HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2023 No. 088

Natural Language Processing Support for Evaluating Similarity in Physical Therapy Laws, Rules, and Regulations

Final Report

Prepared for:	Healthcare Regulatory Research Institute (HRRI) 124 West St. South, 3rd Floor Alexandria, VA 22314	**Prepared under:**	Work Order 14 under Master Services Agreement dated January 9, 2020
Authors:	Jeffrey A. Dahlke	**Date:**	July 27, 2023

Natural Language Processing Support for Evaluating Similarity in Physical Therapy Laws, Rules, and Regulations: Final Report

Table of Contents

Table of Contents (Continued)

List of Tables

List of Figures

Natural Language Processing Support for Evaluating Similarity in Physical Therapy Laws, Rules, and Regulations: Final Report

Background

The Healthcare Regulatory Research Institute (HRRI) partnered with the Human Resources Research Organization (HumRRO) to evaluate the feasibility of computer-automated methods for detecting similarities and differences among the laws, rules, and regulations (LRRs) governing the practice of physical therapy (PT) across the Federation of State Boards of Physical Therapy's (FSBPT) member jurisdictions. Specifically, HRRI asked HumRRO to explore whether natural language processing (NLP) techniques could help to automate comparisons of LRRs across jurisdictions.

NLP is a discipline closely related to computational linguistics with roots in the interdisciplinary use of artificial intelligence and computer science principles to develop computerized models of human language. NLP researchers aim to use such models to understand and analyze text produced by humans, especially in high-volume contexts where systematic manual analyses of the text would be impractical. NLP methods are quite varied, and NLP tools exist to automate a wide variety of tasks, such as language recognition, language parsing, structural/syntactical analysis, representation of text in terms of quantitative dimensions, semantic analysis (of words, sentences, or longer pieces of text), sentiment analysis, and topic modeling, just to highlight a subset of use cases. Such a diverse collection of techniques makes it possible to automate many facets of text-based analyses and streamline tasks that would take human analysts many hours of labor to accomplish without automation. Beyond the diversity of NLP techniques available to support automation, users of these techniques can also select the degree to which a task is automated (e.g., should an NLP tool automate all aspects of an analysis, or only part of a process while leaving room for human intervention?).

HRRI developed an interest in cross-jurisdictional LRR comparisons after learning about NLP-based research HumRRO conducted to identify related occupations for the O*NET program (Dahlke et al., 2022). In that research, the semantic similarity between pieces of text describing O*NET occupations represented two thirds of the information that contributed to a relatedness composite, which was then used to construct a list of related occupations for each O*NET occupation. Theoretically, the types of pairwise comparisons HumRRO used to develop that related occupations framework could be extended to analyses of LRRs across jurisdictions.

HRRI's desire to explore similarities in LRRs across jurisdictions developed out of FSBPT's licensure assessment program. In addition to administering the National Physical Therapy Examination (NPTE®), which is required for licensure as a physical therapist or physical therapy assistant in the US (APTA, 2023; FSBPT, 2023c), FSBPT also offers jurisprudence exams and Jurisprudence Assessment Modules (JAMs®) for a subset of jurisdictions. More than half of jurisdictions in the US require licensees to pass a jurisprudence exam, and FSBPT develops, maintains, and administers jurisprudence exams for Arizona, California, the District of Columbia, Florida, and Nebraska (FSBPT, 2023b). Currently, each jurisdiction that requires licensees to pass a jurisprudence exam relies on an exam that is tailored to the PT-related LRRs in that jurisdiction, and the jurisdiction-specific nature of the exams makes scores non-transferrable among jurisdictions. FSBPT also offers JAMs—online assessments that licensees can take to meet requirements in some jurisdictions when renewing their licenses—for Georgia, Hawaii, Kansas, New Hampshire, New Jersey, Ohio, Oregon, and Texas (FSBPT, 2023a).

Identifying areas of overlap and nuance between the LRRs from different jurisdictions could help FSBPT to develop micro-assessments for practitioners interested in changing jurisdictions or practicing in multiple jurisdictions. Such micro-assessments could help practitioners to understand the similarities and differences between the jurisdictions' LRRs. NLP methods have potential for helping to identify these types of similarities and differences, and the goal of the present research was to evaluate the feasibility of using NLP tools to guide the detection of relevant overlaps and nuances among LRRs from a representative sample of jurisdictions.

We used LRR information provided by HRRI to conduct feasibility analyses in which we used NLP approaches to automate the detection of similar text between jurisdictions. The remainder of this report is dedicated to describing the methods, results, and implications of our feasibility analyses.

Method

Jurisdictions and Text Sources Sampled for Analysis

HRRI designated a set of five jurisdictions that would be appropriate for use in this feasibility study: Alaska, Arizona, Kansas, Oklahoma, and Texas. These jurisdictions represent a diverse subset of US states, and Table 1 shows how the jurisdictions vary in terms of membership in the physical therapy compact, board structure type, and coverage by FSBPT's jurisprudence exams or JAMs.

Table 1. Overview of Jurisdictions Included in this Research

Jurisdiction	Member of Physical Therapy Compact?	Board Structure (A-E)	Has FSBPT-Developed Jurisprudence Exam or JAM?
Alaska	No	C	No
Arizona	Yes	A	Yes
Kansas	Yes	D	No
Oklahoma	Yes	E	No
Texas	Yes	A	Yes

HRRI supplied the relevant LRRs for these jurisdictions, and some jurisdictions had two sources from which we extracted LRRs. We have outlined the file(s) we used for each jurisdiction in Table 2.

Table 2. Sources and Types of Jurisdiction-Level Information

Jurisdiction	File Name	Type of Information
Alaska	Alaska PT-OT Statutes and Regulations	Statutes/Laws
		Regulations
Arizona	Arizona Revised Statutes; Title 32 Chapter 19 (Laws) (revised 6-1-19)	Statutes/Laws
	Arizona (Rules) Code TITLE 4 Chapter 24	Rules
Kansas	Kansas Statutes Annotated and Kansas Administrative Regulations	Statutes/Laws
		Regulations
Oklahoma	Oklahoma PT+LAW_11.2021	Statutes/Laws
	Oklahoma PTRULES_09.2020	Rules
Texas	Texas PT-Practice-Act-2021	Statutes/Laws
	Texas PTrules_2022.03	Rules

As a supplement to jurisdictional information, HRRI also supplied a copy of FSBPT's 7th edition of the Model Practice Act (MPA) for inclusion in our analyses. The MPA contains recommended language for states to adopt when revising their physical therapy act legislation, and its inclusion in our NLP analyses was meant to provide an additional nexus for identifying sets of related LRRs.

Analysis Procedures

We used semantic similarity analyses to automatically detect potentially related LRRs across jurisdictions, followed by a human review at the end of the process to (a) screen out false positives (i.e., LRRs erroneously identified as related) and (b) evaluate the quality of the machine-detected relations. We describe our analysis process below, along with preparatory steps we took to organize and clean the text we used as inputs to our analyses. We used a combination of the *R* and Python computing languages in this process; we used Python to convert text to quantitative representations and we used *R* for all other analytic procedures, including text preparation and cleaning.

Step 1: Organize and Parse Text

First, we organized and prepared the individual LRRs for NLP processing. Specifically, we extracted all relevant text from the supplied files, then cleaned, parsed, and organized the text in a way that would be compatible with our NLP analyses.

Before describing the steps in our text preparation process, it is important to define the nomenclature we use to describe the specificity of the text. From most general to most specific, we use the terms *source*, *file*, *document*, and *document element* (or simply *element*). A *source* is the jurisdiction that produced a collection of text, and we also consider the MPA a source. A *file* is a specific PDF- or Word-formatted collection of text (as listed in Table 2). A *document* is the text from an individual LRR, while a *document element* is a distinct text segment from a document that represents either a single component from a bulleted list or a non-list piece of text that is separated from other document elements by line breaks. The level of analysis in this research typically focuses on either document elements or documents.

Step 1.1: Convert All Files to Plain Text Format. All files we received for the jurisdictions were in PDF format, and the MPA was in Word's "docx" format. Neither of these formats is ideal as an input to NLP analyses, so, before we could begin operating on the text, we needed to convert the files to a format that *R* could easily read in. We copied the content from all files into plain text files that were compatible with our analysis strategy.

