# Toward Implementation Science: A Case Study Using LA-MPS to Research Argument Elements at Scale

Vern R. Walker[1,*,†], Stephen R. Strong[2,†]

[1] Maurice A. Deane School of Law, Hofstra University, Hempstead, New York 11549, United States
[2] Apex, North Carolina, United States

**Abstract**

Access to justice can increase only if AI tools developed in laboratory settings are implemented at scale to address real-world problems. This paper urges the development of implementation science for AI and law—a set of principles and methods to facilitate and study the transfer of AI techniques and tools to real-world use cases. The paper contributes to this development by reporting a case study using LA-MPS, an innovative software application (Legal Apprentice, or "LA") for reading, searching, and annotating legal decisions, which currently consists of three integrated web applications: Marker, Pad, and Search. LA-MPS is open-source, free, and adaptable to different legal domains, and it is deployable both locally (edge-centric) and through cloud servers. The case study uses decisions issued by the U.S. Board of Veterans' Appeals (BVA) that adjudicate disability benefits. The case study simulates the workflow of standard legal research, conducted on a large dataset of unread but automatically annotated legal decisions (10,003 BVA decisions, containing 1,360,230 sentences). Two primary experiments were conducted. First, we used semantic auto-labeling to filter out a subset of 449 decisions (100,514 sentences) that deal with post-traumatic stress disorder (PTSD), and we evaluated auto-labeling accuracy using a stratified random sample (25 decisions, containing 5,529 sentences). Second, we conducted semantic searches on the set of 449 decisions to identify scenarios in which non-VA evidence prevailed over conflicting VA evidence. The case study is designed to demonstrate the feasibility of implementing currently available machine learning (ML) models at scale, to employ scientific methods to compare results at scale with laboratory results, and to evaluate the real-world results based on practical usefulness. The paper discusses the generalizability of these methods for implementation science. The software code and case study datasets are made available to the public.

**Keywords**

implementation, dissemination, argument mining, legal decision annotation, sentence rhetorical role, automated semantic analysis

## 1. Introduction

Access to justice can increase only if artificial intelligence (AI) tools developed in laboratory settings are implemented at scale to address real-world problems. To provide high-quality representation for clients, legal aid organizations should provide their advocates with online research tools using up-to-date primary sources [1, 2].

The biomedical field has long recognized a similar need to translate results from laboratory and clinical research into interventions and treatments that can improve human health at scale [3]. After decades of research on the challenges that impede such translation, the biomedical field has evolved "translational science," involving operational principles and evidence-informed best practices [3]. The field of AI and law is evolving along a similar path.

This paper reports on a case study that uses innovative software for conducting standard legal research at scale, and it suggests several principles for implementation research generally in AI and law. First, contributions to implementation science should demonstrate the feasibility of a use case at scale using normal user workflows and having concern for normal constraints, such as ensuring data privacy and minimizing cost. Second, statistical methods should be used to evaluate results obtained at scale, and to compare those results to those obtained in laboratory settings. Third, evaluation of case study results should include not only quantitative metrics, but also practical costs of error in real-world settings. The case study reported here respects each of these three principles.

The case study involves research on the reasoning patterns found in past decisions issued by the U.S. Board of Veterans' Appeals (BVA), a very large corpus of plaintext documents. The BVA issued 99,721 decisions in fiscal year 2021 alone [4]. On average, therefore, the BVA issued over 1,900 decisions per week. The vast majority of those (over 96%) involved veterans' claims for compensation [4]. Our case study conducted research on the reasoning in a very large batch of BVA decisions (10,003 decisions, consisting of 1,360,230 sentences).

*Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal*

[*] Corresponding author.

[†] These authors contributed equally.

vern.r.walker@hofstra.edu (V. R. Walker); stephen.strongiot@gmail.com (S. R. Strong)
https://sites.hofstra.edu/vern-walker/ (V. R. Walker)

CEUR Workshop Proceedings (CEUR-WS.org)

We conducted two primary experiments. First, we used semantic auto-labeling to filter out a subset of 449 decisions (100,514 sentences) that deal with post-traumatic stress disorder (PTSD), and we evaluated auto-labeling accuracy using a stratified random sample (25 decisions, containing 5,529 sentences). Second, we conducted semantic searches on the set of 449 decisions to identify and investigate scenarios in which non-VA evidence prevailed over conflicting VA evidence.

In the next three sections of this paper, we describe our case study: Section 2 describes the datasets and predictive model employed; Section 3 provides an overview of the software components; and Section 4 reports the results of the two primary experiments. After that, Section 5 discusses some implications for developing implementation science for AI and law, and Section 6 reviews recent, related work.

The major contributions of this paper are:

- a case study evaluating the feasibility of standard attorney workflows for argument mining at scale;
- open-source, adaptable, and free software code for a suite of integrated web applications designed to assist legal practitioners in performing some main tasks involved in legal research;
- a set of curated BVA decisions that add to well-known, gold-standard datasets of decisions for claims involving PTSD; and
- a discussion of best practices to promote the implementation of AI tools at scale in law.

## 2. Components for the Case Study

This section discusses the datasets and the predictive model used in the case study.

### 2.1. The Three BVA Datasets

The LLT Dataset. To establish a baseline and train a machine learning (ML) classifier for the case study, we used a dataset of BVA decisions annotated and made publicly available by the Research Laboratory for Law, Logic and Technology (LLT Lab) at the Maurice A. Deane School of Law at Hofstra University (the "LLT Dataset") [5, 6, 7].[1] That dataset consists of 50 decisions that adjudicate disability claims filed by veterans for service-related PTSD, issued from 2013 through 2017. This dataset is very well-studied in the AI and law research community (see Related Work, Section 6.2 below).

**Table 1**
Frequency of Sentences in the LLT Dataset and in the Analysis Sections of the SRS Dataset, by Rhetorical Type

| Rhetorical Type | LLT Freq | SRS Freq |
|---|---|---|
| Finding Sentence | 490 | 246 |
| Evidence Sentence | 2,419 | 1,762 |
| Reasoning Sentence | 710 | 279 |
| Legal-Rule Sentence | 938 | 616 |
| Citation Sentence | 1,118 | 739 |
| Other Sentences | 478 | 361 |
| Total | 6,153 | 4,003 |

The LLT Dataset labels the six rhetorical roles in legal reasoning that sentences might play: Finding Sentences (primarily stating a finding of fact); Evidence Sentences (primarily stating the content of the testimony of a witness, stating the content of documents introduced into evidence, or describing other evidence); Reasoning Sentences (primarily reporting the trier of fact's reasoning underlying the findings of fact, which often involves an assessment of the credibility and probative value of the evidence); Legal-Rule Sentences (primarily stating one or more legal rules in the abstract, without stating whether the rule conditions are satisfied in the case being decided); Citation Sentences (referencing legal authorities or other materials, and usually containing standard notation that encodes useful information about the cited source); and Other Sentences (not fitting into any of the previous 5 categories). The number of sentences labeled for sentence type in the LLT Dataset is 6,153, with the frequencies of sentence types shown in Table 1.

