From Manual Coding to Artificial Intelligence: Towards Automating the Text Annotation Process in Social Sciences

Jérémy Gilbert

Antoine Lemor

Shannon Dinan

Abstract

As political scientists increasingly turn to large text corpora to study parties, discourse, and media coverage, the need for efficient and reliable annotation methods becomes critical. This article compares traditional dictionary-based approaches with several large language models (LLMs) for annotating political text across three analytical dimensions: thematic classification, party references, and sentiment. Using a corpus of 40,000 sentences drawn from Canadian media headlines and House of Commons debates, each method is evaluated against a manually coded gold standard to assess its precision, recall, and overall consistency. The study also explores the feasibility of using these annotations to train a BERT classifier, with the broader goal of building a robust and scalable annotation pipeline for political text analysis.

Introduction

Textual analysis holds a central place in political science, both for understanding party discourses and for evaluating media coverage of public issues. However, given the abundance of texts and the limitations of manual coding, it is appropriate to explore automated methods. While dictionary-based approaches and crowdsourcing have long represented the standard methods in the field, the emergence of machine learning, and notably the use of large language models (LLMs), opens up new perspectives for enhancing the quality and efficiency of annotation.

In recent years, political scientists and social science researchers more broadly have increasingly relied on text analysis to study party positions, media bias, policy debates, or public opinion formation. The ability to accurately and systematically annotate political text is thus becoming a key methodological challenge, especially as research shifts toward larger corpora and more complex dimensions of meaning.

In this context, this study aims to compare different annotation methods derived from both LLMs and dictionary-based classification across three analytical dimensions (thematic, referential, and sentimental) applied to two complementary corpus: media headlines and Canadian parliamentary debates.

Analysing Political Text

Text has long been a subject of research interest for political scientists. Whether examining political party platforms (Benoit et al. 2016; Laver and Garry 2000; Laver, Benoit, and Garry 2003; Lowe et al. 2011; Slapin and Proksch 2008), transcriptions of debates (Lauderdale and Herzog 2016; Laver, Benoit, and Garry 2003; Tremblay-Antoine et al. 2024) and political speeches (Albaugh et al. 2013; Bilbao-Jayo and Almeida 2018; Dunmire 2012), or media articles (Lawlor and Tolley 2017; Shapiro et al. 2020; Young and Soroka 2012), text offers a rich resource that enables a complete and nuanced analysis of the subject under study. Its unparalleled advantage lies in its abundance, variety, and general accessibility. However, textual analysis remains a rigorous and time-consuming process. Manually coding hundreds or thousands of texts demands substantial resources.

This challenge typically leads to the employment of several research assistants to code the texts over long hours, followed by a validation process among the coders. In addition, human biases and the limitations of sustained attention over extended periods inevitably result in errors. Such errors are normal and fall within the acceptable threshold for a well-conducted study. It should not be assumed that using an alternative methodology reduces biases and errors. Nonetheless, the financial and temporal constraints of manual coding make it difficult to conduct studies on very large text corpus.

Manual annotation

To enable the annotation of a large quantity of data, alternative methods have become popular in the field. A hybrid method is commonly employed to supplement human resources: crowdsourcing.

The idea of crowdsourcing originates from the theory of the wisdom of crowds. This concept is based on the notion that groups of individuals, even if each is only modestly informed, can produce remarkably accurate estimates or decisions when their contributions are aggregated (Surowiecki 2005). This theory was illustrated by the anecdote of Francis Galton: during a fair in 1906, Galton observed that the average of spectators' estimates of an ox's weight was surprisingly close to the actual weight, even though no individual had guessed it precisely.

Applied to the social sciences, crowdsourcing harnesses these principles by assigning complex tasks, such as text coding, to a large number of people. Numerous online platforms, such as Amazon Mechanical Turk, allow tasks to be distributed among hundreds of non-expert coders in exchange for compensation, with their work subsequently aggregated to yield robust results. By combining the efforts of a few experts to validate critical tasks with non-specialized contributions, this hybrid method enables the analysis of larger volumes of content while maintaining high accuracy (Benoit et al. 2016). This approach makes it possible to manually analyze data volumes that would otherwise be inaccessible.

Crowdsourcing, however, requires a triage process to retain only those coders who perform well. To that end, it is customary to have experts code a small subset of texts known as the gold standard. The success rate of online coders on this gold standard is then used to select or reject them for the full task. A major shortcoming of this method lies in its incentive structure: coders are typically compensated based on the volume of annotations, which encourages high output but diminishes the quality of the work performed. Furthermore, results on private platforms such as Amazon Mechanical Turk have declined in quality over recent years, notably due to the use of private servers for fraudulent purposes (Kennedy et al. 2020). Although alternative platforms exist, the use of crowdsourcing now entails greater barriers and yields uncertain results. In light of these challenges, it may now be more interesting to use the gold standard not as a filter for human coders, but as a benchmark for training and evaluating automated annotation methods.

Automating the analysis

The alternative to manual coding is to automate the analysis process. The benefits are clear: it significantly reduces the time and costs associated with human coding, allowing researchers to analyze much larger volumes of text. However, automation should not be seen as a replacement for human interpretation. Instead, it is a complementary tool that enhances analytical capacity—provided that the outputs are rigorously validated (King 1995). In other words, automation enables us to go further and faster, but not without maintaining the safeguards of careful human oversight.

Moreover, one must consider the object of analysis. Text and human language are exceedingly rich and varied, characteristics that allow for the examination of multiple elements within the same group of words. A highly popular trend in the literature is to determine the positions of political parties on various issues or ideological axes. Such studies have been conducted repeatedly, drawing on electoral platforms or speech transcripts to extract party positions (Lauderdale and Herzog 2016; Lowe et al. 2011; Slapin and Proksch 2008). Another element often analyzed is tone (Shapiro et al. 2020; Van Atteveldt, Van Der Velden, and Boukes 2021; Young and Soroka 2012). Language uniquely enables the written expression of more positive or more negative sentiments. For example, analyzing the tone used by a party can help determine whether that party addresses certain issues more negatively than others.

