# Automating the Classification of Finding Sentences for Linguistic Polarity

Vern R. WALKER[a,1], Stephen R. STRONG[b], and Vern E. WALKER[c]
[a] *Maurice A. Deane School of Law at Hofstra University*
[b] *Apprentice Systems, Inc.*
[c] *The Legal Semantics Project*

**Abstract.** Automatically labeling important information in legal decisions issued from high-volume, fact-intensive governmental claims processes can help increase access to justice for claimants, provided adequate machine learning models can be developed using very small datasets. This paper reports on an experiment to determine whether it is possible to automatically label the linguistic polarity (positive or negative) of sentences that state findings of fact on the legal issues presented. The results support the hypothesis that the language used to state affirmative and negative findings of fact is sufficiently regular that predictive models can be adequately trained on a very small amount of labeled data – at least trained "adequately enough" for some valuable use cases. The results also suggest that the predictive errors of such models are not due to the linguistic or logical complexity of the sentence structure. These promising results suggest directions for further improvement in predictive modeling. The sentence-polarity datasets and annotation protocols used in the experiment will be publicly available.

**Keywords.** Sentiment polarity, linguistic polarity, finding sentence, finding of fact, machine learning, neural network, small data.

## 1. Introduction

Society needs to increase access to justice for governmental claims processes that have complex evidentiary records, require expert fact-finding, and occur in high volumes, but which do not provide resources for legal representation. Examples are claims for benefits or compensation for disabilities (by military veterans or other disabled persons). Automated semantic labeling of significant information from past claims decisions, such as the findings of fact in those decisions, could help inform arguments made in future claims. Machine learning (ML) models could help label the text of such decisions, provided it is economically feasible to develop, train and test such models, and if those models perform the annotation task with sufficient accuracy.

Earlier research established that ML models could be trained on very small datasets to identify the sentences in decisions that primarily state findings of fact, with an

[1] Corresponding Author: Vern R. Walker, Maurice A. Deane School of Law, Hofstra University, Hempstead, New York 11549, USA; E-mail: vern.r.walker@hofstra.edu.

Proceedings of the 2020 Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL)

accuracy that is adequate for some valuable use cases [1]. The question we address here is whether ML models can then label those sentences for their sentiment polarity – i.e., whether the findings involved were positive or negative with respect to a legal issue addressed. This paper reports the results of an experiment conducted to test two hypotheses:

**Hypothesis I**: that the language used to state affirmative and negative findings of fact in legal decisions is sufficiently regular that ML models can be trained, on a very small amount of labeled data, to have adequate predictive performance for at least some use cases; and

**Hypothesis II**: that the logical and linguistic complexity of the structure of the sentences stating those findings of fact would be a major cause of the prediction errors generated by such ML models.

Our experimental results support the first hypothesis, but they do not support the second. The paper is organized into the following sections. Section 2 describes the datasets developed for the experiment. Section 3 describes the baseline ML model we used to make predictions of sentence polarity. Section 4 reports results of the experiment, and Section 5 uses the results to evaluate our two hypotheses. Section 6 discusses prior related work, and Section 7 concludes with a brief discussion of future work.

## 2. The Datasets for the Experiment

We tested these hypotheses on two labeled datasets[2] that we developed by annotating two datasets published by the Research Laboratory for Law, Logic & Technology (LLT Lab) at the Maurice A. Deane School of Law at Hofstra University.[3] While the LLT Lab annotated its datasets only for the rhetorical roles of sentences, we added annotations for linguistic polarity to those sentences that are labeled as "Finding Sentences".

### *2.1. The LLT Lab's Two Initial BVA PTSD Datasets*

The LLT Lab has annotated and made publicly available two sets of fact-finding decisions issued by the U.S. Board of Veterans' Appeals (BVA). The first set consists of 50 decisions issued from 2013 through 2017 (the "first BVA PTSD dataset"). Those decisions were arbitrarily selected from adjudicated disability claims filed by veterans for service-related post-traumatic stress disorder (PTSD) [1]. PTSD is a mental health problem that some people develop after experiencing or witnessing a traumatic event, such as combat or sexual assault. If the claimant is dissatisfied with the decision of the Regional Office of the U.S. Department of Veterans Affairs, she may file an appeal to the BVA, which is an administrative appellate body that has the authority to decide the facts of each case based on the evidence [2, 3]. In deciding the appeal, the BVA must review all of the evidence related to the claim, and it must provide a written statement of the reasons or bases for its findings of fact and conclusions. *Caluza v. Brown*, 7 Vet. App. 498, 506 (1995), *aff'd*, 78 F.3d 604 (Fed. Cir. 1996).

