From LLMs to information cascades

Annotate, measure and detect media dynamics at scale

Antoine Lemor, Ph.D.

Postdoctoral researcher

Université de Sherbrooke · RFICS · CIRST · CAPP

CAPP / Centre d'analyse des politiques publiques

RFICS*

CIRST Centre interuniversitaire de recherche sur la science et la technologie

UdS Université de Sherbrooke

Outline

1. **Text analysis in social sciences**
   material, extraction, classification
2. **LLM Tool**
   an end-to-end solution
3. **Application — CCF**
   media cascades on climate change

PART I

Text analysis

Material, extraction, classification

Text as foundational material

Material

Written
Documents, archives, manuscripts

Oral
Speeches, interviews, conversations

Visual
Images, graphics, videos

Text
malleable

transformable material

Extraction

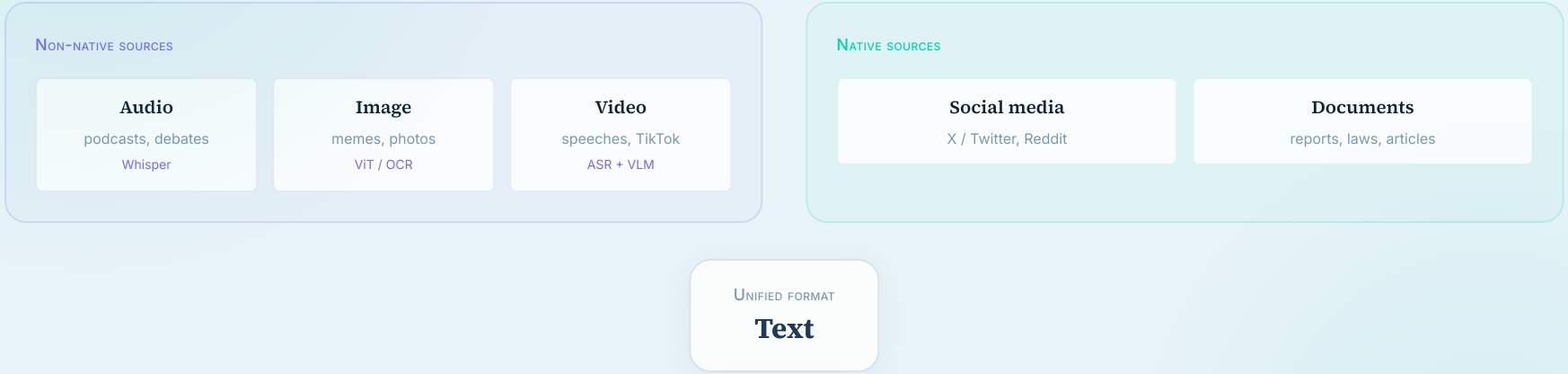
Information
Facts, data, knowledge

Meaning
Significations, interpretations

Structures
Recurrences, relations, trends

Text is a malleable material that both humans and machines can read

What is a text in social sciences?



Text unifies native and transformed sources into structured data, ready for computational analysis.

Why classify, and what to extract?

Content

Themes, topics, concepts

Climate change, economy, public health

Position

Stance, for or against

Support, criticism, neutrality toward a policy

Emotion

Sentiment, expressed affect

Anger, hope, fear, joy, sadness

Entities

Actors, places, organizations

UN, Canada, Guterres, WHO

Text analysis turns unstructured language into quantitative, analyzable data

LLM Tool: challenges and answers

Challenge	Response
Scale Thousands of documents, limited budget and time	**Batch processing** High-speed annotation, free and open source
Complexity Different languages, text types, technical skills	**50+ models, 100+ languages** Multilingual, adaptable, LLM consensus
Quality Inter-annotator agreement, reproducibility, validation	**Validation Lab** State-of-the-art metrics, full traceability
Ethics Sensitive data, commercial APIs, data protection	**100% local** Ollama, no data leaves your machine

LLM Tool: a complete, accessible answer for social science research (quantitative and qualitative)

PART II

LLM Tool

An integrated solution for automated text analysis

What is LLM Tool?

