complexity, with *Scientific Frame* (F1=0.890) and *Environmental Frame* (F1=0.871) leading this category. The hierarchical classification strategy proved particularly effective, as evidenced by consistently higher recall scores for detection categories (mean recall=0.891) that triggered conditional sub-category evaluation. Notably, the rarest category in the validation set, *Disruption of military operations*, reached perfect precision on the French subset where two positive examples were available; the broader *Post-disaster military assistance* category, by contrast, had too few positives to support a reliable evaluation in either language and was flagged as exploratory in the reliability tiers (Supplementary Table S10).

To assess temporal stability, the gold-standard sentences were grouped by publication decade and the pooled macro F1 was recomputed within each period (Fig. 9). Across the well-populated decades (1990s–2020s), pooled macro F1 ranges from 0.861 to 0.882—a maximum inter-decade spread of 2.1 points—with no monotonic degradation over time; the 1980s estimate (0.848) rests on only three gold-standard sentences and is shown for completeness. This temporal consistency supported the models' generalisation capacity across the full 48-year corpus span, and the per-decade computation is fully reproducible from the deposited gold standard and annotations (`Results/Scripts/2_temporal_f1_validation.py`).

Inter-coder assessment of the gold standard. Inter-coder reliability assessment was completed with a second independent annotator evaluating the same 1,000 randomly selected sentences from the validation set. The protocol followed a structured approach: during the first 600 sentences, the expert annotator provided training to the second coder, explaining category definitions and answering questions without directly intervening in the coding process. This training phase allowed the second coder to progressively learn the annotation framework while maintaining independence in coding decisions. After sentence 600, the annotation process became completely blind, with no communication between coders. The resulting comparison between the second annotator's labels and the expert gold-standard annotations provided an explicit evaluation of the reliability and reproducibility of the gold standard.

Given that the authors' main interest lay in the level of reliability once the codebook had been learned, the metrics for the blind coding phase (sentences 601–1000) are reported as the primary benchmark. Table 14 summarises the agreement between the second annotator and the expert gold standard across all 65 annotation categories in this phase. These values indicated moderate agreement on Cohen's κ and substantial agreement on Gwet's AC1, while Krippendorff's α exceeded the commonly used 0.667 threshold for acceptable reliability. In other words, the gold-standard annotations produced by the expert could be reproduced to a reasonably high degree by an independent trained annotator, even in a setting with 65 multi-label categories and often dense and multi-topic climate discourse.

However, several limitations should be kept in mind. The moderate κ values reflected the inherent difficulty of climate discourse annotation at the sentence level, where many sentences contain subtle framings and overlapping categories. Chance-corrected indices such as Cohen's κ are also known to be sensitive to category imbalance—the well-documented *kappa paradox*, whereby high raw agreement can coexist with low κ when a category's prevalence is very low or very high[42,43]. Prevalence imbalance is pronounced in this setting, and the substantially higher Gwet's AC1—an index specifically designed to resist this paradox—suggested that part of the apparent disagreement in κ was driven by prevalence rather than by systematic confusion between coders; the categories concerned are flagged `paradox_kappa` in the reliability lookup. A further limitation concerns the training data itself: as discussed in Phase 3, all 4,000 training sentences were coded by a single expert,