Step 1.2: Clean Text to Eliminate PDF Artifacts. After we converted all files to plain text format, we reviewed these files to identify and remove formatting artifacts introduced during the conversion process. Text copied from PDF files can include non-ASCII characters, unnecessary line breaks, irrelevant text from headers and footers, and unintentionally concatenated text where spaces between words are missing. We made corrections for such formatting artifacts before proceeding with the text.

Step 1.3: Parse Text into Documents and Document Elements. We parsed the cleaned text into documents, using the table of contents (TOC) from each file as our guide. The section titles listed in the TOCs from all files used in this research corresponded to distinct LRRs and, as these section titles were also present directly before the text for their respective LRRs,

we used the section titles from the TOCs as delimiters to divide each file's text into individual LRR-level documents.

We also subdivided each document into document elements, using line breaks as our guide; each line break within a document marked the beginning of a new element. When a document included elements presented as a list, we recorded each element's level of hierarchical organization within that list (i.e., we coded the first-level bullets as 1, the second-level bullets [bullets beneath a first-level bullet] as 2, and so on). When a document contained elements that were not part of a list, we coded them at being at level 0 of the organizing hierarchy. This organizing hierarchy allowed us to reintroduce the proper level of indentation for each document element when we rendered the documents for review later in our analysis process.

Step 1.4: Remove Enumeration from Document Elements' Text. After partitioning lists of text into elements, we performed a follow-up cleaning step to remove leading text from elements that defined the elements' positions in list structures (i.e., bullet-point symbols, letters, and numerals used to separate and identify list components). For example, we removed leading text such as "-", "A.", and "(1)" because this text does not contribute to the semantic meaning of an element and, if left in, might have diluted the meaning of quantitative representations of the text. We reintroduced these position indicators when we rendered the documents for review.

Step 1.5: Exclude Documents Not Suitable for Analysis. As a final step in our text-preparation process, we excluded text that was not appropriate for analysis. We excluded introductory text from each source because it did not contain direct information about LRRs, as well as LRRs that were marked as having been repealed or revoked because they are no longer applicable to the practice of physical therapy. We also removed LRR's titles from the text that we analyzed, but reintroduced these titles when we rendered the documents for review. At the direction of HRRI, we dropped (a) LRRs that were specific to occupational therapy (all of which came from Alaska), (b) text that dealt with the disposition of funds (this only impacted two LRRs from the MPA; one of which was a blank statute and the other of which was marked as an optional statute), and (c) text describing the Physical Therapy Licensure Compact (Physical Therapy Compact Commission, 2021). The PT Licensure Compact text is common across jurisdictions that have adopted the Compact, such that identifying relations between identical sets of Compact language would not contribute to the utility of our NLP similarity analyses and would have had no bearing on conclusions about the feasibility of these analyses.

Step 2: Convert Elements to Quantitative Embeddings

We used a pre-trained sentence transformer model to develop numeric representations of document elements for us in subsequent analyses. Sentence transformers are NLP models that express text in terms of scores on quantitative dimensions derived from prior analyses performed on large and diverse collections of text. These numeric values are known as embeddings and characterize the semantic content of text in ways that support empirical comparisons among pieces of text.

We used the "all-distilroberta-v1" model (Sentence-transformers/all-distilroberta-v1, 2023) from the "sentence-transformers" Python package (Reimers & Gurevych, 2019) to develop element-level embeddings; it is based on the "DistilRoBERTa base" transformer model (Sanh et al., 2019) and was fine-tuned using more than 1 billion sentence pairs. This model is among the top-performing sentence-level transformers and, in past research, we have found this model to be a good all-around option for detecting relations among pieces of text. The embeddings generated by this model summarize the input text on 768 numeric dimensions.

Step 3: Evaluate the Similarity of Documents

We used the numeric embeddings for element-level pieces of text to evaluate the similarity of text from different sources, and we used those element-wise similarity metrics to compute aggregated document-wise similarity metrics. These similarity metrics formed the basis of our procedure for clustering LRRs. We describe our similarity and clustering analyses in the subsections that follow.

Step 3.1: Compute Cosine Similarity for Pairs of Documents. Cosine similarity is a metric for quantifying the alignment between two vectors of coordinates in a multidimensional space. It is the most common metric for comparing the similarity between vectors of word-, sentence-, or paragraph-level embeddings in NLP research. Cosine similarity values can be very useful for identifying commonalities among collections of text that reveal deeper associations among the subjects of those texts. For example, Dahlke et al. (2022) used cosine similarities among the text-based descriptions of occupations as a major component of the procedure HumRRO developed to identify related occupations in the O*NET program's occupational taxonomy.

Cosine similarity is computed as the inner product of two vectors, divided by the product of the vectors' respective norms. The cosine similarity between two vectors called **x** and **y** with *k* dimensions each would be computed as follows (first in vector algebra, then in scalar algebra):

$$Cosine\ Similarity = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \times \|\mathbf{y}\|} = \frac{\sum_{i=1}^{k} x_i \times y_i}{\sqrt{\left(\sum_{i=1}^{k} x_i^2\right) \times \left(\sum_{i=1}^{k} y_i^2\right)}}$$

Cosine similarity metrics are interpreted much like Pearson correlation coefficients, and the result from applying the above formula will be a value ranging from -1 (indicating that the vectors are complete opposites) to +1 (indicating that the vectors are identical). Results between -1 and +1 express varying degrees of similarity, and a result of 0 would indicate that there is no relationship between the vectors.

When applied to embeddings produced by sentence transformer models, as in the case of this research, a cosine similarity coefficient expresses the semantic similarity between two pieces of text. A cosine of +1 would indicate that the two pieces of text express identical ideas, a cosine of -1 would indicate that the pieces of text express exactly opposite ideas, and a cosine of 0 would indicate that the pieces of text express unrelated ideas. In NLP research, we look for high-magnitude positive cosine similarity coefficients when searching for pieces of text with highly similar content. We computed cosine similarity coefficients among all pairs of document elements we extracted from LRRs.

Step 3.2: Compute Match Scores for Pairs of Documents. After computing cosines among all document elements, we used these element-wise relations to compute a document-wise similarity metric we call a "match score." The analytic procedure behind our match score metric was informed heavily by the logic of the BERTScore algorithm developed by Zhang et al. (2020) to compare sentences based on word-level similarity. BERTScore was designed to evaluate the similarity between a candidate sentence and a reference sentence as a function of (a) the highest-value cosine similarity coefficient between each word from the reference sentence and all words of the candidate sentence and (b) the weight associated with each word from the reference sentence (specifically, the inverse document frequency [IDF] weight).

As context for how BERTScore influenced our match score metric, we present a mock example of how BERTScore is computed in Table 3 (for another example; see Zhang et al.'s Figure 1). This example compares two sentences: "Cats are great pets" (the reference sentence) and "Felines make good companions" (the candidate sentence). A human reader can easily discern that these sentences convey similar ideas, but the BERTScore algorithm needs to quantify their similarity based only on word-wise similarity (as indexed by cosines) and the relative rarity with which the words from the reference sentence are used in other sentences (as indexed by IDF weights). A BERTScore similarity value is computed by (a) identifying which word from the candidate sentence is the most similar to each of the words from the reference sentence (indicated by the green highlights in Table 3), (b) multiplying the cosine similarity value for each reference-sentence word's most similar counterpart from the candidate sentence by the reference-sentence word's IDF weight, and (c) dividing the sum of the cosine-IDF products by the sum of the IDF weights from the reference-sentence words. The BERTScore value in this example is .72, indicating considerable similarity between the sentences. BERTScore can reach a maximum value of 1.00, which only occurs when all the words from the reference sentence appear in the candidate sentence.

Table 3. Example BERTScore Analysis

Reference Sentence	Candidate Sentence				Max Cosine	IDF Weight	Cosine-IDF Product
	Felines	make	good	companions			
Cats	.82	.21	.19	.44	.82	6.6	5.41
are	.15	.27	.23	.24	.27	1.2	0.32
great	.08	.20	.74	.16	.74	2.4	1.78
pets	.50	.25	.17	.68	.68	4.5	3.06
Sums for BERTScore Calculation:						**14.7**	**10.57**
BERTScore = 10.57 / 14.7 = 0.72							

BERTScore is a type of soft similarity metric for comparing word usage, where the "soft" aspect refers to its relative leniency in scoring word-level similarity as a function of semantic overlap in word meanings as opposed to the extent to which sentences include the exact same words. We extended this logic to shift the focus of the analysis from comparing sentences in terms of their word-level similarity to comparing documents in terms of their element-level similarity. As noted above, we call the scores produced by our BERTScore-inspired procedure "match scores."