The FS-PTSD Dataset. To investigate the automatic labeling ("auto-enrichment") of legal decisions at scale, we started with a set of decisions downloaded from the BVA website, the first 10,003 decisions issued in 2018 (the year after the decisions included in the LLT Dataset). These decisions are sufficiently anonymized by the BVA before they are issued. We automatically converted these plain-text decisions to LSJson format (see Section 3.1 below), which resulted in a total of 1,360,230 sentences. We then used a predictive model trained on the LLT Dataset (see Section 2.2 below) to auto-label these sentences for the 6 sentence types from the LLT Dataset.

Because the enrichment pipeline auto-labeled all 1,360,230 sentences, we could use the semantic auto-labeling to filter a subset of decisions of interest [8]. To create a manageable dataset for evaluation, we filtered the 10,003 auto-enriched decisions for Finding Sentences that contained the word "PTSD", and we generated a count of the number of such sentences per decision. We then collected the set of all decisions that contained four or more such sentences (the "FS-PTSD Dataset," cases = 449, sentences = 100,514).[2] We indexed these LSJson files in Elasticsearch (see Section 3.4).

[1] The dataset is available at: https://github.com/LLTLab/VetClaims-JSON/BVA Decisions JSON Format.

[2] The FS-PTSD Dataset is available at: https://github.com/LegalApprentice.

The SRS Dataset. To evaluate the accuracy of the auto-enrichment process for this use case, we drew from the FS-PTSD Dataset a stratified random sample of 25 decisions (the “SRS Dataset,” consisting of 5,529 sentences), stratifying on the number of hits per decision.[3] The SRS Dataset was stratified as follows: 8 decisions with 4 hits (i.e., decisions containing 4 Finding Sentences that contain the word “PTSD”), 6 decisions with 5 hits, 6 decisions with 6 or 7 hits, and 5 decisions with 8 or more hits.

To focus on the BVA’s reasoning, we evaluated the predictive accuracy of automatic sentence typing only for sentences in the analysis or discussion sections of the BVA decisions (i.e., those sentences occurring within the document section headed as “REASONS AND BASES FOR FINDINGS AND CONCLUSIONS”). These document sections have no set internal structure, and they contain the rationale for any conclusions reached by the BVA. In total, there were 4,003 such sentences within the SRS Dataset, with the frequencies of sentence types shown in Table 1.

## 2.2. The ML Predictive Model Used for Auto-Enrichment

We wanted to test the usefulness of relatively simple predictive models, which could be trained and deployed without cloud computing in situations with security, privacy, or other legal concerns. Moreover, we envision user experimentation with new semantic tags for argumentation roles (see Sections 3.2.B and D below), requiring economical auto-enrichment of large subsets of decisions. The software we developed (see Section 3 below) can use legal documents that have been semantically enriched using any predictive models.

For this case study, we used the basic neural network (NN) model reported in [9]. When we trained and tested it on the LLT Dataset, we obtained the performance measures on the test data (30%) shown in Table 2. The confusion matrix for the test set is shown in Table 3 (columns display actual types, rows display predicted types). We discuss the model performance for the SRS Dataset in the case study in Section 4.1. On a laptop, it took about 24 hours to convert the 10,003 decisions to LSJson and to auto-enrich the resulting 1,360,230 sentences.

The present case study is not designed to improve model performance, but to test the practical utility of using auto-enrichment at scale. When newer models are developed, they are not always substantial improvements from a practical standpoint. For example, researchers have reported that a pre-trained RoBERTa model had outperformed other models on a different dataset of legal decisions, which were labeled with different semantic information to perform a very different task [10]. We tested their RoBERTa setup on the LLT Dataset and obtained test results that were quite similar to those for our NN model, with the exception of better results for Reasoning Sentences (compare Tables 2 and 4, row 5). For our case study, the pre-trained RoBERTa model would have made little if any difference (see Section 4 below).

**Table 2**
Performance Measures for the NN Model on the Test Data of the LLT Dataset, by Sentence Type

| | Precision | Recall | F1-score |
|---|---|---|---|
| Citation Sents | 0.98 | 0.98 | 0.98 |
| Legal-Rule Sents | 0.89 | 0.88 | 0.89 |
| Evidence Sents | 0.88 | 0.95 | 0.91 |
| Finding Sents | 0.75 | 0.79 | 0.77 |
| Reasoning Sents | 0.66 | 0.51 | 0.58 |
| Other Sents | 0.81 | 0.77 | 0.79 |

**Table 3**
Confusion Matrix for the NN Model on the Test Data of the LLT Dataset, by Sentence Type

| | C | L-R | E | F | R | O | All |
|---|---|---|---|---|---|---|---|
| C | 337 | 3 | 0 | 0 | 1 | 3 | 344 |
| L-R | 4 | 243 | 2 | 3 | 13 | 9 | 274 |
| E | 1 | 5 | 651 | 13 | 60 | 10 | 740 |
| F | 0 | 12 | 5 | 129 | 21 | 6 | 173 |
| R | 2 | 11 | 24 | 15 | 116 | 7 | 175 |
| O | 1 | 1 | 4 | 3 | 17 | 114 | 140 |
| All | 345 | 275 | 686 | 163 | 228 | 149 | 1846 |

**Table 4**
Performance Measures for the pre-trained RoBERTa Model on the Test Data of the LLT Dataset, by Sentence Type

| | Precision | Recall | F1-score |
|---|---|---|---|
| Citation Sents | 1.00 | 0.98 | 0.99 |
| Legal-Rule Sents | 0.76 | 0.97 | 0.85 |
| Evidence Sents | 0.91 | 0.94 | 0.93 |
| Finding Sents | 0.78 | 0.90 | 0.84 |
| Reasoning Sents | 0.84 | 0.60 | 0.70 |
| Other Sents | 0.76 | 0.66 | 0.71 |

# 3. LA-MPS Overview

Legal Apprentice is a suite of web applications that provide user interfaces (UIs) for reading, annotating, and querying sets of semantically enriched legal documents. Legal Apprentice currently has three separate web applications that communicate with each other: LA-Marker,

[3] The SRS Dataset is available at: https://github.com/LegalApprentice.

