Automatic Identification and Empirical Analysis of Legally Relevant Factors

Morgan Gray
Intelligent Systems Program,
University of Pittsburgh
Pittsburgh, USA
mag454@pitt.edu

Jaromir Savelka
School of Computer Science,
Carnegie Mellon University
Pittsburgh, USA
jsavelka@andrew.cmu.edu

Wesley Oliver
Duquesne University,
Thomas R. Kline School of Law
Pittsburgh, USA
oliverw@duq.edu

Kevin Ashley
Intelligent Systems Program,
University of Pittsburgh
Pittsburgh, USA
ashley@pitt.edu

ABSTRACT

This research addresses how to automatically identify certain factors in the texts of legal decisions and analyze their role in courts' decisions. It focuses on drug interdiction auto stop cases in which courts decide whether police officers have reasonable suspicion to detain a motorist. It illustrates how the methods to identify factors automatically can support empirical legal research in the domain and how machine learning methods of different accuracy and interpretability can be harnessed to explain case outcomes in terms legal professionals can understand.

CCS CONCEPTS

• **Applied computing** → **Law**; Annotation; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

text classification, factors, empirical legal analysis, legal text analysis, machine learning

ACM Reference Format:
Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. 2023. Automatic Identification and Empirical Analysis of Legally Relevant Factors. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3594536.3595157

1 INTRODUCTION

Traffic stops are a frequent police-citizen encounter across the United States of America. These common occurrences sometimes result in more than a traffic ticket. Police officers regularly attempt to interdict the trafficking of drugs during routine traffic stops. When officers observe facts that lead them to suspect drug trafficking, they are permitted to detain until a trained drug dog can confirm or dispel the suspicion. If the officer's suspicion is not reasonable, however, the detention may violate the motorist's constitutional rights, an issue the motorist may raise as a defense if prosecuted. In assessing the reasonableness of police officers' suspicion, courts consider a variety of factors.

Below, we present the results of two experiments that take a detailed look into the identification of these factors and how they can be used for empirical legal analysis. Overall, we present evidence that these factors can be identified with a high degree of accuracy using state-of-the-art transformer language models. Furthermore, we apply a variety of machine learning techniques to evaluate the soundness of our list of factors of suspicion and illuminate their role in courts' decisions. By combining the outputs of these models, whose levels of interpretability vary, we illustrate how a system could explain outcomes of cases in terms that legal professionals can understand.

2 RELATED WORK

According to Rempell [21, p. 2], "a factor is a consideration a decision maker must or may take into account to determine an outcome." In law "factors are a foundational and ubiquitous concept . . . "[21, p. 3]. "[T]hey can be prescribed in a statute or regulation, or created by courts," and play a role in diverse areas of law, including assessing spousal support, determining violations of the right to a speedy trial, determining consumer confusion as to the source of goods in trademark infringement, determining works made for hire and copyright fair use [21, p. 2f], [4, p. 1584f], [5, 6] and others.

In the AI and Law field, factors have been defined as stereotypical patterns of facts that tend to strengthen or weaken a plaintiff's argument in favor of a legal claim [2]. Factors have been regularly employed in computational models of case-based argument to represent relevant case facts in a generalized way [1, 7]. Programs that model argumentation with legal rules, cases, and underlying values have also employed factors [11, 12, 14]. Such systems could contribute more effectively to legal practice if a program could automatically identify factors in textual description of facts in case opinions or problem scenarios [2].

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595157

Previous research in AI and Law has made some progress in automatically identifying factors in opinion texts. Wyner and Peters [26] developed an annotation pipeline to extract information related to factors from trade secret legal opinions. Brüninghaus [3] applied machine learning algorithms (C4.5, Naïve Bayes and a k-nearest neighbor approach) to identify factor-related sentences in case summaries prepared by law students. In supervised learning, Falakmasir [13] trained a Support Vector Machine to classify trade secret misappropriation opinions by applicable factors. In [16], the authors employed rules to automatically extract factor values from divorce cases, augmented them using word embeddings, employed them to predict outcomes and explained the outcomes in terms of induced rules. Branting and colleagues trained a machine learning program called SCALE (semi-supervised case annotation for legal explanations) to identify factual finding tags comprising pairs of issues in WIPO (World Intellectual Property Organization) domain name dispute cases and associated factors commonly observed in such cases [8]. For example, the PriorBizUse factor (i.e., "Bona fide business use of Domain Name or demonstrable preparations to do so, prior to notice of the dispute") is related to "No Rights or Legitimate Interests" (NRLI), a required element of a claim in a WIPO case. By automatically identifying factual finding tags in case texts, SCALE predicted case outcomes. In future work, the team expected that the issue and factor labels could help to explain its outcome predictions in terms of reasons that legal professionals would understand [8].

In our approach, supervised machine learning is applied to automatically identify factors of suspicion in drug interdiction auto stop cases, arguably a more challenging task than identifying factors in the divorce cases illustrated in [16] or the WIPO domain name cases of [8]. Auto stop cases written by state and federal judges across the country are more stylistically diverse than WIPO domain name arbitration cases written by arbitrators. The fact situations encountered in auto stop cases are also likely to be more factually diverse than in the WIPO cases and unlikely to be amenable to the rule-based extraction approach of [16].

Unlike [8] we apply a transformer language model, RoBERTa [18], which has been pretrained on an extensive text corpus. We then fine tune the model in applying it on a training set of auto stop cases. In [15], the researchers' multi-label approach to automatically classify factors of suspicion in auto stop cases achieved an average f-1 score of 0.63. They noted that the classifier struggled with classes having low numbers of cases, specifically those with a test sample size of $n<11$. As explained below, in this work we attempt to address that problem by employing a single-label multi-class approach.

We explore using a pipeline similar to the SCALE project that employs factors to explain case outcomes. This is distinct from prior work (e.g., [10]) that sought to predict outcomes from the *full text* of legal decisions. As Branting points out, without factual finding tags, "such systems have very limited inherent explanatory capability." [8]. Unlike the SCALE project, in explaining decided cases, we employ a combination of interpretable ML models such as decision trees, which are known to be intuitively understandable [17], and case-based techniques to illustrate positive and negative examples.