We show a worked example of our process for computing match scores in Table 4. In this example, we compare a four-element document called X to another four-element document called Y. Part A of Table 4 shows the pairwise cosines between both documents' elements, and Parts B and C show how we can use those cosines to compute BERTScore-like similarity metrics comparing documents X and Y. The BERTScore method differentiates between "reference" and "candidate" text, such that one only compares the words from a reference sentence to the words from a candidate sentence. However, we make no such distinction between the documents we are comparing, so Parts B and C of Table 4 show comparisons in both directions: Part B compares the elements of X to the elements of Y, and Part C compares the elements of Y to the elements of X. In each case, we identify which element of the opposing document that is most similar to each element of the focal document (indicated by green-shaded cells), record the cosine similarity metrics associate with those element-wise relations (see the "Max Cosine" values in grey), and average them (see the "Conditional Match Score" values in yellow). Each of conditional match score tells half the story about how similar the documents are, either relating X to Y or relating Y to X, so we average these two conditional

values to get an overall symmetrical index of similarity, as shown in Part D of Table 4. In this example, the overall match score between documents X and Y is .7625.

Table 4. Example Match Score Analysis

(A) Example Cosine Similarity Matrix for Documents X and Y					
Document X Element ID	**Document Y Element ID**				
	Y1	**Y2**	**Y3**	**Y4**	
X1	.83	.22	.47	.10	
X2	.54	.18	.79	.36	
X3	.25	.90	.18	.55	
X4	.11	.23	.04	.51	
(B) Conditional Match Score Relating Document X to Y					
Document X Element ID	**Document Y Element ID**				**Max Cosine**
	Y1	**Y2**	**Y3**	**Y4**	
X1	.83	.22	.47	.10	**.83**
X2	.54	.18	.79	.36	**.79**
X3	.25	.90	.18	.55	**.90**
X4	.11	.23	.04	.51	**.51**
Conditional Match Score					**.7575**
(C) Conditional Match Score Relating Document Y to X					
Document X Element ID	**Document Y Element ID**				**Conditional Match Score**
	Y1	**Y2**	**Y3**	**Y4**	
X1	.83	.22	.47	.10	
X2	.54	.18	.79	.36	
X3	.25	.90	.18	.55	
X4	.11	.23	.04	.51	
Max Cosine	**.83**	**.90**	**.79**	**.55**	**.7675**
(D) Overall Match Score (Average of Conditional Values)					
Overall Match Score					**.7625**

Our process for computing match scores is very similar to the BERTScore approach, but with three critical differences. First, and most obviously, we are analyzing text at a higher level of complexity; rather than using word-level similarity as the basis for evaluating sentence-level similarity, we are using element-level similarity as the basis for evaluating document-level similarity. Second, we are interested in making symmetric comparisons of similarity between documents rather than treating documents differently depending on whether they are

considered reference or candidate documents. The symmetry in our process requires evaluating similarity twice—such that the documents essentially take turns acting as the reference and candidate text—then averaging the results. Third, we have no analogue to IDF weights for our document elements, so our conditional match scores are based purely on cosines. Our process for computing match scores means that, if two documents each consist of only a single element, the match score for those documents will reduce to their cosine similarity coefficient.

Our match score metric functions as a soft similarity index for evaluating the overlap between complex multi-element documents using only element-wise relations as inputs. This aspect of our approach is quite important, as it ensures we only need to apply sentence transformers to sentence-like pieces of text and do not need to use the transformers in a way other than intended by the developers (e.g., by applying them to concatenated long-paragraph-like collections of text). We describe the potential problems associated with alternative strategies for comparing complex documents in the Appendix.

Step 3.3: Cluster Documents Based on Match Scores. We used a network-based clustering approach to group documents into sets as a function of their patterns in pairwise similarity metrics. This clustering approach operates by (1) compiling a matrix of similarity metrics, (2) applying a threshold similarity value to dichotomize the similarity metrics into linked vs. non-linked relations, (3) applying transitivity adjustments to account for documents that are indirectly connected via their mutual associations with other documents, and (4) using transitivity-adjusted linkages to form mutually exclusive clusters of documents.

We present an example application of this approach to a set of seven hypothetical documents in Table 5. In part A of Table 5, we show an example matrix of match scores, like those we described in Step 3.2. These match scores reveal substantial variation in the levels of similarity across pairs of documents, as they range from 0.05 to 0.90 (we exclude documents' relations with themselves, as those values are always 1.00, by definition). Studying this matrix can suggest pockets of similar documents that might be treated as clusters, but the boundaries between one cluster and another are challenging to delineate with clarity.

Part B of Table 5 shows a simpler representation of pairwise relations for our example documents, where documents are either linked with an X or they are not. We achieved this dichotomization of linkages by applying a similarity threshold of 0.75 to all cells of the match score matrix, such that any pairwise relations equal to or greater than 0.75 were coded as an X. When a meaningful threshold value is used (e.g., one that has been determined via expert judgment or empirical experimentation), this dichotomization approach helps to focus attention on linkages that are the most likely to reveal valid commonalities among documents.

Finally, Part C of Table 5 shows how we applied a transitivity adjustment to the dichotomous linkage indicators to produce mutually exclusive clusters of documents. By "transitivity," we mean that relations among certain documents that are not directly linked to one another might be implied by their mutual linkages with other documents. For example, document 4 is linked to documents 5 and 6 after applying our 0.75 threshold, but documents 5 and 6 are not linked to each other. Clusters need to represent mutually exclusive groupings, implying that—because document 4 is linked to documents 5 and 6—documents 5 and 6 must also be associated somehow, even if they share non-overlapping ideas with document 4. This type of transitive association is what makes our clustering approach a network-based strategy. We adjusted for transitivity in our example by implicitly linking documents 5 and 6 so that the trio of documents 4, 5, and 6 form a cluster. As a counterpart to that cluster, the cluster of documents 1, 2, and 3

does not require a transitivity adjustment, as all three documents were mutually connected based on their initial pairwise linkages.

Document 7 in our example did not show high levels of similarity with any other documents and, therefore, did not get included in either of the clusters we constructed. When documents are left out of all clusters like this, it means that their overlap with the other documents fell short of the similarity threshold that guided the grouping process and they do not have an empirical match within the collection of documents used in the analysis.

Table 5. Example Network Clustering Analysis

(A) Example Match Score Matrix

Document ID	Document ID						
	1	2	3	4	5	6	7
1	1.00	.76	.78	.55	.40	.15	.33
2	.76	1.00	.77	.69	.59	.23	.31
3	.78	.77	1.00	.38	.45	.05	.37
4	.55	.69	.38	1.00	.87	.72	.20
5	.40	.59	.45	.87	1.00	.90	.18
6	.15	.23	.05	.72	.90	1.00	.29
7	.33	.31	.37	.20	.18	.29	1.00

(B) Pairwise Linkages After Applying Threshold of .75

Document ID	Document ID						
	1	2	3	4	5	6	7
1	X	X	X				
2	X	X	X				
3	X	X	X				
4				X	X		
5				X	X	X	
6					X	X	
7							X

(C) Clusters Based on Transitivity-Adjusted Linkages

Document ID	Document ID						
	1	2	3	4	5	6	7
1	X	X	X				
2	X	X	X				
3	X	X	X				
4				X	X	X	
5				X	X	X	
6				X	X	X	
7							X

HumRRO has used this clustering method with great success in other contexts; most notably, when identifying groups of items in an exam's item bank that are too similar to co-occur on an exam form. These items—known as "enemy items"—often have similarities that can be detected using semantic analysis methods, like what we have described in this report. When a well-informed similarity threshold is used to group enemy items in an item bank, our method for transitivity-adjusted clustering greatly reduces the amount of human review necessary to correctly identify enemy items. We aim to realize similar levels of efficiency in our analyses and subsequent reviews of LRRs.

