Expanding Your Vocabulary: Topic Integration Using the Segments-as-Topics (SAT) Approach

Roy Gardner[2,3], Matthew Martin[1,2], Ashley Moran[1,2], Zachary Elkins[1,2], and Andres Cruz[1,2]

[1]Department of Government, University of Texas at Austin
[2]Comparative Constitutions Project
[3]PeaceRep, University of Edinburgh

May 2, 2024

Abstract

Topic discovery and integration are essential to maintain vocabularies—the set of concepts underlying a textual corpus. We present a three-stage methodology combining automation and human expertise to assess candidate topics, which we call the segments-as-topic (SAT) approach. To develop the methodology, we use a vocabulary created by the Comparative Constitutions Project (CCP) that tracks more than 330 topics in a corpus of national constitutions. In the (1) SAT generation stage, we formulate topics that are distinct from existing topics, then use a sentence-level semantic similarity model to search for constitution sections (segments) that are similar in meaning to each topic. Domain experts collaborate on the formulation of the topic text until a formulation is identified that produces a set of search results that match the intent of the topic. Once a sufficient number of constitution sections have been matched, the (2) topic expansion stage of the methodology uses the sections themselves to find additional semantically similar sections. These sections are assessed and are either added to the section set or rejected. The process is repeated until no further new sections are found at which point the section set constitutes the definitive set of sections for the topic. Finally, in the (3) validation stage, a panel of scholars decides whether to accept the topic into the CCP vocabulary, after which matching constitution sections are automatically tagged with the topic. Several new topics have been added to the CCP vocabulary with these methods, some of which we present here to illustrate our process and results. The methodology provides researchers with a systematic way to expand existing vocabularies.

Keywords—Topic integration; vocabularies; constitutions

1 Introduction

"The limits of my language means the limits of my world," as the Austrian philosopher Ludwig Wittgenstein (1922) observed more than a century ago. This statement captures a fundamental truth about the relationship between language and reality—our language not only reflects but also shapes our understanding of the world around us. Social scientists must be particularly mindful of this axiom when trying to bridge the gap between an evolving corpus and the conceptual schema underlying their expertise. Vocabularies, as repositories of concepts within textual corpora, serve as the scaffolding upon which knowledge is constructed in a given realm. If we then interpolate Wittgenstein's dictum, the limits of our vocabularies means the limits of our domain.

The importance of vocabularies for knowledge production cannot be overstated. As Cruz et al. (2023, p. 3) note, "by identifying a universe of ideas, a comprehensive [vocabulary] helps to provide a larger perspective on smaller samples of ideas from different populations." A vocabulary (related to taxonomy, schema, ontology), however, is only as comprehensive as the curators are responsive to changes in the conceptual landscape. New topics constantly emerge from diverse sources such as additions to corpora, paradigm shifts within domains, or the cross-pollination of ideas from disparate fields. Robust vocabularies must undergo constant scrutiny and evolution in order to stay up to date. Domain experts, as topographers of conceptual landscapes, are uniquely positioned to spearhead topic integration.

This paper presents a methodology that synthesizes automated classification of text using vocabularies and systematic topic integration by domain experts. This approach enables us to formulate, evaluate, and incorporate new topics into an existing vocabulary. We employ a comprehensive vocabulary curated by the Comparative Constitutions Project (CCP), which tracks over 330 topics within a corpus of national constitutions (Elkins and Ginsburg 2007). A semantic similarity model forms the basis of our framework, allowing us to construct sentence-level encoding vectors for candidate topics, existing topics, and constitution sections without the need for any pre-processing of the text (Cruz et al. 2023; Gardner 2023). We develop an approach to topic curation that uses semantic similarity tools to gather an initial set of segments that form the conceptual baseline of

the topic. We then leverage these segments to find additional relevant segments in our corpus. Together, we refer to this as the segments-as-topic (SAT) approach.

Our methodology has three stages: (1) SAT generation; (2) SAT expansion; and (3) validation of final results. We develop an iterative approach whereby an initial set of segments, gathered using a relatively high semantic similarity threshold, gradually blossoms into a larger set of constitution segments assessed by domain experts at each stage. This process ultimately yields a set of constitution segments that reify the conceptual intent of a candidate topic. Importantly, we also measure the semantic similarity of candidate topics to current CCP topics in order to prevent duplication.

Human expertise plays an indispensable role in our methodology. Rather than relegating domain experts to *post hoc* validation, they are involved in the refinement of candidate topics at every stage. Generation, expansion, and validation necessarily require experts to accept or reject segments matched to topics under consideration in order for the SAT to continue to find relevant segments in the corpus. This collaborative process refines a topic's segments until virtually all of the constitution segments matched to a topic are consistently rejected. Final results are then evaluated by a panel of scholars. Upon validation, the corpus undergoes automated tagging, providing seamless integration of the newly ratified topics into the existing fabric of the CCP vocabulary. Below we discuss a set of candidate topics that have gone through these steps.