If successful, our project could substantially improve empirical legal studies of factors like those in [4–6, 25]. For example, Beebe investigated 331 federal district court opinions "from 2000-2004 that made substantial use of a multi-factor test for the likelihood of consumer confusion" [4, p. 1584] as to the source of goods in trademark infringement cases. He manually classified the opinions recording, for each factor of the multi-factor test, whether the factor was "found to favor [a/no] likelihood of confusion or otherwise not to favor [no/a] likelihood of confusion." He then applied simple classification trees to 192 opinions involving preliminary injunctions or bench trials [4, p. 1603]. Based on the classification trees, Beebe concluded that "judges determine the test outcome based on a limited number of core factors and then adjust the rest of the factor outcomes to accord with that result." [4, p. 1587]. Shao, et al. applied decision trees with factors in child custody law to identify the three most significant factors [25]. Rissland and Friedman [22] applied decision trees using factor tests to model the state of the law concerning good faith in bankruptcy cases over time. They developed metrics to characterize the degree of change in the decision tree structures and identify changes in the related legal concepts.

Our results suggest that, instead of manually identifying factors in case texts, empirical legal studies like [4], [25] and [22] could employ text analytics to automatically identify factors in much larger numbers of cases, improving their machine learning models of case outcomes.

3 DATA

The experiments below employ the same annotated corpus of auto stop cases introduced in [15]. This corpus of data contains a collection of legal opinions on point to the issue of whether a police officer has suspicion to detain a motorist on the grounds of reasonable suspicion. Generally, United States jurisprudence prohibits the prolonged detention of motorists, beyond what a traffic stop would normally require, absent a showing of suspicion. The corpus has been annotated to identify factors that are considered by officers in making the determination of whether suspicion is present. A court's conclusion as to whether suspicion was found was also annotated. In prior experiments we employed (and here employ) the corpus of 211 state and federal auto stop cases, of which the courts found reasonable suspicion was present in 63% and not present in 37%. The same factors/types have also been employed here. The corpus was annotated using the Gloss annotation environment developed at the University of Pittsburgh [23]. As shown in Table 1, there are 20 substantive factors of suspicion, divided into six categories. The sixth category contains a type for annotating any "other" factors and two outcome types to indicate if the court found that reasonable suspicion was or was not present.[1]

In [15], the researchers employed a multi-label full-sentence annotation scheme, where each sentence [24] was assigned all factors that applied. If multiple factors were expressed in the sentence, the sentence was assigned multiple labels. The annotation guidelines and process reported in [15] achieved a moderate level of inter-annotator agreement (0.57). As explained below, in this work we have employed a single-label approach, in which those sentences

[1]The type "6U Possibly Off Point" is not included in this representation because it is simply an administrative label used by annotators to indicate the case may be irrelevant to the auto stop project.

with multiple labels have been broken into parts, each assigned a corresponding single label.

Table 1: Factors of Suspicion.

1 Occupant Appearance or Behavior	2 Occupant Status
1A Furtive Movement	2E Motorist License
1B Physical Appearance of Nervousness	2F Driver Status
1C Nervous Behavior	2G Refused Consent
1D Suspicious or Inconsistent Answers	2H Legal Indications of Drug Use
	2I Motorist's Appearance Related to Drug Use
3 Travel Plans	**4 Vehicle**
3J Possible Drug Route	4L Expensive Vehicle
3K Unusual Travel Plans	4M Vehicle License Plate or Registration
	4N Unusual Vehicle Ownership
5 Vehicle Status	**6 Other Annotation Labels**
5O Indicia of Hard Travel	6T Other
5P Masking Agent	6V Suspicion Found? - No
5Q Vehicle Contents Suggest Drugs	6W Suspicion Found? - Yes
5R Suspicious Communication Device	
5S Suspicious Storage	

4 EXPERIMENTS

Our two experiments aim to assess the feasibility of a pipeline to automatically identify factors in auto stop cases and, based on those factors, to explain case decisions about whether reasonable suspicion is satisfied. The first experiment aims to improve the performance, reported in [15], of classifying factors of suspicion. The second experiment aims to use interpretable machine learning models to explain the outcome of auto stop cases concerning whether reasonable suspicion is present in a case based on the automatically identified factors.

4.1 Classification of Factors

In the first experiment, we employed a multi-class approach to automatically identify factors in case texts. The approach requires that each individual data point bears a single label. As a result, sentences that bore multiple labels were converted to single-label, sub-sentence annotations.

For example, the sentence, "[The driver] also lied to the trooper regarding his criminal history," would be broken into two parts. The first, "[The driver] also lied to the trooper," captures the part of the sentence describing a suspicious answer. The second, "regarding his criminal history," describes the driver's criminal history.

The use of a single-label approach distinguishes this work from [15], where the researchers had employed a multi-label approach. If a sentence was an instance of two or more factors, they assigned it multiple labels. As shown in figure 1, the vast majority of labeled sentences bore only a single label. A smaller fraction bore two labels, and minuscule amounts bore three labels or more.

We hypothesized that converting sentences with multiple labels into sub-sentences with single labels would increase the number of training instances for each label and decrease the possibility of confusing the classification model. This is supported by the finding that the amounts of *n*-label sentences containing more than one label contained a considerable numbers of unique sets of labels. Specifically, 25% of 2-label, 64% of 3-label, 85% of 4-label, 80% of 5-label 40% of 7-label sentences were unique label combinations. Given the high numbers of unique label combinations, we were concerned about how well the classifier could cope with the complexity.

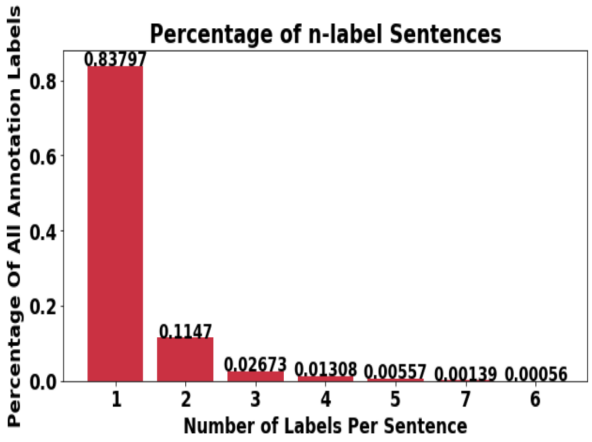

Figure 1: This figure shows the percentage of *n*-label sentences in the data set prior to the data set being re-annotated to a single-label approach. As shown in the chart, the vast majority of sentences bear a single label (84%). The other 16% of the sentences bear two or more labels, with a vast majority of this portion bearing only two.