Open-source platform

Automatic text annotation powered by AI

Turns any text corpus into structured data along your own categories — without writing a line of code.

Accessible

Guided interface, zero programming

Flexible

10 to 10M documents, quanti and quali

Secure

100% local, no external API

Rigorous

Traceability, reproducibility

Methodological validation

Macro F1 = 66.73

Lemor, Dinan, Gilbert (2025)

Continuous evolution

Image, audio (soon)

ViT, Whisper integration

How does LLM Tool work?

• Parameters traced from start to finish •

A complete protocol: from raw data to a validated classification

Mode 1 — The Annotator

The Annotator: guided interface for LLM-based classification

Main function

Uses language models (OpenAI, Ollama) to automatically classify your texts along your category schema.

Capabilities

- Advanced prompt engineering features
- Handles diverse text types
- Runs 100% locally with Ollama (free)
- Auto-repairs format errors
- Exports to CSV, Excel, review platforms

Ideal use

Quickly annotate a small corpus, or build training data.

Mode 2 — Annotator Factory

Main function

Chains annotation, data preparation, model training and deployment in a single automated process.

Automated steps

- Annotates a sample with an LLM
- Prepares data (balancing, cleaning)
- Splits train / validation / test
- Selects and deploys the best model
- Applies the model(s) to your whole corpus

Ideal use

A custom model for a large corpus in a few hours.

Annotator Factory: end-to-end model creation

Mode 3 — Training Arena

Training Arena: trains and compares model families

Main function

Trains up to 50 different model types and automatically picks the best one for your data.

Capabilities

- BERT, RoBERTa, DeBERTa, CamemBERT (FR), XLM-R
- Auto-detects the language of your texts
- Compares performance with cross-validation
- Keeps only the best model
- GPU-optimized

Ideal use

Get the optimal model for your training corpus.

Mode 4 — BERT Annotation Studio

Main function

Applies your trained BERT models to very large document collections at high throughput.

Performance

- High-speed GPU annotation
- Auto-optimizes available memory
- Confidence score per prediction
- Handles millions of documents
- Auto-resumes if interrupted

Ideal use

Apply your model to large volumes of text.

BERT Annotation Studio: applies trained models at scale

Mode 5 — Validation Lab

Validation Lab: agreement metrics and quality control

Main function

Systematically evaluates annotation quality by computing inter-annotator agreement and surfacing problematic cases.

Validation tools

- Cohen's Kappa, Krippendorff's Alpha
- Flags unreliable annotations (score < 0.5)
- Identifies problematic categories
- Stratified samples for human review
- Detailed reports

Ideal use

Ensure scientific quality before publication or final analysis.

Demonstration

LLM Tool in action

llm-tool.com

Why use LLM Tool for your research?

Accessibility

- **Zero programming** — guided interactive interface
- **Free option** — Ollama, fully local
- **Multilingual** — 100+ languages
- **Multi-platform** — macOS, Linux, Windows

Scientific rigor

- **Reproducibility** — full traceability
- **Validation** — inter-annotator agreement metrics
- **Transparency** — full parameter export
- **Open source** — public code and docs