Our approach reveals the synergy of natural language processing (NLP) and human acumen, producing topics that are not only semantically similar but also resonant among domain experts. We seek to empower users (scholars, practitioners, citizens) in formulating candidate topics that could make a valuable addition to their own vocabularies. In other words, these tools are designed for application far beyond the constitutional domain. By combining automation and human expertise in topic integration, we open the door to new conceptual frontiers awaiting exploration.

1.1 Measuring Semantic Similarity

Sentence-level semantic similarity measures the degree to which two or more natural language sentences or clauses convey similar meaning. This approach has been applied to a range of tasks including text search (Farouk 2018) and machine translation (Yang et al. 2019). Among the methods used to calculate sentence-level similarity, sentence-sequence representations of sentences show

significant promise (Cruz et al. 2023; Gardner 2023). Sentence-sequence representations account for both the meaning of individual words and their sequential relationships within a sentence (Aggarwal 2022). This approach captures the context of the natural language in which words appear, allowing the model to recognize subtle differences in meaning that arise from word order or phrasing. These representations are particularly effective in comparing sentences that convey similar ideas but use different vocabulary or structure.

Here we employ version 4 of Google's Universal Sentence Encoder (USE v4)[1] to generate high-dimensional numerical representations of sentences, referred to as encoding vectors or embeddings (Cer et al. 2018). Sentences are represented as discrete points in a 512-dimension semantic space, and the distance between two points is used to measure the divergence in the meaning of the corresponding text.

The semantic similarity score σ of two text segments a and b is measured as the inverse of the angular distance between the encoding vectors of the segments. This distance measure performs better on average than cosine similarity (Cer et al. 2018).

$$\sigma(a,b) = 1 - \frac{\arccos\left(\dfrac{\boldsymbol{a}\cdot\boldsymbol{b}}{||\boldsymbol{a}||\,||\boldsymbol{b}||}\right)}{\pi} \tag{1}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the encoding vectors of a and b respectively.

The inverse of this distance produces a semantic similarity score ranging from 0.0 to 1.0. A score of 1.0 indicates that two sentences are identical in meaning and comprise the same words in the same order. As the meanings of the sentences diverge, the similarity score decreases; a score of 0.0 indicates completely distinct meanings.

USE models facilitate efficient and accurate computation of sentence-level encoding vectors, enabling large-scale semantic similarity tasks across multi-language datasets with minimal text pre-processing (Cruz et al. 2023; Gardner 2023). Our choice of the version 4 USE model over other, slightly more accurate models, was determined by performance in standard benchmarks, and speed of computation. In our own tests, we found that USE version 4 was 70 times faster

[1] https://www.kaggle.com/models/google/universal-sentence-encoder/frameworks/tensorFlow2/variations/universal-sentence-encoder/versions/2?tfhub-redirect=true

than USE version 5 when generating encoding vectors, and 30 times faster than SBERT models. However, our methodology is based on the ability to perform sentence-level semantic similarity and is therefore independent of the model generating the vectors. Assessment of other models is on-going and if a better model is found that meets our selection criteria then it will be adopted.

This efficiency is particularly valuable in legal and constitutional contexts, where precise wording can lead to divergent interpretations, impacting the rights and duties within these documents (Goldsworthy 2007). By comparing sections within a corpus, researchers can evaluate their semantic alignment, which is crucial for interpreting legal texts and tracking their evolution over time, given the expressive function of law (Sunstein 1996). Thus, our use of the USE model for measuring sentence-level similarity is particularly well-suited for identifying sections in different constitutions that address similar topics.

2 Methods

2.1 Data Sources

Our document corpus comprises 192 in-force constitutions. CCP has segmented these constitutional texts according to their hierarchical structure (sections, subsections, etc.). We flatten the hierarchy, ignore titles and headers, and analyze only those segments containing the substantive content of constitution sections. Altogether, 192 national constitutions provide a total of 163,102 constitution segments.

To ensure that our new topics are semantically dissimilar from current CCP topics, we use an extended set of 334 CCP topics where each topic comprises a label field and a longer description field. To prepare topics for encoding, we generate grammatical (or near-grammatical) multi-sentence text comprising the label followed by the description. For example, the text segment for the topic on the "right to strike" is structured as follows: *Right to strike. Grants individuals and groups the right to cease work for a period of time, in an effort to exert pressure on their employer.*

2.2 Text Processing

We have an indexed set of constitution segment identifiers:

$$S = \{s_1, s_2, \ldots, s_N\} \tag{2}$$

and the encoding vectors obtained from the USE v4 model for each segment's text:

$$\boldsymbol{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N\} \tag{3}$$

We have a set of indexed identifiers for CCP topics:

$$T = \{t_1, t_2, \ldots, t_M\} \tag{4}$$

and the encoding vectors obtained from the USE v4 model for each topic's text:

$$\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_M\} \tag{5}$$

Segment text is stored in a dictionary where the key is a segment identifier. A segment identifier also identifies a segment's constitution and provides access to constitution metadata. Similarly, topic text and category data are stored in a dictionary where the key is a topic identifier.