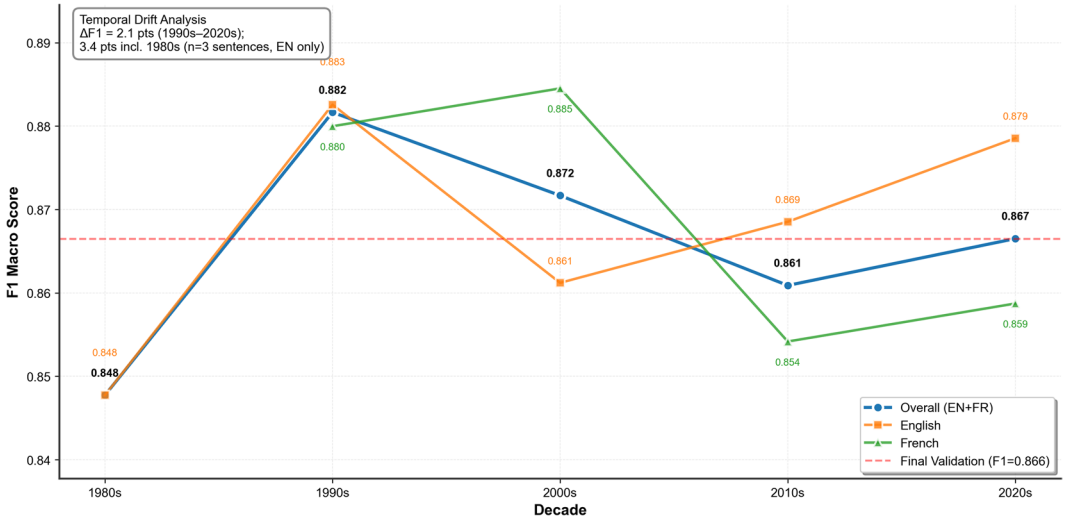

Fig. 9 Temporal validation: pooled macro F1 of the deployed models against the gold standard, by publication decade of the validated sentences. Variation across the well-populated decades (1990s–2020s) stays within 2.1 points, with no monotonic degradation; the 1980s point rests on three sentences only.

	Cohen's κ	Gwet's AC1	Krippendorff's α
Blind coding phase (601–1000, $n = 400$)	0.596	0.894	0.698

Table 14. Inter-coder reliability metrics for the blind coding phase (sentences 601–1000).

with the second coder reserved for the independent gold-standard validation. This design is consistent with the standard recommendation of stabilising the codebook through one expert before validating on a held-out sample[38], but it nevertheless means that the model behaviour reflects, at training time, the interpretive choices of a single coder. The per-category reliability tiers (Supplementary Table S10) and the blind-phase agreement metrics (Supplementary Table S9) are designed to make these choices auditable; future releases of the CCF will extend the multi-coder protocol to a stratified subsample of the training data. A second limitation concerns the three pre-trained NER models (Supplementary Table S6), whose precision, recall, and F1 figures are reported from the original-model documentation rather than from a CCF-specific gold standard. A systematic in-domain NER evaluation against a manually coded sample is planned for a subsequent release of the database; users are accordingly advised to treat the named-entity rollup as a high-recall extraction layer rather than as a benchmarked classifier.

To better understand how reliability evolved as the second annotator internalised the codebook, agreement metrics in successive blocks of 100 sentences were computed. This analysis showed a pattern consistent with a learning curve for a complex codebook: both Cohen's κ and Krippendorff's α increased over time, with higher values in the later blocks than in the early ones. Figure 10 illustrates this progression across the 1,000 sentences, with a continuing improvement after the transition to blind coding at sentence 600. The upward trajectory of both indices across the 100-sentence blocks indicated that, once training was complete, the annotation framework could be applied consistently. This reinforces the view that the expert-defined gold standard is reproducible by trained coders.

Validation of the database enrichment. Three integrity checks were run at the end of the enrichment pipeline to confirm that the derived tables remained exactly synchronised with the underlying sentence-level annotations.

First, the article-level aggregates were cross-validated against the source sentence table: for each of the 65 `prop_X` columns, the SQL identity `prop_X = SUM(COALESCE(X, 0)) / COUNT(*)` was verified to hold within floating-point tolerance (10^{-9}) for every one of the 283,964 articles, and the `top_frame` column was confirmed to equal the column code with the highest `prop_X` in every non-empty row. Second, the per-article entity rollup was checked by re-extracting the union of all distinct entities from the `ner_entities` JSON column for a stratified random sample of 1,000 articles and comparing the result element-wise against the JSONB arrays stored in `CCF_article_entities`; all 1,000 articles matched exactly. Third, the BGE-M3 embeddings were verified to have an L2 norm of 1.0 ± 10^{-3} on a random sample of 10,000 rows—the expected behaviour of vectors produced with `normalize_embeddings = True`—and an HNSW round-trip query (retrieve the top-1 nearest neighbour of an arbitrary vector and confirm that it is the vector itself, or an exact textual duplicate) returned the probe at cosine distance below 10^{-3} in 99 of 100 spot-checks with the search