Built for social science researchers, by social science researchers

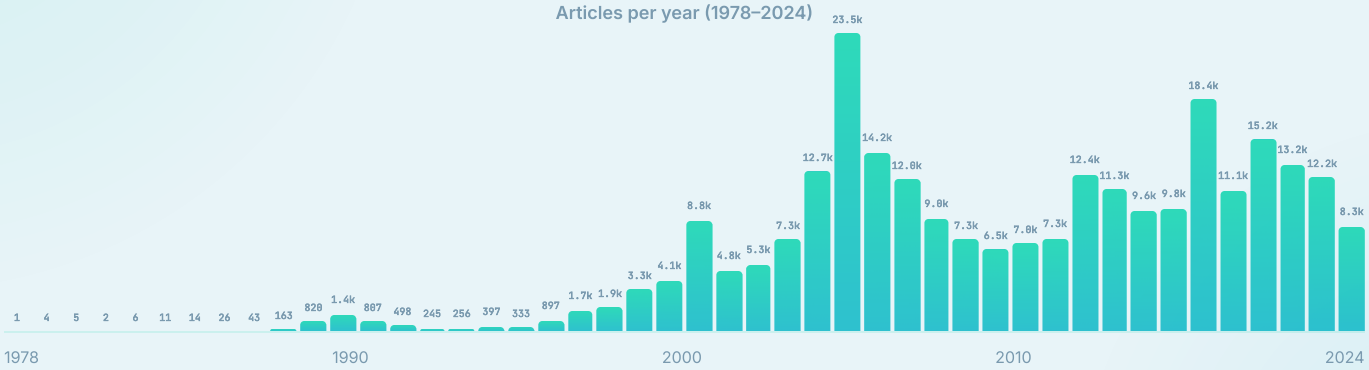
PART III

Application — CCF

Canadian Climate Framing — media cascades

The CCF database: Canadian Climate Framing

266K	20	9.2M	65
Articles	Newspapers	Annotated sentences	Categories
1978 – 2024	EN + FR	BERT / CamemBERT	hierarchical

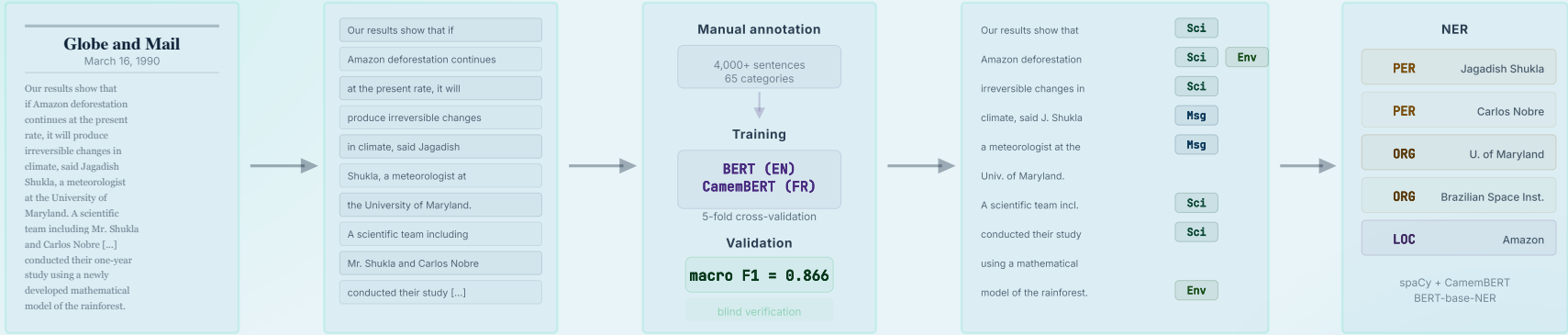


The largest ML-annotated climate discourse corpus in Canada

Macro F1 : 0.866 – Lemor, Pillod, Taylor (2025)

Sentence-level annotation: the pipeline

Each sentence of each article is classified by trained models



65 CATEGORIES ACROSS 10 FAMILIES

Economic 6 Costs, benefits, impacts, sector footprint	**Political** 5 Policy, debate, positioning, opinion	**Scientific** 5 Discovery, debate, skepticism, defense	**Environmental** 3 Habitat loss, species loss	**Health** 3 Negative impacts, co-benefits
Security 5 Refugees, conflict, military, disruption	**Justice** 6 Winners/losers, responsibility, vulnerability	**Cultural** 5 Art, events, Indigenous practices, footprint	**Messengers** 10 Scientists, officials, activists, experts	**Events + Solutions** 11 Weather, elections, protests, mitigation

+ Emotional tone (3) · Geographic focus (1) · Urgency (1) · NER (PER, ORG, LOC)

67 hierarchical categories

8 thematic frames + complementary dimensions

Sentence: *"The PM announced..."* → Main category: **Politics** → Sub-category: **Political action** → × 67