In order to utilize the multi-class approach, we developed a set of guidelines for systematically dividing the sentences with two or more labels into parts and assigning a single label to each part. The guidelines were drafted with the intent to be implemented in further large scale annotation tasks. Following the guidelines three legal experts converted the multi-label sentences to single labels. Two experts assigned the labels to each of the sub-sentence annotations. To ensure compliance with the annotation guidelines, a third expert then reviewed the assignments.

Using these annotations, we fine-tuned a roBERTa base model [19], using a total of 48,390 sentences representing no-type and 4,601 representing a single type. Training occurred over 15 epochs, with evaluation occurring at each epoch.

Table 2: Sample of Data Frame

Case Name	Features	Conclusion
usa_v_powell	1, 1, 0, 0, 1, ...0, 0, 0, 0, 0	1
state_v_haar	1, 0, 1, 0, 1 ...1, 0, 1, 0, 0	0
usa_v_smith	0, 0, 1, 0, 1... 0, 0, 0, 0, 0	0
usa_v_johnson	1, 1, 0, 1, 1, ...1, 0, 0, 0, 0	1
usa_v_walton	0, 1, 0, 1, 1, ...0, 0, 0, 0, 1	1

4.2 Explaining Outcomes

The second experiment tested how a machine learning model could explain cases' outcomes based on the factors of suspicion. We use this experiment to gauge the performance of a machine learning model under ideal conditions using annotated data.

The inputs were cases, each represented by the subset of the 20 substantive factors of suspicion that the automated classifier had identified in the case. The two possible outcome labels were "6V

Suspicion Found? - No" and "6W Suspicion Found? - Yes". These indicated the model's factor-based prediction regarding the court's conclusion as to whether or not the police officer had reasonable suspicion that drugs were present in the automobile. We show that it is possible to reliably predict the outcome of the cases given the factors, thereby providing strong evidence of a sound representation of the selected domain in terms of the factors we identified.

As illustrated in Table 2, the cases, factors, and outcomes were structured as a data frame, with each row representing a single case. The 20 factors were coded as a vector of binary features with each factor in its own column.[2] If a feature was present in a case it was coded as 1; if not it was coded as 0. The outcomes were coded as 1 if reasonable suspicion was present and 0 if it was not. The determination on the presence of a factor and the outcome was based on the gold standard annotations. The data frame illustrated in Table 2 has dimensions of 206 x 22.

We checked for correlation between the inputs, but detected no issues. Because the features are binary variables, a tetrachoric correlation matrix was used to assess correlation between the inputs. As shown in Figure 2 there were no features with a negative correlation of less than -0.85 or a positive correlation greater than 0.85.

We trained eight different models to predict whether reasonable suspicion was present in a case based on the factors we identified. The data was divided into training and testing data using an 80:20 split. The same split was to train and test each model. Training was performed using cross-validation with 10-folds, repeating the procedure 3 times. The best parameters were found using a grid search. Model predictions were assessed based on their accuracy on the test set. The models we chose ranged in interpretability from easily interpretable Decision Trees to difficult/impossible to interpret methods like Neural Networks. The importance of model interpretability is discussed below and was an important consideration in these experiments. Where possible, the importance of each variable in a model was calculated. Variable importance was calculated using the Caret library available in R.[3] Essentially, this calculation estimates what variable or variables were most important to the accuracy of the model in an attempt to reduce the lack of interpretability of a model [9].

To get an idea of whether the models are capable of discriminating between the classes we compare the models to two naive baselines. The first baseline predicts based on the frequency of the outcome labels in the data. The second baseline model randomly selects an outcome label with uniform probability for each label.

5 RESULTS AND DISCUSSION

5.1 Classification Task

The results of the first experiment's classification task are shown in the classification report in Table 3. There was a noticeable improvement from the results reported in [15]. The overall average f-1 score increased from 0.63 to 0.83, an increase of 0.20, with individual categories increasing an average of 0.18.

[2]The ordering of the columns associated with the features is: 1B, 4N, 3J, 2G, 2H, 1C, 3K, 4M, 1D, 5P, 6T, 5Q, 5S, 2E, 1A, 5R, 2F, 5O, 2I

[3]Specifics about how the variable importance for each model is calculated can be found here: https://topepo.github.io/caret/variable-importance.html

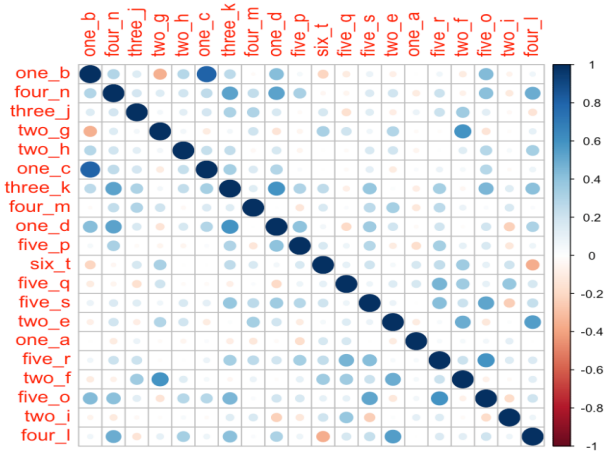

Figure 2: This figure shows the correlation between the inputs. The color blue indicates that there is a positive correlation between the inputs. The larger and more blue the circle, the more positively correlated the features are. The red circles indicate negative correlation. The larger and more red the circle, the more negatively the features are correlated.

In general, moving from the multi-label approach to a single-label multi-class approach significantly improved the results. In addition, in [15] it was noted that classes with low support were predicted with poor quality; classes with *n<11* samples performed worse. The current experiment shows that where the number of samples in the test set is 11 or greater, f-1 scores did not drop below 0.70 and remained consistently above 0.70.[4] In addition, as a result of the single-label approach, three factors of suspicion now have 11 or more supporting instances (4M, 5R, and 5O).