Supervised and Non-supervised Methods

A single sentence can therefore be examined from multiple perspectives. With this multiplicity comes a variety of methods. Just as human coders use different coding schemes depending on the analytical goal, automated text analysis provides a variety of methods tailored to the object and purpose of the analysis (Grimmer and Stewart 2013). One key distinction lies between supervised and unsupervised methods, depending on whether the categories of interest are known in advance. In the absence of predefined categories, unsupervised clustering methods such as Latent Dirichlet Allocation (LDA) group texts based on word co-occurrence patterns, allowing researchers to identify emerging themes.

When the categories are known in advance, the classification process is said to be supervised. Among the simplest and most widely used supervised approaches in political science is dictionary-based text analysis. This method relies on predefined lists of words associated with specific themes or concepts. Texts are scanned for matches with these word lists, and the frequency of each term is used to estimate the presence or prominence of the corresponding category.

One of the key advantages of this method is its ease of use and relatively low cost. Off-the-shelf dictionaries are available for a range of topics in political science, such as the Lexicoder Topic Dictionary (LTD), which classifies public policy issues (Albaugh et al. 2013).

As a way to demonstrate what dictionary-based methods can achieve in practice, we conducted an analysis of the tone used in media headlines mentioning Canadian political parties. The corpus consists of all headlines published between May 2024 and April 2025 by twelve major news outlets, totaling several thousand entries. Among these, we focused on those that explicitly referenced political parties. This kind of analysis, difficult to carry out manually at such a scale, can be implemented efficiently using automated methods. While dictionaries are often used to classify content thematically, this example shows how they can also be applied to the study of political actors in media coverage. Figure 1 presents the standardized tone of party mentions across outlets.

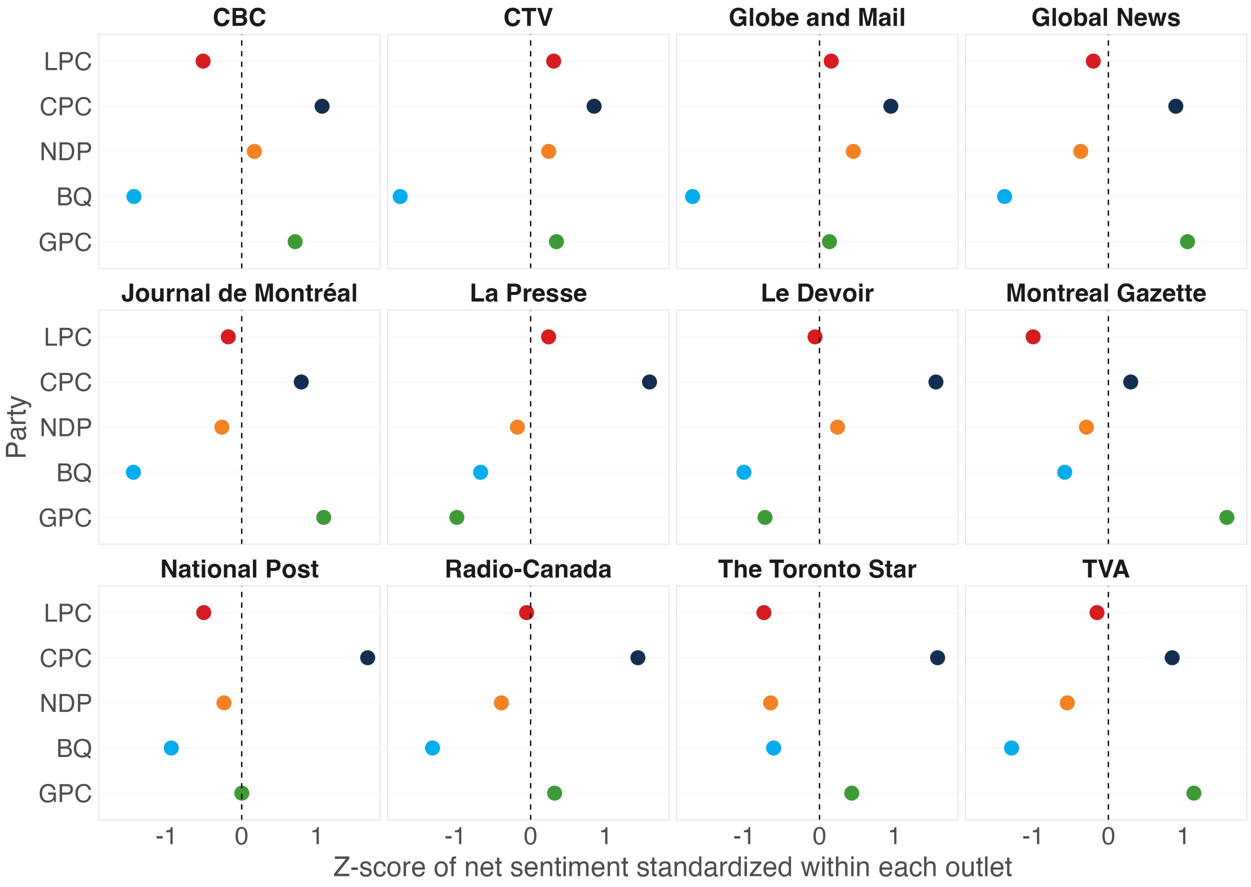

Figure 1: Standardized Tone of party Mentions by Outlet

Nevertheless, dictionary-based methods present important limitations. They are highly sensitive to context, and certain words may carry different meanings depending on how they are used. Moreover, dictionaries often struggle to capture the evolving vocabulary of political discourse. In the context of Canadian politics, new political figures, such as Mark Carney, may emerge as central actors. If their names are not included in the dictionary, important references to parties or political developments may go undetected. This highlights the need to continuously update dictionaries to reflect the changing language and content of political and media debates.

These limitations underscore the importance of validating dictionaries to ensure their relevance to the domain under study. Validation can include comparisons with human-coded data or with results from other annotation methods. Even dictionaries designed to classify similar content may yield different outputs when applied to the same corpus (Young and Soroka 2012). When used without validation, the results of a dictionary-based analysis—however detailed or convincing they may appear—remain uncertain in terms of their accuracy and representativeness. While dictionary methods allow researchers to produce large-scale analyses efficiently, their ease of implementation and versatility also explain their continued use in studies focused on thematic or emotional trends. However, these advantages must be weighed against the need for careful validation, especially given that the accuracy of dictionary-based results is generally considered relatively low (Van Atteveldt, Van Der Velden, and Boukes 2021; Widmann and Wich 2023).