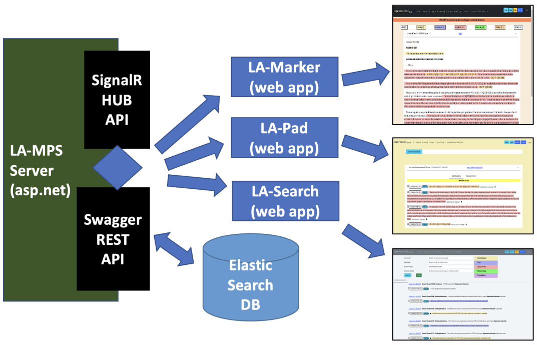


**Figure 1:** LA-MPS Architecture Schematic

LA-Pad, and LA-Search (collectively, "LA-MPS").[4] The LA-MPS architecture (see Figure 1) is designed to be flexible and scalable. LA-MPS is also containerized, meaning that it can run entirely on a single computer or in the cloud, or by connection through a local area network behind a firewall for security. Communication among the Marker, Pad, and Search web applications uses SignalR, a free and open-source software library. This section briefly describes the data structure used in LA-MPS, and then each of the three web applications.

### 3.1. The LSJson Format

The LA-MPS software stores, processes, and exchanges data in JSON format. The format that we call Legal Semantic JSON ("LSJson") is a lightweight, extensible, data-exchange format for capturing the text and the semantic information associated with legal documents. LSJson stores the original decision string of characters, metadata about the decision, details about any predictive models used to auto-enrich the document, and added semantic information about the sentences and paragraphs of the decision. Decisions from a legal tribunal must first be converted from their original formats into LSJson. We have successfully converted original decision files from plain-text, HTML, and PDF formats.

The conversion package for LA-MPS is separate from the rest of the code, and any adequately accurate conversion package can be used, whether rule-based or based on machine learning. Researchers have demonstrated that general systems for detecting sentence boundaries perform much worse on legal documents when compared to their performance on news articles data sets [11], and they showed that a general conditional random field (CRF) model trained on the legal data performed significantly better. Later research confirmed that the CRF model is the most practical approach, with a neural network model not performing significantly better than the CRF model [12]. Occasional errors in converting individual documents can be corrected using the Editor Mode of LA-Marker (see Section 3.2(F) below).

### 3.2. LA-Marker

LA-Marker (or simply Marker) is a web application with a UI for viewing and annotating individual legal documents (see Figure 2). It is written using the open-source Angular web framework from Google.

Marker has six main modes for working with individual legal documents (see Figure 2, at (1)). The following briefly describes each mode.

A. Viewer. Viewer is a "read-only" user mode. The user can read the decision text as annotated with semantic information, but the user cannot modify the annotations.

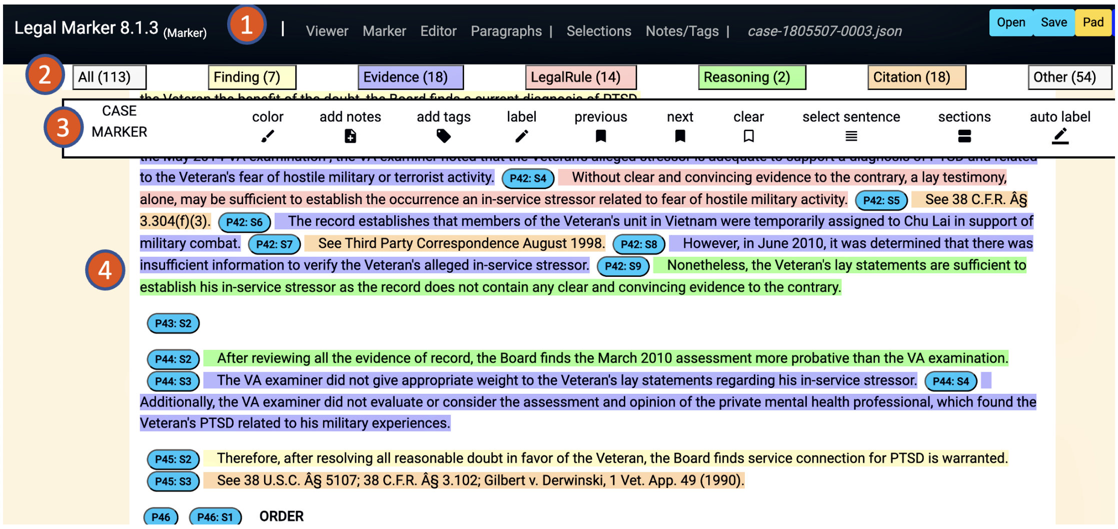
B. Marker. Marker is a user mode that allows a user to read the text and its semantic annotations, and to add or edit annotations (Figure 2, at (3)). The Marker Mode has 5 primary functionalities:

- Displaying the semantic information stored in the LSJson file of the document (e.g., document-level metadata, or sentence-level types, notes, and tags).
- Filtering the document by six sentence types (discussed in Section 2.1): Finding, Evidence, Legal Rule, Reasoning, Citation, and Other (Figure 2, at (2)). Filter buttons display lists of sentences of the selected type in the order in which they occur in the document (Figure 2, at (4)).
- Manually adding or editing the type (role) of a sentence. If an automatic classifier has been used to predict a sentence's type, all possible types are displayed in buttons showing their predicted classification scores. Selecting a button manually assigns a rhetorical role to the sentence.
- Manually adding or editing sentence-level notes or tags, as well as paragraph-level notes or tags (see Figure 2, at 3)).
- Selecting sentences or paragraphs for further action (see the Selections Mode below).

C. Paragraphs. The Paragraphs Mode (see Figure 2, at (1)) displays entire paragraphs in order of priority. The lists are currently sorted by an "interest-score" devised to prioritize paragraphs that might contain entire arguments. The user can add or edit paragraph-level notes or tags, or select specific paragraphs for further action (see Selections Mode below).

D. Notes/Tags. Notes/Tags is a user mode that displays, in grid format, all the notes or tags (sentence-level or paragraph-level) that are present in the document. A database of notes or tags can be exported as a CSV file.

[4] The software code is open source and free, and it is available at: https://github.com/LegalApprentice.



the Veteran the benefit of the doubt, the Board finds a current diagnosis of PTSD

the May 2011 VA examination", the VA examiner noted that the Veteran's alleged stressor is adequate to support a diagnosis of PTSD and related to the Veteran's fear of hostile military or terrorist activity. P42: S4 Without clear and convincing evidence to the contrary, a lay testimony, alone, may be sufficient to establish the occurrence an in-service stressor related to fear of hostile military activity. P42: S5 See 38 C.F.R. Â§ 3.304(f)(3). P42: S6 The record establishes that members of the Veteran's unit in Vietnam were temporarily assigned to Chu Lai in support of military combat. P42: S7 See Third Party Correspondence August 1998. P42: S8 However, in June 2010, it was determined that there was insufficient information to verify the Veteran's alleged in-service stressor. P42: S9 Nonetheless, the Veteran's lay statements are sufficient to establish his in-service stressor as the record does not contain any clear and convincing evidence to the contrary.