5.1.1 Classification Task Error Analysis. The factor classification model made 243 errors out of 9,678 sentences in the test set. The confusion matrix in Figure 3 plots focuses on errors. The y-axis represents the true labels assigned by the annotations; the x-axis represents the predicted label given by the model. It shows that the actual label "no_type" was frequently "wrongly" predicted by the model, which assigned some factor of suspicion. Of the errors made by the model, 43 of 243, or roughly 18%, involved the classifier confusing one factor for another. In the overwhelming majority of errors, 200/243 or 82%, the model assigned a factor label to a sentence which the human annotators had left unlabeled. In effect, the human-assigned label was no_type. This is shown in Figure 3 on the line of the y-axis associated with no_type, i.e., where the model predicted that a factor was not described in a sentence. This shows that a significant number of no_type labels were predicted as belonging on to a type. However, on the line of the x-axis associated with n_t (the abbreviation for no type) we also see that many sentences belonging to a type were also predicted as not belonging to a type.

[4]The exception to this in Table 3 is for the factor 2F Driver Status, which has an f-1 score of 1.00, however, the support for this category is only a single case and therefore at least moderately suspect.

Figure 3: This figure shows the confusion matrix of the errors made by the classifier. The y-axis are represents the actual label and the x-axis represents the predicted label. The key to the right of the plot shows the colors assigned to each block of the matrix based on the number of errors between actual and predicted labels of each type. The bottom row of the x-axis shows where the actual label was no_type and the model predicted another label. The last column (farthest to the right) on the y-axis shows where the model predicted no_type, despite being labelled as belonging to a type.

Table 3: Classification Report: Multi-Class

	P	R	F1	n
no_type	0.99	0.99	0.99	8764
3J Possible Drug Route	0.90	0.94	0.92	47
5S Suspicious Storage	0.94	0.91	0.92	32
1B Physical Appearance of Nervousness	0.88	0.92	0.90	126
1D Suspicious or Inconsistent Answers	0.86	0.87	0.87	127
4N Unusual Vehicle Ownership	0.87	0.88	0.87	75
5P Masking Agent	0.86	0.86	0.86	28
2E Motorist License or Identification	0.94	0.79	0.86	19
6T Other	0.86	0.86	0.86	14
5O Indicia of Hard Travel	0.90	0.82	0.86	11
5R Suspicious Communication Device	0.88	0.82	0.85	17
6W Suspicion Found? - Yes	0.82	0.85	0.83	53
2G Refused Consent	0.78	0.88	0.83	49
6V Suspicion Found? - No	0.83	0.83	0.83	30
3K Unusual Travel Plans	0.88	0.76	0.82	38
2H Legal Indications of Drug Use	0.83	0.79	0.81	57
1C Nervous Behavior	0.78	0.82	0.80	96
5Q Vehicle Contents Suggest Drugs	0.77	0.73	0.75	41
4M Vehicle License Plate or Registration	0.64	0.90	0.75	10
2I Appearance Related to Drug Use or Sale	0.75	0.71	0.73	17
1A Furtive Movement	0.79	0.62	0.70	24
4L Expensive Vehicle	0.33	1.00	0.50	1
2F Driver Status	1.00	1.00	1.00	1
accuracy			0.97	9678
macro avg	0.84	0.86	0.84	9678
weighted avg	0.98	0.97	0.97	9678

Upon further investigation, it appears that some of these "errors" supposedly made by the model were actually the model correcting errors occasionally made by manual annotation. Of the 243 "errors", the classifier had correctly assigned a label to a sentence that had been mislabelled as no_type. For example, this sentence was mis-annotated as belonging to no_type:

> Officer Keeler testified that he reviewed the agreement while filling out the citation and noted that the car was not rented to Goss.

The classifier correctly predicted the label as 4N Unusual Vehicle Ownership. Upon investigation of sentences labelled as no_type, where the model predicted the sentence as a type, it appears that the model reliably caught flaws in manual annotation and predicted the correct label.

Since we plan to use factors to explain case outcomes and to use factor-based predictions of case outcomes to validate the factors we identified, it was important to annotate only factors on which the court relied even though this would complicate the machine learning process and lead to false positives. The annotation guidelines direct one to annotate sentences describing "factors the court relies upon in reaching its result." As a result, in certain situations, a sentence in an opinion may appear to be an instance of a factor, but it should not be annotated.

For instance, the court may have employed a sentence to describe a factor in another case that it has cited. For example, a court may say something like

> In *Smith v. State*, the trooper noted that there were rolling papers in the vehicle. In this case, the trooper also noticed rolling papers in the vehicle.

Only the second sentence should be annotated because it discusses a factor in the case at hand. The first sentence looks very much like the second, however, leading the model to make an error.

Alternatively, a sentence may describe something the officer observed after making the determination of reasonable suspicion. Since the court should not rely on this factor in analyzing whether the officer's suspicion was reasonable, the sentence should not be annotated. For example, the classifier labelled the following sentence as an instance of factor 5Q, Vehicle Contents Suggest Drugs.

> The officer then observed rolling papers consistent with those used by drug users as well as tobacco spread around the interior of the car.

Since rolling papers are a type of paraphernalia used to consume drugs, this sentence appears of a type indicating that the vehicle may contain drugs. The officer, however, did not make this observation until after making the determination of suspicion. Legally, such *post hoc* observations are irrelevant.

Another source of classification errors is the fact that even though sentences were annotated at the sub-sentence level, the model needs to make predictions at both the sentence and sub-sentence level. A sentence in the test set may have parts to which different labels should be assigned. Since the classifier lacks a mechanism to break a sentence down into parts and predict their labels, the classifier would be forced to choose a single label for a sentence describing more than one factor. For example, the following sentence should have been assigned two labels, factors 2G, Legal Indications of Drug Use, and 1D, Suspicious or Inconsistent Answer:

> Blake testified that Dion's "drug trafficking history, which he obviously lied about," contributed to Blake's rising suspicions.

The model, however, only predicted one of the labels, factor 2G. A majority of the errors where the classifier mistook one label for another appeared to be due to the sub-sentence annotation scheme. Of all 43 such errors, 30 involved sentences with multiple labels.