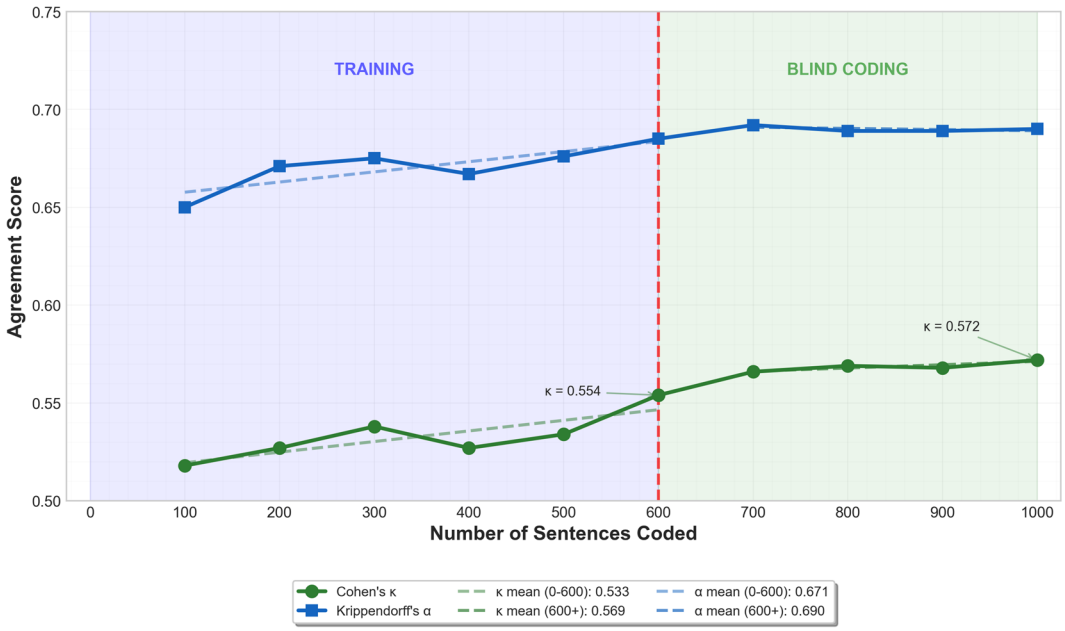

Fig. 10 Inter-coder reliability progression across 1,000 annotated sentences, computed in successive blocks of 100 sentences.

parameter `hnsw.ef_search` raised to 400—the expected behaviour of an approximate index, whose recall is governed by the query-time search width.

Usage Notes

Start from the reliability tier. The recommended workflow for any substantive analysis is to open `CCF_reliability_tiers` first and restrict the analysis to Tier A categories (27 of 65), unless lower-tier categories are explicitly treated as exploratory. Because categories are *columns* in the annotation tables, the lookup is used to select columns rather than to join rows. In SQL, against the restored PostgreSQL edition:

```
-Tier-A category codes
SELECT code FROM "CCF_reliability_tiers" WHERE tier_overall = 'A';
- Example: average attcl frame (a Tier-A category)
- across articles, by year
SELECT EXTRACT(YEAR FROM f.date) AS year, AVG(a.prop_political_frame)
FROM "CCF_article_aggregates" a
JOIN "CCF_full_data" f USING (doc_id)
GROUP BY year ORDER BY year;
```

The same selection in Python, against the Parquet mirror:

```
import pandas as pd
tiers = pd.read_parquet("CCF_reliability_tiers.parquet")
keep = ["prop_" + c for c in tiers.loc[tiers.tier_overall.eq("A"),
"code"]]
agg = pd.read_parquet("CCF_article_aggregates.parquet",
columns=["doc_id", *keep])
```

Language coverage of two categories. *Post-disaster military assistance* and *disruption of military operations* are annotated on French sentences only (no English model could be trained; see Phase 3): their values are `NULL` on English rows of `CCF_processed_data` and coalesced to zero in the aggregates. Analyses of these two categories should be restricted to the French subset.

Versioning and updates. The database is refreshed approximately every six months through the CCF Observatory's continuous extraction and annotation pipeline. For reproducible work, cite and pin the *version* DOI of the release used (this paper describes v2.1.0); use the concept DOIs to always obtain the latest release. Release notes in each deposit's README document the corpus growth between versions; the validation layer (training data, gold standard, reliability tiers) is frozen and carries over unchanged between releases unless explicitly re-run and re-documented.