For each of the 50 BVA decisions in the first dataset, the LLT Lab extracted those sentences that address the factual issues related to the claim for PTSD, or for a closely-

[2] The two labeled datasets and our annotation protocol for labeling Finding Sentence polarity are available at: https://github.com/vernrwalker/FindingSentencePolarity .

[3] Available at: https://github.com/LLTLab/VetClaims-JSON .

related psychiatric disorder [1]. The LLT Lab then classified each of these sentences for its rhetorical role. The types of rhetorical role were: Finding Sentence (primarily stating a finding of fact by the BVA on a legal issue), Evidence Sentence (primarily stating evidence), Reasoning Sentence (primarily stating the BVA's reasoning underlying the findings of fact); Legal-Rule Sentence (primarily stating a legal rule in the abstract); Citation Sentence (primarily stating a citation to legal authorities or other materials); and "Other" Sentences (playing other rhetorical roles). [4]

The LLT Lab has subsequently annotated and published a second set of BVA PTSD decisions (the "second BVA PTSD dataset"). This second set consists of 25 decisions issued during 2018 and 2019. The LLT Lab annotated the PTSD-related sentences in these decisions for the same rhetorical roles as those used in the first set.

### *2.2. The Rhetorical Role of a "Finding Sentence" in the Datasets*

A Finding Sentence is a sentence that primarily states an authoritative finding, conclusion or determination of the trier of fact, regarding a legal issue presented in the case. In these datasets, the trier of fact is the BVA, which determines whether various legal conditions have been satisfied in the case "as a matter of fact".

Identifying the Finding Sentences within a BVA decision is important for a number of reasons. First, in order to understand the merits of the decision or to conduct an appeal, a lawyer or party must be able to unambiguously determine the factual findings of the Board. Second, because the Finding Sentences state the major conclusions of the BVA, any summary of the decision would likely include the major findings of fact made by the Board. Third, if we can map a Finding Sentence to the corresponding legal issue(s) that it is about, we can then link the evidence assessment in the decision to particular legal issues. Fourth, Finding Sentences anchor a larger segment of text for purposes of argument mining [1, 4]. Thus, automatically and accurately identifying Finding Sentences is an important analytical task.

In analyzing the content and nature of a Finding Sentence, we employ attribution theory [4]. Stated in general, attribution theory tries to identify attribution relations within texts. Such relations typically have three main elements: the attribution cue, the attribution subject, and the attribution object. The "attribution cue" is the word or phrase that signals the attribution, and which provides the lexical grounds (warrant) for our making the attribution. The "attribution subject" is the actor to whom we attribute acceptance of a proposition as being true. The "attribution object" is the propositional content of a sentence or clause that we attribute to the attribution subject, and which states what the actor accepts as true.

In order to be classified as a Finding Sentence, the text of the sentence and/or its context must warrant our attributing at least one proposition (the finding-attribution object, or simply "finding") to the finder of fact (in these datasets, the BVA) as an authoritative determination of fact in the case. The finding-attribution cue is the portion of text that warrants our attributing the proposition to the BVA. For example, in the Finding Sentence "*the Board finds that the Veteran has PTSD*," the finding-attribution subject and cue "*the Board finds that*" warrants our attributing to the Board the finding that "the Veteran has PTSD". When manually annotating a sentence to be a Finding Sentence, an important check is determining that it contains at least one finding-attribution relation.

### *2.3. Two Experimental Datasets: Labeling the Linguistic Polarity of Finding Sentences*

The legal function of a finding of fact is to authoritatively determine whether a legal requirement has been satisfied or not. Logically speaking, a finding can either have **positive polarity** with respect to a legal issue (i.e., the legal requirement has been satisfied) or **negative polarity** (the legal requirement has not been satisfied). From a linguistic perspective, the English language encodes positive or negative polarity using words of "affirmation" and "negation". If a Finding Sentence expresses positive polarity with respect to the legal issue, then it states that the rule requirement has been satisfied in the case. If the text expresses negative polarity, then it states that the rule requirement has not been satisfied.