5.2 Using Machine Learning to Predict Outcomes and Assess Factors

Table 4 shows a break down of the performance of each model, on gold standard annotations, in predicting case outcomes based on the factors in our list. The best performing models were tied, with the neural network and the random forest models predicting with 0.975 accuracy. All models clearly outperformed the naive baseline models.

These models predicted the outcome of reasonable suspicion with a high degree of accuracy. The results show that the 20 factors of suspicion listed in Table 2 can be used to predict the outcome of suspicion with roughly 98% accuracy. Given the cross-validation training and testing procedure and the accuracy on the test set, likelihood is very low that the models overfit the data.

Further analysis shows which factors are the most important in predicting the outcome. As noted, variable importance metrics can be calculated for many of the models, including all of the best performing models, that is, the models with an accuracy of greater than 0.90.

First, where applicable, the importance of each input for a particular model was determined by calculating its variable importance. As noted by [9] this can help disambiguate a model's behavior by shedding light on what variables were important in reaching a prediction. Two factors are very important to the analysis of reasonable suspicion. The first concerns whether the motorist gave a suspicious answer; the second is whether the motorist was stopped on a route known for drug trafficking or coming to or from a city/area known for drug trafficking. In almost every model, with the exception of the neural network, factor 1D, Suspicious or Inconsistent Answers, was identified in the top three most important variables to the model. It was the most important in the GLM and ElasticNet models, the second most important in the XGBoost model, and the third most important in the random forest and decision tree models. Next, factor 3J, Possible Drug Route, was identified as an important factor in all but one model, random forest. This was the most important factor for the XGBoost model and the second most important factor for the GLM, ElasticNet, the neural network, and the decision tree.

A third factor, 4N Unusual Vehicle Ownership, was the most important feature in the neural network, the second most important feature in the random forest model, and the second most important in the decision tree model. As discussed below, this feature is particularly important to the decision tree and can be used to better understand predictions made by the model.

Analysis also shows the factors that *are not* important. A hallmark of the ElasticNet model is that it "turns off" unimportant features in a model by pushing coefficients to 0. The ElasticNet model that we trained used pure Lasso Regression to push the coefficients for 11 factors of suspicion to 0: 1C Nervous Behavior, 1B Physical Appearance of Nervousness, 2G Refused Consent, 3K Unusual Travel Plans, 4M Vehicle License Plate or Registration, 5Q Vehicle Contents Suggest Drugs, 5S Suspicious Storage, 2E Motorist License or Identification, 1A Furtive Movement, 2F Driver ~~Status~~, and 5O Indicia of Hard Travel.

Some of the factors that were pushed to 0 reflect legal requirements or courts' legal observations. For example, the law prohibits using a motorist's refusal to consent to a vehicle search against the motorist in determining whether reasonable suspicion is present.[5] In addition, some courts state that nervousness has limited utility in determining whether suspicion is reasonable, because many people are nervous when stopped by the police[6].

The "turning off" of other factors, however, does not seem consistent with legal requirements. According to Lasso regression, Factor 5Q, Vehicle Contents Suggest Drugs, was unimportant. This factor encompasses the smell of drugs, drugs in plain view, and paraphernalia in plain view. Normally, the fact that drugs and/or paraphernalia were in plain view in the automobile would determine the presence of reasonable suspicion but would also provide probable cause to search a vehicle [7] Although the dichotomous feature representation shows that the presence of factors alone is a representation of the factors worthy of accuracy at roughly 98%, clearly the binary coding fails to capture important insights like

[5] "Courts generally hold that refusal to consent cannot establish or-according to some courts-even support reasonable suspicion." *State v. Gomez*, 275 P.3d 1073, 1077 (Utah App. 2012).

[6] "We have repeatedly held that nervousness is of limited significance in determining reasonable suspicion and that the government's repetitive reliance on the nervousness of either the driver or passenger as a basis for reasonable suspicion …" *United States v. Fernandez*, 18 F.3d 874, 179 (10th Cir. 1994)

[7] *Walter v. State*, 28 S.W.3d 538 (Tex. Crim. App. 2000)

Table 4: Model Performance

Model Name	Accuracy on Test Set	3 Most Important Variables	Model Parameters
Generalized Linear Model	0.902	1D, 3J, 5P	Binomial Family
Elastic Net	0.902	1D, 3J, 5P	$\alpha = 1$, $\lambda = 0.0258$
Neural Network	0.975	4N, 3J, 5P	Hidden Layers = 5, Decay = 0.1
Random Forest	0.975	5S, 4N, 1D	$m_{try} = 11$
XGBoost	0.951	3J, 1D, 5S	$\eta = 0.05$, Iters = 200, Depth = 5
Decision Tree	0.801	1D, 4N, 3J,	Complexity Parameter = 0.07
kNN	0.83	NA	k=14, Distance = Jaccard
Weighted kNN	0.829	NA	Weight = Optimal
Most Frequent Label Baseline	0.64	NA	NA
Random Label Baseline	0.46	NA	NA

the relationship between reasonable suspicion and probable cause. This anomalous result begs further investigation.

Our tentative conclusions as to the importance or lack of importance of particular factors of suspicion are subject, of course, to the limited number of cases in our corpus, 211 in all. [8] Our ability to draw such conclusions, even tentatively, illustrates, however, the kinds of empirical legal analyses that can be applied to ever larger numbers of cases once factors can be identified automatically.

The results of our two experiments confirm the feasibility of a pipeline to automatically identify the factors we selected to represent auto stop cases and, based on those factors, to predict whether reasonable suspicion is satisfied. The level of accuracy in predicting reasonable suspicion outcomes validates our selection of factors to annotate. Although under the law any factor can be relevant, it appears that certain patterns of commonly recurring facts are important. Presumably, if the fact patterns identified by this research, the level of granularity needed to identify the factors of suspicion, and the grouping of the factors were a poor estimate of the law, the predictive accuracy of these models would be considerably lower.