Data availability

The CCF Database v2.1.0 is openly available on Zenodo under a CC-BY 4.0 licence, without embargo or access restriction, in two editions sharing identical schemas: the canonical PostgreSQL edition (DOI https://doi.org/10.5281/zenodo.22019287) and the Apache Parquet mirror (DOI 10.5281/zenodo.22019288). The concept DOIs https://doi.org/10.5281/zenodo.20346363 and https://doi.org/10.5281/zenodo.20346372 always resolve to the latest release.

Code availability

All custom code used to build, validate, enrich, and export the database—the annotation pipeline (`Scripts/Annotation/`, scripts 01–17), the enrichment and export pipeline (`Scripts/Database_creation/`), and the manuscript's table- and figure-generation scripts—is openly archived on Zenodo (DOI https://doi.org/10.5281/zenodo.22019289 [46]; concept DOI https://doi.org/10.5281/zenodo.21921214 always resolves to the latest version) and, for convenience, bundled with both data editions as `ccf_code_and_paper.tar.gz`. A development mirror is maintained at https://github.com/antoinelemor/CCF-canadian-climate-framing. Pre-existing third-party models and libraries used by the pipeline (BERT-base-cased, CamemBERT-base, the AugmentedSocialScientist fork, `dslim/bert-base-NER`, `Jean-Baptiste/camembert-ner`, spaCy `fr_core_news_lg`, BAAI `bge-m3`, `pgvector`, DuckDB) are identified, with citations, in the Methods section.

Received: 6 January 2026; Accepted: 10 September 2026;
Published: xx xx xxxx

References

1. Scheufele, D. A. & Tewksbury, D. Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models. *Journal of Communication* **57**, 9–20 https://doi.org/10.1111/j.0021-9916.2007.00326.x (2007).
2. Walgrave, S. & Van Aelst, P. The Contingency of the Mass Media's Political Agenda Setting Power: Toward a Preliminary Theory. *Journal of Communication* **56**, 88–109 https://doi.org/10.1111/j.1460-2466.2006.00005.x (2006).
3. Cohen, S.*Folk Devils and Moral Panics: The Creation of the Mods and Rockers* (MacGibbon & Kee, London, 1972).
4. Hoynes, W. Media and the Construction of Social Problems. In Treviño, A. J. (ed.) *The Cambridge Handbook of Social Problems*, vol. 1, 477–496 https://doi.org/10.1017/9781108656184.027 (Cambridge University Press, Cambridge, 2018).
5. Bastos, M. T. Shares, Pins, and Tweets: News readership from daily papers to social media. *Journalism Studies* **16**, 305–325 https://doi.org/10.1080/1461670X.2014.891857 (2015).
6. Casero-Ripolles, A. Impact of COVID-19 on the Media System. Communicative and Democratic Consequences of News Consumption during the Outbreak (2020).
7. Lulham, N., Warren, F. J., Walsh, K. & Szwarc, J. Canada in a changing climate: synthesis report. Tech. Rep., Natural Resources Canada (2023).
8. Natural Resources Canada. Energy Fact Book, 2025-2026. Tech. Rep., Natural Resources Canada (2025).
9. Harrison, K. The Struggle of Ideas and Self-Interest in Canadian Climate Policy. In Harrison, K. & Sundstrom, L. M. (eds.) *Global Commons, Domestic Decisions: The Comparative Politics of Climate Change*, 336 (MIT Press, 2010).
10. Hatch, C. *et al*. What Do Canadians Really Think About Climate Change? A Summary of Public Opinion Research for Communicators. Tech. Rep., Re.Climate (2025).
11. Young, N. & Dugas, E. Comparing Climate Change Coverage in Canadian English- and French-Language Print Media: Environmental Values, Media Cultures, and the Narration of Global Warming. *The Canadian Journal of Sociology / Cahiers canadiens de sociologie* **37**, 25–54 https://doi.org/10.29173/cjs9733 (2012).
12. King, N. *et al*. How do Canadian media report climate change impacts on health? A newspaper review. *Climatic Change* **152**, 581–596 https://doi.org/10.1007/s10584-018-2311-2 (2019).
13. Swarnakar, P. & Modi, A. NLP for Climate Policy: Creating a Knowledge Platform for Holistic and Effective Climate Action. https://doi.org/10.48550/arXiv.2105.05621 (2021).
14. Schäfer, M. S. & Hase, V. Computational methods for the analysis of climate change communication: Towards an integrative and reflexive approach. *Wiley Interdisciplinary Reviews: Climate Change* **14**, e806 https://doi.org/10.1002/wcc.806 (2023).
15. Hirsbrunner, S. D. Computational methods for climate change frame analysis: Techniques, critiques, and cautious ways forward. *Wiley Interdisciplinary Reviews: Climate Change* **15**, e902 https://doi.org/10.1002/wcc.902 (2024).
16. Nicholls, T. & Culpepper, P. D. Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities. In Theocharis, Y. & Jungherr, A. (eds.) *Computational Political Communication* (Routledge, 2024). Num Pages: 23.
17. Otmakhova, Y., Khanehzar, S. & Frermann, L. Media Framing: A typology and Survey of Computational Approaches Across Disciplines. In Ku, L.-W., Martins, A. & Srikumar, V. (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15407–15428 https://doi.org/10.18653/v1/2024.acl-long.822 (Association for Computational Linguistics, Bangkok, Thailand, 2024).
18. Entman, R. M. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication* **43**, 51–58 https://doi.org/10.1111/j.1460-2466.1993.tb01304.x (1993).
19. Luo, Y., Card, D. & Jurafsky, D. Detecting Stance in Media on Global Warming. In Cohn, T., He, Y. & Liu, Y. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3296–3315 https://doi.org/10.18653/v1/2020.findings-emnlp.296 (Association for Computational Linguistics, Online, 2020).
20. Meddeb, P., Ruseti, S., Dascalu, M., Terian, S.-M. & Travadel, S. Counteracting French Fake News on Climate Change Using Language Models. *Sustainability* **14**, 11724 https://doi.org/10.3390/su141811724 (2022).
21. Frermann, L., Li, J., Khanehzar, S. & Mikolajczak, G. Conflicts, Villains, Resolutions: Towards models of Narrative Media Framing. In Rogers, A., Boyd-Graber, J. & Okazaki, N. (eds.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8712–8732 https://doi.org/10.18653/v1/2023.acl-long.486 (Association for Computational Linguistics, Toronto, Canada, 2023).
22. Badullovich, N., Stede, M., Aktaş, B., Mirzakhmedova, N. & Saint-Dizier, P. Manual and automatic paragraph-level analysis of climate change framing in academic journal editorials. *Scientometrics* **131**, 353–387 https://doi.org/10.1007/s11192-025-05494-w (2026).
23. Canada, N. Circulation Report: Daily Newspapers. Tech. Rep., Newspapers Canada (2015).
24. Lawrence, R. G. Framing Obesity: The Evolution of News Discourse on a Public Health Issue. *Harvard International Journal of Press/Politics* **9**, 56–75 https://doi.org/10.1177/1081180X04266581 (2004).
25. Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431 https://doi.org/10.18653/v1/E17-2068 (Association for Computational Linguistics, Valencia, Spain, 2017).