Our experience with clustering aggregated sets of text (where, in this case, aggregation represents our combination of element-wise cosine similarity coefficients into match score metrics) has suggested that a threshold of around 0.80 can produce a useful clustering of documents. However, we found that a threshold of 0.80 was likely too high for use in this context because it identified very few pairs of similar documents (18 pairs of documents from different sources). We experimented with alternative thresholds that might be a better fit for this research by identifying a larger—yet manageable—set of document pairs for review. We found that decreasing the threshold to 0.70 was an overcorrection because it identified more document pairs than we could accommodate within the scope of this research (925 pairs of documents from different sources), but a threshold of 0.75 offered a good balance between under- and over-flagging potentially related document pairs (103 pairs of documents from different sources). We used a threshold of 0.75 to perform our final document-level clustering analysis.

Alternative Approaches Considered for Grouping Text. Before adopting the clustering approach described above, we evaluated alternative methods for grouping text. We experimented with traditional cluster analyses applied to embeddings (both element-wise embeddings and document-level embeddings that represented averages of the embedding vectors from their associated elements) and context-sensitive topic models via the BERTopic algorithm (Grootendorst, 2022). We encountered challenges with both alternative approaches: The cluster analyses produced implausibly large clusters that appeared to hinge on unimportant features of the text, and the topic models produced topics that were difficult to interpret and that sometimes highlighted unimportant features of the text. Topic models and traditional clustering algorithms also each involve a substantial stochastic element, whereas our network-based approach does not. These challenges prompted us to forego model-based text-grouping strategies in favor of the similarity-based, network-oriented clustering strategy described above.

Step 3.4: Cluster Elements Outside of Document Clusters Based on Cosine Similarity Coefficients. After we completed our document-level clustering, we compiled a list of element pairs with high levels of similarity that were not subsumed within the document-wise pairs from our cluster analysis. Our aim in doing this was to produce a supplemental set of potentially related text segments that could complement our review of document pairs. This would allow us to identify strong element-wise correspondences between documents that failed to achieve sufficiently high match scores to be included in a cluster.

We did this by applying the clustering method from our document-level clustering analysis to all element-level cosine similarity metrics, but used a different similarity threshold to develop the clusters. Based on insights gained through HumRRO's previous use of the transformer model we applied in this research, we set the cosine similarity threshold for this element-level clustering analysis at 0.90. This cosine similarity threshold is higher than the match score threshold we used to cluster documents because cosines between sentence-like pieces have a stronger tendency to take on large values than aggregated similarity indices (e.g., match

scores). Differences in typical similarity levels for individual pairs of sentences versus aggregates of multiple sentences are due to the dilution of semantic signals that occurs when sentences expressing different ideas are combined. Sentence-wise and aggregate similarity metrics are both useful; they simply require different benchmarks to interpret.

After we clustered the elements, we filtered out element pairs that were either (a) from the same jurisdiction (e.g., we omitted pairs of text elements that both described LRRs from Alaska, as our focus is on comparing text across jurisdictions) or (b) from document pairs that were grouped together in our document-level cluster solution.

Step 4: Review Documents Flagged as Similar

We used the results of our analyses to inform a review of empirically similar pairs of text. Two HumRRO researchers reviewed all pairs of text flagged as similar and recorded whether they considered each pair to be a valid match. In addition to determining whether documents/elements were similar enough to be matched, the reviewers summarized notable nuances/differences between the documents/elements they identified as similar. In the event of a disagreement between the two reviewers about whether a pair of text represented a valid match, a third HumRRO researcher reviewed the text and rendered a tie-breaking decision.

Results

The key results of our analyses were distributions of similarity metrics and decisions reached by our reviewers. We summarize these metrics and decisions in the subsections that follow. The document-level text we analyzed is presented in Part A of the accompanying Supplement (“LRR Text”).

Descriptive Statistics for Similarity Metrics

The foundational component of our text-matching analyses was the collection of cosine similarity coefficients we computed between pairs of document elements. These values determined our document-level match score results and served as the inputs to our supplemental element-level clustering analysis. We have summarized the descriptive statistics for distributions of cosines between all pairs of text sources in Table 6 , including cosines between elements that came from the same source.

While cosine similarity values represent the building blocks of our analyses, match scores are the more import values for understanding relations among documents (i.e., LRRs). We have summarized the descriptive statistics for distributions of match scores between all pairs of text sources in Table 7, including match scores between documents that came from the same source. The complete matrix of match scores among LRRs is presented in Part B of the accompanying Supplement (“Match Score Matrix”).

The distributions of similarity metrics in Tables 6 and 7 were the inputs to the element- and document-level clustering analyses we used to identify categorical relations among pieces of text. We describe our reviews of these cluster-based relations next.

Table 6. Descriptive Statistics for Cosine Similarity Coefficients for Document Element Pairs from All Pairs of Sources

Source X	Source Y	# Elements X	# Elements Y	# Element Pairs	M	Mdn	SD	Min	Max
MPA	*MPA*	*375*	*375*	*70,125*	*0.303*	*0.269*	*0.195*	*-0.126*	*1.000*
MPA	Alaska	375	383	143,625	0.262	0.242	0.164	-0.164	0.971
MPA	Arizona	375	925	346,875	0.242	0.215	0.156	-0.153	1.000
MPA	Kansas	375	363	136,125	0.250	0.222	0.170	-0.140	0.910
MPA	Oklahoma	375	517	193,875	0.251	0.224	0.167	-0.159	0.995
MPA	Texas	375	1,257	471,375	0.227	0.202	0.150	-0.176	1.000
Alaska	*Alaska*	*383*	*383*	*73,153*	*0.281*	*0.261*	*0.169*	*-0.210*	*1.000*
Alaska	Arizona	383	925	354,275	0.240	0.220	0.148	-0.198	0.972
Alaska	Kansas	383	363	139,029	0.245	0.222	0.159	-0.168	0.893
Alaska	Oklahoma	383	517	198,011	0.240	0.215	0.159	-0.149	0.969
Alaska	Texas	383	1,257	481,431	0.227	0.208	0.144	-0.203	0.898
Arizona	*Arizona*	*925*	*925*	*427,350*	*0.234*	*0.214*	*0.144*	*-0.151*	*1.000*
Arizona	Kansas	925	363	335,775	0.225	0.202	0.148	-0.193	0.971
Arizona	Oklahoma	925	517	478,225	0.215	0.192	0.147	-0.178	0.971
Arizona	Texas	925	1,257	1,162,725	0.219	0.202	0.136	-0.158	1.000
Kansas	*Kansas*	*363*	*363*	*65,703*	*0.241*	*0.212*	*0.167*	*-0.153*	*1.000*
Kansas	Oklahoma	363	517	187,671	0.226	0.197	0.160	-0.165	0.943
Kansas	Texas	363	1,257	456,291	0.210	0.190	0.142	-0.181	0.923
Oklahoma	*Oklahoma*	*517*	*517*	*133,386*	*0.230*	*0.200*	*0.162*	*-0.142*	*1.000*
Oklahoma	Texas	517	1,257	649,869	0.202	0.181	0.144	-0.170	0.958
Texas	*Texas*	*1,257*	*1,257*	*789,396*	*0.214*	*0.196*	*0.139*	*-0.160*	*1.000*

Note. Results for element pairs from the same source exclude (a) duplicated pairs of elements and (b) elements paired with themselves (i.e., # Element Pairs = [# Elements X] × [# Elements Y – 1] / 2). The number of element pairs for elements from different sources is equal to the product of the number of elements from each source. Results for same-source element pairs are in *italics*.