P43: S2

P44: S2 After reviewing all the evidence of record, the Board finds the March 2010 assessment more probative than the VA examination.
P44: S3 The VA examiner did not give appropriate weight to the Veteran's lay statements regarding his in-service stressor. P44: S4 Additionally, the VA examiner did not evaluate or consider the assessment and opinion of the private mental health professional, which found the Veteran's PTSD related to his military experiences.

P45: S2 Therefore, after resolving all reasonable doubt in favor of the Veteran, the Board finds service connection for PTSD is warranted.
P45: S3 See 38 U.S.C. Â§ 5107; 38 C.F.R. Â§ 3.102; Gilbert v. Derwinski, 1 Vet. App. 49 (1990).

P46 P46: S1 ORDER

**Figure 2:** View of LA-Marker opening BVA Decision #1805507, indicating: (1) the 6 main modes; (2) the 7 filter buttons; (3) the marking tools available to the user; and (4) the decision text, color-coded for sentence rhetorical role.

E. Selections. The Selections Mode allows a user to gather selected sentences and paragraphs, so the user can send the selections to LA-Pad for further annotation (discussed in Section 3.3 below).

F. Editor. In working with documents in LSJson format, the user might occasionally find errors in segmentation or boundary identification (e.g., of sentences or paragraphs). The user can correct such errors in the LSJson file itself, not merely in the display.

## 3.3. LA-Pad

LA-Pad (or simply Pad) is a web application with a UI for gathering sentences and paragraphs from different decisions (either via Marker or via Search), grouping them into user-defined sets, and annotating those groups with notes or tags. Pad is designed to simulate a traditional "legal pad," which a lawyer might use to gather notes on a topic of interest as her research proceeds. Pad files can be saved and reopened later, to add further research. Figure 3 shows a group of sentences sent either from Marker (see Section 3.2) or from Search (Section 3.4), in the UI for grouping and annotating selected items. The contents displayed in Figure 3 will be explained in Section 4.2, where we discuss the results of the case study.

## 3.4. LA-Search

LA-Search (or simply Search) is a web application with a UI for searching a large set of documents that have been semantically enriched. Search is configured to use Elasticsearch, a search engine based on the Lucene library, but it could be configured to use any search engine. Figure 4 shows a view of Search which displays a selection of sentences that can be sent to Pad for further semantic enrichment (see Section 3.3). The content of the example in Figure 4 will be discussed in Section 4.2 below, where we discuss the results of the case study.

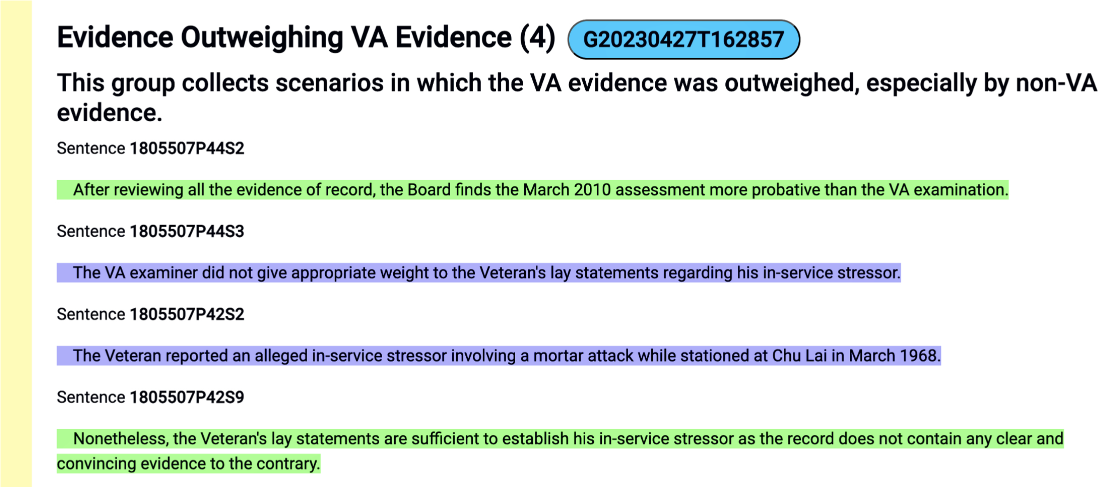
# 4. Evaluation of the Case Study, Using LA-MPS

Using LA-MPS, we conducted a case study simulating normal attorney workflows. We evaluated the accuracy of the auto-labeling in the SRS Dataset for reading annotated decisions, and we performed typical legal research tasks using the auto-enriched FS-PTSD Dataset. This section discusses our results.

## 4.1. Experiments to Evaluate Auto-Labeling Accuracy, Using LA-Marker

A normal workflow in argument mining is efficiently reading a legal decision to identify the elements of reasoning patterns from the evidence to the findings of fact. We used LA-Marker to read annotated decisions and to assess the likelihood and practical significance of errors in sentence auto-labeling.

Quantitative Performance Metrics. The evaluation of the adequacy of the auto-labeling occurred at two levels in the case study. First, in the SRS Dataset, all 25 decisions in fact decided claims involving PTSD. Nine of



Evidence Outweighing VA Evidence (4) G20230427T162857

This group collects scenarios in which the VA evidence was outweighed, especially by non-VA evidence.

Sentence **1805507P44S2**

After reviewing all the evidence of record, the Board finds the March 2010 assessment more probative than the VA examination.

Sentence **1805507P44S3**

The VA examiner did not give appropriate weight to the Veteran's lay statements regarding his in-service stressor.

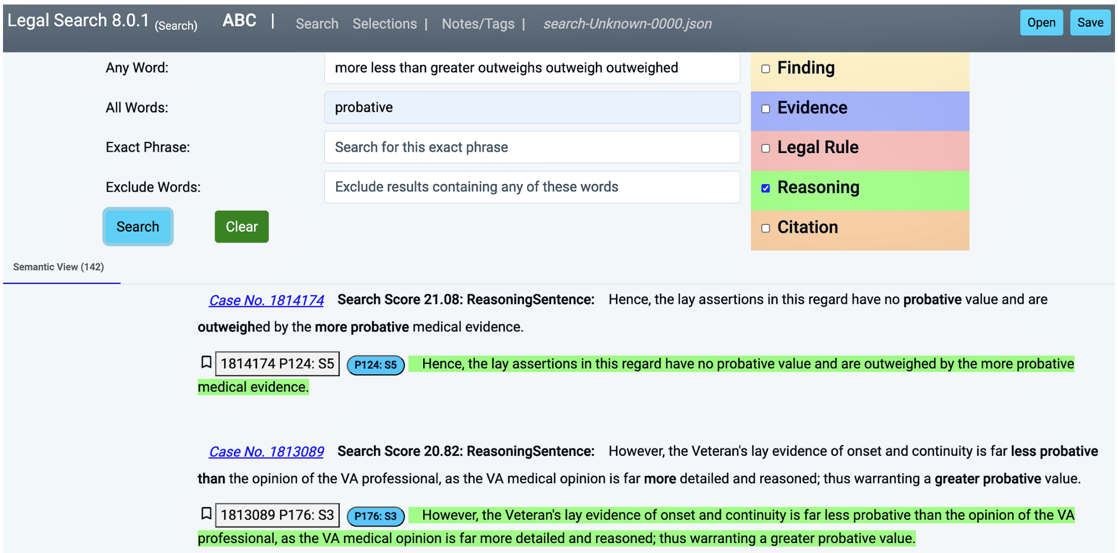
Sentence **1805507P42S2**

The Veteran reported an alleged in-service stressor involving a mortar attack while stationed at Chu Lai in March 1968.