2.3 SAT Generation

SAT generation is an iterative process in which initial textual formulations of a topic are tested against the corpus of constitution segments to produce a set of search results. The output of the process is a set of constitution segments that match the meaning of a topic formulation and which provides the seed segments of the topic. The identifiers of any unmatched segments are assigned to the rejected set. SAT generation involves three steps: (1) measuring the semantic similarity of the candidate topic to existing topics in the vocabulary under expansion; (2) measuring the semantic similarity of the candidate topic to text segments in the corpus, in this case constitution segments; and (3) clustering the results from step two to streamline the acceptance or rejection of segments for the seed set, which is then used for SAT expansion.

2.3.1 Measuring semantic similarity to current topics

If a topic formulation is too close in meaning to a current topic, then this may indicate that the topic already exists in the current version of the CCP vocabulary. In such cases, the formulation of the candidate topic (i.e., label and description), must be revised to increase the formulation's semantic distance from existing topics, thus ensuring that segments that match the candidate topic are not already captured by existing topics.

To measure the semantic similarity of a candidate topic to current topics, a vector $\boldsymbol{u}$ is computed that contains the semantic similarity scores σ between the candidate topic c and each of the 334 current topics in T:

$$\boldsymbol{u} = \{\sigma(c, t_1), \sigma(c, t_2), \ldots, \sigma(c, t_M)\} \tag{6}$$

If any similarity score in $\boldsymbol{u}$ exceeds a threshold of 0.7, the candidate topic is considered too similar to existing topics and is either rejected or requires further refinement to minimize semantic overlap. If no similarity score in the distribution surpasses 0.7, then the candidate topic is considered sufficiently distinct and can be further evaluated for inclusion in the expanded vocabulary.

2.3.2 Measuring semantic similarity to constitution segments

Having passed the preliminary test above, the formulation of the candidate topic is used to find semantically similar text segments in the constitutions comprising our corpus. This step involves computing similarity scores between the candidate topic and every text segment in order to identify relevant matches.These are then analyzed to determine their relevance to the topic under consideration. Users begin to construct the seed set for the SAT, ensuring that it is grounded in actual constitutional segments that accurately reflect the conceptual intent of the candidate topic.

A vector $\boldsymbol{v}$ is computed that contains the semantic similarity scores between the formulation of the candidate topic c, and each of the constitution segments in S:

$$\boldsymbol{v} = \{\sigma(c, s_1), \sigma(c, s_2), \ldots, \sigma(c, s_N)\} \tag{7}$$

A threshold θ_{search} is applied to the similarity scores in $\boldsymbol{v}$ to obtain the search results R – the set of above-threshold segment identifiers and their semantic similarity scores:

$$(s_n, \boldsymbol{v}_n) \begin{cases} \in R & if\ \boldsymbol{v}_n \geq \theta_{search} \\ \notin R & if\ \boldsymbol{v}_n < \theta_{search} \end{cases} \tag{8}$$

At this stage, there is no need for a rigid threshold, though employing a relatively conservative threshold can help avoid overwhelming the user with too many matching segments. It can also improve the quality of the results gathered for the seed set in terms of semantic similarity. We initially used a threshold of 0.63 but now prefer a more conservative threshold to better manage the results at this stage, which we discuss in greater detail below.

2.3.3 Clustering search results

The search results in R are clustered to facilitate the user's decision-making process regarding the acceptance or rejection of segments for inclusion in the seed set. A similarity matrix is constructed where both the rows and columns of the matrix map onto the set of segment identifiers in R:

$$R = \{r_1, r_2, \ldots, r_K\} \tag{9}$$

and their encoding vectors:

$$\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_K\} \tag{10}$$

The similarity matrix $\boldsymbol{W}$ is generated by computing the semantic similarity of every pair of search result segments. A threshold $\theta_{cluster}$ is applied to convert the similarity matrix $\boldsymbol{W}$ into a binary-valued matrix $\boldsymbol{H}$ as follows:

$$h_{m,n} = \begin{cases} 0 & if\ w_{m,n} < \theta_{cluster} \\ 1 & if\ w_{m,n} \geq \theta_{cluster} \end{cases} \tag{11}$$

The matrix $\boldsymbol{H}$ represents an undirected graph where the value 1 indicates an above threshold

connection between a pair of search result segments. The connected components of the graph are found using the Python SciPy API[2]. Each connected component identifies a cluster of semantically similar search results. Search results that do not belong to a connected component are referred to as singletons. Search results are organized into cluster and singleton sets before presentation to the user. Clustering groups of semantically similar text segments together makes it easier for the user to identify patterns and determine which segments best capture the essence of the candidate topic. This approach further aids the decision-making process, allowing for more efficient and informed inclusion or rejection of segments for the seed set.