Table 7. Descriptive Statistics for Match Score Coefficients for Document Pairs from All Pairs of Sources

Source X	Source Y	# Docs X	# Docs Y	# Doc Pairs	M	Mdn	SD	Min	Max
MPA	*MPA*	*24*	*24*	*276*	*0.465*	*0.476*	*0.115*	*0.085*	*0.825*
MPA	Alaska	24	50	1,200	0.407	0.419	0.117	-0.022	0.826
MPA	Arizona	24	66	1,584	0.420	0.431	0.107	0.083	0.974
MPA	Kansas	24	38	912	0.420	0.436	0.124	0.026	0.782
MPA	Oklahoma	24	44	1,056	0.432	0.440	0.107	0.072	0.759
MPA	Texas	24	145	3,480	0.368	0.377	0.115	-0.015	0.823
Alaska	*Alaska*	*50*	*50*	*1,225*	*0.407*	*0.422*	*0.131*	*-0.033*	*0.970*
Alaska	Arizona	50	66	3,300	0.384	0.394	0.110	0.056	0.811
Alaska	Kansas	50	38	1,900	0.396	0.411	0.125	0.007	0.771
Alaska	Oklahoma	50	44	2,200	0.409	0.418	0.114	-0.049	0.841
Alaska	Texas	50	145	7,250	0.331	0.329	0.120	-0.057	0.799
Arizona	*Arizona*	*66*	*66*	*2,145*	*0.398*	*0.405*	*0.098*	*0.107*	*0.947*
Arizona	Kansas	66	38	2,508	0.387	0.398	0.112	0.064	0.785
Arizona	Oklahoma	66	44	2,904	0.392	0.403	0.104	0.028	0.823
Arizona	Texas	66	145	9,570	0.352	0.353	0.105	0.011	0.739
Kansas	*Kansas*	*38*	*38*	*703*	*0.422*	*0.432*	*0.138*	*0.052*	*0.829*
Kansas	Oklahoma	38	44	1,672	0.424	0.430	0.117	0.056	0.838
Kansas	Texas	38	145	5,510	0.330	0.325	0.118	-0.052	0.783
Oklahoma	*Oklahoma*	*44*	*44*	*946*	*0.444*	*0.442*	*0.095*	*0.081*	*0.936*
Oklahoma	Texas	44	145	6,380	0.334	0.334	0.117	-0.030	0.738
Texas	*Texas*	*145*	*145*	*10,440*	*0.328*	*0.329*	*0.114*	*-0.032*	*0.854*

Note. Results for document pairs from the same source exclude (a) duplicated pairs of documents and (b) documents paired with themselves (i.e., # Doc Pairs = [# Docs X] × [# Docs Y – 1] / 2). The number of document pairs for documents from different sources is equal to the product of the number of documents from each source. Results for same-source document pairs are in *italics*.

Summary of Review Outcomes

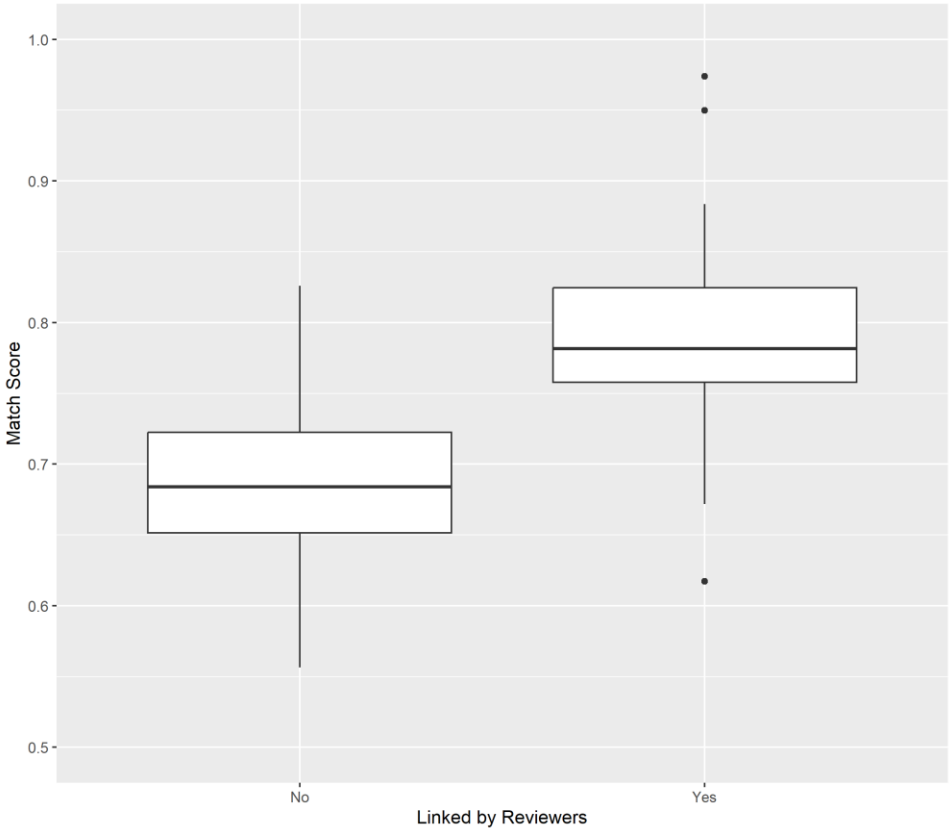
Our cluster analysis results identified pairs of text that were seemingly similar to each other, and we reviewed these pairs to determine whether they were similar enough to be linked. We reviewed document- and element-level pairs separately and, as noted in the Method section, we limited the scope of our reviews to pairs of documents from different sources because commonalities between pieces of text from the same source would not support the types of insights that were of interest to HRRI. For our element-level review, we also screened out pairs of elements that were subsumed within the pairs of documents we reviewed, as reviewing elements pairs individually while also reviewing them in context within their respective documents would be redundant.

Document-Level Relations

Our first review focused on relations between pairs of documents (LRRs) identified via our match score-based clustering analyses. Our cluster solution identified a total of 34 clusters, and 16 of those clusters contained pairs of LRRs that were eligible for review because the documents came from different sources. We identified a total of 103 document pairs for review.

Of the 103 pairs of documents, the two primary reviewers agreed on 89 decisions (86.4%) and, after the third reviewer provided tie-breaking decisions, 35 pairs (34.0%) were judged to be valid matches. This is a rather low rate of linkages between the pairs of documents, so it can be helpful to compare the similarity metrics between the linked and non-linked pairs to see how much the distributions overlap. Figure 1 shows a boxplot comparing distributions of match scores by reviewer decisions, revealing that, while match scores for linked document pairs tend to be larger in magnitude, the distributions still have considerable overlap. The point-biserial correlation between match scores and reviewers' linkage decisions was 0.61, which reinforces the trend depicted in Figure 1. The text pairs we reviewed, along with reviewers' decisions, are presented in Part C of the accompanying Supplement ("Document Review Outcomes").

Figure 1. Boxplot for Document-Level Match Score Values by Reviewer Decisions



Note. The bottom of a box represents the first quartile (the 25th percentile), the horizontal line within a box represents the second quartile (the median), the top of a box represents the third quartile (the 75th percentile), the whiskers represent ranges equal to 1.5 times the inter-quartile range (i.e., the distance between the first and third quartiles) below the first quartile and above the third quartile, and individual points beyond the whiskers represent outliers.

It is possible that clusters could vary in the rates at which they accumulate documents that have substantive connections with each other, so we examined how reviewers' decisions were distributed across clusters. Table 8 provides a summary of review outcomes by document cluster. This summary gives the frequencies of document pairs (the total number reviewed, as well as the frequency with which they were linked or not linked by reviewers), the percentage of document pairs that were linked, and the mean match scores for document pairs that were reviewed, linked, and not linked. These results reveal that clusters with more pairs of documents that were eligible for review tended to contain a smaller proportion of valid matches.

Only two clusters contributed multiple pairs of documents to the review and, of these, reviewers considered 24.7% of pairs from the cluster with 81 pairs to be valid linkages and they considered 50.0% of pairs from the cluster with 8 pairs to be valid linkages (the overall true positive rate for pairs from these two clusters was 27.0%). The remaining 14 clusters each contributed one pair of documents to our review, and, of those, the reviewers considered 12 of the pairs (85.7%) to be valid linkages. This pattern of results suggests that the similarity threshold we used to develop the clusters functions rather well at identifying pairs of similar documents but might have difficulty when developing larger clusters; this is because our transitivity adjustments allow for more false positives in the pairwise connections among documents. The fact that a single cluster contributed most pairs of documents is not necessarily a concern in this research, as the clustering process was primarily a way to cultivate document pairs for review and we did not interpret the clusters as part of our feasibility evaluation.