The high accuracy also validates the annotation scheme that we have developed for identifying the factors. It enables one to identify language describing factors both to train a factor classifier and to predict courts' determinations of reasonable suspicion. This indicates that using automatically identified factors as inputs to a model could be viable. In developing the factor annotation scheme we tried defining the factors at various levels of granularity, ranging from a high of 66 different factors to a low of 14 (12 substantive and 2 conclusion factors). Ultimately, we decided on 23 factors of suspicion, with 21 substantive labels and 2 conclusion labels. The choice represents a compromise across a number of considerations. If the granularity were too high, there would be too much noise in the classes. If too fine, overlapping features of factors may be overlooked, which could also introduce noise. With too many features, models are prone to over-fitting and do not generalize well. Moreover, as the number factors increases, annotation becomes more burdensome, and the need for training data increases due to more classes.

One can compare the predictive accuracy of our models with those of full-text prediction approaches. For example, Medvedeva reported achieving average accuracy of 79% in categorizing judgements according to whether or not the court found a violation of 9 articles of the European Convention on Human Rights. [20]. We note that the 206×22 matrix used to represent the name of each case, the factors of each case, and the outcomes of each case is significantly smaller than other methods. This is a large reduction from the thousands of words that would need to be represented if using a text-based methods. The binary factor representation avoids the need for large structures to represent words such as bag-of-words representations or word embeddings. The relatively small 206×22 matrix is easy to understand, simple to create, and is computationally inexpensive to operate. Furthermore, using such representations in machine learning models reduces training time. This representation does not affect the efficacy of model performance in predicting the reasonable suspicion outcomes. Perhaps more importantly, as discussed below, it enables explaining case outcomes in ways that deep learning models cannot.

6 EXPLAINING CASE OUTCOMES

The models' calculations and estimations can be used to explain case outcomes in complementary ways. We focus particularly on the decision tree and the distance used to calculate the nearest neighbors in the *k*NN model. These models are more interpretable than methods like neural networks, random forests, or XGBoost.

Our decision tree, based on the CART method, splits the features into regions. For each region of the tree, it searches across every value of every feature, assesses the loss at that split, and recursively performs this procedure, choosing the split that gives the lowest error. In other words, a split, in our tree represents where the model has chosen that the presence or absence of a particular factor results in a prediction with the lowest amount of error as compared to the presence or absence of any other factor. Since the tree generated by the model, shown in Figure 4, represents how all predictions were made, the model's prediction for a particular case is readily interpretable.

When classifying an incoming data point, the tree begins at the root node, and proceeds to apply the tests at each node through the rest of the tree until it reaches a terminal node with the predicted classification. In Figure 4 factor 4N, Unusual Vehicle Ownership, is at the root node of the tree. Thus, the model first asks whether factor 4N is present in the input case. If 4N is present, the left branch is followed, reaching a terminal node. The model predicts

[8] Five cases were dropped from this experiment because the court's legal conclusion was unclear.

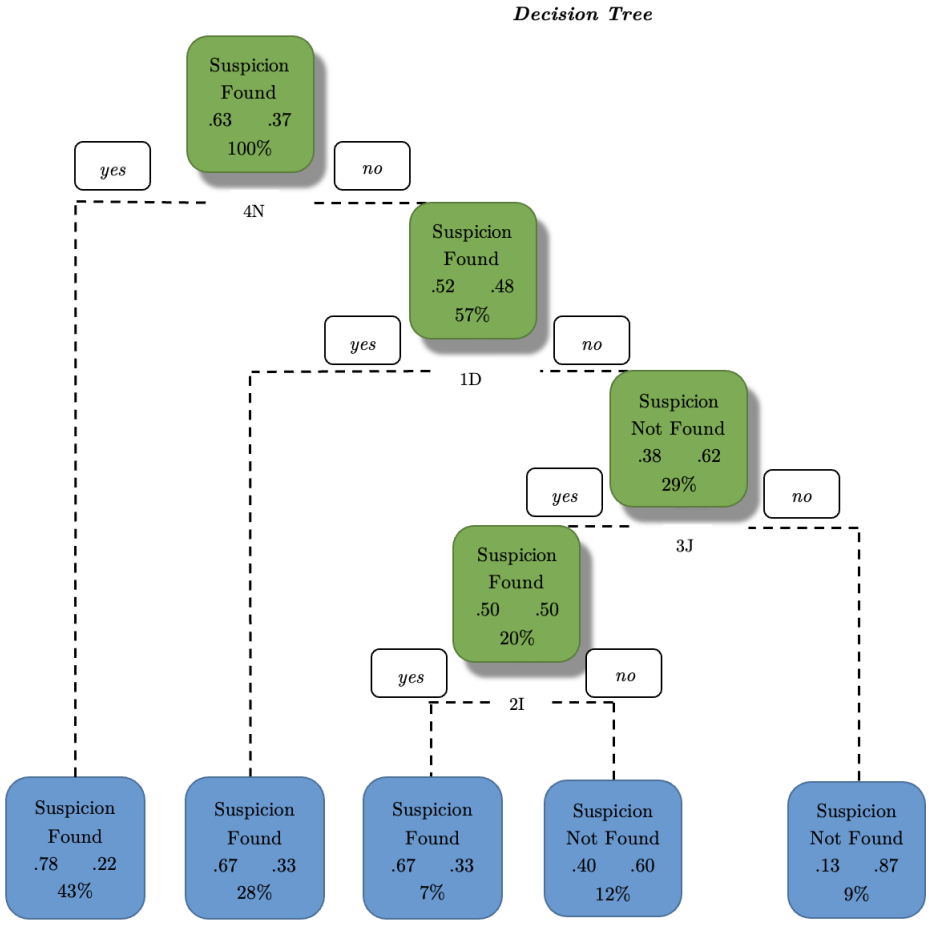

Figure 4: This was the decision tree generated by the training data using a complexity parameter of 0.07. The nodes pictured in green show where the decision tree assesses whether a factor is present. The factors considered by the tree are 4N Unusual Vehicle Ownership, 1D Suspicious or Inconsistent Answers, 3J Possible Drug Route, and 2I Appeareance Related to Drug Use or Sale. Depending on the presence of the factors, which are assessed by the green nodes, the model eventually stops at a terminal node (blue) and will predict the outcome an outcome of Suspicion Found or Suspicion Not Found based on a probability.

that reasonable suspicion is found with a probability of 0.78. If 4N is not present the right branch is followed and the model asks if factor 1D, Suspicious or Inconsistent Answers, is present. If yes, the left branch leads to a terminal node and a prediction of reasonable suspicion found with a probability $p = 0.67$. The decision tree provides both a prediction and an explanation in terms of a rule readily induced from the branches based on the answers to the factor tests.