26. Joulin, A. *et al*. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).
27. Grundmann, R. Using large text news archives for the analysis of climate change discourse: some methodological observations. *Journal of Risk Research* **25**, 395–406 https://doi.org/10.1080/13669877.2021.1894471 (2022).
28. Grundmann, R. Climate change as a wicked social problem. *Nature Geoscience* **9**, 562–563 https://doi.org/10.1038/ngeo2780 (2016).
29. Stede, M. & Patz, R. The Climate Change Debate and Natural Language Processing. In Field, A. *et al*. (eds.) *Proceedings of the 1st Workshop on NLP for Positive Impact*, 8–18 https://doi.org/10.18653/v1/2021.nlp4posimpact-1.2 (Association for Computational Linguistics, Online, 2021).
30. Ford, J. & King, D. Coverage and framing of climate change adaptation in the media: A review of influential North American newspapers during 1993–2013. *Environmental Science & Policy* **48**, 137–146 https://doi.org/10.1016/j.envsci.2014.12.003 (2015).
31. Trumbo, C. Constructing climate change: claims and frames in US news coverage of an environmental issue. *Public Understanding of Science* **5**, 269–283 https://doi.org/10.1088/0963-6625/5/3/006 (1996).
32. Boykoff, M. T. *Who Speaks for the Climate?: Making Sense of Media Reporting on Climate Change* (Cambridge University Press, 2011).
33. Downs, A. Up and down with ecology-the issue attention cycle. *Public Interest* **28**, 38–50 (1972).
34. Liu, X., Lindquist, E. & Vedlitz, A. Explaining Media and Congressional Attention to Global Climate Change, 1969-2005: An Empirical Test of Agenda-Setting Theory. *Political Research Quarterly* **64**, 405–419 https://doi.org/10.1177/1065912909346744 (2011).
35. Moser, S. C. Reflections on climate change communication research and practice in the second decade of the 21st century: what more is there to say? *Wiley Interdisciplinary Reviews: Climate Change* **7**, 345–369 https://doi.org/10.1002/wcc.403 (2016).
36. Salama, S. & Aboukoura, K. Role of Emotions in Climate Change Communication. In Leal Filho, W., Manolas, E., Azul, A. M., Azeiteiro, U. M. & McGhie, H. (eds.) *Handbook of Climate Change Communication: Vol. 1: Theory of Climate Change Communication*, 137–150 https://doi.org/10.1007/978-3-319-69838-0_9 (Springer International Publishing, Cham, 2018).
37. Nabi, R. L., Gustafson, A. & Jensen, R. Framing Climate Change: Exploring the Role of Emotion in Generating Advocacy Behavior. *Science Communication* **40**, 442–468 https://doi.org/10.1177/1075547018776019 (2018).
38. Krippendorff, K.*Content Analysis: An Introduction to Its Methodology* (SAGE Publications, Thousand Oaks, CA, 2019), 4th edn.
39. Do, S., Ollion, t & Shen, R. The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy. *Sociological Methods & Research* **53**, 1167–1200 https://doi.org/10.1177/00491241221134526 (2024).
40. Lemor, A. antoinelemor/AugmentedSocialScientistFork (2025).
41. Chen, J. *et al*. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2318–2335 https://doi.org/10.18653/v1/2024.findings-acl.137 (Association for Computational Linguistics, Bangkok, Thailand, 2024).
42. Feinstein, A. R. & Cicchetti, D. V. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* **43**, 543–549 https://doi.org/10.1016/0895-4356(90)90158-L (1990).
43. Byrt, T., Bishop, J. & Carlin, J. B. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* **46**, 423–429 https://doi.org/10.1016/0895-4356(93)90018-V (1993).
44. Lemor, A., Pillod, A., Taylor, M. & Nadeau, R. CCF Database (PostgreSQL edition): A sentence-level corpus of 283,964 Canadian climate-change articles from 22 news outlets (1978–2026). https://doi.org/10.5281/zenodo.22019287 (2026).
45. Lemor, A., Pillod, A., Taylor, M. & Nadeau, R. CCF Database (Apache Parquet mirror): column-oriented edition of the Canadian Climate Framing corpus (1978–2026). https://doi.org/10.5281/zenodo.22019288 (2026).
46. Lemor, A., Pillod, A., Taylor, M. & Nadeau, R. CCF – Canadian Climate Framing: Code, documentation and reproducible pipeline. https://doi.org/10.5281/zenodo.22019289 (2026).

