Political Propaganda and Trade Policy — Identifying Sovereignty Threat Events through Unsupervised Machine Learning Methods

Yung-Chun Chen

M.S. Computational Social Science, University of California, San Diego

May, 2024

All data and source code have been uploaded to my GitHub@Yung-Chun. This article presents a comprehensive framework for detecting sovereignty threat events and conducts basic data analysis. For the complete code structure, please refer to the appendix.

Identifying Sovereignty Threat Events through Unsupervised Machine Learning Methods

Introduction

This project aims at developing a data processing framework for exploring the evolving relationship between national governments and international entities.

A central tension within international relations scholarship is the role of the sovereign state: is it a fundamental component of international society, or is the sovereign state losing its autonomy due to the forces of globalization? According to the concept of the sovereign-state system, a sovereign state must possess domestic autonomy and receive recognition and acceptance from other nations and international organizations (Thomson 1995). The United States used to be recognized as a global unilateral power. However, in the context of globalization, there are transnational forces that undermine state independence (Barkin 2023). As members of the international community, sovereign states may lose some of their power over independent policy-making as they adapt to global politics, contributing to a shift towards a more multilateral approach in international relations.

As a result, this project employs machine learning methods to analyze how the sovereignty of the United States is impacted by the World Trade Organization (WTO), representing a nexus of global countries. The project is structured into two primary components: event extraction and threat detection. By applying topic modeling techniques, we gain a more systematic understanding about the focuses of WTO. Combining this with the feature sovereignty threat, we can then identify so-called “threat events.” This enables us to make predictions of how the U.S. government responds to international pressures and communicates these issues to the American public on social media platforms.

Research Questions

This project intends to address the following questions using unsupervised machine learning methods:

1. **Topic Diversity:** What are the different topics covered in WTO articles? This explores the breadth and diversity of content within the WTO’s publications, identifying the main themes and areas of focus over time.
2. **Sovereignty Threats:** Does the textual content of WTO publications suggest any sovereignty threats to the U.S.? If so, what aspects of sovereignty are implicated? This inquiry examines whether WTO texts contain narratives or explicit statements that could be seen as challenging or undermining the sovereignty of the U.S.

3. **Threat Level:** Are some topics viewed as more threatening than others? This question aims to assess the severity of specific threats, analyzing their nature both locally and globally within the dataset. It includes consideration of both general and specific threat characteristics.

Event Extraction

There may be multiple events over time related to a particular topic. In this project, I employ BERTopic for event extraction. BERTopic (Grootendorst 2022) is a topic modeling framework that diverges from traditional Bayesian network approaches by utilizing a density-based clustering algorithm. The framework entails several crucial steps: firstly, leveraging Sentence-BERT (SBERT), a BERT-based sentence transformer, to generate embeddings for subsequent clustering. Secondly, after dimensionality reduction using UMAP, documents are clustered using HDBSCAN. Thirdly, all documents within a cluster are treated as a single document for representation, employing tokenization and class-based TF-IDF (c-TF-IDF) to unify the representation for a group of clustered documents. SBERT combines the advantages of BERT and SentenceTransformer. BERT, as described by Devlin et al. (2019), captures contextual information bidirectionally within sentences and allows for bidirectional understanding of text through fine-tuning with feature-based approaches such as MNLI, NER, and SQuAD. SentenceTransformer focuses on sentence-level embeddings, as opposed to word-level embeddings in original BERT, making it more efficient and suitable for semantic understanding using machine learning techniques like clustering and topic modeling.

This framework provides flexibility through customizable processes for different use cases and allows fine-tuning of representations without retraining the entire model. For instance, users can opt for different tokenizers or transformers and alternative clustering algorithms. Moreover, they can choose whether to remove stopwords before computing embeddings or during representation generation. In this project, stopwords are removed before computing the embeddings. After clustering the topics, we employ dynamic topic modeling to pinpoint specific events occurring at distinct times. This method allows us to track how topics evolve and identify time-specific events within the dataset.

As a result, compared to classic topic modeling tools such as Latent Dirichlet Allocation (LDA), BERTopic offers several advantages. It excels in representing the semantic content of documents over the bag-of-words concept, processes large corpora with time efficiency and accuracy, and eliminates the need to assign numerous topic numbers for optimal classification, as it operates as an unsupervised method. Furthermore, by assigning each document to a single topic instead of distributing latent topics among documents, BERTopic aligns more closely with the project's objectives and avoids ambiguity when mapping documents to other features.

Threat Detection

Branch & Stockbruegger (2023) provide a comprehensive definition of sovereignty as the supreme authority or power held by a state within its territory, encompassing both internal and external dimensions. Internally, sovereignty pertains to a state's authority over domestic affairs, including law enforcement and service provision. Externally, it relates to a state's independence in interactions with other states and international actors. Despite its significance, sovereignty is not absolute and can face limitations, such as those imposed by international law or interventions in cases of human rights abuses or threats to peace and security.

Building upon this understanding, this project delineates sovereignty threat into four definitions, aiming to capture the implications of threatening messages within WTO publications. Broadly, sovereignty threat against the United States is defined as "anything that might hinder what the United States intends to do," deconstructed into distinct elements for analysis and interpretation.

1. isDisappointment: Whether the article shows the United States' disappointments.
2. isComplain: Whether the article contains complaints about the United States from other countries.
3. isCritic: Whether the article contains criticism of the United States from the WTO.
4. isAffect: Whether the article implies that there is a need for a policy change within the United States.

While large language models have been applied in sentiment analysis, as demonstrated by Fatouros et al. (2023) and Zhang (2023), our project takes a different approach. In our case, positive or negative sentiment doesn't directly align with the definitions of sovereignty threat. Traditional sentiment analysis primarily focuses on differentiating emotional tones within documents. However, our project is centered on understanding the semantic context that indicates sovereignty threats. As such, we aim to capture nuanced expressions related to sovereignty threats rather than simply categorizing sentiments as positive or negative. Therefore, I conduct prompt engineering through OpenAI API and the LangChain framework for annotation, offering a more efficient alternative to manual labeling by researchers.

GPT-4 is an advanced artificial intelligence language model developed by OpenAI, designed to understand and generate human-like text based on the input it receives. For this project, I utilize the gpt-4-0125-preview version, which was trained on data up to December 2023, to perform multiple tasks related to threat detection. LangChain is an open-source framework that connects various large language models (LLMs). It provides semi-standardized structures for application development, enhancing manageability. LangChain includes several features, such as assigning roles for specific prompts, creating prompt formatting templates, and parsing the responses from LLMs into a fixed format. With the concept of "chaining", we are able to interact with LLMs by creating a sequence of prompts comprising elementized steps and obtaining responses in a structured manner.

Methods

Data Collection and Data Cleansing

I developed a web crawler using the Python Selenium package to archive all content from the News and Events section of the website of the WTO. As of January 2024, the crawler extracted 10,110 unique URLs, including articles, audio and video content. To increase the efficiency, interpretability and accuracy of topic modeling, I only dealt with textual data and applied language detection using spaCy. Additionally, for further dynamic topic modeling analysis, only articles with complete format of day, month, and year will be processed. After filtering for English articles with standardized date formats, 8,296 remained, across from 1991 to 2023. Each entry in the dataset includes details such as titles, dates, abstracts, full content, and outbound links within the articles. The primary objective is to prepare the data for Dynamic Topic Modeling, enhancing the understanding of evolving topics in WTO communications. All scripts and datasets are available on my GitHub repository at https://github.com/Yung-Chun/WTO-News-and-Events-Archive-Crawler.