0.866 Macro F1

67 BERT / CamemBERT models
1 per category · 63 available

Thematic frames

8 frames → 40 sub-categories

Economy
- Negative impacts
- Positive impacts
- Costs of action
- Benefits of action
- Sectoral footprint

Health
- Negative impacts
- Positive impacts
- Co-benefits
- Sector footprint

Security
- Climate refugees
- Resource conflicts
- Military assistance
- Military disruption

Justice
- Winners/losers
- Responsibility
- Vulnerability
- Unequal access
- Intergenerational

Politics
- Political action
- Political debate
- Positioning
- Public opinion

Science
- Debate
- Discovery
- Skepticism
- Science advocacy

Environment
- Habitat loss
- Species loss

Culture
- Art
- Events
- Indigenous practices
- Cultural footprint

Complementary dimensions

4 dimensions → 27 sub-categories

Messengers 10
- Health
- Economy
- Security
- Legal
- Cultural
- Nat. sci.
- Soc. sci.
- Activist
- Public official

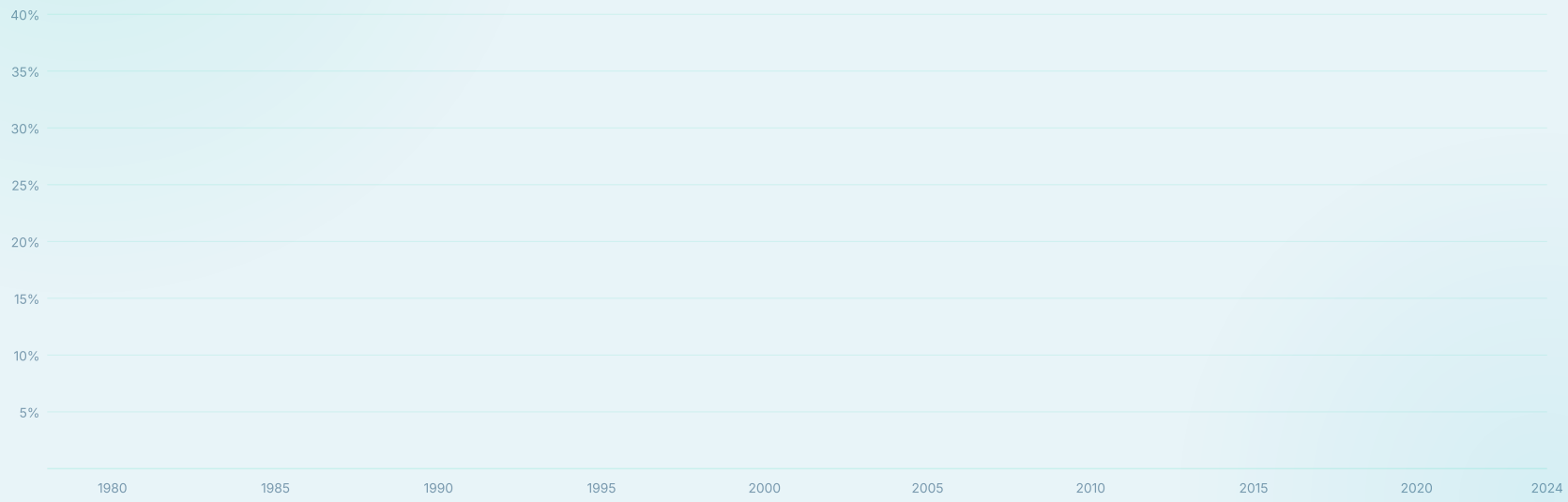
Events 9
- Extreme weather
- Meeting
- Publication
- Election
- Policy
- Judicial
- Cultural
- Protest

Solutions 3
- Mitigation
- Adaptation

Tone + Other 5
- Positive
- Negative
- Neutral
- Canada
- Urgency

Evolution of thematic frames (1978–2024)