Sentence **1805507P42S9**

Nonetheless, the Veteran's lay statements are sufficient to establish his in-service stressor as the record does not contain any clear and convincing evidence to the contrary.

**Figure 3:** View of LA-Pad showing sentences selected for grouping arguments of the type ”Evidence Outweighing VA Evidence.”



Any Word: more less than greater outweighs outweigh outweighed

All Words: probative

Exact Phrase: Search for this exact phrase

Exclude Words: Exclude results containing any of these words

□ Finding

□ Evidence

□ Legal Rule

☑ Reasoning

□ Citation

*Case No. 1814174* **Search Score 21.08: ReasoningSentence:** Hence, the lay assertions in this regard have no **probative** value and are **outweigh**ed by the **more probative** medical evidence.

1814174 P124: S5 P124: S5 Hence, the lay assertions in this regard have no probative value and are outweighed by the more probative medical evidence.

*Case No. 1813089* **Search Score 20.82: ReasoningSentence:** However, the Veteran's lay evidence of onset and continuity is far **less probative than** the opinion of the VA professional, as the VA medical opinion is far **more** detailed and reasoned; thus warranting a **greater probative** value.

1813089 P176: S3 P176: S3 However, the Veteran's lay evidence of onset and continuity is far less probative than the opinion of the VA professional, as the VA medical opinion is far more detailed and reasoned; thus warranting a greater probative value.

**Figure 4:** View of LA-Search showing the first two of 142 sentences retrieved from the FS-PTSD Dataset on a query including the word ”probative” and comparative words, and also limited to auto-labeled Reasoning Sentences.

these either granted or denied claims to establish a service connection for PTSD, and 14 either granted or denied claims for a particular disability rating for PTSD (e.g., a claim to set the disability rating at 70 percent). Disability ratings (expressed as a percentage) are assigned to veterans based on the severity of their disability. Disability ratings are used to determine a disability compensation rate, and they help determine eligibility for other veterans benefits. A few decisions addressed a variety of other PTSD-related claims (e.g., claiming disability for sleep apnea secondary to service-connected PTSD). In sum, there were no false positives in the SRS Dataset at the level of whole decisions.

Second, to evaluate the accuracy of the auto-enrichment, a practicing attorney with expertise in legal reasoning constructed for each of the 25 decisions in the SRS a confusion matrix, and we calculated the performance of the predictive model on the SRS Dataset. The performance measures for the auto-labeling of sentences in the SRS are shown in Table 5, and the overall confusion matrix is shown in Table 6 (columns display actual types, rows display predicted types). The results are re-

**Table 5**
Performance Measures for the NN Model on the Analysis Sections of the SRS Dataset, by Sentence Type

| | Precision | Recall | F1-score |
|---|---|---|---|
| Citation Sents | 0.97 | 0.96 | 0.96 |
| Legal-Rule Sents | 0.93 | 0.57 | 0.71 |
| Evidence Sents | 0.90 | 0.94 | 0.92 |
| Finding Sents | 0.84 | 0.72 | 0.78 |
| Reasoning Sents | 0.71 | 0.53 | 0.61 |
| Other Sents | 0.56 | 0.96 | 0.71 |

**Table 6**
Confusion Matrix for the NN Model on the Analysis Sections of the SRS Dataset, by Sentence Type

| | C | L-R | E | F | R | O | All |
|---|---|---|---|---|---|---|---|
| C | 707 | 11 | 4 | 2 | 2 | 3 | 729 |
| L-R | 2 | 353 | 1 | 14 | 10 | 0 | 380 |
| E | 3 | 99 | 1665 | 19 | 57 | 8 | 1851 |
| F | 0 | 8 | 11 | 178 | 12 | 4 | 213 |
| R | 2 | 35 | 9 | 15 | 149 | 0 | 210 |
| O | 25 | 110 | 72 | 18 | 49 | 346 | 620 |
| All | 739 | 616 | 1762 | 246 | 279 | 361 | 4003 |

markably similar to the results in the test set of the LLT Dataset (compare Tables 2 and 5). For BVA decisions, these results were achieved with a very small training set and a relatively small training vocabulary.

In particular, the false positive rate for Finding Sentences in the SRS Dataset (precision = 0.84, for a false positive rate = 0.16) was lower than we observed when we trained the model on the LLT Dataset (false positive rate = 0.25). Finding Sentences are critical in identifying reasoning because they state the conclusions of the tribunal. The precision measures for Evidence Sentences, Legal-Rule Sentences, and Citation Sentences were quite high (0.90, 0.93, and 0.97, respectively). Reasoning Sentences had comparable F1-scores in the LLT and the SRS datasets, and they had the least accuracy of all sentence types.

These results at scale were consistent with data-centric analysis using the LLT Dataset [13], which suggested that this classification system of sentence types would be robust as the number of labeled BVA decisions increases. The present case study provides some confirmation that a data-centric analysis can be indicative of robustness of a classification system at scale.

Moreover, many of the prediction errors we observed in the SRS Dataset are perhaps avoidable through retraining the model by combining the LLT Dataset with a curated SRS Dataset.[5] For example, while the LLT Dataset focused on claims to establish service connection for PTSD in the first instance, many of the SRS decisions were claims involving a particular disability rating for PTSD. Within the SRS decisions to establish a rating, one recurring error was auto-labeling as Evidence Sentences the detailed regulatory criteria and symptoms for assigning a particular percent of disability. Because such sentences report symptoms from regulations, they should be labeled as Legal-Rule Sentences, not Evidence Sentences. Such classification errors occurred for 55 sentences in the SRS Dataset. If the model is re-trained on a dataset that includes such sentences as part of the gold standard, then such auto-labeling errors might be reduced or avoided.