2.4 SAT Expansion

SAT expansion, as the name implies, increases the segment set for a candidate topic, a process that is repeated until no further segments can be found in the corpus. This iterative process consists of two main steps: (1) identifying new constitution segments using the seed set obtained during the SAT generation stage; and (2) clustering these results using the same method as in SAT generation. Accepted segments are added to the topic's segment set, while the remaining segments are assigned to the rejected set. This process is repeated until the SAT no longer expands, meaning no further semantically similar segments are found in the corpus.

2.4.1 Finding new constitution segments

A semantic similarity matrix $\boldsymbol{U}$ is created with topic segments in rows and constitution segments in columns. After creating $\boldsymbol{U}$, the next step is to evaluate the similarity scores between the candidate topic's segments and the constitution segments. This involves analyzing the matrix to identify which constitution segments (discounting the candidate topic's own segments and the rejected set) are semantically similar to the candidate topics. A threshold θ_{topic} is applied to the similarity scores in $\boldsymbol{U}$ to obtain the topic search results W – the set of above threshold segment identifiers:

$$s_n \begin{cases} \in W & if\ u_{m,n} \geq \theta_{topic} \\ \notin W & if\ u_{m,n} < \theta_{topic} \end{cases} \tag{12}$$

[2] scipy.sparse.csgraph.connected_components

Segments with similarity scores above the predefined threshold are considered relevant and are further examined for inclusion in the expanded topic set.

2.4.2 Clustering topic search results

Using the method described in SAT generation above, search results are clustered, and both clusters and individual segments are presented to the user for evaluation. The user reviews these results and accepts segments that align with the topic's meaning. Accepted segments are then incorporated into the topic's segment set, while those not fitting the criteria are placed in the rejected set.

This find and cluster procedure is repeated until no new relevant segments can be added, signaling that the topic expansion is complete. Following the procedure offers several key advantages. Firstly, it allows for tracking of rejected segments, ensuring that only those truly reflecting the topic's core are included. Secondly, it supports systematic refinement and validation of the candidate topic, improving the robustness and reliability of the expanded vocabulary. Lastly, by documenting user decisions and segment selections, the process is transparent and reproducible.

2.5 Topic Validation

A completed SAT represents the final set of segments for a topic, in our case sections of constitutions. Completed SATs are evaluated by a panel of scholars who assess their relevance and accuracy. Domain experts lead the SAT expansion stage to determine which segments should be accepted or rejected. In the final stage, they decide whether to integrate the topic into the vocabulary, considering its potential impact on the vocabulary hierarchy and its contribution to the conceptual framework. While the topic should have undergone thorough review before this stage, further revisions may still be necessary to ensure its substantive value. This may involve revising the segment set or reconsidering the topic's placement within the vocabulary hierarchy. Throughout the process, no results are accepted at face value; rigorous scrutiny is maintained at every step. Once the final revisions are complete, the new topic is formally integrated into the vocabulary, with all changes meticulously documented to ensure transparency. Below we discuss how the SAT approach is applied to expand the CCP vocabulary.

3 Results

We present results obtained by applying our new methodology to create a topic related to the rights and duties of parents toward their children (hereinafter referred to as the parents topic). The parents topic has recently been integrated into the CCP vocabulary, and can be viewed on the Constitute website.

3.1 SAT Generation

The SAT generation stage involves the iterative formulation of topic text designed to capture the intent of the parents topic. Each formulation of the topic text is tested by finding a set of semantically similar constitution sections, which are ultimately accepted or rejected. The final formulation of the parents topic reads as follows:

> "Grants parents certain rights or duties regarding their children. May include the duty to provide for one's legal children, or the right to make decisions about their education, upbringing, and other aspects of their lives."

The process of SAT generation necessitates manipulation of the search and cluster thresholds. For the parents topic, a final search threshold of 0.63 was used, which most effectively balanced the ratio of accepted to rejected segments and ensured that the user did not have too many results to evaluate. This threshold, however, is likely not conservative enough, for reasons discussed below. With a search threshold of 0.63, the SAT generation stage produced 203 results divided into two sets: the accepted set comprising 120 segments which constituted the SAT set S, and the rejected set R comprising 83 segments.