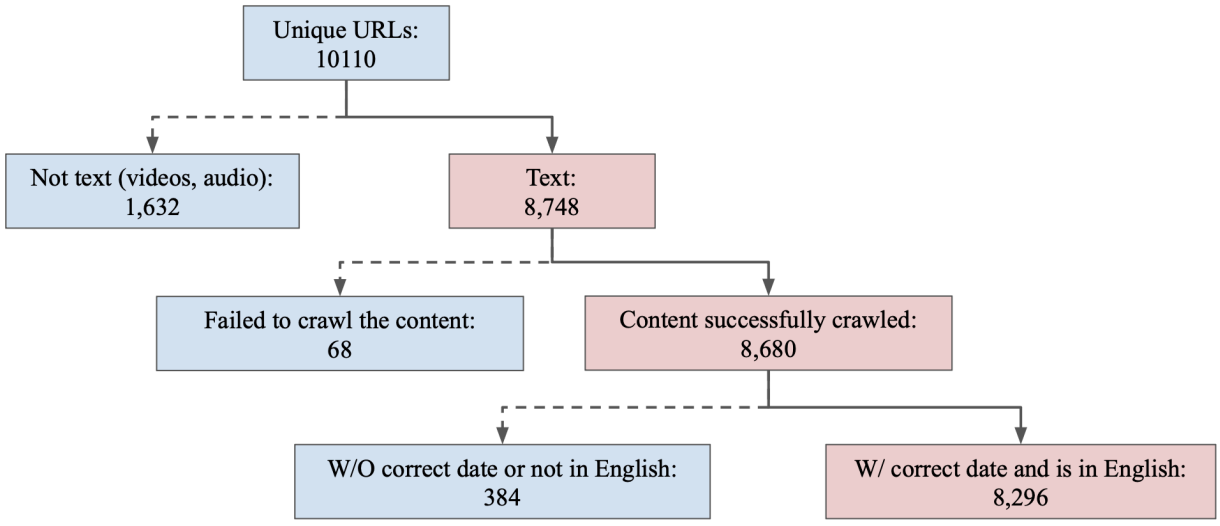

Figure 1. Distributions in the Data Cleansing Process

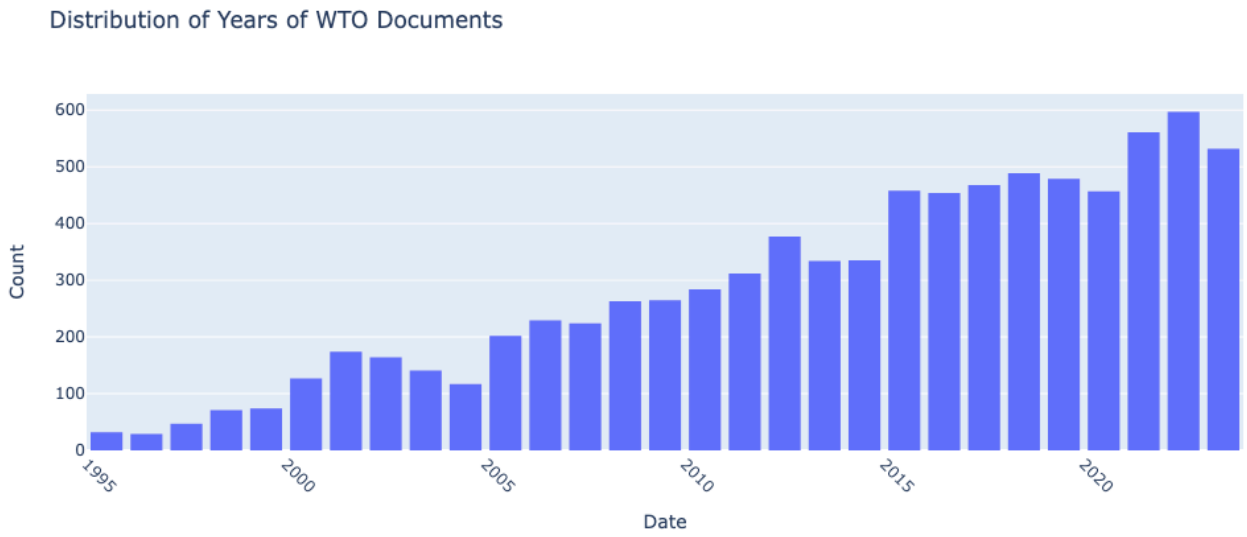

Figure 2. Distribution of Years of WTO Documents

Preprocessing

The preprocessing stage is crucial in preparing the dataset for topic modeling and involves several steps designed to standardize and cleanse the data. These steps include:

1. Removing punctuation
2. Converting to lowercase
3. Lemmatization
4. Removing stopwords and irrelevant terms. (Note: Commonly used words that do not contribute to the semantic meaning, such as “the”, “is”, and “in”, are removed alongside specific terms that might skew the representation of the text’s content. These include words like “download”, “pdf”, “word”, “share”, and any URLs, as they are often not relevant to the core analysis.)
5. Eliminating Digits and Extra Spaces. (Note: All numbers are removed to focus purely on textual data. Additionally, multiple spaces are reduced to a single space to maintain text consistency and readability.)

Embeddings

There are several models in SBERT (Reimers 2019). With the consideration of both performance and time efficiency, this project uses all-MiniLM-L6-v2 as the transformer model. all-MiniLM-L6-v2 is one of the latest models fine-tuned with multiple datasets, including over 1 billion sentence pairs. For more details, see SBERT.net and HuggingFace/sentence-transformers/all-MiniLM-L6-v2.

Topic Modeling

In this study, the BERTopic was utilized to effectively cluster similar articles, thereby generating interpretable thematic representations for distinct events. Central to optimizing BERTopic's performance were two hyperparameters: the minimum topic size (min_topic_size) and the similarity threshold for zero-shot classification (zeroshot_min_similarity). An experiment was conducted by varying min_topic_size from 2 to 10 and adjusting zeroshot_min_similarity in 0.05 increments, ranging from 0.7 to 0.9. This procedure yielded 45 distinct parameter configurations, each offering a unique perspective on the dataset. In addition, the model allows a topic for unclassified documents, avoiding the ambiguity caused by forcibly merging scattered documents to coherent topics.

Evaluation metrics—coherence (Newman et al. 2010), perplexity (Griffiths & Steyvers 2004), and the total number of topics generated—were calculated for each configuration. Coherence measures the meaningfulness of topics generated by a model based on the semantic similarity among the top words within each topic. It evaluates the strength of association between these high-ranking words by analyzing their co-occurrence within a reference corpus. A high coherence score indicates that the words frequently appear close together in texts, suggesting that the topic likely represents a cohesive concept or theme. Perplexity, on the other hand, quantifies the likelihood of the test data under the model, with lower values indicating a more accurate fit of the model to the data. It is calculated as the exponential of the negative average log-likelihood of the test set. A topic model with lower perplexity demonstrates a better capacity to predict unseen documents, thereby generalizing more effectively from the training data to new instances.

The results were visualized through line charts to identify the "elbow points," which signify the most effective balance between model complexity and performance. Interestingly, the relationship between number of topics and the minimum topic size generated demonstrated remarkable stability, providing a consistent benchmark across various similarity thresholds that all tested thresholds converged at a minimum topic size of 4. However, the relationship between coherence and the minimum topic size, as well as the relationship between perplexity and the minimum topic size do not provide a significant converged point for parameter choosing.

Based on these findings, a BERTopic configuration with a minimum topic size of 4 and a similarity threshold of 0.7 was selected. This means that each classified topic should contain at least four documents, and the similarity between each document should exceed 0.7. This configuration ensures a pragmatic balance, maximizing the number of discernible topics while maintaining interpretability.