One difficulty with labeling an entire sentence with a single linguistic polarity is the grammatical and logical complexity of some Finding Sentences. While every Finding Sentence must warrant our attributing at least one finding to the trier of fact, some Finding Sentences are structurally complex. They might state multiple findings (some positive, some negative), or they might state additional propositions that are not findings (e.g., those stating evidence or legal rules). For example, the following Finding Sentence from BVA decision Number 1715225 consists of multiple clauses, each with a different semantic role: "*As the Board has found the March 2014 VA examination to be the most probative medical evidence of record, and the VA examiner did not find such disorder to be present diagnostically or to have been present in service, the Board finds that service connection for such a disorder is not warranted.*" The first dependent clause of this complex sentence ("*As the Board has found the March 2014 VA examination to be the most probative medical evidence of record*") states the reasoning of the Board in evaluating the probative value of the evidence. The second dependent clause ("*as … the VA examiner did not find such disorder to be present diagnostically or to have been present in service*") is a statement of the evidence attributed to a VA examiner. The main clause is the actual finding of fact, and it states a negative finding of fact ("*service connection for such a disorder is not warranted*"). Observing such complex sentences in our datasets led us to formulate Hypothesis II – that the logical and linguistic complexity of sentence structure could be a major cause of prediction errors generated by trained ML models.

**The Training-Testing Dataset.** To test our two hypotheses, we assigned a polarity value to each Finding Sentence from the first LLT dataset (50 decisions, containing 487 Finding Sentences). We first analyzed complex Finding Sentences into finding-attribution relations, and then we assigned a polarity value to each finding-attribution. These annotations were made, first, by one of two annotators who have advanced degrees in English grammar and composition, and who used protocols (guidelines) for annotating the attribution elements and the polarity of finding attributions.[4] Second, all annotations were reviewed by a legal expert, who has practiced and taught law in the United States for over 30 years.

For this experiment, the legal expert then used the attribution polarities to assign a single polarity to each Finding Sentence as a whole, using the following guidelines. If the Finding Sentence contains only a single finding, then the polarity of the finding becomes the polarity of the Finding Sentence. If the Finding Sentence contains more than

[4] We wish to acknowledge the contribution of Amy Nee, a member of The Legal Semantics Project, whose annotations of finding attributions helped lay the foundation for the experiment reported here, but who did not participate in annotating polarity.

one finding, but all of those findings have the same polarity, then that polarity is assigned to the Finding Sentence. If the Finding Sentence has more than one finding and the findings have different polarity values, then we balance two potentially competing considerations. First, other things being equal, we classify the sentence polarity as negative if the sentence contains at least one negative finding, because we want the predictive ML model to learn to identify negative-polarity cues as features. However, on occasion, although one or more negative findings are present, the principal finding is positive. In such a situation, we assigned that sentence a positive polarity, on the theory that we do not want the predictive model to misclassify the polarity of a finding that is critical to the decision.

**The Testing-Only Dataset.** We used the second LLT dataset (25 decisions, 136 Finding Sentences) as an additional test of the performance of the predictive model. This dataset contains no sentences that were used to train the model. To each Finding Sentence, our legal expert used the guidelines in the previous paragraph to assign a sentence-level value for polarity. This ensured that the labeling of Finding Sentences was consistent between the two datasets.

Table 1 shows the number of Finding Sentences in each of the two datasets, along with the frequencies of positive and negative sentence polarity in each dataset.

**Table 1.** Two datasets in the experiment, the Training-Testing Dataset and the Testing-Only Dataset, with their frequencies of Finding Sentences (FSs), divided into positive FS polarity, and negative FS polarity.

| **Training-Testing Dataset (50 Decisions)** | | | **Testing-Only Dataset (25 Decisions)** | | |
|---|---|---|---|---|---|
| **Freq Finding Sentences** | **Freq Positive-Polarity** | **Freq Negative-Polarity** | **Freq Finding Sentences** | **Freq Positive-Polarity** | **Freq Negative-Polarity** |
| 487 | 194 | 293 | 136 | 83 | 53 |

## 3. Design of the Predictive Model

To test the two hypotheses stated in Section 1 using the datasets in Section 2, we designed a basic neural network (NN) model to predict the linguistic polarity of Finding Sentences, on the basis of only the sentence text. We call this the baseline model, because our objective was to create a model whose predictive power would derive as much as possible from the legal language used in each Finding Sentence.