With the development of machine learning, annotation techniques have evolved beyond rule-based methods, making it possible to classify text through supervised learning. These approaches rely on the manual coding of a sample of the corpus, which serves as training data for an algorithm tasked with classifying the remainder. In contrast to dictionaries, supervised methods are generally more capable of capturing semantic nuance and context, making them a more flexible option for complex classification tasks. While several algorithms are commonly used for this purpose—such as support vector machines or logistic regression (Pranckevičius and Marcinkevičius 2017; Salman and Al-Jawher 2024; Shyrokykh, Girnyk, and Dellmuth 2023)—this article does not evaluate these approaches directly. Instead, we focus on a more recent and promising category of models: large language models, which integrate supervised learning at a much larger scale and offer new possibilities for annotation.

LLMs' arrival

In 2022, ChatGPT was popularized and generated considerable excitement around the use of artificial intelligence. ChatGPT is a large language model (LLM) trained on an enormous amount of human-generated web content, designed to produce responses that resemble those of a human. Although it was ChatGPT that introduced the daily use of such models to the general public, it is part of a continuously evolving group that includes both proprietary models and others that are freely accessible. LLMs are designed to predict the probability of a sequence of words (Linegar, Kocielnik, and Alvarez 2023). These models rely on complex architectures such as transformers, which enable them to process and generate text fluidly and contextually by predicting the likelihood of word sequences based on their training. Transformers, introduced by Vaswani et al. (2017), revolutionized natural language processing by allowing models to capture contextual relationships across long sequences of text. The responses provided by LLMs are more nuanced than those of previous models, and the tasks they perform are correspondingly more complex.

The use of artificial intelligence for completing human tasks animates debate on ethical concerns. Of course, LLMs have their own biases, being trained on human generated data. LLM biases have been studied extensively in the last decade. In social sciences, reaserchers must be conscious that the models they are using most likely display political (McGee 2023; Rozado 2023; Van den Broek 2023) and gender biases (Dong et al. 2023; Gilardi et al. 2022; Zack et al. 2023). These biases, combined with the ethical concern of the possibility that LLMs produce work otherwise accomplished by a human raises important considerations when using them for annotation. Human oversight and transparency in their use is therefore crucial (Ferdaus et al. 2024; Talboy and Fuller 2023).

While its rise in popularity has sparked both enthusiasm and apprehension regarding the potential consequences of its use, a new body of literature is rapidly emerging that examines the ability of LLMs to perform human-like tasks. Considering that text analysis is time-consuming and costly when performed by experts, that the reliability of crowdsourced coders is variable, and that dictionary-based methods lack flexibility and require constant adaptation to context, the prospect of rigorous, reliable, high-volume analysis at low cost via artificial intelligence is fueling a new wave of research in the field.

Early studies on the subject already show that various versions of ChatGPT can annotate and classify texts more effectively than Turkers (Gilardi, Alizadeh, and Kubli 2023), and even outperform domain experts (Törnberg 2023; Zhang et al. 2024). Although the results seem to indicate enhanced performance by LLMs, few studies in social sciences have yet examined their capabilities in text analysis, given the wide array of existing methods.

Do, Ollion, and Shen (2022) propose a relatively simple method for social science researchers that is effective for text annotation. Groups with varying levels of expertise manually annotate phrases and words to classify texts by category. Once the annotation is validated by the expert group, the sample is used to train a BERT model on the English corpus (Devlin et al. 2019) and a camemBERT model on the French corpus (Martin et al. 2020)—non-generative models that are employed to perform specific tasks. This type of model is trained by masking a portion of the sentence and predicting the missing word. It represents a new supervised learning annotation procedure that builds on the foundations of machine learning. This marks a significant breakthrough in the use of LLMs for text annotation in the social sciences.

With literature beginning to demonstrate the potentially superior capabilities of LLMs compared to human annotation, and with the development of supervised learning methods such as those proposed by Do, Ollion, and Shen (2022), the complete automation of the text annotation process is becoming conceivable. This article focuses on the annotation of content from media headlines and Canadian parliamentary debates. It raises the following questions:

- **Q1: How does the annotation produced by various LLMs and dictionaries compare to expert manual annotation?**
- **Q2: Is it possible to build a fully automated text annotation process that trains a BERT model using a previously validated automatic annotation method?**

To address these questions, an annotation process is conducted on data from media headlines and parliamentary debates across three analytical dimensions. The objective is to identify the issue categories present in the text (thematic dimension), references to political parties (referential dimension), and the tone of the sentences within the corpus (sentimental dimension). These three dimensions represent distinct annotation processes upon which the different methods can be compared.

Data and Method

Two text corpus are employed: Canadian media headlines and transcripts of the Question Period from the Canadian House of Commons.

Media Headlines are collected as part of the Radar+ project, an academic initiative developped with the help of Infoscope[1]. The front pages from 13 Canadian media outlets have been gathered since May 2024. These outlets were selected based on readership figures, representing 13 of the most widely read newspapers and providing a comprehensive perspective on media coverage in Canada. The 13 media outlets are:

> *Le Devoir*, *La Presse*, *National Post*, *Le Journal de Montréal*, *TVA Nouvelles*, *CBC News*, *CTV News Global News*, *The Globe and Mail*, *The Toronto Star*, *Vancouver Sun*, *Montreal Gazette*

Data Collection

To collect the content of media front pages, Radar+ functions as a web scraper. Since the target data are in the form of unstructured text, it is first necessary to identify the relevant information, then extract and store it. The Radar+ scraper accomplishes these steps by traversing the HTML code of the identified newspapers' websites. Every 10 minutes, the front pages from the 13 media outlets are collected.

The parliamentary corpus is collected from the transcripts of debates available on the Canadian House of Commons website, covering the same period as the media corpus. It is scraped similarly, when new transcription gets published on the House of Commons website.

Analysis Dimensions

The corpus is annotated based on three analytical dimensions: thematic, referential, and sentimental. These dimensions were selected to test a range of annotation tasks, from topical classification to entity recognition and tone detection. Each represents a distinct type of analytical output commonly found in political text analysis, allowing us to evaluate how well different annotation methods

[1] https://www.infoscope.ca/

perform across varied linguistic and conceptual challenges.