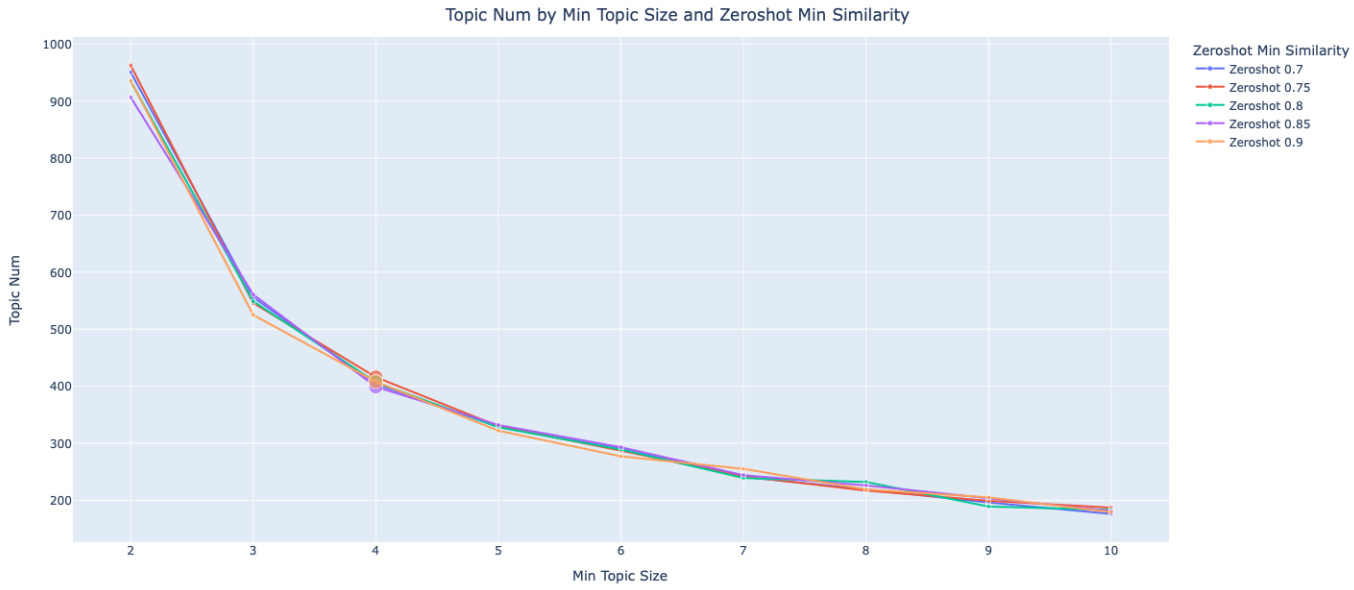

Figure 3. Number of Topics by Minimum Topic Size

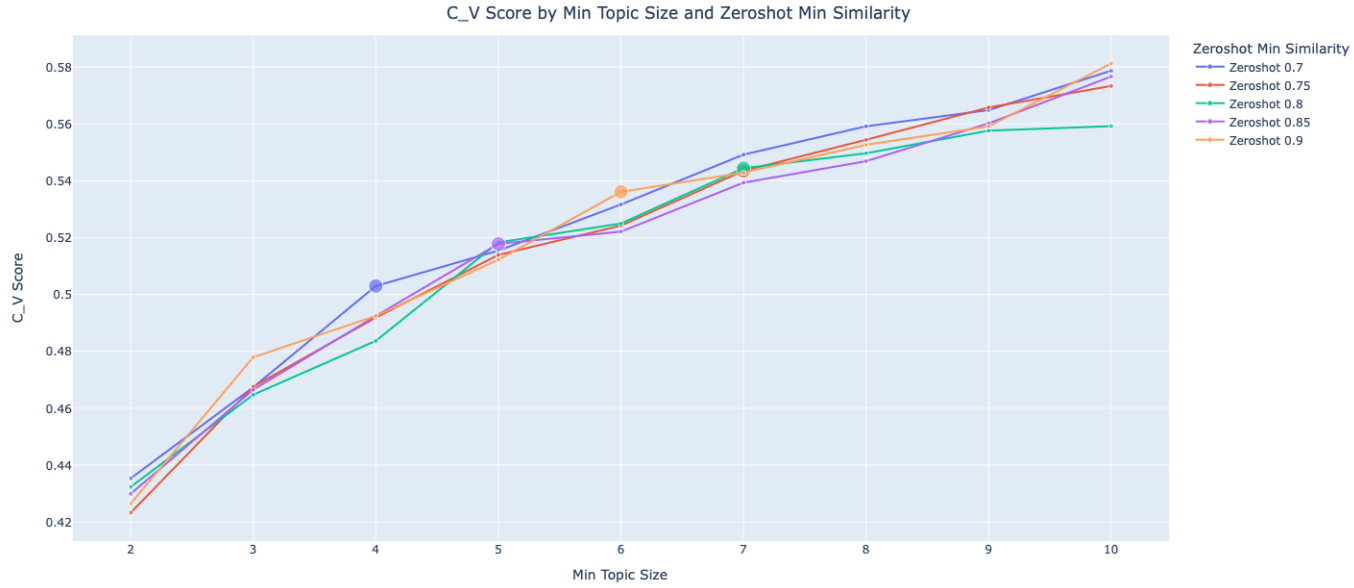

Figure 4. Coherence Score by Minimum Topic Size

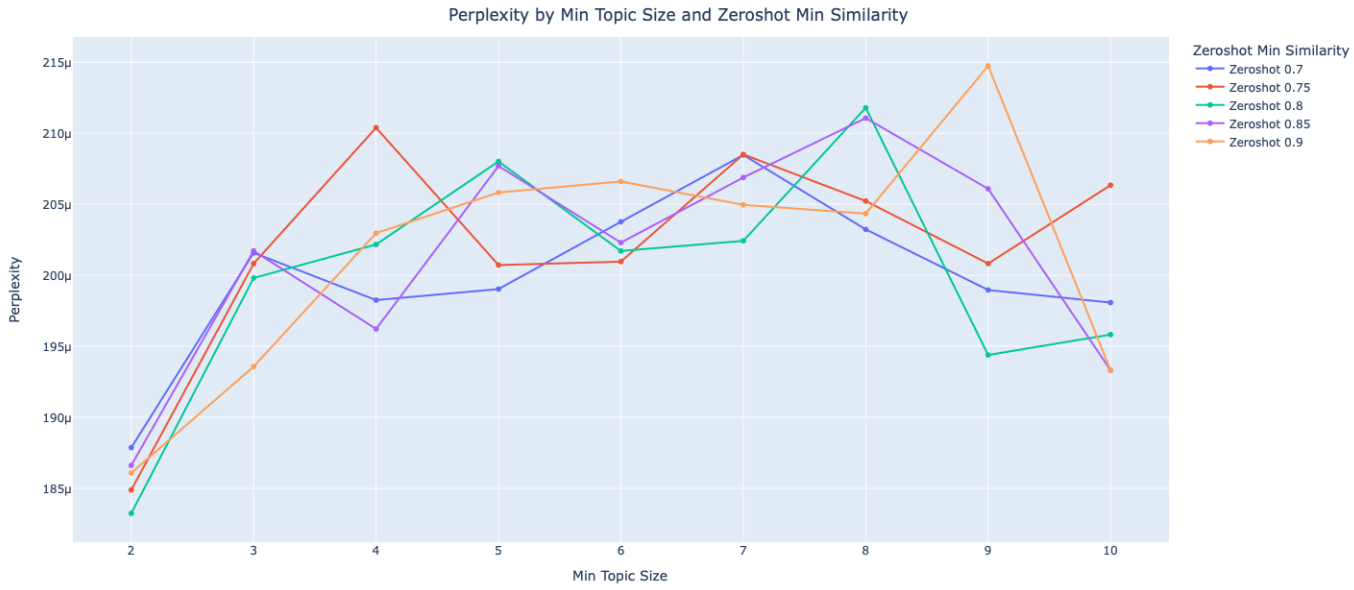

Figure 5. Perplexity Score by Minimum Topic Size

Prompt Engineering

The input message comprises both a system message and a human message. Studies have demonstrated that role-play prompting can significantly enhance accuracy in reasoning (Kong et al., 2024) and argument extraction (Hu et al. 2023) when coupled with specific domain knowledge. Therefore, I designate the system role as a political economics researcher, with a focus on international relations. Regarding the human message, in addition to the article processed in each loop and the breakdown of the definition of sovereignty threat, I construct response schemas to ensure that responses from the OpenAI API conform to a predefined dictionary structure, facilitating easy storage of results as structured output.

In this study, I employ zero-shot learning techniques (Larochelle et al. 2008; Brown et al. 2020) to construct the human message for exploratory purposes. This approach enables the model to recognize and generate responses without being provided with explicit examples beforehand. For each article, the model is tasked with answering questions as outlined in the human message and categorizing the responses according to predefined keys in the response schema. The use of zero-shot learning allows the model to adapt to a wide range of tasks without the need for task-specific data. This flexibility significantly reduces the time and costs associated with collecting and annotating large datasets for each new task.

Table 1 shows the content of the system message and the human message, while Table 2 presents the defined structure for the responses. For more technical insights into chaining input data, prompts, and parsing schemas, please refer to my GitHub repository at https://github.com/Yung-Chun/WTO-Threat-Detection.