3.2 SAT Expansion

A semantic similarity matrix is generated with the SAT segments in rows and the complete set of constitution segments in columns. A threshold of 0.72 is applied to the matrix and the set of above threshold constitution segments, known as the SAT-found set F, are obtained from the similarity matrix. In other words, F contains those constitution segments that are semantically similar to one or more segments of S. The relationship between S and F is shown in Figure 1. It

can be seen that $S \subset F$.

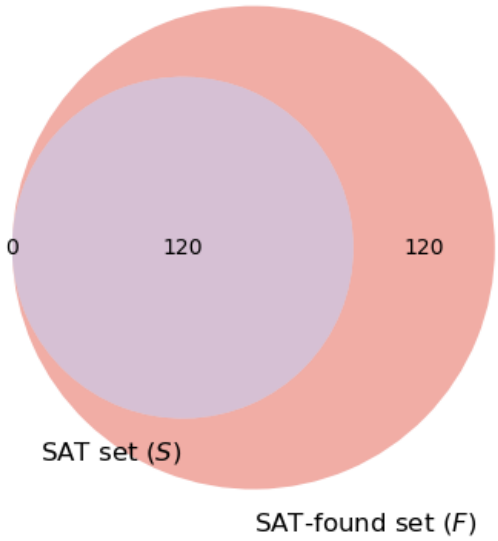

Figure 1: Venn diagram showing the relationship between SAT segments S and SAT-found segments F. The SAT set S is a subset of the SAT-found set F.

Next, we identified the segments A in F that were not in S, i.e., $A = F \setminus S$. The set A comprised 120 segments (see pink partition of Figure 1). We then identified the segments B in F that were not in the rejected set R, i.e., $B = F \setminus R$. The set B comprised 236 segments and is shown in the Venn diagram in Figure 2.

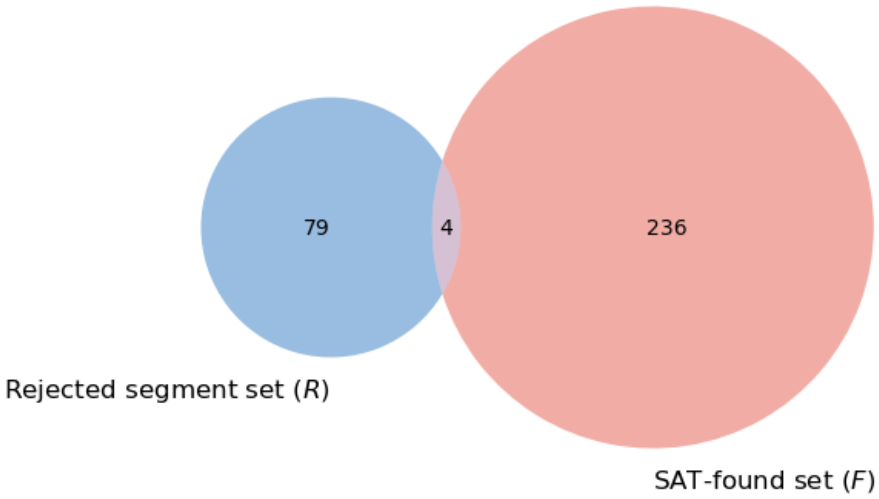

Figure 2: Venn diagram showing the relationship between SAT segments S and rejected segments R.

Finally, we found the intersection $A \cap B$ which comprises 116 segments shown in the Venn diagram in Figure 3. These 116 segments were then clustered and presented to the user for analysis.

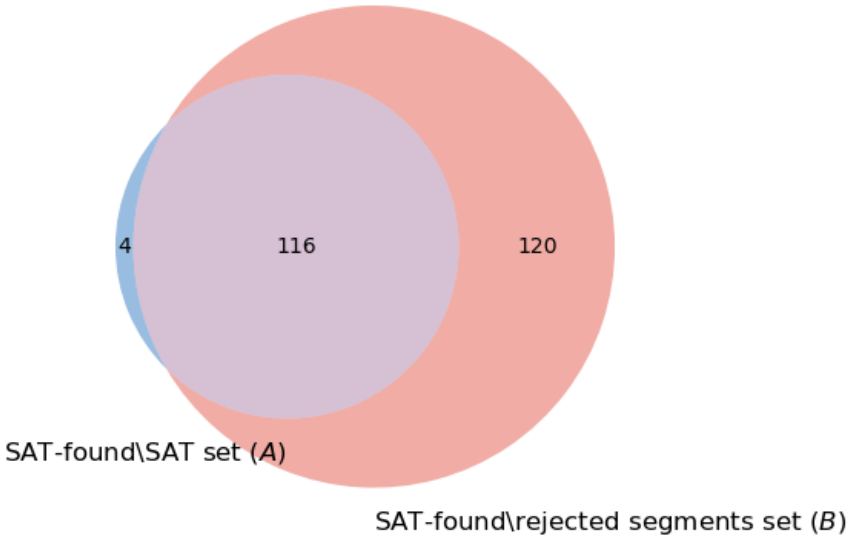

Figure 3: Venn diagram illustrating the search space of 116 segments in the intersection of A and B, i.e., $(F \setminus S) \cap (F \setminus R)$.