We used the preprocessing tokenizer in Keras, which disregards punctuation and represents sentences as space-separated sequences of words (maximum vocabulary = 3000 words). The Numpy matrix was constructed using "term frequency – inverse document frequency" (*tf-idf*). The architecture of the NN model consisted of a Keras sequential model with five layers. The activation layer after the first hidden layer was rectified linear unit (*relu*), and the activation function on the output layer was *softmax*. (The softmax activation function returns the likelihood of the predicted classification as a vector.) There was also a dropout layer ($p = 0.5$) between the activation layer and the output layer, to help prevent overfitting.

We conducted standard training and testing of the NN model using the Training-Testing Dataset. For training, we fit the model (optimizer = adam, loss = categorical_crossentropy) to the training data (training = 70%, with validation split = 10%; epochs = 20). For testing, we obtained performance measures on the test data (30%)

of the dataset. We also used the Testing-Only Dataset to observe the performance of the NN model on a separate set of sentences that were not used in training the model.

## 4. Results of the Experiment

We report here the results of the trained model, and we discuss them in Section 5.

### *4.1. The Training-Testing Dataset*

Table 2 shows the confusion matrix for the predictive model, tested against the test data of the Training-Testing Dataset. The dataset contains 487 Finding Sentences, of which 30% (147) were used for test data. Model accuracy = 88.4 %. For predictions of positive polarity, precision = 0.90, recall = 0.83, and F1 = 0.86. For predictions of negative polarity, precision = 0.88, recall =0.93, and F1 = 0.90. Out of 147 Finding Sentences, the model incorrectly predicted positive polarity for 6 sentences, and it incorrectly predicted negative polarity for 11 sentences. A 10-fold cross-validation confirmed a mean accuracy of 89%, with a standard deviation less than 1%.

**Table 2.** Confusion matrix for the predictive model on the test data of the Training-Testing Dataset.

| | **Actual Positive** | **Actual Negative** | **Totals** |
|---|---|---|---|
| **Predicted Positive** | 52 | 6 | 58 |
| **Predicted Negative** | 11 | 78 | 89 |
| **Totals** | 63 | 84 | 147 |

### *4.2. The Testing-Only Dataset*

Table 3 shows the confusion matrix for the predictive model, tested against the sentence data of the Testing-Only Dataset. The dataset contains 136 Finding Sentences. Model accuracy = 87.5 %. For predictions of positive polarity, precision = 0.96, recall = 0.83, and F1 = 0.89. For predictions of negative polarity, precision = 0.78, recall =0.94, and F1 = 0.85. Out of 136 Finding Sentences, the model incorrectly predicted positive polarity for 3 sentences, but it incorrectly predicted negative polarity for 14 sentences. We examine typical error sentences in Section 5.2.

**Table 3.** Confusion matrix for the predictive model on the Finding Sentences of the Testing-Only Dataset.

| | **Actual Positive** | **Actual Negative** | **Totals** |
|---|---|---|---|
| **Predicted Positive** | 69 | 3 | 72 |
| **Predicted Negative** | 14 | 50 | 64 |
| **Totals** | 83 | 53 | 136 |

## 5. Analysis and Discussion

### *5.1. Hypothesis I: Adequate Model Performance for Some Use Cases*

Our results support Hypothesis I – that language used to state affirmative and negative findings of fact is sufficiently regular that we can train ML models, on a small amount of labeled data, to have adequate performance for at least some important use cases.

We trained the baseline model on only 340 labeled Finding Sentences (70% of 487 Finding Sentences), including the 10% used for ML validation. Despite the simplicity of the model and the paucity of the training data, the results were encouraging and robust. Overall accuracy was 87.5% in the Testing-Only Dataset and 88.4% in the test data of the Training-Testing Dataset (mean of 89% in a 10-fold cross-validation). These consistent results suggest that the model was not overfit to the training data, and that the model will perform well on new BVA data.