Thematic dimension

The thematic annotation of the text corpus is carried out using the public policy issue categories established by the Comparative Agendas Project (CAP), an international initiative for classifying public policies. The CAP has categorized public sphere issues under 21 thematic categories:

> *Law and Crime*, *Immigration*, *Technology*, *Macroeconomics*, *Labor*, *Transportation*, *Housing*, *Domestic Commerce*, *Foreign Trade*, *Public Lands*, *Agriculture*, *Environment*, *Energy*, *International Affairs*, *Defence*, *Governments and Governance*, *Culture and Nationalism*, *Rights, Liberties, Minorities, and Discrimination*, *Education*, *Health*, *Social Welfare.*

These 21 categories also include 220 subcategories for added granularity (Bevan 2019). Our study focuses on the 21 top-level categories, as analyzing 220 subcategories would require a much larger number of manually coded sentences to reach statistical significance.

This dimension tests whether a sentence can be accurately classified under one of these predefined public policy themes. It evaluates an annotation method's ability to extract and interpret the main issue discussed in a sentence, which may involve abstract or general vocabulary, indirect references, or overlapping policy areas.

The issue dictionary used for the dictionary-based annotation is derived from the Lexicoder Topic Dictionary (LTD), which was itself developed based on CAP categories (Albaugh et al. 2013). The Lexicoder has been translated into French and supplemented to cover terms that were not current at the time of the dictionary's creation [2].

Referential dimension

The text is annotated to mark references to Canadian political parties. The dictionary used for this task includes terms referring to parties and their parliamentarians [3]. The goal of this task is

[2] The translation and enhancement of the Lexicoder were carried out by the Leadership Chair in the Teaching of Digital Social Sciences (CLESSN).

[3] This dictionary was designed by the CLESSN through a double-coder process.

to identify whether a sentence refers to a political actor or group, rather than to a policy topic or to the sentiment expressed.

This dimension is methodologically distinct from thematic classification, as it evaluates the ability of annotation methods to recognize named entities and context-specific references to political organizations.

Sentimental Dimension

The third dimension focuses on the tone of the sentences in the corpus. This analysis aims to determine whether a sentence expresses a positive, neutral, or negative sentiment. The Lexicoder Sentiment Dictionary (Young and Soroka 2012) and its translated version (Duval and Pétry 2016) are employed for dictionary-based annotation. As with the other dimensions, sentiment analysis engages a distinct annotation task, one that is different from identifying thematic content or political references.

Sentiment analysis in political texts is particularly challenging, as tone can be implicit, sarcastic, or embedded within issue framing. This task therefore evaluates whether annotation methods can detect broad affective orientation in a politically charged context.

Sample

To test the accuracy of the annotations produced by the different methods, a sample consisting of 40,000 sentences extracted from both corpus is utilized. Half of the sample comes from media front pages, and the other half is derived from the transcripts of parliamentary debates. Additionally, half of the sentences are in French while the remaining half are in English. The objective is to determine whether there is a difference in model performance based on the nature of the corpus and the language. Parliamentary debates contain a considerable amount of empty content and procedural sentences, which are expected to affect the annotation. The sentences are selected randomly, provided they contain a minimum of 5 words, in order to reduce noise.

The sample size is determined to ensure sufficient representation of each issue. Issues that occur less frequently in media front pages, such as *Agriculture*, require annotation over a larger number of sentences to ensure that the comparative statistical analysis between the models is robust.

Annotation Procedure

Automatic Annotation

The efficacy thresholds of different models and methods are compared. Each method annotates the sample sentences across the three analytical dimensions. The ChatGPT API and five freely accessible LLMs—Deepseek R1, Nemotron, Mixtral 8x22B, Llama 3.3, and Qwen 2.5—are employed to perform the annotation task. These models include both proprietary (ChatGPT) and open-source models, selected for their performance and accessibility at the time of writing, and because they can be run on the team's local infrastructure.

To convey the task to the models, a unified prompt was drafted. It defines the role of the LLM and specifies the annotation task across all three dimensions. To ensure clarity and structure, the prompt includes several sections: first, it defines the LLM's role; next, it specifies the 21 CAP issue categories and the list of Canadian political parties; then, it provides detailed annotation instructions along with an example; and finally, it outlines the expected output in a standardized JSON format [4].

This structure allows the LLM to simultaneously annotate each sentence for the presence of public policy issues, references to political parties, and overall sentiment. The JSON format was chosen because it supports the return of multiple values within a dimension, such as several issues or party references, and also allows the use of "null" values when no relevant information is present. This unified approach improves consistency across tasks and facilitates the structured comparison of results.

The same corpus is also submitted to dictionary-based methods. As established earlier, this approach involves counting the occurrences of words in the corpus that correspond to expressions contained in a dictionary. To annotate the sample, the Lexicoder Topic Dictionary, the in-house dictionary for political parties, the Lexicoder Sentiment Dictionary, and their translated versions are used.

[4] See the complete prompt in the appendix.

Manual Annotation

A sub-sample of the corpus is annotated manually using a five-coder procedure. The same three analytical dimensions—thematic, referential, and sentimental—are applied, following the same label structure as used in the LLM prompt and the dictionary-based methods. This manual annotation serves as the reference category for the first research question (Q1), which aims to evaluate how the annotations produced by different methods compare to expert human judgment.

The coding is conducted using Doccano, an open-source annotation platform hosted locally. This interface allows for sentence-by-sentence annotation in a blind, standardized environment. Coders do not see each other's answers and annotate each sentence independently, ensuring unbiased coding across the sample.

Each of the five coders is assigned a partially overlapping portion of the sub-sample: 20% of the sentences are shared across all coders, while the remaining 80% are unique to each individual. This setup allows for the calculation of inter-coder agreement based on the overlapping portion. Before proceeding with the individual annotations, coders must reach a minimum Krippendorff's alpha of 0.80 on the shared subset. This threshold ensures a sufficient level of consistency in the coding scheme and interpretation of labels across annotators, thereby reinforcing the reliability of the manually coded reference set.

This annotated sub-sample constitutes the gold standard against which all automated methods are evaluated.