From the viewpoint of empirical legal studies, the decision tree in figure 4 also provides some significant information about the auto stop legal domain, at least based on the 211 cases in our corpus. The presence of factor 4N, Unusual Vehicle Ownership resolves 43% of the training data. If factor 4N is not present, the presence of factor 1D resolves another 28% of cases. Similar observations apply with respect to factors 3J, Possible Drug Route, and 2I, Motorists Appearance Related to Drug Use. Achieving this kind of insight about a legal domain motivates empirical legal scholars such as Prof. Beebe to identify factors and apply models like decision trees. With the ability to apply text analytics to identify factors automatically, researchers can process many more cases and strengthen their conclusions.

In order to further explain a case outcome, we employ the distance metric of the kNN model to calculate similarity between cases. We trained the kNN model using Jaccard Distance. We focus on the model's metric of dissimilarity. For vectors of dichotomous features, dissimilarity is calculated as

$$\frac{f_{01} + f_{10}}{f_{11} + f_{01} + f_{10}}$$

Let f represent a factor, $case_x$ represent a single observation in the data frame, i.e., a legal opinion represented by dichotomous inputs, and $case_y$ represent a different case represented in the same way. In the formula above, f_{01} represents the scenario where an individual factor was not present in $case_x$ but was present in $case_y$.

The reverse is true for f_{10}, in this situation a factor was present in case $case_x$ but was not in $case_y$. Lastly, f_{11} represents the situation where a factor was present in both $case_x$ and $case_y$. Take for example, the following vectors:

v_0 = [1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
v_1 = [1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

These vectors have a dissimilarity of 0.0, because they are identical. As dissimilarity between the vectors increases, so does the score. Thus, the following vectors have a dissimilarity score of 0.625

v_0 = [1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
v_1 = [1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0]

Measuring dissimilarity between the cases represented as dichotomous vectors is useful because we can compute the distance between cases based on the factors that are and are not present, and use that similarity to identify cases with similar facts and outcomes to support a lawyer's argument. Conversely, lawyers would also be interested in knowing about cases that have similar facts, and different outcomes. Certainly, this high level representation of facts does not replace a lawyer's skill in reading, understanding, comparing, and contrasting cases to formulate legal arguments. Nonetheless, these empirical insights can serve as a useful guide in understanding auto stop cases.

By reporting a combination of predictions from the neural network and random forest models, which are hard to explain, as well as the decision tree, which indicates which factors are important, and the dissimilarity metric between cases, we can better explain case outcomes.

6.0.1 Example Explanation. In this example, we focus on the explanation of a single case using the methods described above. Assume that six factors apply to the focal case: 1B, Physical Appearance of Nervousness, 1C, Nervous Behavior, 1D, Suspicious or Inconsistent Answers, 4N, Unusual Vehicle Ownership, 5O, Indicia of Hard Travel, and 5P, Masking Agent. The corresponding vector representation of the focal case is:

[1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]

Both the random forest and the neural network models predicted an outcome of reasonable suspicion for this case. The neural network predicted that reasonable suspicion was present in this case with a probability 0.99%. The random forest model, predicted reasonable suspicion present with a probability of 0.95%. Both models are highly confident in these factor-based predictions, lending credibility that the factors are a sound representation.

In order to gain some insight about the individual factors, we next refer to the decision tree. An examination of the feature vector shows that factor 4N, Unusual Vehicle Ownership, is present. According to the decision tree, we know that this factor is important to the suspicion analysis and if present, the decision tree, will explain the presence of reasonable suspicion with a probability of 0.78.

Finally, comparing the current fact situation with the most similar cases may lead to a better understanding of on what a fact situation's outcome depends. We can measure the Jaccard dissimilarity between the case of interest and all other cases. The resulting scores are ranked from most similar to dissimilar. Then the value of each score is determined by the following criteria:

Score = 0.0: The case is directly on point.
Score = 0.1-0.3: This case is not directly on point but is similar to the case of interest.
Score = >0.3: This case is likely to be unhelpful.

The top five cases with scores of 0.3 or less are reported, along with a comparison of the factors and the conclusion reached. If there are fewer than five cases with acceptable scores, then only cases meeting the scoring criteria are reported. Applying this procedure to the given case yields:

Rank = 1:
Case Name: United States v. Anguiano
Score: 0.2
Similar Factors: '1B', '4N', '1C', '1D', '5O'
Dissimilar Factors: '6T*', '5P-'
Outcome: Suspicion Found
Rank = 2:
Case Name: State v. Myles
Score: 0.3
Similar Factors: '1B', '4N', '1C', '1D'
Dissimilar Factors: '3K*', '5O-', '5P-'
Outcome: Suspicion Found

These outputs provide the name of the similar case, the similarity score, the factors that were similar between the cases, the factors that were different, and the outcome of the case. Dissimilar factors that are present in the case of interest, but not present in the similar case are represented with a "-"; those that are present in the similar case but not in the case of interest are identified with a "*". In this example, there were five similar cases. The rest of the cases (not shown) had a score of 0.3.

Here, the most similar cases both happen to have findings of reasonable suspicion. Factors 1B, Physical Appearance of Nervousness, 1C, Nervous Behavior, 1D, Suspicious or Inconsistent Answers, and 4N, Unusual Vehicle Ownership, seem to lie at the core of the two similar cases. (Indeed, the decision tree focused on the importance of 4N and 1D.) The presence in the focal case of 5P, Masking Agent, or 5O, Indicia of Hard Travel, does not seem to make a difference to the outcome, nor does the absence of 3K, Unusual Travel Plans or 6T, Other. Nearby cases might well have had the opposite outcome, however, which might reduce one's confidence in a prediction. Such counterexamples can provide information about factors whose presence or absence is significant and result in the opposite outcome.

6.0.2 Prediction with Automatically Identified Factors. In Section 5.1, machine learning was employed to predict and explain case outcomes based on manually-identified factors. Suppose automatically identified factors were employed to predict and explain the outcome of the focal case in the example of the previous subsection.