Practical Cost of Error. The confusion matrix for the SRS Dataset also suggests that even when false positives occur, they are often of little practical importance given our use case. For example, when a sentence is mislabeled as a Finding Sentence, it may actually be a Reasoning Sentence (see Table 3, row 4; Table 6, row 4). The practical cost of such an error is likely to be low because a search result involving a Reasoning Sentence instead of a Finding Sentence could still produce an instructive example of reasoning on the search topic.

In addition, it became clear in the case study that in a LA-MPS work environment, many sentence-level typing errors are in practice “harmless errors” because they are visually obvious. When reading a decision in which an entire paragraph is devoted to reciting evidence or legal rules (as is often the case), an isolated classification error for a single sentence or two stands out visually, and the error is quickly recognized and discounted mentally. Another example involves sentences auto-labeled as Other Sentences, a category with much lower precision than other categories (see Tables 5 and 6, row 6). In LA, Other Sentences are color-coded with white background. When reading a decision in Marker, such sentences tend to stand out visually and be mentally re-classified by the user. Thus, the efficiency with which LA-Marker allows an auto-enriched decision to be read and understood is not much affected by visually obvious errors.

In sum, this part of the case study confirms our hypothesis that a predictive model trained and tested on a relatively small dataset (the LLT Dataset) can retain its level of performance when used to auto-enrich at scale. Also, some errors can be visually discounted in practice when reading decisions with semantic color-coding.

## 4.2. Experiments Involving Search, Using LA-Search and LA-Pad

In addition to reading and annotating decisions efficiently, it is important to identify the relevant decisions to read. For this aspect of the case study, we focused on resolving conflicting evidence.

Indexing the 100,514 sentences from the FS-PTSD

[5] The curated SRS Dataset is available at: https://github.com/LegalApprentice.

Dataset into Elasticsearch showed that 5,010 sentences were auto-labeled as Reasoning Sentences (RSs). We filtered these RSs for the word “probative” (resulting N = 510), and then queried further for sentences containing one or more comparative words (e.g., “more”, “less”, “than”, “outweigh”), retrieving a total of 142 RSs (see Figure 4). In reviewing these sentences, a practicing attorney with expertise in legal reasoning noted that a recurring factor cited by the Board in weighing the comparative probative value of conflicting evidence was access to, and review of, the complete evidence in a veteran’s record and claims file–a factor that might sometimes favor opinions by VA experts over non-VA experts, because of ease of access to complete records.

By narrowing the search further to sentences also containing the word “VA”, we retrieved a set of 60 RSs (occurring in 46 unique decisions) to analyze further. Eight of the 60 RSs appeared on their face to assign more probative weight to non-VA evidence than to VA evidence, and review of the contexts of those 8 using LA-Marker provided 3 examples of scenarios in which the non-VA evidence prevailed. All 3 dealt with the issue of proving a causal nexus between a current diagnosis and some event or injury that occurred during active military service. In one case (#1805507), the VA examiner failed to consider other evidence in the record, including the expert opinion of the private mental health professional. In a second case (#1803222), the VA examiner’s reasoning failed to take into account or was inconsistent with extensive other evidence. In the third case (#1806062), the Veteran’s private physician provided more explanation of the intervening causal steps in the nexus, and supplied supporting scholarly research.

Using Pad, we grouped and reorganized selected sentences from multiple decisions, naming the type of argument ”Evidence Outweighing VA Evidence” (see Figure 3). We have begun to build a library of such argument patterns.

The legal research conducted in this case study would have been practically impossible without auto-enrichment and semantic search. The FS-PTSD Dataset was drawn from 10,003 decisions using the auto-labeling for Finding Sentences. Even within the 449 decisions of the FS-PTSD Dataset, the word “probative” occurs in 294 of the decisions, so a query at the level of whole decisions would not narrow the search substantially. LA-MPS and auto-enrichment enabled efficiently locating examples of the targeted reasoning within the decisions.

# 5. Discussion: Toward Implementation Research

This section briefly discusses some lessons from our case study for three principles of implementation research.

## 5.1. Feasibility at Scale

It can be a challenge to identify all the normal workflows involved in legal work at scale. Our case study investigated reading annotated sentences in context and filtered for sentence types, as well as searching for relevant decisions to review. Other workflows might include question-answering (employing chatbots), document summarization, document drafting, and predictive-factor extraction.

It would be worthwhile to identify all the security, privacy, and other regulatory constraints on working with legal data at scale. Our case study simulated a scenario in which the labeled data, model development, and auto-enrichment would all be local. Depending upon the constraints in a particular implementation, cloud storage and computing could be used instead.

## 5.2. Statistical Evaluation

Implementation at scale places a premium on the *validity* of the gold-standard data and of the results, not merely on the *reliability.* Given the resource-intensive nature of generating data for training and testing, it can be more efficient to employ non-experts to manually label data or to evaluate the predictive results, and to rely upon measures of reliability (consistency) among labelers for quality assurance. In real-world implementation, however, the validity (accuracy) of the labeling is what is important—i.e., whether the text is correctly classified. Our case study experiments employed a practicing attorney with expertise in legal reasoning. What are needed are best practices for evaluating validity at scale, as efficiently as possible.

Accuracy is always a concern when transferring predictive models from laboratory settings to auto-enrichment at scale, especially when the training and testing has been done on a small dataset that the model might have overfit. Random sampling from a population at scale can provide some reassurance about continued accuracy after model transfer. What are needed are best practices for using random samples to estimate statistics like precision, recall, and F1-scores for data populations at scale.

Making both open-source code and labeled data publicly available is important for facilitating widespread implementation at scale. But also, when the objective is the real-world validity of the results, replication and verification are necessary to achieve a consensus on accuracy at scale.

## 5.3. Practical Usefulness

Implementing a use case at scale poses the challenge of evaluating its practical usefulness. For example, a moderate level of performance, with precision on the order of 0.75 and similar recall, might be adequate for some practical use cases—such as in the use case reported in this paper, mining illustrative examples of arguments that have been successful in certain evidential situations. In our case study, we have tried to evaluate the results using both standard quantitative measures and a practical assessment of the cost of expected errors.

# 6. Prior Related Work

This section surveys recent related work in three areas: research on auto-enrichment and evaluation at scale, research using BVA datasets, and other recent innovative applications.

## 6.1. Auto-Enrichment and Evaluation at Scale

There have been many experiments at scale on individual tasks related to legal research. To the best of our knowledge, none have deployed integrated web applications to assist standard attorney workflows in reading, searching, and annotating legal decisions enriched at scale for semantic information, the workflows assisted by LA-MPS. We discuss here several recent studies that have evaluated auto-enriched samples drawn from large sets of legal decisions. We discuss them from the perspective of the three principles of implementation research discussed in Section 5.