BERT Model Training

The first phase of the approach involves comparing the efficacy of several automated annotation methods (based on the ChatGPT API, five local LLMs, and the dictionary-based approach) for each of the three analytical dimensions. This comparison allows for the identification of the method (or methods) whose annotations most closely approximate the gold standard obtained through manual annotation (**Q1**).

Once the most effective method has been identified for each dimension, an annotated training dataset is constructed using this optimal approach. Based on this dataset, the fine-tuning procedure for a BERT model is initiated, following a process similar to that described by Do, Ollion, and Shen (2022). The model is first trained on a dedicated corpus and then evaluated on a test corpus for which manual annotation serves as the reference.

When the BERT model achieves a satisfactory level of performance, it is used to annotate the entire corpus, thus establishing a fully automated annotation pipeline. To validate the feasibility of an automated pipeline, several BERT models are trained using the different employed methods to verify whether the BERT models perform as effectively when trained on manual annotation as they do when trained on automated annotation.

Once the pipeline is validated, it accepts a prompt that initially generates a sample of annotations. This sample is then used to fine-tune the BERT model. Finally, the trained model proceeds to annotate the complete corpus. This process addresses our second research question (**Q2**) by demonstrating the feasibility of an automated annotation pipeline whose quality approximates that of manual coding. The entire process is illustrated in Figure 2.

Results

The process explained above is currently underway. Preliminary analyses have been conducted using two LLMs, Deepseek R1 and Nemotron, as well as with the dictionaries. 200 sentences from the corpus were manually annotated by five coders to obtain initial performance metrics. The results below are from these preliminary analyses.

To assess overall model performance, we examine aggregated measures of precision and recall. Precision reflects how often the model's predictions are correct. Recall, on the other hand, measures how well the model identifies all relevant labels. The micro-average gives equal weight to each prediction, offering a global view of performance, whereas the macro-average calculates metrics per category and then averages them, ensuring that rare labels are not overshadowed by more frequent ones. As shown in Figure 3, Deepseek achieves the highest micro-precision (68%) and overall performs best in terms of accuracy, while Nemotron demonstrates stronger recall, particularly in the macro-average (63%), suggesting it is better at capturing less frequent categories. In contrast, the dictionary-based approach underperforms on both metrics compared to LLMs.