Table 5: Comparison of gold standard and automatically identified feature vectors

POSITION:	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]
ROBERTA:	[0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
GOLD:	[1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0]

A visual comparison of the gold standard annotated vector shows that the classifier missed three factors located in position 0, 5, and 17. At this point, machine learning models used to predict and explain an outcome would be based on similar, but different subsets of factors. Based on the automatically classified factors, all models would still correctly predict the outcome of suspicion. As observed above, the roBERTa model identified factors 4N, Unusual Vehicle Ownership, and 1D, Suspicious or Inconsistent Answers, which we know are important variables, as determined by the variable importance calculations, and are important in reducing error in the decision tree. The roBERTa model failed to identify factors that happened to have been "turned-off" by regularization (1B, and 1C). Although, there is no necessary link between which factors the roBERTa model will classify and those that will be important in the prediction models, it seems that what we learned in the previous subsection on explaining outcomes still holds.

7 FUTURE WORK

A limitation of the classifier discussed above is that we do not include a mechanism to break up test sentences into multiple parts, therefore, a test sentence that describes multiple-factors, using our current approach will only receive a single label. We hypothesize that by using parsing methods, we may be able to effectively break up sentences. We seek to improve the automatic identification of factors and to identify other models that may have interpretable usefulness.

8 CONCLUSION

Our results provide meaningful evidence that automatic identification of auto-stop factors is feasible. Moving forward, based on the factor identification experiment, we will attempt to increase the amount of training data, improve the text analytic techniques, and apply them to ever larger numbers of auto stop cases. As we increase the number of annotated cases, we will revisit our tentative conclusions that judicial determinations of reasonable suspicion depend on a relatively small subset of commonly used factors and report our methods and results to the empirical legal research community.

Analysis of the machine learning experiments indicates that less interpretable methods to predict outcomes of auto stop cases can achieve high levels of accuracy while less accurate but more interpretable models can shed light on these predictions. We show that by combining these methods, a system can explain case outcomes in terms that legal professionals can understand.

Potentially, one could imagine a system that would allow a human user to enter factors and obtain a prediction of an outcome should the auto-stop search be contested before a court. We leave consideration of this possibility and how to achieve it for future work.

ACKNOWLEDGMENTS

We acknowledge support for this research from a Pitt Momentum Funds Teaming Grant and Scaling Grant and from the Center for Text Analytic Methods in Legal Studies at the University of Pittsburgh.[9]

[9]https://www.law.pitt.edu/center-text-analytic-methods-legal-studies

REFERENCES

[1] Kevin D Ashley. 1990. *Modeling Legal Arguments: Reasoning with Cases and Hypotheticals.* MIT Press.
[2] Kevin D Ashley. 2017. *Artificial Intelligence and Legal Analytics.* Cambridge U. Press.
[3] Kevin D Ashley and Stefanie Brüninghaus. 2009. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law* 17, 2 (2009), 125–165.
[4] Barton Beebe. 2006. An empirical study of the multifactor tests for trademark infringement. *Calif. L. Rev.* 94 (2006), 1581.
[5] Barton Beebe. 2007. An empirical study of US copyright fair use opinions, 1978-2005. *U. Pa. L. Rev.* 156 (2007), 549.
[6] Barton Beebe. 2020. An Empirical Study of US Copyright Fair Use Opinions Updated, 1978-2019. *NYU J. Intell. Prop. & Ent. L.* 10 (2020), 1.
[7] Trevor Bench-Capon. 2017. HYPO's legacy: introduction to the virtual special issue. *Artificial Intelligence and Law* 25, 1 (2017), 205–250.
[8] L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and explainable legal prediction. *Artificial Intelligence and Law* 29, 2 (2021), 213–238.
[9] Leo Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32. https://link.springer.com/article/10.1023/a:1010933404324#citeas
[10] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. *arXiv preprint arXiv:1906.02059* (2019).
[11] Alison Chorley and Trevor Bench-Capon. 2005. AGATHA: Using heuristic search to automate the construction of case law theories. *Artificial Intelligence and Law* 13, 1 (2005), 9–51.
[12] Alison Chorley and Trevor Bench-Capon. 2005. An empirical investigation of reasoning with legal cases through theory construction and application. *AI and Law* 13, 3 (2005), 323–371.
[13] Mohammad Falakmasir and Kevin Ashley. 2017. Utilizing Vector Space Models for Identifying Legal Factors from Text. In *JURIX 2017*, Vol. 302. IOS Press, 183–192.
[14] Matthias Grabmair. 2016. Modeling Purposive Legal Argumentation & Case Outcome Prediction using Argument Schemes in the Value Judgment Formalism. Language: English.
[15] Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. 2022. Toward Automatically Identifying Legally Relevant Factors. In *Legal Knowledge and Information Systems.* IOS Press, 53–62.
[16] Jiajing Li, Guoying Zhang, Longxue Yu, and Tao Meng. 2019. Research and design on cognitive computing framework for predicting judicial decisions. *Journal of Signal Processing Systems* 91 (2019), 1159–1167.
[17] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490
[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[20] Masha Medvedeva. 2022. Identification, Categorisation and Forecasting of Court Decisions.
[21] Scott Rempell. 2022. Factors. *Buff. L. Rev.* 70 (2022), 1755. http://dx.doi.org/10.2139/ssrn.4095435
[22] Edwina L Rissland and M Timur Friedman. 1995. Detecting change in legal concepts. In *Proceedings of the 5th international conference on Artificial intelligence and law.* 127–136.
[23] Jaromir Savelka and Kevin D Ashley. 2018. Segmenting US Court Decisions into Functional and Issue Specific Parts.. In *JURIX.* 111–120.
[24] Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. 2017. Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues* 58 (2017), 21.
[25] Hsuan-Lei Shao, Robert B Leflar, and Sieh-Chuen Huang. 2022. Factors Determining Child Custody in Taiwan after Patriarchy's Decline: Decision Tree Analysis on Family Court Decisions. *Asian Journal of Comparative Law* (2022), 1–17.
[26] A. Wyner and W. Peters. 2010. Towards annotating and extracting textual legal case factors.. In *SPLeT-2012.* 36–45.