Acknowledgements

The authors thank Maëlle Rouyer for her valuable work in the development of this project. The authors are also grateful to the Centre interuniversitaire de recherche sur la science et la technologie (CIRST) for its valuable support and guidance, including access to its research infrastructure. Earlier versions of this paper were presented at the annual meetings of the Canadian Political Science Association, the Prairie Political Science Association, and the American Political Science Association, as well as at the Université de Montréal (special thanks to Vincent-Arel Bundock, Evelyne Brie and Jean-François Godbout). The authors wish to acknowledge the anonymous reviewers and editor for their feedback throughout the development of this paper.

Author contributions

Antoine Lemor designed and implemented the complete data architecture, including data extraction, cleaning, parsing, and database construction. Developed all machine learning models and methodological approaches. Wrote all code and drafted the methodology and results sections. Alizée Pillod developed the theoretical and conceptual foundations underlying the annotation framework. Designed the complete annotation scheme and performed all manual annotation. Drafted the theoretical background and literature review sections. Matthew Taylor contributed to conceptual development, analysis and training design. Reviewed and edited the complete manuscript.Richard Nadeau contributed to the conceptual development of the project.

Funding

Funding for this research was provided by the Fonds de recherche du Québec (FRQ grant number 337081) and the Canadian Social Sciences and Humanities Research Council (SSHRC grant number RNH01687).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-08330-9.

Correspondence and requests for materials should be addressed to A.L., A.P., M.T. or R.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.