Table 1. Prompts

Message	Content
System Message	You are a political economics researcher, focusing on international relations.
Human Message	This is the article: {article} If the article shows the United States' disappointments, please type "yes", else type "no" for the key isDisappointment. If the article contains complaints about the United States from other countries, please type "yes", else type "no" for the key isComplain. If the article contains criticism of the United States from the WTO, please type "yes", else type "no" for the key isCritic. If the article implies that there is a need for a policy change within the United States, please type "yes", else type "no" for the key isAffect. And then give a reason in one sentence to all keys. {format_instructions}

Table 2. Response Schemas

Key Name	Value Description
isDisappointment	Response with only yes or no.
isDisappointment_reason	One-sentence reason for the response.
isComplain	Response with only yes or no.
isComplain_reason	One-sentence reason for the response.
isCritic	Response with only yes or no.
isCritic_reason	One-sentence reason for the response.
isAffect	Response with only yes or no.
isAffect_reason	One-sentence reason for the response.

Results

Event Extraction

In this study, despite the deterministic nature of the HDBSCAN algorithm, over 400 distinct topics were initially identified from an analysis involving more than 8,000 documents. Managing such a large number of topics poses significant challenges, especially given HDBSCAN's parameter sensitivity. Notably, the minimum topic size parameter frequently results in the fragmentation of similar topics into multiple smaller clusters. This fragmentation highlights a critical trade-off: the desire to preserve small yet potentially insightful topics against the practical need for consolidation.

To address this, topic merging is employed as a strategic approach to combine similar topics into larger, more manageable groups. This process is enhanced through the application of Hierarchical Topic Modeling (HTM), which quantitatively assesses similarities between topic pairs and organizes them into a hierarchical structure. Expert knowledge from the relevant domain is then integrated to guide the final merging decisions by examining the representations of the topics. This approach not only facilitates the identification of merge candidates but also supports a more informed consolidation process that aligns with researchers' needs. An example of a subset of these topics is provided in Figure 6, enabling researchers to visually evaluate and determine the optimal sets of topics for merging, based on their thematic similarities and contextual relevance.

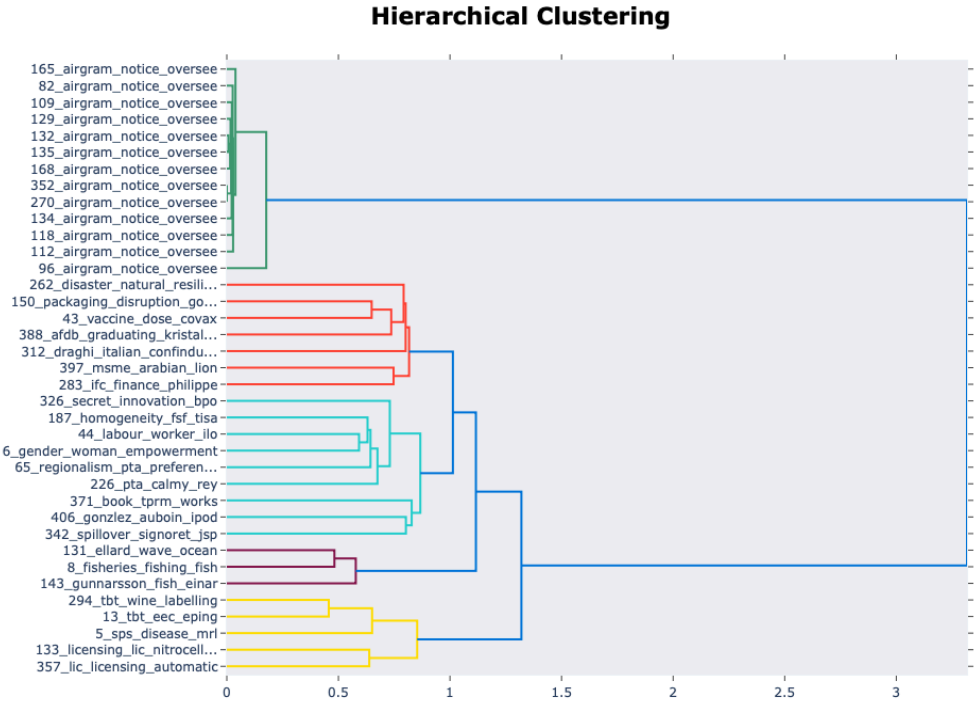

Figure 6. Example of the Hierarchical Topics

This topic merging strategy successfully consolidated the initial clusters into 156 well-defined topics across the dataset. Initially, the topic modeling yielded nearly 400 topic groups, each comprising fewer than 50 documents. By employing a topic merging strategy, informed by expert domain knowledge, the number of these smaller topic groups was significantly reduced to approximately 120. This strategic consolidation effectively diminished the prevalence of less substantiated topics, thereby enhancing the thematic solidity of the results.

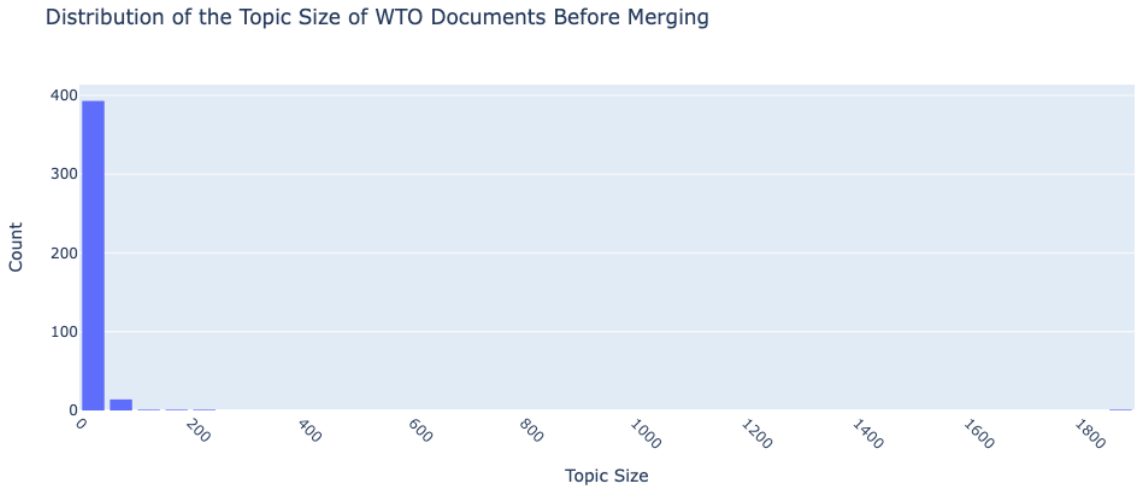

Figure 7. Distribution of the Topic Size of WTO Documents Before Merging

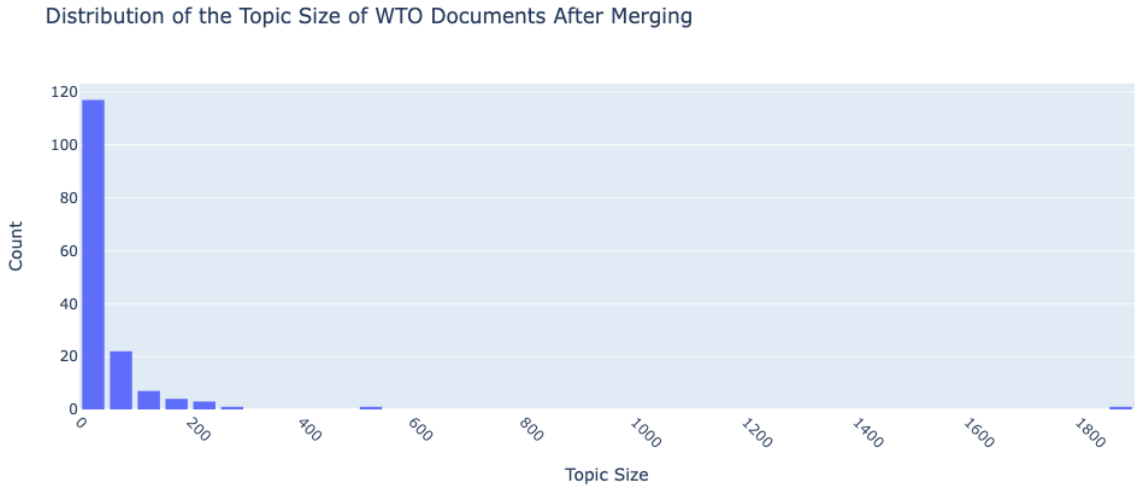

Figure 8. Distribution of the Topic Size of WTO Documents After Merging

After merging the topics, the coherence improved from 0.50297 to 0.60809. Although there was a slight increase in perplexity, it was not significant. Thus, we can conclude that topic merging not only enhances coherence but also maintains the overall solidity of the model.