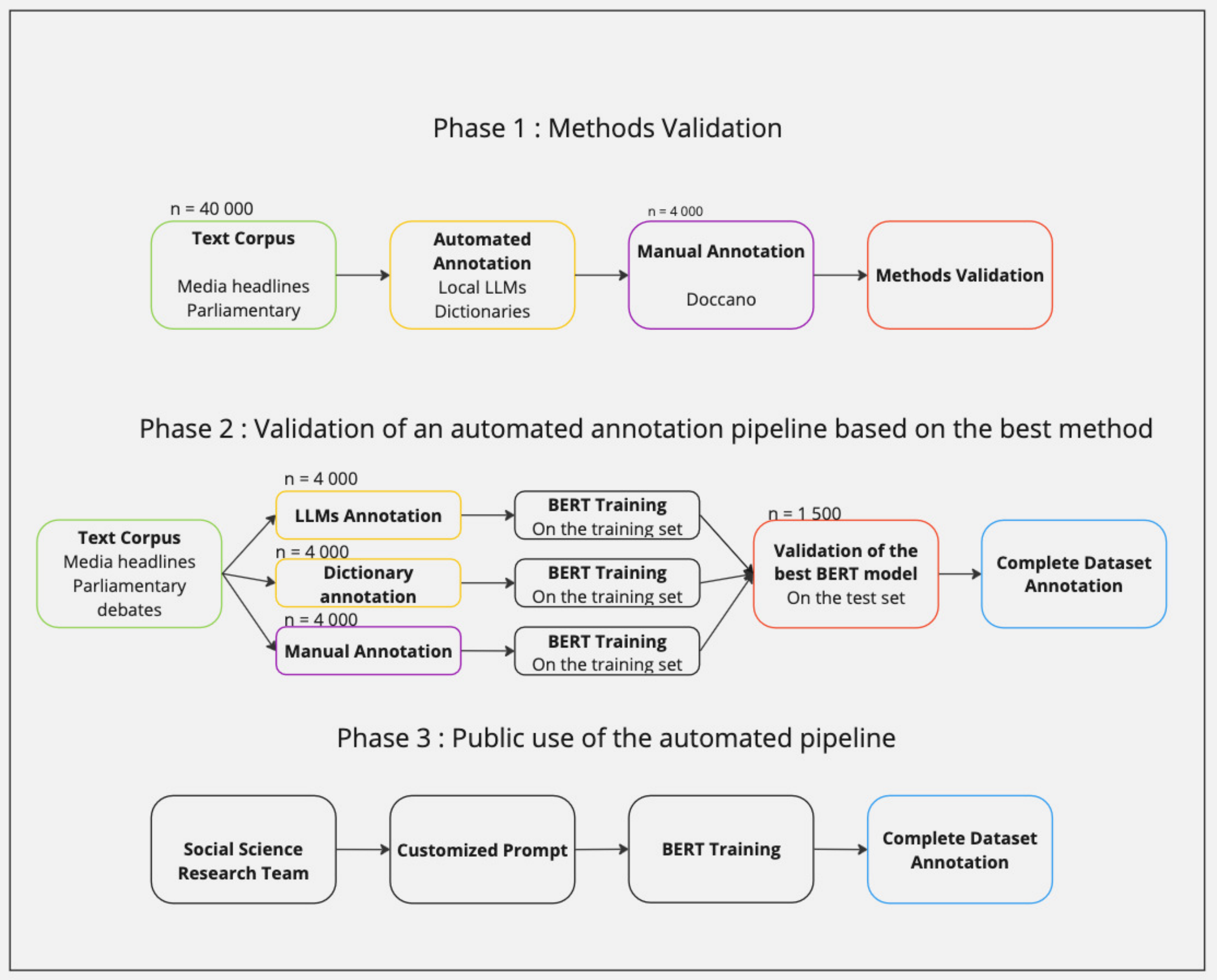

Figure 2: Automated pipeline validation process

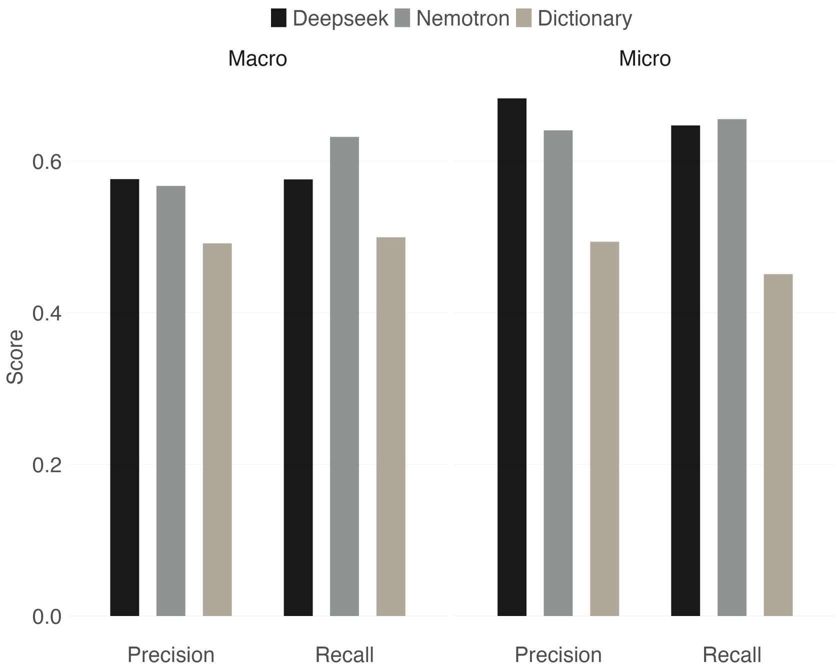

Figure 3: Aggregated metrics

A disaggregated analysis allows for observing the models' performance differences by issue. Figure 4 presents the distribution of F1 scores across issues. The F1 score is the harmonic mean of precision and recall, combining both into a single measure. In this case, it is particularly useful because we aim to identify issues in the most exact and comprehensive way. Since both precision and recall matter, the F1 score offers a more balanced measure of overall performance.

Results show substantial variation in performance across issues, which can partly be explained by the low frequency of some categories. We expect this variation to diminish as more sentences are manually annotated. Nonetheless, interesting patterns are observable. Deepseek achieves the highest overall F1 score, with particularly strong performance on well-annotated issues such as Energy, Environment, Labor, and Education. In contrast, Nemotron performs better than Deepseek on issues that are more difficult to identify, such as Macroeconomics, Rights & Discrimination, and Public Lands. This suggests that Nemotron is more consistent across issue types, whereas Deepseek excels where annotations are clearer.

The choice of the best model to train the BERT classifier depends on the research team's goals. If the objective is to classify a broad range of issues, Nemotron may be the better choice due

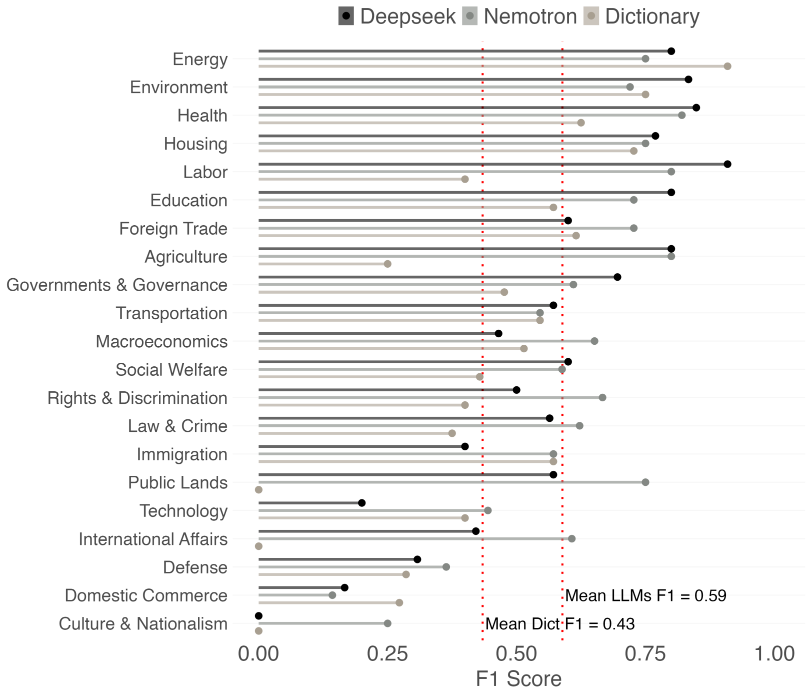

Figure 4: F1 scores Distribution by issue

to its stability across categories. If the focus is on maximizing performance for specific themes, Deepseek appears to be more effective. This illustrates the importance of testing multiple models and annotation strategies. In building an automated classification system, the optimal training method should always align with the specific aims of the research.

As expected, the issue dictionary, with a mean F1 score of 0.43, performs below both LLMs. However, it still shows potential. It achieved its highest performance on the Energy issue, with an F1 score of 0.91. This highlights how a well-crafted dictionary, tailored to the context of the corpus, can be highly effective. At the same time, it also underscores how much care and precision are required to create a comprehensive, adaptable, and up-to-date dictionary across multiple issues. This is where LLMs have a major advantage: they require no domain-specific setup and can capture subtle nuances in language, even though they remain less transparent than dictionary-based methods.

Figure 5 presents the F1 score for each method at annotating the sentiment of the sentences. Once again, the sentiment dictionary underperforms when compared to the LLMs, although it presents slightly better results for negative sentences than for neutral and positive ones. Deepseek presents a higher F1 for each of the three sentiments, suggesting it is better than Nemotron for tone analysis.

Both LLM are better with positive sentences.