Table 8. Summary of Document Clusters Evaluated by Reviewers

Cluster ID	Frequency of Document Pairs			% Linked Within Cluster	Mean Match Score		
	Reviewed	Linked	Not Linked		Reviewed	Linked	Not Linked
1	81	19	62	23.5	0.700	0.765	0.681
2	8	4	4	50.0	0.746	0.763	0.728
7	1	0	1	0.0	0.783	---	0.783
9	1	0	1	0.0	0.755	---	0.755
13	1	1	0	100.0	0.786	0.786	---
14	1	1	0	100.0	0.950	0.950	---
15	1	1	0	100.0	0.796	0.796	---
16	1	1	0	100.0	0.823	0.823	---
17	1	1	0	100.0	0.863	0.863	---
18	1	1	0	100.0	0.791	0.791	---
19	1	1	0	100.0	0.809	0.809	---
20	1	1	0	100.0	0.826	0.826	---
21	1	1	0	100.0	0.781	0.781	---
22	1	1	0	100.0	0.835	0.835	---
23	1	1	0	100.0	0.884	0.884	---
30	1	1	0	100.0	0.799	0.799	---

Note. Cluster ID values are arbitrary and were assigned by sequencing clusters in order of decreasing size.

We used transitivity adjustments in our clustering approach to help cast a wider net for identifying legitimate patterns of overlap among documents than would be possible if we simply applied a similarity threshold for the pairwise relations among documents. Adjusting for transitivity in our cluster solutions was meant to identify document pairs that fell short of our threshold value of 0.75 but still demonstrated high levels of mutual similarity with other documents. However, the results in Table 8 suggest that our transitivity adjustments might have been undesirable when applied to LRRs.

When we did away with the transitivity adjustments and only examined the pairs of documents from our review that had match scores of 0.75 or greater, we found a much higher rate of detection for related documents. Ignoring transitivity limited our focus to 38 of the 103 document pairs (36.9%) originally identified for review. Of the 38 pairs of documents that met or exceeded the 0.75 similarity threshold, the two primary reviewers agreed on 31 decisions (81.6%) and, after the third reviewer provided tie-breaking decisions, 28 pairs (73.7%) were judged to be valid matches. These 28 linkages represented 80.0% of the 36 linkages we identified in the complete review activity.

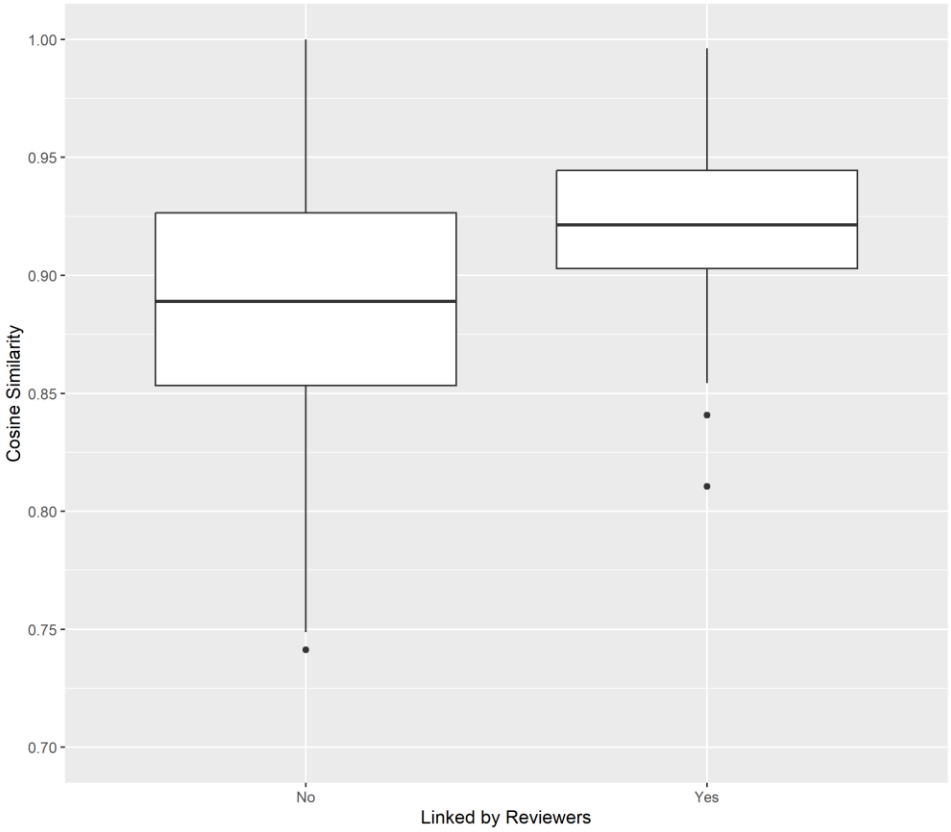
The differences between our patterns of findings with and without applying transitivity adjustments suggests that, while adjusting for transitivity does help to identify a larger number of related documents, it does so at the expense of nearly tripling the number of document pairs identified for review. When we used pairwise similarity as the sole basis for flagging potentially similar pairs of documents, the set of flagged pairs contained a higher percentage of valid matches, and we could still identify most of the same pairs of related documents as in the larger review activity.

Element-Level Relations

Our second review focused on relations between pairs of elements identified via our cosine-based clustering analyses. Our analysis identified a total of 394 clusters, and 58 of those clusters contained pairs of elements that were eligible for review because (a) they came from different sources and (b) they were not subsumed within the pairs of LRRs that we reviewed from the document-level clusters. We identified a total of 166 element pairs for review.

Of the 166 pairs of elements, the two primary reviewers agreed on 139 decisions (83.7%) and, after the third reviewer provided tie-breaking decisions, 66 pairs (39.8%) were judged to be valid matches. This rate of linkages was slightly higher than in our review of document pairs, but still rather low. Figure 2 shows a boxplot comparing distributions of cosines by reviewer decisions, revealing a much weaker distinction between the distributions of cosines than we saw for the distributions of match scores (cf. Figure 1). The point-biserial correlation between cosines and reviewer's linkage decisions was 0.32, which is consistent with a weak but positive association between cosines and reviewers' judgments about whether element pairs were valid matches. The text pairs we reviewed, along with reviewers' decisions, are presented in Part D of the accompanying Supplement ("Element Review Outcomes").

Figure 2. Boxplot for Element-Level Cosine Similarity Values by Reviewer Decisions



Note. The bottom of a box represents the first quartile (the 25th percentile), the horizontal line within a box represents the second quartile (the median), the top of a box represents the third quartile (the 75th percentile), the whiskers represent ranges equal to 1.5 times the inter-quartile range (i.e., the distance between the first and third quartiles) below the first quartile and above the third quartile, and individual points beyond the whiskers represent outliers.

Given that we observed a marked difference in the rates at which reviewers linked pairs of documents as a function of cluster size in our document-level review, we computed a cluster-level summary of element-wise relations in Table 9 to see if a similar trend was operating in our element-level review. This cluster-level examination of the element pair review pool revealed that, although pairs of elements from clusters that only contributed one pair the analysis to review were linked at a higher rate than pairs from other clusters (20 of 42 [47.6%] compared to 46 of 124 [37.1%]), the difference was much smaller than what we observed in the document-level review.

Table 9. Summary of Element Clusters Evaluated by Reviewers

Cluster ID	Frequency of Element Pairs			% Linked Within Cluster	Mean Cosine		
	Reviewed	Linked	Not Linked		Reviewed	Linked	Not Linked
319	42	11	31	26.2	0.857	0.879	0.849
431	25	4	21	16.0	0.877	0.943	0.864
457	20	14	6	70.0	0.919	0.930	0.893
801	7	0	7	0.0	0.934	---	0.934
3128	4	0	4	0.0	0.976	---	0.976
373	4	4	0	100.0	0.911	0.911	---
325	3	0	3	0.0	0.969	---	0.969
262	3	3	0	100.0	0.928	0.928	---
548	2	2	0	100.0	0.935	0.935	---
486	2	2	0	100.0	0.929	0.929	---
607	2	2	0	100.0	0.920	0.920	---
816	2	0	2	0.0	0.909	---	0.909
1158	2	0	2	0.0	0.904	---	0.904
490	2	2	0	100.0	0.902	0.902	---
466	2	2	0	100.0	0.901	0.901	---
806	2	0	2	0.0	0.883	---	0.883
265	1	0	1	0.0	1.000	---	1.000
264	1	1	0	100.0	0.996	0.996	---
752	1	1	0	100.0	0.995	0.995	---
750	1	1	0	100.0	0.994	0.994	---
281	1	0	1	0.0	0.990	---	0.990
544	1	1	0	100.0	0.987	0.987	---
334	1	0	1	0.0	0.972	---	0.972
257	1	1	0	100.0	0.971	0.971	---
266	1	1	0	100.0	0.964	0.964	---
1109	1	0	1	0.0	0.958	---	0.958
540	1	0	1	0.0	0.951	---	0.951
330	1	1	0	100.0	0.949	0.949	---
858	1	1	0	100.0	0.948	0.948	---
255	1	0	1	0.0	0.946	---	0.946
527	1	1	0	100.0	0.944	0.944	---
590	1	0	1	0.0	0.943	---	0.943
986	1	0	1	0.0	0.936	---	0.936
3090	1	0	1	0.0	0.932	---	0.932
764	1	1	0	100.0	0.930	0.930	---
187	1	0	1	0.0	0.928	---	0.928