In the first pass we found 30 segments in the intersection set $A \cap B$ that were judged to match the intent of the parents topic. These 30 segments were added to the SAT set S to bring the total of SAT segments to 150, and the remaining 86 segments were added to the rejected set R. The process described above was repeated with the updated S and R sets in order to create new values of the sets A and B. As can be seen in Figure 6 the search set is reduced to 13 segments. Within this set, one additional segment was found that was added to S to bring the total number of segments to 151. At this point the SAT expansion process was terminated.

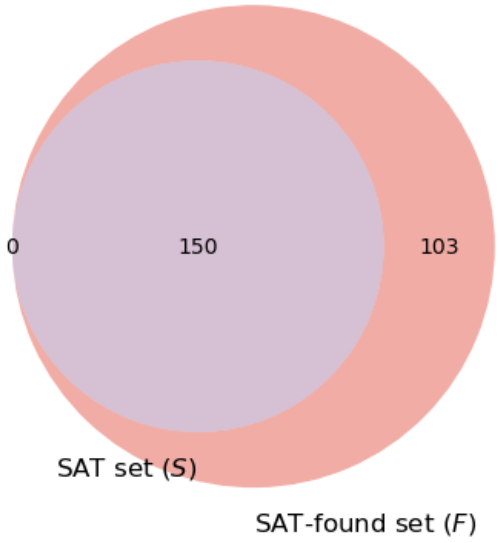

Figure 4: Venn diagram showing the relationship between expanded SAT segments S and SAT-found segments F. The expanded SAT set S is a subset of the SAT-found set F.

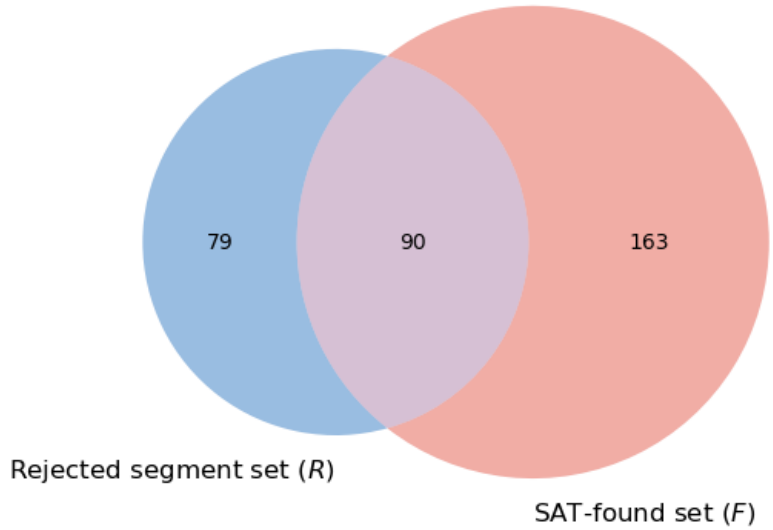

Figure 5: Venn diagram showing the relationship between expanded SAT set S and the expanded rejected set R.

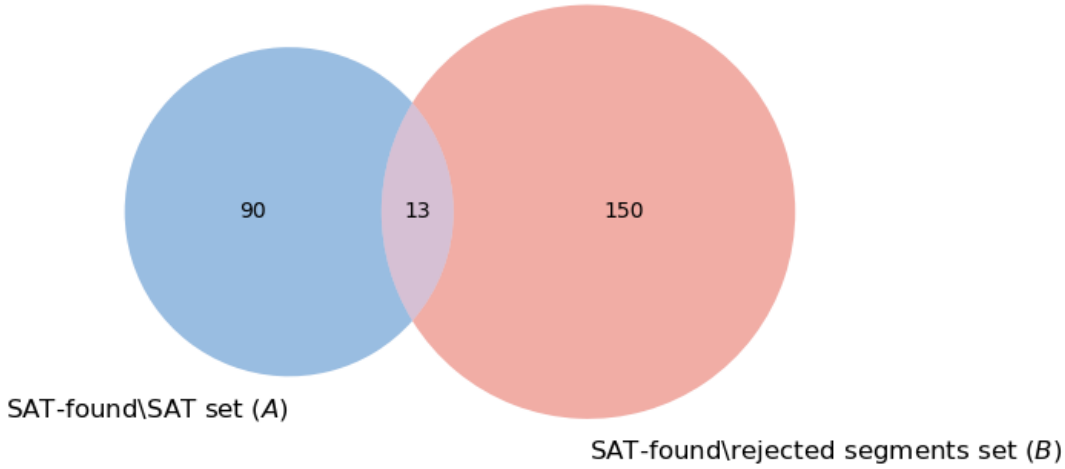

Figure 6: Venn diagram illustrating the search space of 13 segments in the intersection of A and B, i.e., $(F \setminus S) \cap (F \setminus R)$