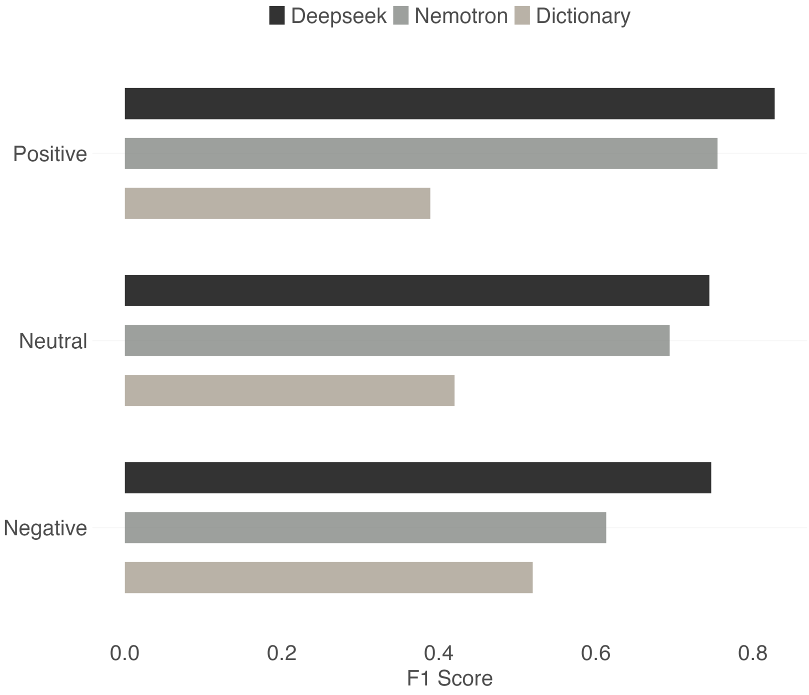

Figure 5: F1 Score distribution by Sentiment

These preliminary analyses confirm the importance of testing multiple methods and models. While the manually coded corpus is not yet large enough to draw firm conclusions about performance differences across issues, the results already reveal that models and methods vary depending on the specific task. Beyond the goal of building a fully automated annotation pipeline for social scientists, we hope this study provides guidance to researchers seeking to determine which automated method is best suited to train a BERT classifier for their study within this framework.

References

Albaugh, Quinn, Julie Sevenans, Stuart Soroka, and Peter John Loewen. 2013. "The Automated Coding of Policy Agendas: A Dictionary-Based Approach." In. Antwerp.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–95. https://doi.org/10.1017/S0003055416000058.

Bevan, Shaun. 2019. "Gone Fishing: The Creation of the Comparative Agendas Project Master Codebook." In *Comparative Policy Agendas*, edited by Frank R. Baumgartner, Christian Breunig, and Emiliano Grossman, 1st ed., 17–34. Oxford University PressOxford. https://doi.org/10.1093/oso/9780198835332.003.0002.

Bilbao-Jayo, Aritz, and Aitor Almeida. 2018. "Automatic Political Discourse Analysis with Multi-Scale Convolutional Neural Networks and Contextual Data." *International Journal of Distributed Sensor Networks* 14 (11): 155014771881182. https://doi.org/10.1177/1550147718811827.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of naacL-HLT* 1 (2).

Do, Salomé, Étienne Ollion, and Rubing Shen. 2022. "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy." *Sociological Methods & Research* 53 (3): 1167–1200. https://doi.org/10.1177/00491241221134526.

Dong, Xiangjue, Yibo Wang, Philip S. Yu, and James Caverlee. 2023. "Probing Explicit and Implicit Gender Bias Through LLM Conditional Text Generation." November 1, 2023. https://doi.org/10.48550/arXiv.2311.00306.

Dunmire, Patricia L. 2012. "Political Discourse Analysis: Exploring the Language of Politics and the Politics of Language." *Language and Linguistics Compass* 6 (11): 735–51. https://doi.org/10.1002/lnc3.365.

Duval, Dominic, and François Pétry. 2016. "L'analyse automatisée du ton médiatique : construction et utilisation de la version française du *Lexicoder Sentiment Dictionary*." *Canadian Journal of*

Political Science 49 (2): 197–220. https://doi.org/10.1017/S000842391600055X.

Ferdaus, Md Meftahul, Mahdi Abdelguerfi, Elias Ioup, Kendall N. Niles, Ken Pathak, and Steven Sloan. 2024. "Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models." June 1, 2024. https://doi.org/10.48550/arXiv.2407.13934.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120. https://doi.org/10.1073/pnas.2305016120.

Gilardi, Fabrizio, Theresa Gessler, Maël Kubli, and Stefan Müller. 2022. "Social Media and Political Agenda Setting." *Political Communication* 39 (1): 39–60. https://doi.org/10.1080/10584609.2021.1910390.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Kennedy, Ryan, Scott Clifford, Tyler Burleigh, Philip D. Waggoner, Ryan Jewell, and Nicholas J. G. Winter. 2020. "The Shape of and Solutions to the MTurk Quality Crisis." *Political Science Research and Methods* 8 (4): 614–29. https://doi.org/10.1017/psrm.2020.6.

King, Gary. 1995. "Replication, Replication." *Political Science & Politics* 28 (3): 444–52.

Lauderdale, Benjamin E., and Alexander Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24 (3): 374–94. https://doi.org/10.1093/pan/mpw017.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (02). https://doi.org/10.1017/S0003055403000698.

Laver, Michael, and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44 (3): 619. https://doi.org/10.2307/2669268.

Lawlor, Andrea, and Erin Tolley. 2017. "Deciding Who's Legitimate: News Media Framing of Immigrants and Refugees." *International Journal of Communication* 11: 967–91.

Linegar, Mitchell, Rafal Kocielnik, and R. Michael Alvarez. 2023. "Large Language Models and Political Science." *Frontiers in Political Science* 5 (October): 1257092. https://doi.org/10.3389/fpos.2023.1257092.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. "Scaling Policy Preferences from Coded Political Texts." *Legislative Studies Quarterly* 36 (1): 123–55. https://doi.org/10.1111/j.1939-9162.2010.00006.x.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. "CamemBERT: A Tasty French Language Model." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–19. https://doi.org/10.18653/v1/2020.acl-main.645.

McGee, Robert W. 2023. "Is Chat Gpt Biased Against Conservatives? An Empirical Study." SSRN Scholarly Paper. Rochester, NY. February 15, 2023. https://doi.org/10.2139/ssrn.4359405.

Pranckevičius, Tomas, and Virginijus Marcinkevičius. 2017. "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification." *Baltic Journal of Modern Computing* 5 (2). https://doi.org/10.22364/bjmc.2017.5.2.05.

Rozado, David. 2023. "The Political Biases of ChatGPT." *Social Sciences* 12 (3): 148.