Table 3. Coherence and Perplexity of Topic Modeling

	Coherence	Perplexity
Before topic merging	0.50297	0
After topic merging	0.60809	0.00017

Table 4. displays examples of the five largest topics out of 156 and their representations. The algorithm automatically assigns index numbers to topics, with unclassified documents labeled as -1. Topic -1 represents the largest group, comprising 1859 documents, with 219 of them containing threatening messages. It is important to note that the “unclassified” topic may simply indicate that these documents do not meet our thresholds, suggesting they lack solidity or scalability to be classified as a distinct topic. Following this, the second largest topic pertains to climate and environment, followed by an academic activity called Regional Trade Policy Course (RTPC), the Dispute Settlement Body (DSB) panel, and the Trade Policy Review Body (TPRB).

Table 4. Top Five Largest Topics

Topic	Name	isThreat	Size	1-Gram Representation	2-Gram Representation
-1	-1_pandemic_gatt_shall_panel	219	1859	['pandemic', 'gatt', 'shall', 'panel', 'request', 'geneva', 'dsb', 'appellate', 'ministers', 'political']	['panel', 'request', 'general', 'multilateral', 'pandemic', 'dispute', 'system', 'need', 'trading system', 'director']
0	0_climate_environmental_food_energy	90	537	['climate', 'environmental', 'food', 'energy', 'crisis', 'security', 'carbon', 'change', 'war', 'environment']	['climate', 'environmental', 'food', 'climate change', 'energy', 'crisis', 'change', 'security', 'food security', 'environment']
1	1_university_rtpc_essay_academic	0	283	['university', 'rtpc', 'essay', 'academic', 'chairs', 'author', 'young', 'award', 'professor', 'course']	['university', 'rtpc', 'academic', 'essay', 'chairs', 'chairs programme', 'course', 'young', 'award', 'author']
2	2_dsb_appellate_ruling_panel	77	231	['dsb', 'appellate', 'ruling', 'panel', 'dispute', 'body', 'dumping', 'settlement', 'establishment', 'dsu']	['dsb', 'appellate body', 'appellate', 'panel', 'ruling', 'dispute', 'body', 'settlement', 'dispute settlement', 'status report']
3	3_concluding_tprb_corrigenda_revisions	0	222	['concluding', 'tprb', 'corrigenda', 'revisions', 'minutes', 'independently', 'frequency', 'interval', 'vary', 'reviews']	['available week', 'review report', 'chairperson concluding', 'government report', 'week meeting', 'concluding remark', 'concluding', 'tprb', 'secretariat report', 'basis review']

Threat Detection

Eventually, 7598 documents (91.59%) were successfully processed by GPT-4, while 698 documents (8.41%) encountered an OutputParserException Error requiring further resolution. Despite these failed cases, it is found that 11.73% of the publications exhibit threatening messages, encompassing dissatisfaction, complaints, criticism from international entities, or domestic policy changes. Specifically, 2.24% express disappointment with WTO policies, 6.53% contain complaints about the United States from other countries, 0.68% feature criticism of the United States from the WTO, and 7.24% suggest a need for policy changes within the United States. Each definition corresponds to a specific threat level, with a single document earning 1 score if it meets any definition. Publications not meeting any definition are categorized as level 1 risk (88.27%). Consequently, 7.52% are at level 2, 3.47% at level 3, 0.72% at level 4, and 0.01% at level 5. This analysis offers a breakdown of insight into the diverse levels of threat perception evident in the publications, helping to assess and rank potential sovereignty threats. Figure 10. also shows the distribution of threat and each aspect of it over time.

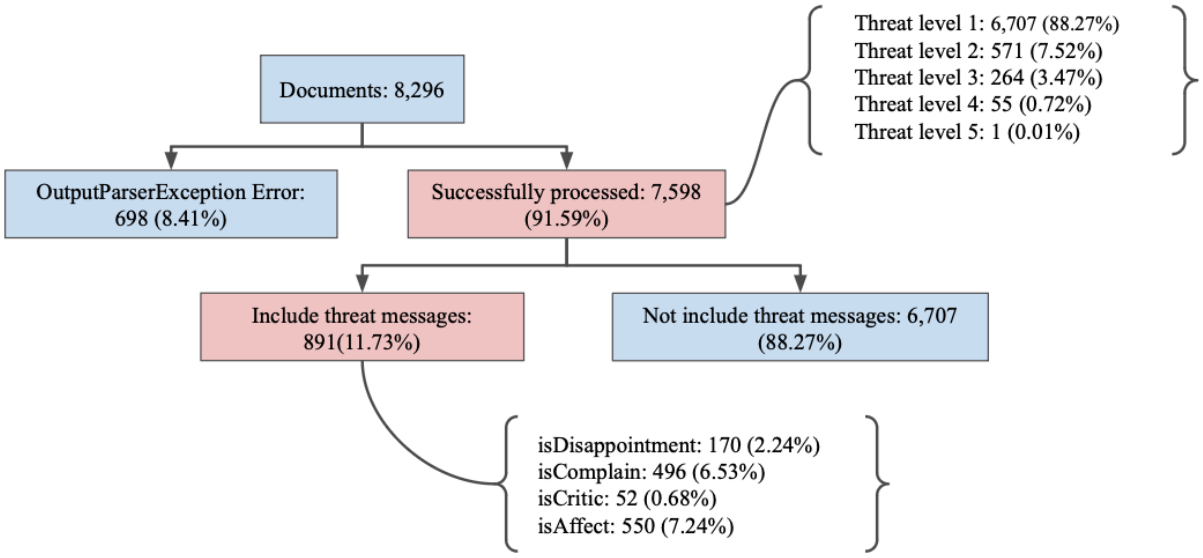

Figure 9. The Distribution of Threat Detection Result on the Document Level

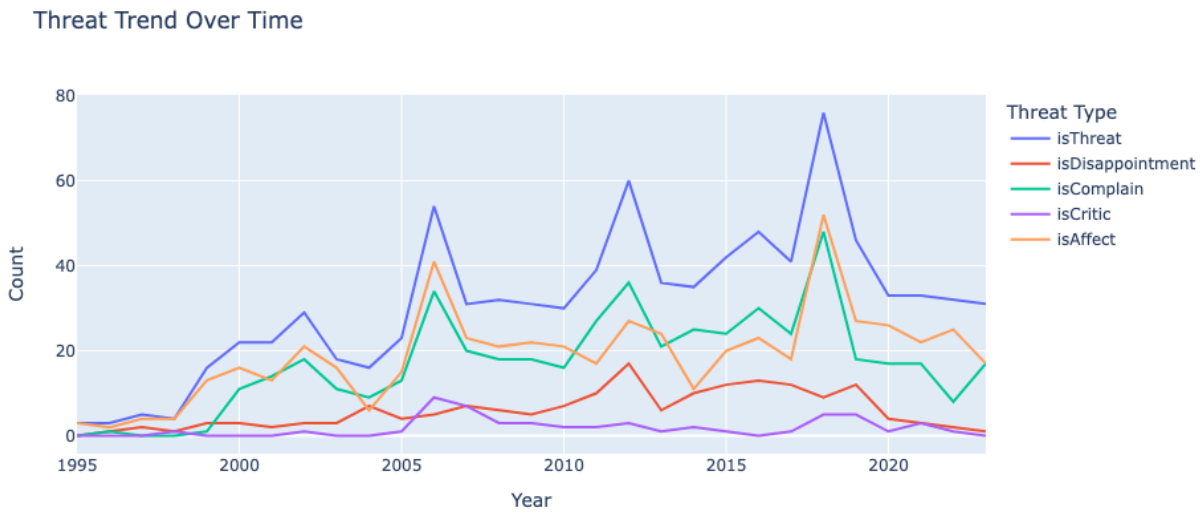

Figure 10. The Distribution of Threat Type Overtime

To assess the effectiveness of annotations by GPT-4 for each component of sovereignty threat, I applied quota sampling to select 20 observations that satisfy the definition, and another 20 observations that satisfy neither. In total, there are 100 samples. This method ensures that we have sufficient data for both positive and negative classifications when evaluating the model's performance. Oversampling was also applied to ensure balanced classification of the ground truth (manual labels) and prevent distortions in accuracy measurements. The accuracy, precision, recall, and F-1 score for each defined component of sovereignty threat are detailed below.