Two studies have used the extensive corpus of full-text BVA decisions (over 1 million decisions from 1999 to 2017 [14]) to conduct research at scale. One advantage of using this corpus is that decisions are anonymized before they are published.

One study simulated a BVA staff attorney drafting an opinion in a new case, and the study developed a tool to recommend a legal citation to a published judicial decision, statute, or regulation [14]. From over 1 million BVA decisions, researchers filtered a subset of 324,309 BVA cases that raised a single issue and had complete metadata. They split those cases into training, validation, and test sets (72%, 18%, 10%, respectively). Two neural models (a Bi-directional Long Short Term Memory model and a fine-tuned RoBERTa-based model) performed comparably and better than other methods, using sequences of words in the draft opinion as context to predict the next citation. They used recall at 1, 5, and 20 as the quantitative metric (the proportion of data instances where the correct next citation is among the model's top 1, 5, or 20 predictions). They considered recall@5 to simulate a "typical user, who benefits from a small number of recommendations that can quickly be examined for the most appropriate." Neural model training for extended periods continuously improved up to a recall@5 of 83.2%, which they considered "acceptable performance." They also performed a qualitative error analysis on a sample of 200 erroneous predictions, concluding that "even incorrect predictions may still be useful." One next step, they concluded, was "to evaluate the usefulness of the models trained here with expert users."

The second study simulated drafting extractive summaries that could enable readers to make an informed decision about whether to read the full decision [15]. From nearly 1 million BVA decisions, researchers filtered about 35,000 single-issue decisions that dealt with service connection for PTSD, from which they randomly sampled 112 cases for their experiments (92 for training and validation, 20 for testing). The task was to generate case summaries that were between 6-10 sentences long, in which 2-6 sentences should summarize the BVA's reasons and the evidence considered. They first extracted "predictive sentences," and then they trained a random forest classifier to classify them as either "Reasoning/EvidentialSupport" or "Other." For 954 training sentences and 341 testing sentences, their classifier reached 0.85 precision, 0.77 recall, and 0.81 F-score. They used the sentence classification and Maximal Marginal Relevance (MMR) to select the variable number of Reasoning/Evidential Support sentences for the summary. Although their ROUGE-1 and ROUGE-2 scores were only 0.269 and 0.102, respectively, they had evidence from human drafted summaries that the value of ROUGE scores as metrics were of limited use for evaluating summaries of legal opinions. They also conducted extensive qualitative error analysis, from which they hypothesized that "sentences involving evidential reasoning" might be useful for identifying more details in automated summaries.

Other recent large-scale research includes: exploring court data, using more than a quarter-million case dockets in HTML format and ontology-leveraged tools [8]; predicting the outcome of motions on the basis of court administrative data and complaint documents, using 184,125 civil cases from the State of Connecticut Judicial Branch to draw a sample of 7904 motions to strike, and testing 6 auto-classification models [16]; predicting verdict labels on the basis of the pre-verdict text of a decision, using a corpus of 544,857 court decision documents in French for landlord-tenant disputes in Quebec, Canada, and CamemBERT (a BERT model pretrained on French material) [17].

## 6.2. Research Using BVA Datasets

In addition to the two BVA studies at scale discussed in Section 6.1, researchers have employed the small LLT Dataset of 50 BVA decisions in various studies.

Some have tested methods and tools to perform tasks that could be relevant to enriching BVA decisions at scale. Recent studies include: training a general conditional random field (CRF) model to detect sentence boundaries

[11, 12]; classifying Finding Sentences for their linguistic polarity (i.e., whether the finding is positive or negative on the legal issues presented) [9]; evaluating an annotation system (CAESAR) based on the hypothesis that sentences that are semantically similar often have the same rhetorical type [18]; investigating how changes in the size of the dataset, the train/test splits, and human labeling accuracy affect the performance of a trained deep learning classifier [13]; and testing whether a small set of labelled data could train deeper and more accurate predictive models (obtaining the highest accuracy with a Bidirectional Long Short-Term Memory (Bi-LSTM) model for classifying sentence rhetorical roles) [19].

Others have investigated methods and tools for performing related tasks, such as: assessing the performance of different explainability methods (XAI), after using a convolutional neural network for classifying sentences for rhetorical role [20]; identifying factors that the tribunal considers when assessing the credibility or trustworthiness of individual items of evidence [21]; investigating computer-assisted text classification using Boolean matching rules (CASE) [22]; and examining the ability of pre-trained language models to generalize beyond the legal domain and dataset they were trained on (finding that the performance of an SVM and a RoBERTa model trained to classify sentences as “fact” or “non-fact” was “surprisingly high,” despite being trained on datasets from different domains and jurisdictions) [23].

### 6.3. Innovative Applications

This paper introduces an innovative suite of AI web applications, LA-MPS, to assist attorneys and judges when reading, searching, and annotating semantically enriched legal decisions. Some notable innovative applications in recent years could assist legal practice with some aspects of this workflow.

Recent applications for annotation include: the Scribe web application for annotation of French court decisions, which facilitates a collaborative workflow between annotators and developers [24]; a LegAi annotation editor, which supports annotating legal texts with the LegAi higher language, with the goal of constructing formal knowledge bases that can support efficient reasoning [25]; an annotator assistant that allows users to create, update, and delete annotations suggested by an algorithmic annotator for named entity recognition (NER) [26]; and a prototype interface CAESAR (Computer-Assisted Enhanced Semantic Annotation & Ranking) that allows assigning the same rhetorical type to sentences that are semantically similar [18].

Other use cases include: an intelligent tutoring system for analyzing legal decisions, employing a cognitive computing framework that matches various ML capabilities to the proficiency of the user [10]; a legal support system with a natural UI connected to an Abstract Dialectical Framework (ADF) to predict admissibility before the European Court of Human Rights, and to present an explanation of the prediction to the user [27]; and a platform that allows users to explore data and drive analysis by leveraging an ontology configuration, with natural language statements in the UI [8].

## 7. Conclusion and Future Work

LA-MPS is ready to be deployed at scale for BVA research, and for adaptation to other legal domains. It should be especially useful in legal areas with a high volume of decisions that assess multiple kinds of evidence and employ complicated reasoning, but where mining that reasoning is difficult.

Although the case study reported here examined the usefulness of auto-labeling only for sentence rhetorical role, we have already auto-enriched all Finding Sentences in the 10,003 cases for their linguistic polarity (positive or negative) [9], and for the legal issue addressed by the sentence. This should assist creating a library of legal reasoning patterns at greater granularity (e.g., successful and unsuccessful arguments on specific legal issues).

The LA-MPS environment built on LSJson is also extendable by adding new web applications. Currently under development for LA-MPS is a fourth integrated web application, LA-Draw. This application will enable the user to create graphic conceptual networks connecting terms, sentences, paragraphs, or decisions.