Salman, Ahmed Hussein, and Waleed Ameen Mahmoud Al-Jawher. 2024. "Performance Comparison of Support Vector Machines, AdaBoost, and Random Forest for Sentiment Text Analysis and Classification." *Journal Port Science Research* 7 (3): 300–311. https://doi.org/10.36371/port.2024.3.8.

Shapiro, Adam H., Moritz Sudhof, Stanford University, Daniel Wilson, and Federal Reserve Bank of San Francisco. 2020. "Measuring News Sentiment." *Federal Reserve Bank of San Francisco, Working Paper Series*, March, 01–49. https://doi.org/10.24148/wp2017-01.

Shyrokykh, Karina, Max Girnyk, and Lisa Dellmuth. 2023. "Short Text Classification with Machine Learning in the Social Sciences: The Case of Climate Change on Twitter." Edited by Nebojsa Bacanin. *PLOS ONE* 18 (9): e0290762. https://doi.org/10.1371/journal.pone.0290762.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22. https://doi.org/10.1111/j.1540-5907.2008.00338.x.

Surowiecki, James. 2005. *The Wisdom of Crowds*. Knopf Doubleday Publishing Group. https://books.google.com?id=_t2KDQAAQBAJ.

Talboy, Alaina N., and Elizabeth Fuller. 2023. "Challenging the Appearance of Machine Intelligence: Cognitive Bias in LLMs and Best Practices for Adoption." *arXiv Preprint arXiv:2304.01358*.

Törnberg, Petter. 2023. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning." April 13, 2023. https://doi.org/10.48550/arXiv.2304.06588.

Tremblay-Antoine, Camille, Steve Jacob, Yannick Dufresne, Patrick Poncet, and Shannon Dinan. 2024. "An Open Window into Politics: A Structured Database of Plenary Sessions of the European Parliament." *European Union Politics* 25 (3): 605–22. https://doi.org/10.1177/14651165241239637.

Van Atteveldt, Wouter, Mariken A. C. G. Van Der Velden, and Mark Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." *Communication Methods and Measures* 15 (2): 121–40. https://doi.org/10.1080/19312458.2020.1869198.

Van den Broek, Merel. 2023. "ChatGPT's Left-Leaning Liberal Bias." *University of Leiden.* https://www.universiteitleiden.nl/binaries/content/assets/algemeen/bb-scm/nieuws/political_bias_in_chatgpt.pdf.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems.*

Widmann, Tobias, and Maximilian Wich. 2023. "Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text." *Political Analysis* 31 (4): 626–41. https://doi.org/10.1017/pan.2022.15.

Young, Lori, and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29 (2): 205–31. https://doi.org/10.1080/10584609.2012.671234.

Zack, Travis, Eric Lehman, Mirac Suzgun, Jorge A. Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W. Bates, and Raja-Elie E. Abdulnour. 2023. "Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare." *medRxiv*, 2023–07. https://www.medrxiv.org/content/10.1101/2023.07.13.23292577.abstract.

Zhang, Yazhou, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024. "Pushing The Limit of LLM Capacity for Text Classification." February 16, 2024. https://doi.org/10.48550/arXiv.2402.07470.

Appendix

Prompt

You are a text annotator. Analyze and summarize the main themes addressed in these one-sentence excerpts from media articles or parliamentary debate transcripts using the following categories to structure the output in JSON. You must write exclusively the JSON with no other explanatory text. The categories must be clear, and appropriate values should be used:

Expected Keys : - `"themes"` : "macroeconomics" if the sentence relates to macroeconomic issues, "rights_liberties_minorities_discrimination" if it relates to rights, freedoms, discrimination, minorities, or freedom of expression, "health" if it relates to health care or medical coverage, "agriculture" if it relates to agriculture, food standards, or fisheries, "labor" if it relates to labor, unions, or pensions, "education" if it relates to education, "environment" if it relates to environmental issues, "energy" if it relates to energy, "immigration" if it relates to immigration, "transportation" if it relates to transportation or infrastructure, "law_and_crime" if it relates to law, crime, or family issues, "social_welfare" if it relates to social protection, "housing" if it relates to housing or urban affairs, "domestic_commerce" if it relates to domestic commerce, "defense" if it relates to defense, "technology" if it relates to space, science, technology, or communications, "foreign_trade" if it relates to international trade, "international_affairs" if it relates to international relations or foreign aid, "governments_governance" if it relates to government operations or intergovernmental relations, "public_lands" if it relates to public lands, water management, or territorial issues, "culture_nationalism" if it relates to culture or nationalism, "null" if the sentence does not clearly relate to any of these themes.

- ' "political_parties" ' : "LPC" if the sentence refers to the Liberal Party of Canada, its MPs or candidates, "CPC" for the Conservative Party of Canada, "BQ" for the Bloc Québécois, "NDP" for the New Democratic Party, "GPC" for the Green Party, "PPC" for the People's Party, "CAQ" for Coalition Avenir Québec, "PLQ" for the Quebec Liberal Party, "PQ" for the Parti Québécois, "QS" for Québec solidaire, "PCQ" for the Conservative Party of Québec, "null" if no political party is clearly referenced.
- ' "sentiment" ' : "positive" if the sentence has a positive tone toward any of the 21 "themes",

political parties, or “specific_themes”, “neutral” if the tone is neutral, “negative” if the tone is negative.

Instructions : - Strictly follow the key structure defined above. Ensure the output conforms to these keys.

- Ensure all keys are present in the JSON, using null where necessary.
- Do not include any keys not defined above.
- Write exclusively in JSON format without additional commentary or explanation.
- Indicate multiple themes if multiple are present.
- Indicate multiple political parties if multiple are referenced.
- Indicate only one sentiment per sentence.

Example annotation for the sentence :

Pierre Poilievre strongly criticized the government’s new policy imposing strict environmental standards on the oil industry, arguing that this excessive regulation would jeopardize business competitiveness and result in significant job losses in producing regions.

Example JSON :

{ “themes”: [“environment”, “energy”, “labor”, “macroeconomics”], “political_parties”: “CPC”, “sentiment”: “negative” }

Follow this structure for each analyzed sentence. Do not add any other comments or details besides the requested JSON structure and specified categories.

Expected JSON Keys { “themes”: ““,”political_parties” : ““,”sentiment” : “ ” }