Table 9. (Continued)

Cluster ID	Frequency of Element Pairs			% Linked Within Cluster	Mean Cosine		
	Reviewed	Linked	Not Linked		Reviewed	Linked	Not Linked
2947	1	1	0	100.0	0.927	0.927	---
483	1	0	1	0.0	0.927	---	0.927
996	1	0	1	0.0	0.926	---	0.926
189	1	0	1	0.0	0.924	---	0.924
2975	1	0	1	0.0	0.924	---	0.924
153	1	1	0	100.0	0.922	0.922	---
1669	1	0	1	0.0	0.921	---	0.921
346	1	1	0	100.0	0.921	0.921	---
506	1	1	0	100.0	0.918	0.918	---
531	1	1	0	100.0	0.918	0.918	---
765	1	1	0	100.0	0.917	0.917	---
857	1	1	0	100.0	0.916	0.916	---
902	1	0	1	0.0	0.916	---	0.916
165	1	0	1	0.0	0.913	---	0.913
1104	1	0	1	0.0	0.913	---	0.913
331	1	0	1	0.0	0.912	---	0.912
1355	1	1	0	100.0	0.912	0.912	---
591	1	0	1	0.0	0.911	---	0.911
328	1	1	0	100.0	0.908	0.908	---
655	1	1	0	100.0	0.907	0.907	---
388	1	0	1	0.0	0.905	---	0.905
3215	1	0	1	0.0	0.904	---	0.904

Note. Cluster ID values are arbitrary, and the values can exceed the number of valid clusters reviewed because each element not linked with another element was assigned to a nominal “cluster” by itself and assigned a cluster ID.

Our document-level analyses revealed that transitivity adjustments had substantial impacts on the number of text pairs flagged for review and the true-positive rate for pairs identified as empirically similar, so we also examined the impact of transitivity adjustments on our review of element pairs. As with our document-level analyses, when we ignored transitivity adjustments and only examined the pairs of elements from our review that had cosine similarity values of 0.90 or greater, we found a higher rate of detection for related elements. Of the 101 pairs of elements that met or exceeded the 0.90 similarity threshold, the two primary reviewers agreed on 83 decisions (82.2%) and, after the third reviewer provided tie-breaking decisions, 54 pairs (53.5%) were judged to be valid matches. These 54 linkages represented 81.8% of the 66 linkages we identified in the complete review activity.

When we limited the set of element pairs flagged for review to only those that had cosines of 0.90 or greater, we reduced the scope of the review activity from 166 to 101 pairs (a reduction of 39.2%) while still identifying most of the same pairs of related elements as in the larger review activity. Consistent with our document-level review, when we used pairwise similarity as

the sole basis for flagging potentially similar pairs of elements, the set of flagged pairs contained a higher percentage of valid matches, but the increase in the true positive rate was lower than the increase we observed in our document-level review.

Summary of Similarities and Nuances in Text Pairs Linked by Reviewers

The most common differences/nuances we observed between pairs of related text had to do with the organization, scope, and detail of the information presented. We summarize these below, and comments about individual text pairs linked by reviewers are available in Parts C and D of the accompanying Supplement ("Document Review Outcomes" and "Element Review Outcomes," respectively).

In terms of organization, some related LRRs presented information in different orders (a rather trivial difference), while some had asymmetric patterns of overlap (e.g., Text X might be fully subsumed by the content of Text Y, but not vice-versa). Instances of asymmetric overlap were often due to differences in how information about physical therapists (PTs) and physical therapy assistants (PTAs) were organized within a collection of LRRs; some sources covered information about PTs and PTAs in separate LRRs, while others covered the information about PTs and PTAs in different subsections of a single LRR.

Regarding scope, some LRRs were more thoroughly segmented into dependent sets, whereas other LRRs were more completely integrated. For example, an LRR from one source might list out all the relevant requirements for PTs/PTAs within that LRR, while a related LRR from another source might simply refer to requirements listed in other LRRs from that source. The former type of LRR is more completely integrated and internally complete, and the latter type is part of a collection of LRRs that is more segmented and introduces more dependencies within the LRR collection.

Differences in level of detail were quite pervasive. These differences were such that, within a pair of text reviewers considered related, one LRR simply provided more comprehensive set of ideas than the other. For instance, an LRR from one source might offer a more elaborate and in-depth definition than an LRR for another source. These types of differences often made it appear that one LRR was subsumed within another, but it could be unclear whether that appearance was merely due to omission on the part of the jurisdiction with the less-detailed LRR. A notable example of this type of nuance occurred for LRRs describing how PTs/PTAs were allowed to use titles and abbreviations. Some jurisdictions' LRRs on this topic stuck very closely to describing proper use of the titles "physical therapist" and "physical therapy assistant" and the abbreviations "PT" and "PTA," while other jurisdictions' related LRRs covered the use of other titles and abbreviations (e.g., "doctor of physical therapy" and "DPT") and/or the use of titles/abbreviations in combination with an indication of one's "retired" status.

Discussion

HRRI requested that HumRRO examine whether it would be feasible to use NLP methods to detect similarities and differences in the LRRs governing the practice of PT across US jurisdictions. We analyzed the LRRs from a diverse set of five sample jurisdictions, as well as those from FSBPT's MPA, via a combination of semantic similarity analyses and cluster analyses. We then conducted follow-up reviews to evaluate the quality of the pairs of text that emerged as empirically similar.

We found that the effectiveness of the empirical methods we used to identify potentially related pieces of text was quite different depending on whether we used a clustering approach or simply screened pairs of text based on their magnitudes of pairwise similarity. We had anticipated that our clustering approach would help to identify pairs of text that had similarity metrics below our threshold values but still had deep commonalities due to their being connected via a network of other similar pieces of text. The clustering approach we applied in this research has performed quite well in other settings, but it was not as effective in this context. The difference appears to stem from the greater complexity of the LRR text than the text we have analyzed with this method in other settings. For example, our clustering approach has performed very well when grouping similar multiple-choice items from large testing programs' item banks and when distilling large collections of unstructured sentence-length documents into meaningful clusters.

Contrary to expectations, the analyses we conducted after reviewers evaluated cluster-based relations shows that, from a signal-detection perspective, it is arguably better to skip clustering and base the empirical detection of potentially similar text solely on their pairwise magnitudes of similarity.

Evaluation of Feasibility

Although we were able to identify pairs of documents that our reviewers considered valid matches and we identified a methodological alteration that can increase the rate at which empirical analyses identify substantively related pieces of text, we would consider the feasibility of using NLP to guide the detection of similar LRRs across jurisdictions to be limited. This conclusion is primarily due to our observation that the amount of up-front effort required to prepare text for analysis is disproportionately large compared to the amount of related text we were able to identify using automated analyses. In other words, the return on investment (ROI) of scaling this approach to additional jurisdictions would be low in terms of a labor-to-yield tradeoff.

Our conclusion about feasibility would be quite different if the input text were available in a consistent, tabular format across jurisdictions that did not require extensive preparation prior to analysis. If, for instance, LRRs were available from jurisdictions a spreadsheet format and the text could be read into an analysis program with little up-front reformatting, NLP analyses could be run rather easily, and it would shift our ROI-based evaluation in their favor. The reality, however, is that we expect the process we used to prepare the text for this study would need to be replicated for any additional jurisdictions and, due to the large amount of customization required for each jurisdiction, the steps we took to prepare and clean the text in this research will not benefit from the efficiency gains that sometimes accompany scaling-up a procedure.