Next steps in the BVA domain include comparative testing of a variety of ML algorithms and large language models, using the combined LLT and SRS curated datasets. The best-performing models could then be used to auto-enrich a larger number of BVA decisions, and to provide an indexed database available for research by the public, by veterans’ representatives, and by lawyers and judges at the BVA. Such a service could be hosted on a cloud computing platform such as Microsoft Azure.

Finally, LA-MPS can also be implemented as the user interface for deploying other tools, such as citation recommendation [14] or auto-generated extractive summaries [15]. In addition, LA-MPS could be used to provide training and quality assurance tools to assist human authors in writing legal decisions, by providing feedback on how well the decision states the tribunal’s reasoning [10].

## References

[1] Legal Services Corporation, Technology Baselines: Technologies that Should Be in Place in a Legal Office Today (Proposed for Public Comment), 2023.

[2] American Bar Association, Standards for the Provision of Civil Legal Aid, Standard 6.7 on Providing

Adequate Resources for Research and Investigation, 2021.

[3] J. M. Faupel-Badger, A. L. Vogel, C. P. Austin, J. L. Rutter, Advancing translational science education, Clin. Transl. Sci. 15 (2022) 2555–2566. doi:10.1111/cts.13390.

[4] Board of Veterans' Appeals, U.S. Department of Veterans Affairs, Annual Report, Fiscal Year 2021, 2021.

[5] V. R. Walker, K. Pillaipakkamnatt, A. M. Davidson, M. Linares, D. J. Pesce, Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning, in: Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019), 2019.

[6] V. R. Walker, J. H. Han, X. Ni, K. Yoseda, Semantic types for computational legal reasoning: Propositional connectives and sentence roles in the veterans' claims dataset, in: Proceedings of ICAIL '17, ACM Digital Library, 2017, pp. 217–226.

[7] V. R. Walker, A. Hemendinger, N. Okpara, T. Ahmed, Semantic types for decomposing evidence assessment in decisions on veterans' disability claims for ptsd, in: Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2017), 2017.

[8] A. Paley, A. L. L. Zhao, H. Pack, S. Servantez, R. F. Adler, M. Sterbentz, A. Pah, D. Schwartz, C. Barrie, A. Einarsson, K. Hammond, From data to information: Automating data science to explore the U.S. court system, in: Proceedings of ICAIL '21, ACM Digital Library, 2021, pp. 119–128.

[9] V. R. Walker, S. R. Strong, V. E. Walker, Automating the classification of finding sentences for linguistic polarity, in: Proceedings of the 2020 Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL), 2020.

[10] H. Westermann, J. Savelka, V. R. Walker, K. D. Ashley, K. Benyekhlef, Toward an intelligent tutoring system for argument mining in legal texts, in: Proceedings of JURIX '22, 2022, pp. 133–142.

[11] J. Savelka, V. R. Walker, M. Grabmair, K. D. Ashley, Sentence boundary detection in adjudicatory decisions in the United States, TAL 58 (2017) 21–45.

[12] G. Sanchez, Sentence boundary detection in legal text, in: Proceedings of the Natural Legal Language Processing Workshop 2019, 2019, pp. 31–38.

[13] H. Westermann, J. Savelka, V. R. Walker, K. D. Ashley, K. Benyekhlef, Data-centric machine learning: Improving model performance and understanding through dataset analysis, in: Proceedings of JURIX '21, 2021, pp. 54–57.

[14] Z. Huang, C. Low, M. Teng, H. Zhang, D. E. Ho, M. S. Krass, M. Grabmair, Context-aware legal citation recommendation using deep learning, in: Proceedings ICAIL '21, ACM Digital Library, 2021, pp. 79–88.

[15] L. Zhong, Z. Zhong, Z. Zhao, S. Wang, K. D. Ashley, M. Grabmair, Automatic summarization of legal decisions using iterative masking of predictive sentences, in: Proceedings of ICAIL '19, ACM Digital Library, 2019, pp. 163–172.

[16] D. J. McConnell, J. Zhu, S. Pandya, D. Aguiar, Case-level prediction of motion outcomes in civil litigation, in: Proceedings of ICAIL '21, ACM Digital Library, 2021, pp. 99–108.

[17] O. Salaün, P. Langlais, K. Benyekhlef, Labels distribution matters in performance achieved in legal judgment prediction tasks, in: Proceedings of ICAIL '21, ACM Digital Library, 2021, pp. 268–269.

[18] H. Westermann, J. Savelka, V. R. Walker, K. D. Ashley, K. Benyekhlef, Sentence embeddings and high-speed similarity search for fast computer assisted annotation of legal documents, in: Proceedings of JURIX '20, 2020, pp. 164–173.

[19] R. Ahmad, D. Harris, M. I. Sahibzada, Understanding legal documents: Classification of rhetorical role of sentences using deep learning and natural language processing, in: Proceedings of the 14th IEEE International Conference on Semantic Computing (ICSC2020), 2020, pp. 464–467.

[20] Łukasz Górski, S. Ramakrishna, Explainable artificial intelligence, lawyer's perspective, in: Proceedings of ICAIL '21, ACM Digital Library, 2021, pp. 60–68.

[21] V. R. Walker, D. Foerster, J. M. Ponce, M. Rosen, Evidence types, credibility factors, and patterns or soft rules for weighing conflicting evidence: Argument mining in the context of legal rules governing evidence assessment, in: Proceedings of the 5th Workshop on Argument Mining, 2018, pp. 68–78.

[22] H. Westermann, J. Savelka, V. R. Walker, K. D. Ashley, K. Benyekhlef, Computer-assisted creation of Boolean search rules for text classification in the legal domain, in: Proceedings of JURIX '19, 2019, pp. 123–132.

[23] J. Savelka, H. Westermann, K. Benyekhlef, Cross-domain generalization and knowledge transfer in transformers trained on legal data, in: Proceedings of the 2020 Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL), 2020.

[24] S. A. Mahmoudi, G. Zambrano, C. Condevaux, S. Mussard, *Scribe*: A specialized collaborative tool for legal judgment annotation, in: Proceedings of JURIX '22, 2022, pp. 290–293.

[25] T. Libal, The LegAi editor: A tool for the construction of legal knowledge bases, in: Proceedings of JURIX '22, 2022, pp. 286–289.

[26] Y.-T. Huang, H.-R. Lin, C.-L. Liu, Toward an inte-

grated annotation and inference platform for enhancing justifications for algorithmically generated legal recommendations and decisions, in: Proceedings of JURIX '22, 2022, pp. 281–285.

[27] K. Atkinson, J. Collenette, T. Bench-Capon, K. Dzehtsiarou, Practical tools from formal models: The ECHR as a case study, in: Proceedings of ICAIL '21, ACM Digital Library, 2021, pp. 170–174.