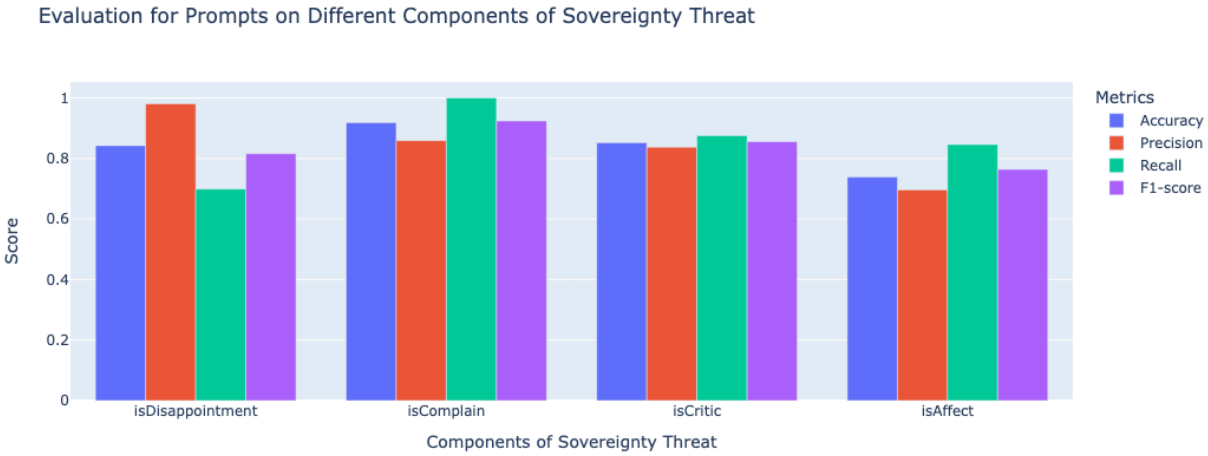

Figure 11. Evaluation for Prompts on Different Components of Sovereignty Threat

The evaluation metrics displayed in the graph reveal distinct performance levels across different components of sovereignty threat. In brief, isComplain stands out with the highest accuracy at

91.79%, indicating that it most reliably identifies complaints. isCritic follows with an accuracy of 85.23% and isDisappointment has a lower accuracy at 84.25%, showing fairly reliable performance. The least accurate, isAffect, at 73.85%, indicates significant room for improvement in accurately detecting affect-related expressions.

The category isComplain excels notably with a recall of 100%, indicating that the model successfully identifies all instances of complaints without missing any, although its precision of 85.90% suggests some false positives are also classified as complaints. This balance yields a high F1-score of 92.41%, reflecting effective detection capabilities.

For isDisappointment, while the precision is exceptionally high at 98.08%, indicating that almost all predictions of disappointment are correct, the recall is 69.86%. This lower recall points to the model missing a significant number of actual disappointments, suggesting a need for tuning the model to capture more of these instances without sacrificing precision.
isCritic shows more balanced metrics, with a precision of 83.70% and a recall of 87.50%. This indicates that while the model is quite reliable in its predictions of criticism, it occasionally includes false positives and misses some true instances of criticism, as reflected in its F1-score of 85.56%.

isAffect has a high recall of 84.62%, demonstrating its capability to capture most instances of affect-related expressions. However, its precision at 69.62% is lower, indicating a considerable number of false positives, which could be due to less precise or ambiguous definitions in the training data.

Overall, while isComplain and isDisappointment demonstrate strong capabilities in specific aspects (recall and precision, respectively), isCritic achieves a reasonable balance between the two metrics. isAffect, however, shows that there is significant room for improvement in precision without compromising its ability to detect true positives, indicating a potential need for further refinement of the model's training parameters and definitions in this category.

By combining all components into isThreat as a general aspect derived from these components, the model's ability to ensure no threats are overlooked (as indicated by the perfect recall) is its strongest attribute. However, the lower precision and moderate accuracy highlight areas for improvement. The model tends to over-classify situations as threats, leading to a higher number of false alarms. This could be problematic in practical applications where false positives could lead to unnecessary actions or resource expenditure. Therefore, refining the model to improve its precision without sacrificing recall could enhance its utility, making it more reliable for detecting genuine threats while reducing false alarms.

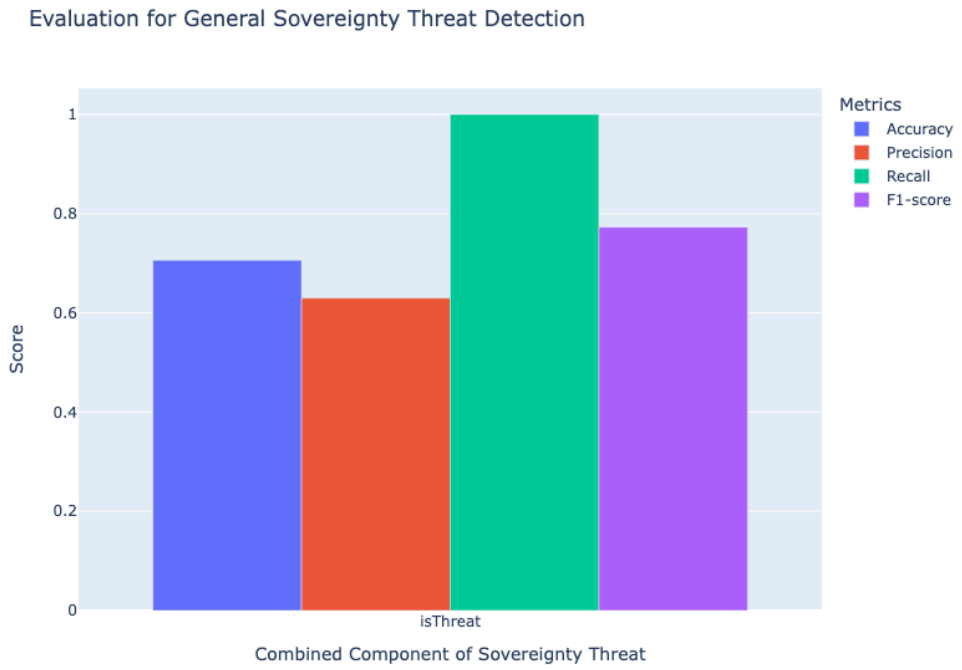

Figure 12. Evaluation for General Sovereignty Threat Detection

Threat Events

After mapping the results of event extraction and threat detection by URLs, we are able to locate threat events on both the topic level and the document level. Given that the sizes of each topic are different, I have created two standardized metrics, local threat rate and global threat rate, to assess the ranking and comparisons among groups. The local threat rate represents the proportion of documents with threatening messages within a particular topic, while the global threat rate is the proportion of documents with threatening messages across all documents. The local and global rates for each aspect of sovereignty are also provided. These metrics provide a clearer understanding of threat prevalence within and across topics, aiding in comprehensive analysis and interpretation of the data. Table 5 and 6 display the examples of five topics of the highest local threat rate and global threat rate.

Table 5. Examples of Top Five Topics of the Highest Local Threat Rate

Topic	Name	Size	LocalThreatRate	isComplainLR	isCriticLR	isAffectLR	isDisappointmentLR
56	56_kb_ab_page_mb	24	1.00000	0.00000	0.00000	0.00000	1.00000
150	150_investigation_dump_glycol_anti	4	1.00000	0.00000	0.75000	0.25000	1.00000
109	109_wt_omnibus_ds176_appropriations	11	0.90909	0.00000	0.90909	0.00000	0.90909
121	121_semi_dump_anti_ad	8	0.87500	0.00000	0.00000	0.00000	0.87500
61	61_ag_ims_rice_milk	21	0.52381	0.04762	0.28571	0.19048	0.52381

Table 6. Examples of Top Five Topics of the Highest Global Threat Rate

Topic	Name	Size	GlobalThreatRate	isComplainGR	isCriticGR	isAffectGR	isDisappointmentGR
-1	-1_pandemic_gatt_shall_panel	219	0.01230	0.00157	0.01784	0.00603	0.01230
0	0_climate_environmental_food_energy	90	0.00084	0.00121	0.01025	0.00048	0.00084
6	6_sps_tbt_animal_disease	83	0.00663	0.00000	0.00325	0.00374	0.00663
2	2_dsb_appellate_ruling_panel	77	0.00820	0.00024	0.00567	0.00374	0.00820
11	11_adjudication_litigation_satisfactory_consultations	46	0.00506	0.00000	0.00289	0.00048	0.00506

By employing dynamic topic modeling, this study explores the trends of specific threat events. For instance, Figures 13-16 display the number of publications for the top five threat topics, sorted by both local and global threat rates. Sorting by the local threat rate helps identify events with a high possibility of containing threatening messages. The data presented show that while these events are indeed threatening, their occurrence is not excessively frequent—they are threatening but occasional. Conversely, sorting by the global threat rate reveals threatening topics that have consistently garnered focus over time and occur frequently.