We based our feasibility evaluation on a consideration of the labor-to-yield implications of conducting these analyses and, while the labor to process new text remains non-trivial, the

yield-based considerations are subject to more variability. In this case, "yield" represents the number of valid pairs of text identified and is a function of (a) the threshold value used to identify empirically similar text and (b) the rate at which valid matches are identified in the pairs of text that exceed the threshold value (a rate that is not constant and will decrease as we lower the threshold). The 0.75 match score threshold we used in this research was based on a cluster analysis perspective, and we chose it to balance the 0.80 value we have found useful in past clustering of aggregated similarity metrics against our goals of obtaining a practical number of text pairs to review and not over-flagging content for review. Lowering the threshold would allow us to find more matches but, as the threshold decreases, the amount of work for reviewers would increase; this is the other consideration in regarding labor.

Given the investment required to prepare text and review pairs of text matched by an NLP analysis, we suspect that HRRI could achieve a better ROI by using alternative methods to identify pieces of related text. We provide more details about this in the Recommendations section.

Recommendations

Based on our feasibility evaluation above and insights we developed by working with the jurisdictions' LRRs, we suspect that HRRI might be better served by developing comparisons between jurisdictions LRRs through more conventional means. The NLP analyses we investigated in this research can identify related text, but they require more resources than their results might be able to justify. In addition to requiring a good deal of up-front text processing and reviewer attention, they also run a high risk of overlooking pairs of related text; every automated process has an error rate, and the false negative error rate for this process is unknown.

Rather than attempting to scale this approach to more jurisdictions, we anticipate that a greater ROI could be achieved by developing a rationally derived crosswalk linking elements from the TOCs for jurisdictions' collections of LRRs and relying on subject matter experts (SMEs) to evaluate the text linked via the crosswalk. Relating sets of LRRs to each other via their TOCs is a much more direct way of developing sets of related text and would require substantially fewer resources than even a partially automated NLP approach. FSBPT has already been doing something similar to this for the five jurisdictions for which it manages jurisprudence exams.

The content outline for each of FSBPT's jurisprudence exams (for Arizona, California, DC, Florida, and Nebraska) maps LRRs onto six major categories, each of which has well-defined subcategory sections (FSBPT, 2023b):

- 1000 Legislative Intent & Definitions
- 2000 Board of Physical Therapy Powers & Duties
- 3000 Licensure & Examination
- 4000 Patient Care Management
- 5000 Disciplinary Actions/Procedures; Unlawful Practice
- 6000 Consumer Advocacy

This content taxonomy would make an excellent starting point for developing a cross-jurisdictional LRR crosswalk, which HRRI could use as the organizing structure for a relational database containing the LRRs from all jurisdictions covered by the crosswalk. Access to such a

relational database could be offered as a service, as it could place the relevant LRRs for specific categories/sections side by side for database users to study and compare.

Aside from its immediate value to users as a repository of LRR text, a relational database built on the structure present in FSBPT's content outlines could provide the starting point for more sophisticated and targeted applications of computational language models. The advent of advanced large language models (LLMs) such as the GPT-3.5 model that underlies the popular ChatGPT application could make it possible for HRRI to develop an approach for comparing LRRs linked within the crosswalk, using the LLM as an engine for summarizing the similarities and differences between the linked pieces of text. This approach would leverage rational linkages among jurisdiction-specific LRRs as a roadmap for directing the attention of an LLM so it can draw relevant comparisons.

The analyses we performed within the scope of this feasibility analysis could only get HRRI part of the way toward the comprehensive relational understanding of LRRs that would be possible with a rationally developed crosswalk. We view the detailed and well-delineated structure for the categories/sections in the jurisprudence exam outlines as a promising starting point for a cross-jurisdictional crosswalk, and we view a rational mapping of LRRs onto that crosswalk a more promising approach than NLP similarity analyses for connecting related LRRs.

References

American Physical Therapy Association. (2023). *Licensure*. APTA. https://www.apta.org/your-practice/licensure

Dahlke, J. A., Putka, D. J., Shewach, O. R., & Lewis, P. (2022). *Developing related occupations for the O*NET program* (HumRRO Report 2022 No. 036). Human Resources Research Organization. https://www.onetcenter.org/dl_files/Related_2022.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. ArXiv:1810.04805 [Cs]. http://arxiv.org/abs/1810.04805

FSBPT. (2023a). *Jurisprudence Assessment Module (JAM) services*. https://www.fsbpt.org/Our-Services/Jurisprudence-Assessment-Module-JAM-Services

FSBPT. (2023b). *Jurisprudence exam*. https://www.fsbpt.org/Secondary-Pages/Exam-Candidates/Jurisprudence-Exam

FSBPT. (2023c). *National exam (NPTE®)*. https://www.fsbpt.org/Secondary-Pages/Exam-Candidates/National-Exam-NPTE

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv. http://arxiv.org/abs/2203.05794

Physical Therapy Compact Commission. (2021). *Physical therapy compact model language*.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks* (arXiv:1908.10084). arXiv. http://arxiv.org/abs/1908.10084

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. http://arxiv.org/abs/1910.01108

Sentence-transformers/all-distilroberta-v1 · Hugging Face. (2023, June 9). https://huggingface.co/sentence-transformers/all-distilroberta-v1

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2021). *Big bird: Transformers for longer sequences* (arXiv:2007.14062). arXiv. http://arxiv.org/abs/2007.14062

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating text generation with BERT* (arXiv:1904.09675). arXiv. http://arxiv.org/abs/1904.09675

Appendix: Implications for Text Structure in Similarity Analyses

The list-type structure of most LRRs we analyzed and the differences in how ideas were structured across jurisdictions' documents make the automated detection of similar LRRs rather challenging. A typical LRR is multifaceted and segmented into bullet-like statements, sometimes with many levels of nesting within a hierarchical bulleted list. This segmentation and hierarchical organization of segments has mixed implications for similarity analyses. On the one hand, segmentation produces sentence-like pieces of text that are suitable for analysis with sentence transformer models; on the other, segmentation into a hierarchical structure makes it quite challenging to properly combine or "roll up" the results of segments that share critical dependencies (e.g., segments that are nested beneath other segments). If the results for segments in a list are combined, how much weight should the dependent segment(s) receive, and how should these weights be allocated when lists have more than two levels of hierarchical organization? If the results for segments are left separate, how can one account for the high rate at which superficial similarities between segments are detected when attempting to find more meaningful similarities?

Further complicating the situation is the fact that not all jurisdictions organize ideas in the same way. What is presented as a bulleted list by one jurisdiction might be presented as a more thoroughly integrated paragraph or set of paragraphs by another. These differences in organization and formatting impact the probability that documents will get matched with each other through an empirical process, even if the ideas they express are quite similar.

One approach to resolving this formatting issue might be to concatenate the text for list-type LRRs into more paragraph-like blocks of text so all the ideas can be synthesized into one quantitative representation. However, this runs the risk of creating blocks of text with more "tokens" (words or word-like text) than sentence transformers can accommodate (most can only handle a couple hundred tokens, and excess tokens are ignored). There are transformers that can accommodate longer pieces of text (e.g., the "Big Bird" model; Zaheer et al., 2021), but they cannot necessarily overcome a more substantive problem: Concatenating lists of segments that express distinct or diverse ideas can dilute the meaning of embeddings computed for the concatenated text, thereby making it challenging to match the text to other documents.

The opposite approach to concatenating lists is to subdivide paragraph-like documents into sentence-like segments. This avoids the challenges associated with analyzing long and potentially heterogeneous pieces of text, but introduces the challenges associated with analyzing smaller and less complete pieces of text that may exhibit nuanced patterns of dependencies (unlike lists, the structural dependencies among sentences within a paragraph are as not clear). Parsing paragraphs into segments also requires reliably determining where one sentence ends and another begins; humans are very good at this, but automated computer-driven parsing procedures can struggle due to the diverse ways in which punctuation symbols can be used, especially in legal documents (e.g., a period followed by a space can indicate the end of a sentence, or it can occur after an abbreviation, or it can be part of the labeling scheme used to identify legal statutes).

In short, there is no perfect, one-size-fits-all approach for accommodating differently formatted text in a similarity analysis; each approach comes with its own mix of virtues and drawbacks. The strategy we used to aggregate element-wise cosines into "match score" values was our attempt at striking a favorable balance, such that we were able to aggregate the similarity metrics for sentence-like pairs of text using a soft-similarity approach.