3.3 SAT Validation

The final SAT was presented to a panel of experts at the Comparative Constitutions Project for validation. Our research team currently consists of two directors, a project manager, a research director, a research associate, and four senior research analysts. Upon review by our full research team, the parents topic was accepted for integration, making it topic number 335 in the CCP vocabulary.

3.4 Automated Tagging

To conclude the process, our corpus underwent automated tagging by the new parents topic. In other words, the topic was applied to the SAT segments in the XML files that constitute our corpus of 192 in-force national constitutions. Using the SAT segments to tag the corresponding constitutions sections thus avoids potential human errors of manual tagging such as failing to tag a discrete SAT segment, or erroneously tagging a segment that does not form part of the SAT.

4 Discussion

Self-assessment has been crucial for the development of our methodology. In fact, the SAT approach developed here is itself the product of an internal validation exercise. Our prior framework rested on an iterative process of topic reformulation and manipulation of search and clustering thresholds to generate an optimized set of results, which introduced some inefficiencies. Most importantly, users of the semantic search tools were responsible for reviewing distinct sets of matches based on new topic text and/or thresholds. If the topic formulation had been more concise, for example, some relevant segments might not have been matched to the parents topic. Conversely, if a less conservative threshold were employed, the user could have been inundated with hundreds of results, making the identification of acceptable segments time consuming. In other words, testing new topic text and threshold combinations is inherently inductive, allowing users to explore possible new topics but with less certainty that all relevant segments in the corpus have been captured. Compared to the SAT approach presented here, there was not a systematic way to narrow the pool of constitution segments for human review. Rather than standing on its own, this process is now part of the SAT generation stage, providing the seed set of segments used to expand the

SAT approach. From here, the process of SAT expansion, through a series of iterations in which users accept or reject segments found by the SAT, gives us greater confidence that our current methodology more effectively identifies relevant segments within a given corpus without losing track of rejected results.

It is important to acknowledge that achieving complete certainty in identifying all sections within a corpus relevant to a particular topic is often impractical, if not impossible. However, the strength of our methodology lies in the systematic and iterative approach we employ, which allows us to gather and evaluate evidence comprehensively. Our methodology involves multiple stages, the formulation of candidate topics in order to generate topic segments, the expansion of topic segment sets, and validation by domain experts and a panel of scholars. At each stage, we leverage both automated tools and human expertise to refine and validate our findings. Thus, while we acknowledge the inherent uncertainty, we are satisfied that our approach yields sufficiently robust and resonant results for practical use and further exploration.

In this way, the validation stage is fundamental for assessing the effectiveness of our methodology. Validation is ultimately a human decision, meaning false positives and false negatives are less of a concern. When using automated classification or tagging, the system may incorrectly identify sections as matching to a particular topic (false positives), or fail to identify segments that actually belong to the topic of interest (false negatives). In contrast, manually tagging sections depends on human expertise at each iteration. The method discussed above necessarily encourages the use of low search thresholds at the topic generation stage in order to harvest accepted *and* rejected results. Users then identify and correct false positives by rejecting segments that do not align with the conceptual intent of the topic, or vice versa for false negatives. These results provide insight into performance, specifically whether a user's formulated topic text is generating results with a satisfactory proportion of matching segments. Because the final set of results is evaluated by a panel of scholars, moreover, a form of inter-coder reliability is institutionalized in our process. If these domain experts conclude that some additional segment should be added to, or removed from, the final set defining a topic, for example, the risk of false positives and false negatives is further mitigated.

Finally, our approach may prove *complementary* to other techniques under certain circumstances. To return to the issue of identifying as many relevant segments in a corpus as possible, there may be scenarios in which a dictionary-type search is useful. For example, in any corpus, including our own, there may exist long, multi-concept sentences that are missed by the semantic search method because the topic-segment semantic similarity score is below threshold. Such segments might be found by a dictionary search.

5 Conclusions

Our methodology has expanded the boundaries of our vocabulary, and thus the scope of our conceptual world. By combining automated text classification and expert-driven topic curation, we have developed the segments-as-topic (SAT) approach that now empowers us to identify and integrate new topics into the Comparative Constitutions Project (CCP) vocabulary with far greater ease. In fact, our team currently has six topics on deck, poised for systematic evaluation and, hopefully, further enrichment of the CCP vocabulary. Importantly, all our team members have access to these tools, giving us equal opportunity to propose new topics for collaborative expert review. By harnessing the individual initiative of our domain experts in this way, we wish to ensure that our vocabulary remains up-to-date and reflective of contemporary constitutional discourse.