More importantly, we believe that even this model performed adequately for at least some valuable use cases. For example, if the use case retrieves similar cases and findings in order to provide the user with examples illustrating how evidence has been successfully argued in similar circumstances, then a precision on the order of 0.78-0.96 for Finding Sentences might well be sufficient. By contrast, if the use case requires an accurate calculation of positive/negative Finding Sentences, in order to estimate a probability of argument success in a new case, then the precision and recall of the baseline model would probably not be adequate. Nevertheless, our results support Hypothesis I, and provide a basis for optimism in our ability to increase access to justice in many governmental claims processes.

### *5.2. Hypothesis II: Effect of Sentence Complexity on Model Performance*

An examination of the predictive errors of the NN model provides no support for Hypothesis II. Sentences with erroneous polarity predictions ranged from extremely simple to very complex.

**Errors in Predicting Positive Polarity.** In the Testing-Only Dataset, only three positive predictions were in error (Table 3). One sentence was relatively simple in structure: "*An in-service stressor sufficient to cause PTSD has not been verified or corroborated by evidence of record*". Clearly the token "*not*" was not sufficient to cause the model to predict negative polarity, although it may have affected the calculated probability (probability = 0.69) that this sentence has positive polarity. The other two sentences with erroneous positive predictions also contained negation words and were more complex grammatically (containing dependent clauses).

To shed additional light on the effect (if any) of sentence structure on the model's errors in predicting positive polarity, we generated model predictions for all 487 sentences in the Training-Testing Dataset, and we examined the resulting 8 errors for positive predictions. These sentences ranged from very simple to very complex, and they did not support Hypothesis II. (We acknowledge that some of these 8 sentences may have been included in the training data for the model, but we are here examining these sentences only to shed light on the sentence structures where errors were predicted, not to evaluate predictive results.)

**Errors in Predicting Negative Polarity.** In the Testing-Only Dataset, there were 14 such sentences (Table 3). The structural complexity of these sentences had a wide

range. For example, this sentence erroneously predicted to be negative contains no obvious negative cues and is structurally simple: "*Service connection for PTSD is warranted*". Other sentences were more complex, but also contained no obvious negative cues – e.g., "*The Board notes that competent medical evidence of record, which includes a positive nexus opinion from a private psychiatrist, reveals that the Veteran's psychiatric conditions are in fact proximately caused/aggravated by the Veteran's, now, service-connected back condition.*". An examination of errors predicted in the entire Training-Testing Dataset showed examples consistent with those in the Testing-Only Dataset.

## 6. Prior Related Work

Our experiment involved automatically identifying the polarity of a Finding Sentence, based on the wording of the sentence. Thus, prior work is directly related if it experimented with automatically assigning argument polarity to legal sentences, and it is indirectly related if it automatically assigned sentiment polarity to other types of text. Our work does not duplicate any published work, to the best of our knowledge.

The general problem space of classifying argumentative text for sentiment polarity has received considerable attention. [5, 6] However, it has not often been studied with the type of legal text discussed in this paper. Researchers have used the sentiment polarity of text spans from the 2016 US presidential debates, in order to calculate the divisiveness of an issue [7]. [8] used textual entailment and semantic textual similarity on a corpus of comment-argument pairs derived from two debate websites, in order to predict the polarity (attack, support) of the comments toward the arguments. [9] used comments on news reports about a controversial medical study, in order to predict the stance of the comments toward the target study (positive/negative). [10] used sentiment polarity of text portions from general webpages that play roles in the Toulmin argument model, in order to generate arguments on debate claims (e.g., "*the death penalty should not be banned*"). [11] used a knowledge-based argument mining approach, based on question-answering techniques, on a set of 21 texts containing arguments on 3 issues, in order to assign polarity to correctly recognized arguments.