Average sentence share per frame, by year

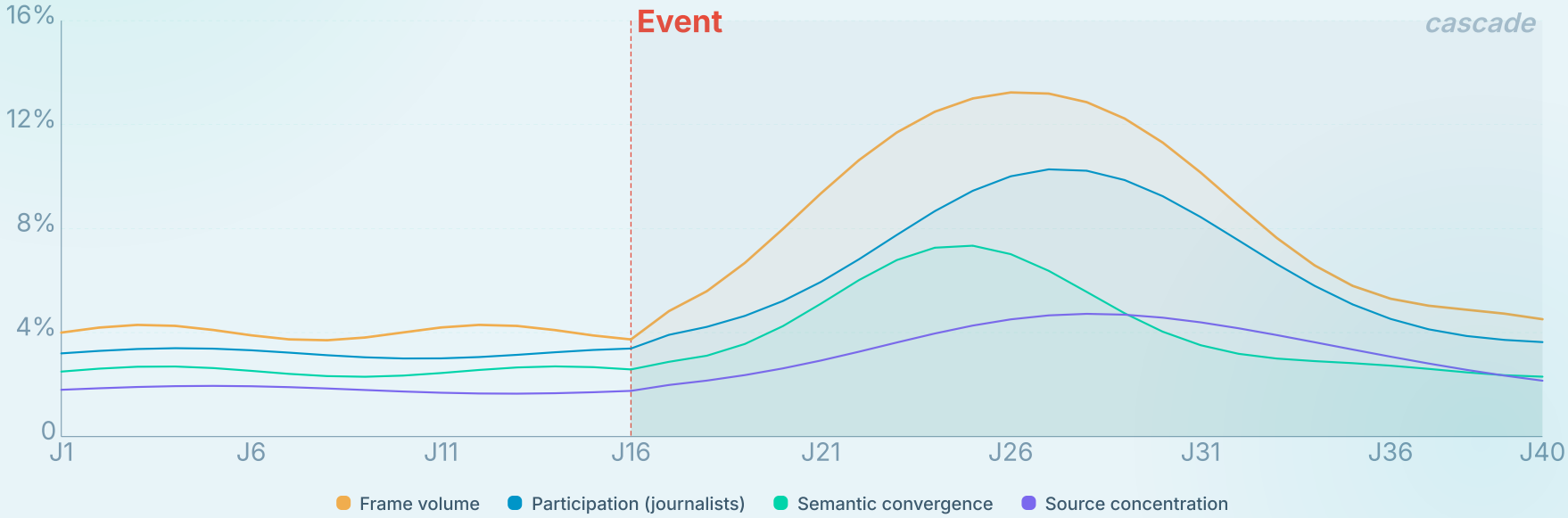


What is a media cascade?

A frame propagating massively across outlets after a triggering event

Media cascade =

a topic that spreads rapidly across media and journalists, with **convergence** of content, an **increase** in coverage volume, and the arrival of **new actors** in the debate



How to detect and measure a cascade?

Five signals, statistical detection, multi-dimensional score

Detecting cascades

Five signals tracked daily

- **Temporal** — *Is the frame rising?*
- **Participation** — *New journalists writing?*
- **Convergence** — *Is the frame dominant?*
- **Source** — *Same messengers across outlets?*
- **Semantic** — *Articles look alike?*

Cascade detected

When the signals jointly exceed the background noise

Strong · Moderate · Weak

Detecting events

Group articles into clusters

- **Annotations** — *What kind of event? (weather, policy, publication...)*
- **Embeddings** — *Are the articles semantically similar?*
- **Temporality** — *Published at the same moment?*
- **Entities** — *Same people, places, organizations?*

Cluster = event

Each cluster groups articles covering the same concrete event

Health cascade — Summer 2023

The CCF platform

1

Explorer

Map of Canada, temporal trends, media profiles

2

Search

Semantic (embeddings), named entities, cross-tabulations

3

Analysis

Media cascades, event clusters, causal impact

Public API

`data.ccf-project.ca`

50+ REST endpoints · bilingual FR/EN · open access

Thank you!

Questions and discussion

Antoine Lemor, Ph.D.

Université de Sherbrooke · RFICS · CIRST · CAPP

antoinelemor.com

llm-tool.com

data.ccf-project.ca