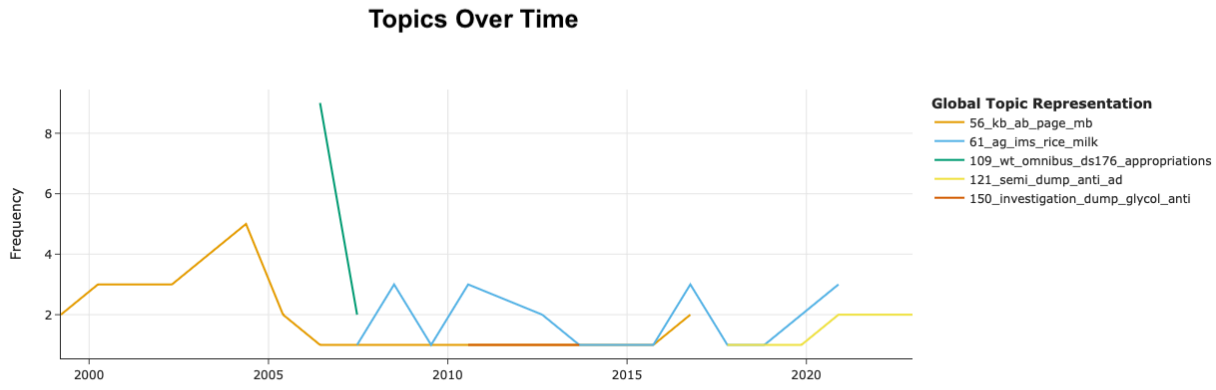

Figure 13. Distribution of Top Five Topics with Local Threat Rate Over Time

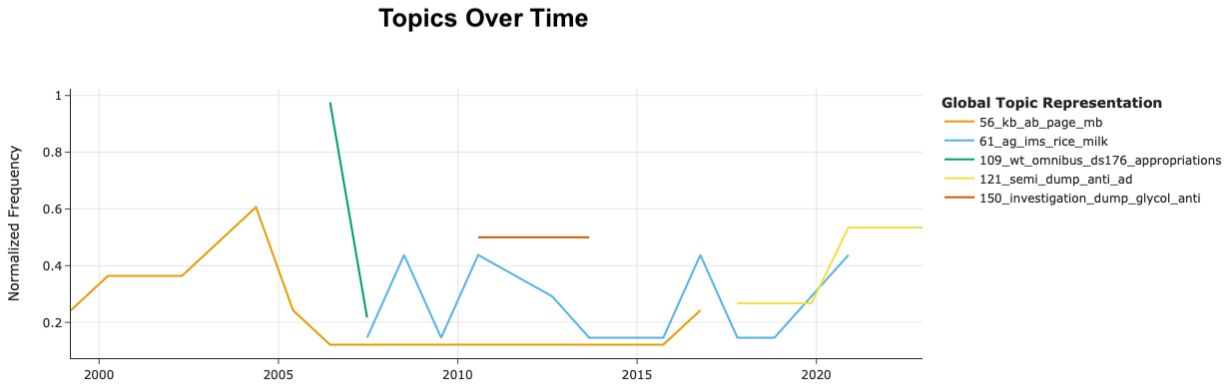

Figure 14. Normalized Distribution of Top Five Topics with Local Threat Rate Over Time

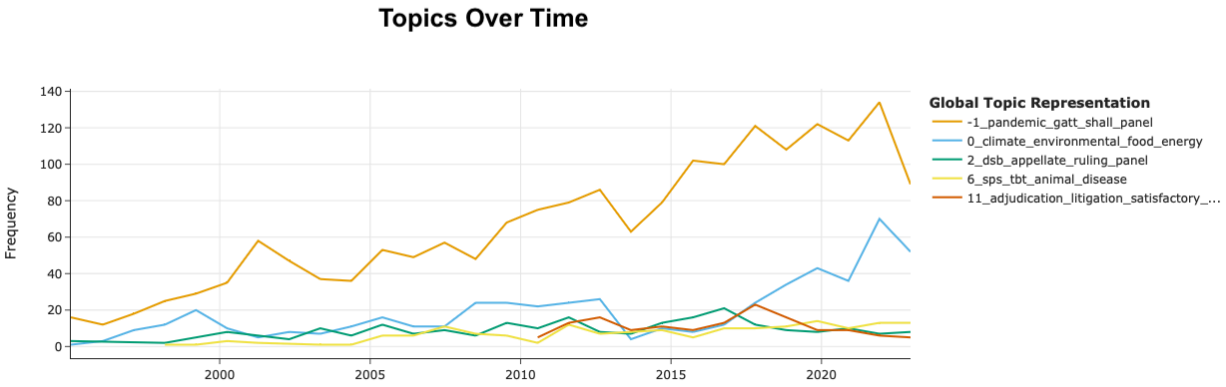

Figure 15. Distribution of Top Five Topics with Global Threat Rate Over Time

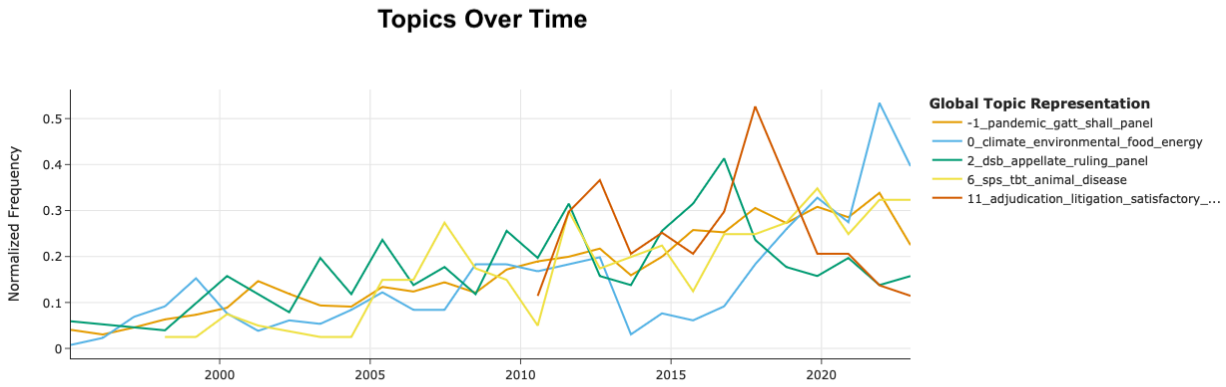

Figure 16. Normalized Distribution of Top Five Topics with Global Threat Rate Over Time

Conclusion

This article outlines a framework for identifying sovereignty threats using machine learning methods. The process handles large volumes of data and delves into the complexities of international relations beyond human capabilities, offering valuable metrics for analysis. For scholars interested in delving deeper into the topic, detailed analyses are available at both the topic and document levels. For topic-level analysis, please refer to the CSV file available at https://github.com/Yung-Chun/WTO-Threat-Detection/blob/main/threat_events_topic_info_merge.csv. Similarly, for document-level analysis, the corresponding CSV file can be found at https://github.com/Yung-Chun/WTO-Threat-Detection/blob/main/threat_events_document_info_merge.csv. These resources offer significant potential for academic inquiry and further investigation into sovereignty threats posed by international entities.

As for further improvement, the primary task should be to refine the prompts of the definitions of sovereignty threats. Firstly, the definitions of isCritic and isAffect need to be more solid and exclusive to reduce ambiguity and improve labeling accuracy. Enhancing model training by explicitly identifying subjects before labeling and providing specific sentences for training can refine the context understanding of the model. Observations from manual labeling validation indicate that the WTO rarely criticizes, acting more as a third-party organization, and instances where WTO is implied to conclude the complaints from other countries should not be labeled as isCritic positive. For isAffect, it is crucial to clarify whether this should only label an article positive if it directly mentions policy changes, and specify the impact on internal versus external sovereignty to refine the scope of isAffect. These steps will collectively enhance the precision and reliability of the model in recognizing and interpreting sovereignty threats.
Furthermore, given that we already have labels based on expert knowledge, we can improve our methods by applying in-context learning (ICL) with few-shot prompting. This approach not only leverages a small number of expertly labeled examples to train the model but also allows for

simultaneous subject identification, enhancing our ability to delve deeper into the nuances of disputes among countries.