Other studies have explored using syntactic analysis of sentence structure to aid in identifying semantic types with a connection to polarity. [12] used automatically learned syntax features from dependency trees of sentences in a biomedical dataset, in order to predict the scope of negations. [13] used the semantic orientation of the words in opinion sentences of newswire articles, as well as part-of-speech tagging, in order to predict sentence polarity (positive, negative, neutral). [14] used a methodology on pairs of sequential sentences from U.S. criminal court opinions that were considered to be discussing the same topic, in order to predict whether there was a "shift-in-view" from the first sentence to the second. In developing their methodology, they investigated whether a difference in sentiment polarity between subordinate clauses of the two sentences would indicate a shift in view, but the precision obtained was very low (0.382).

Bar-Haim et al. distinguished among (A) the positive/negative sentiment of a claim (asserted sentence) toward its sentiment target, (B) the consistent/contrastive relation between two sentiment targets, and (C) the pro/con stance of a claim toward its topic. [15] For example, the claim "*people feel greater dignity when choosing their head of state*" has positive sentiment toward its sentiment target ("*choosing their head of state"*), but that sentiment target is inconsistent with the sentiment target of the topic ("*the*

*monarchy*"). Because the topic itself expresses a negative sentiment toward its own sentiment target, the claim has a "con" stance toward the topic. The dataset of [15] consisted of 2,394 claims in Wikipedia articles for 55 topics. For reasons discussed in [1, 4], the present work on Finding Sentences in BVA decisions is different than that of [15] on a number of points. The logical relations among legal issues and findings of fact replace the consistent/contrastive relations between the sentiment targets of topics and claims, and those logical relations are provided by the legal rules. The closest analogy for the sentence polarity studied in the current paper is the sentiment expressed in a sentence toward its sentiment target. For example, if a legal issue to be proved is that the Veteran has a present disability, then the Finding Sentence "*the Veteran does not have a present disability*" has a negative polarity on that issue. Thus, predicting Finding-Sentence polarity should not require the complexities of predicting stance in [15]. Nevertheless, in future work we may find that the techniques of [15] for automatic production of a sentiment lexicon and for including contextual features will also enhance predictive modeling for Finding-Sentence polarity.

Branting et al. investigated two approaches to explainable outcome prediction for decisions of the World Intellectual Property Organization (WIPO). [16, 17, 18] Their second approach involved the possibility of automatically projecting sentence polarity from a small subset of manually labeled finding sentences (25 WIPO decisions) onto sentences of a much larger corpus (16,024 WIPO decisions, 2.64 million sentences). The authors used word and sentence embeddings to project complex sentence tags (labels) that encoded sentence polarity onto semantically similar sentences. However, this approach proved unsatisfactory for predicting sentence polarity due to the lack of separation in vector space between two sentences that have identical wording except one includes the word "not". The authors concluded: "We hypothesize that separate negation-analysis is needed to handle such sentence pairs but … we defer development of this capability to future work." [16] Although those authors provided such qualitative insights, they did not report separate performance measures on the success of projecting the polarity attribute from the manually annotated subset to the larger corpus. Our paper discusses results for predicting sentence polarity, and it tests an hypothesis about a possible cause of error. Moreover, the WIPO decisions studied in [16] are more highly structured than the BVA decisions studied in this paper, and the vast majority of Finding Sentences in BVA decisions occur in a general discussion section of the decision.

## 7. Conclusion and Future Work

Our ultimate goal is to map findings of fact (stated within Finding Sentences) to their respective legal issues or legal-rule requirements, assemble the evidence and reasoning (arguments) relevant to each finding, and infer the outcome of the case using the logic of the legal-rule tree. Using the model reported here as a baseline, we plan to develop and test more complicated ML models for predicting Finding Sentence polarity, using (e.g.) word embeddings, sentence embeddings, and attention. [16]

In order to extract the needed information from a complex Finding Sentence, it will probably be necessary to parse such a sentence into clauses [12, 14], at least one of which states a finding of fact, and to predict the polarity of each finding clause separately. Such parsing may also be necessary in order to map individual findings to their corresponding legal issues. Such analysis may also help improve the performance of predictive models for identifying Finding Sentences [1].

## References


[1] Walker VR, Pillaipakkamnatt K, Davidson AM, Linares M, Pesce DJ. Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning. Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019); 2019 June 21; Montreal, QC, Canada. Ceur-ws.org; Vol-2385/paper 1.