Lastly, these tools have considerable potential for tasks and domains beyond those of the present study. The possible applications of the SAT approach are many, but it is worth spotlighting a few promising examples. First, lawyers and legal researchers, for instance, often sift through vast amounts of case law to find relevant precedents and legal principles. Our methodology could be used to automate the classification and integration of new case law into existing legal taxonomies, making it easier to identify pertinent sections of case law when conducting research.

Second, our tools could serve contract law and regulatory compliance by systematically tracking changes across versions of legal documents. By analyzing and comparing different versions, we could identify how specific provisions persist, disappear, or emerge over time. This capability would permit legal professionals to monitor the evolution of contractual terms and regulatory requirements, ensuring that all updates and modifications are accurately captured. This not only

aids in maintaining compliance with current legal standards but also provides valuable insights into the historical context and development of legal agreements and regulations.

Lastly, our methodology holds promise for other research projects in development at the CCP. For instance, our tools could be utilized to create a sub-vocabulary specifically focused on topics related to public consultation, providing a deeper understanding of how these concepts are referenced by constitutional drafters. By systematically identifying and categorizing phrases, terms, and sections related to public consultation through semantic similarity analysis and N-gram searches, there is the potential to build a comprehensive sub-vocabulary that captures the diverse ways in which public input is used and discussed by political elites. This specialized vocabulary could enable researchers to track the evolution of public consultation themes across different drafts and versions of constitutional texts, offering deeper insights into the role and influence of public participation in the drafting process.

Our methodology not only serves forward-looking objectives, as discussed above, but also encompasses retrospective goals. Most importantly, our next step is to expand existing topics from the CCP vocabulary that were formulated before we adopted semantic similarity technology. In the past, these topics were manually tagged by the CCP team, meaning we searched through our corpus of 192 constitutions, as well as number of historical and draft constitutional texts, for specific provisions that matched the corresponding topics. Now, our SAT approach has the power to improve the comprehensiveness of our search for relevant (i.e., semantically similar) segments of text. We can identify additional constitutional provisions that may have been overlooked during the manual tagging process. In other words, our methodology allows us to reduce the margin of human error and ensure a more comprehensive exploration of the corpus, maximizing the coverage and depth of our topic integration efforts. In essence, we are not just expanding our vocabulary; we are expanding our horizons.

References

Aggarwal, Charu C. (2022). *Machine Learning for Text*. Second Edition. Springer.

Cer, Daniel et al. (2018). *Universal Sentence Encoder*. arXiv: 1803.11175 [cs.CL].

Cruz, Andrés et al. (Dec. 2023). "Measuring constitutional preferences: A new method for analyzing public consultation data". en. In: *PLOS ONE* 18.12. Ed. by Jerg Gutmann, e0295396. URL: https://dx.plos.org/10.1371/journal.pone.0295396.

Elkins, Zachary and Tom Ginsburg (2007). *Comparative Constitutions Project*. URL: https://comparativeconstitutionsproject.org/.

Farouk, Mamdouh (Dec. 2018). "Sentence Semantic Similarity based on Word Embedding and WordNet". In: *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, pp. 33–37.

Gardner, Roy (2023). "Semantic Analysis to Support Peace Analytics". In: *Peace Analytics Series. PeaceRep: The Peace and Conflict Resolution Evidence Platform, University of Edinburgh*. URL: https://peacerep.org/publication/semantic-analysis-to-support-peace-analytics/.

Goldsworthy, Jeffrey (June 2007). "1Introduction". In: *Interpreting Constitutions: A Comparative Study*. Oxford University Press. eprint: https://academic.oup.com/book/0/chapter/142958195/chapter-pdf/39245089/acprof-9780199226474-chapter-1.pdf. URL: https://doi.org/10.1093/acprof:oso/9780199226474.003.0001.

Sunstein, Cass R. (1996). "On the Expressive Function of Law". In: *University of Pennsylvania Law Review* 144.5. Publisher: The University of Pennsylvania Law Review, pp. 2021–2053. URL: https://www.jstor.org/stable/3312647 (visited on 08/14/2024).

Wittgenstein, Ludwig (May 1922). *Tractatus Logico-Philosophicus*. Trans. by Frank P. Ramsey and Charles Kay Ogden. London, United Kingdom: Kegan Paul, Trench, Trubner & Co. Ltd., pp. 573–588.

Yang, Mingming et al. (July 2019). "Sentence-Level Agreement for Neural Machine Translation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3076–3082. URL: https://aclanthology.org/P19-1296.