As the volume of data increases in the future, we will have the opportunity to extract more topics from unclassified documents. In terms of analytics, incorporating a time window would be beneficial for more precise extraction of specific events. This method enhances our ability to pinpoint and analyze particular events, improving the specificity and relevance of our insights.

Overall, this project establishes the main workflow for generating labels from unstructured textual data and offers insights into creating new metrics for further analysis and sorting. This framework not only facilitates data exploration but also enhances the ability to derive meaningful patterns and conclusions from the information processed.

Reference

Barkin, J. S. (2023). Sovereignty and Globalization. In M. Bukovansky et al. (Eds.), The Oxford Handbook of History and International Relations (pp. 12-23). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-22559-8_2

Branch, J., & Stockbruegger, J. (2023). State, Territoriality, and Sovereignty. In M. Bukovansky et al. (Eds.), The Oxford Handbook of History and International Relations. Oxford Academic. https://doi.org/10.1093/oxfordhb/9780198873457.013.12

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165v4 [cs.CL]. https://doi.org/10.48550/arXiv.2005.14165

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding (Version 2) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., & Kyriazis, D. (2023). Transforming sentiment analysis in the financial domain with ChatGPT. Machine Learning with Applications, 14, 100508. https://doi.org/10.1016/j.mlwa.2023.100508

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Suppl 1), 5228-5235.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2203.05794

Hu, R., Liu, H., & Zhou, H. (2023). Role knowledge prompting for document-level event argument extraction. Applied Sciences, 13(5), 3041. https://doi.org/10.3390/app13053041

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2024). Better zero-shot reasoning with role-play prompting. NAACL 2024. https://doi.org/10.48550/arXiv.2308.07702

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., & Bengio, Y. (2008). Zero-data learning of new tasks. Proceedings of the 25th International Conference on Machine Learning: Workshop on Learning from Unlabeled Data (ICML UDL 2008).

Newman, D., Lau, J.H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 100-108). Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084. https://doi.org/10.48550/arXiv.1908.10084

Thomson, J. (1995). State sovereignty in international relations: Bridging the gap between theory and empirical research. International Studies Quarterly, 39, 213–234.

Zhang, B., Yang, H., Zhou, T., Ali Babar, M., & Liu, X.-Y. (2023). Enhancing financial sentiment analysis via retrieval augmented large language models. In Proceedings of the Fourth ACM International Conference on AI in Finance (pp. 349-356). Association for Computing Machinery. https://doi.org/10.1145/3604237.3626866

Appendix: How to use the code

Data Collection: WTO-News-and-Events-Archive-Crawler

This repository contains a crawler designed to retrieve all archives from the WTO News and Events section. The scripts utilize the Selenium Python package to scrape the website. The data, including articles and their details, is stored in JSON format for easy access and manipulation.

The crawler operates in the following steps:

1. Retrieving Submenu URLs: Execute **get_submenu_URL.py** to collect all submenu URLs under the Archives section (target menu). This script generates a JSON file named **targetMenuUrlDict.json,** which serves as a reference for subsequent steps.
2. Collecting Article URLs: Run **get_article_URL.py** to gather URLs for each article listed under each submenu. For each submenu, the URLs will be saved in a CSV file in the path **f'../WTO_data_article/{menuPathName}'**.
3. Extracting Article Details: Use **get_article_content.py** to scrape and retrieve details from each individual article. If the article is successfully crawled, it will be saved in a JSON file named all_article_content.json, otherwise it is logged as a failed record in **fail_record.json**.
4. Date Correction: Use **correct_date.ipynb** to adjust any discrepancies in dates. Modify the code if you find other patterns for data cleansing. Please note that this repository currently does not offer functionality for label correction.

In the end **all_article_content.json** includes title, raw_date, raw_label, abstract, content, outboundLinks, outboundLinksText, and date under each URL.

Keys	Description	Type
title	The title of the article.	str
raw_date	The date initially crawled from the website.	str
raw_label	The label/category initially crawled from the website.	str
abstract	The abstract of the article.	str
content	The content of the article.	str
outboundLinks	The links (URL) contained in the article.	list
outboundLinksText	The texts associate with the links	list
date	The dates in the correct format.	str

Event Extraction: WTO-Event-Extraction

This repository employs BERTopic framework to apply topic modeling to the publications on the WTO website.

Data Cleansing

Notebook: WTO-data-cleansing.ipynb
Purpose: Filters the articles we need by filtering English articles with the date format "%d %B %Y"
Input:

1. The raw data file **all_article_content.json**

Output:

1. **training_data/wto_eng_docs.json:** Includes the URLs, dates, and content of the articles.

Topic Modeling

Notebook: WTO-BERTopic-topic-modeling.ipynb
Purpose: Explore events by applying BERTopic.
Input:

1. Clean data from **training_data/wto_eng_docs.json**

Preprocessed Data: Stored in **training_data** directory to avoid regeneration.

1. **wto_doc_tokens.json:** Preprocessed and tokenized documents.
2. **wto_embeddings.npy:** Embedding matrix derived from **WTO_doc_tokens.json**.

Model Training Records: Stored in **evaluation** directory.

1. **full_wto_bertopic_param_eva_{curr_time}.json**: Records hyperparameters, coherence, and perplexity values.
2. HTML files visualizing various metrics for evaluation.

Results: Stored in **results** directory.

1. **full_wto_bertopic_probs_{curr_time}.npy:** Probability distributions of topics.
2. **full_wto_bertopic_topic_info_{curr_time}.csv:** Information for each topic.
3. **full_wto_bertopic_document_info_{curr_time}.csv:** Document-specific information.
4. **full_wto_bertopic_hierarchical_topics_{curr_time}.csv:** Recategorized topics via hierarchical modeling.
5. **WTO_HierarchicalTopics.html:** Visualization of hierarchical topic modeling.
6. **WTO_HierarchicalDocuments.html:** Visualization of documents in hierarchical topic modeling.
7. **full_wto_bertopic_merge_group_dup_{curr_time}.csv:** Candidate topics for manual merging.
8. **full_wto_bertopic_merge_group_dup_{curr_time} edit.csv:** Topics selected for merging after expert review.
9. **full_wto_bertopic_document_info_merge_{curr_time}.csv:** Document information post-topic merging.

10. **full_wto_bertopic_topic_info_merge_{curr_time}.csv:** Topic information post-merging.
11. **WTO_IntertopicDistanceMap.html:** Distance map among topics.
12. **full_wto_bertopic_topics_over_time_{curr_time}.csv:** Results from dynamic topic modeling post-merging.
13. **WTO_TopicsOverTime.html:** Visualizes topic evolution over time.

Model Files: Stored as Pickle Files.

1. **WTO_BERTopic_trained.pkl:** Selected model.
2. **WTO_BERTopic_merged.pkl:** Updated model post-merging.

Threat Detection: WTO-Threat-Detection

This repository employs OpenAI's GPT-4 model to analyze publications on the WTO website, focusing on detecting any implications of a potential threat to US sovereignty.

Threat Detection

Notebook: threat_detection.ipynb
Purpose: Identifies sovereignty threat using GPT-4.
Input:

1. **WTO-Event-Extraction/training_data/wto_eng_docs.json**

Output:

1. **threat_detection_output.json:** The structured responses from GPT-4
2. **gpt_errors.log:** Log of errors encountered during document processing.

Results Mapping

Notebook: event_mapping.ipynb
Purpose: Maps the results of event extraction and threat detection. Calculates some metrics of risk.
Input:

1. **threat_detection_output.json**
2. **WTO-Event-Extraction/results/full_wto_bertopic_document_info_merge_{curr_time}.csv**
3. **WTO-Event-Extraction/results/full_wto_bertopic_topic_info_merge_{curr_time}.csv**

Output:

1. **threat_events_document_info_merge.csv:** Threat events based on the document level. Includes the threat level for each document.
2. **threat_events_topic_info_merge.csv:** Threat events based on the topic level. Includes local and global threat rates.