[2] Allen MP. Significant Developments in Veterans Law (2004-2006) and What They Reveal about the U.S. Court of Appeals for Veterans Claims and the U.S. Court of Appeals for the Federal Circuit. University of Michigan Journal of Law Reform; 2007; 40: 483-568.

[3] Moshiashwili VH. The Downfall of Auer Deference: Veterans Law at the Federal Circuit in 2014. American University Law Review; 2015; 64: 1007-1087.

[4] Walker VR, Hemendinger A, Okpara N, Ahmed T. Semantic Types for Decomposing Evidence Assessment in Decisions on Veterans' Disability Claims for PTSD. Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2017); 2017 June 16; London, UK. Ceur-ws.org; Vol-2143/paper 8.

[5] Lawrence J, Reed C. Argument Mining: A Survey. Computational Linguistics; 2019; 45: (4) p. 765-818.

[6] Cabrio E, Villata S. Five Years of Argument Mining: a Data-driven Analysis. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18); 2018 July 13-19; Stockholm, Sweden. IJCAI; p. 5427-5433.

[7] Lawrence J, Reed C. Using Complex Argumentative Interactions to Reconstruct the Argumentative Structure of Large-Scale Debates. Proceedings of the 4th Workshop on Argument Mining; 2017 September 8; Copenhagen, Denmark. ACL; 2017; p. 108-117.

[8] Boltuzic F, Snajder J. Back up your Stance: Recognizing Arguments in Online Discussions. Proceedings of the First Workshop on Argumentation Mining; 2014 June 26; Baltimore, Maryland, USA. ACL; 2014; p. 49-58.

[9] Sobhani P, Inkpen D, Matwin S. From Argument Mining to Stance Classification. Proceedings of the 2nd Workshop on Argumentation Mining; 2015 June 4; Denver, Colorado, USA. ACL; 2015; p. 67-77.

[10] Reisert P, Inoue N, Okazaki N, Inui K. A Computational Approach for Generating Toulmin Model Argumentation. Proceedings of the 2nd Workshop on Argumentation Mining; 2015 June 4; Denver, Colorado, USA. ACL; 2015; p. 45-55.

[11] Saint-Dizier P. Using Question-Answering Techniques to Implement a Knowledge-Driven Argument Mining Approach. Proceedings of the 4th Workshop on Argument Mining; 2017 September 8; Copenhagen, Denmark. ACL; 2017; p. 85-90.

[12] Ren Y, Fei H, Peng Q. Detecting the Scope of Negation and Speculation in Biomedical Texts by Using Recursive Neural Network. Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018. IEEE; 2018; p. 739-742.

[13] Yu H, Hatzivassiloglou V. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003); 2003 July 11-12; Sapporo, Japan. EMNLP, paper W03-1017.

[14] Ratnayaka G, Rupasinghe T, de Silva N, Gamage VS, Perera AS. Shift-of-Perspective Identification within Legal Cases. Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019); 2019 June 21; Montreal, QC, Canada. Ceur-ws.org; Vol-2385/paper 3.

[15] Bar-Haim R, Edelstein L, Jochim C, Slonim N. Improving Claim Stance Classification with Lexical Knowledge Expansion and Context Utilization. Proceedings of the 4th Workshop on Argument Mining; 2017 September 8; Copenhagen, Denmark. ACL; 2017; p. 32-38.

[16] Branting LK, Pfeifer C, Brown B, Ferro L, Aberdeen J, Weiss B, Pfaff M, Liao B. Scalable and explainable legal prediction. Artificial Intelligence and Law; 2020. https://doi.org/10.1007/s10506-020-09273-1.

[17] Branting K, Weiss B, Brown B, Pfeifer C, Chakraborty A, Ferro L, Pfaff M, Yeh A. Semi-Supervised Methods for Explainable Legal Prediction. Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL 2019); 2019 June 17-21; Montreal, QC, CA. ACM; New York, NY, USA; 2019; p. 22-31.

[18] Ferro L, Aberdeen J, Branting K, Pfeifer C, Yeh A, Chakraborty A. Scalable Methods for Annotating Legal-Decision Corpora. Proceedings of the Natural Language Processing Workshop 2019; 2019 June 7; Minneapolis, Minnesota. ACL; 2019. p. 12-20.