European Journal of Law and Economics (2020) 50:451–467
https://doi.org/10.1007/s10657-020-09645-7

Automated fact-value distinction in court opinions

Yu Cao[1] · **Elliott Ash**[2] · **Daniel L. Chen**[3]

Published online: 6 March 2020

Abstract
This paper studies the problem of automated classification of fact statements and value statements in written judicial decisions. We compare a range of methods and demonstrate that the linguistic features of sentences and paragraphs can be used to successfully classify them along this dimension. The Wordscores method by Laver et al. (Am Polit Sci Rev 97(2):311–331, 2003) performs best in held out data. In an application, we show that the value segments of opinions are more informative than fact segments of the ideological direction of U.S. circuit court opinions.

Keywords K40 · Facts versus law · Law and machine learning · Law and NLP · Text data

JEL Classification K40

1 Introduction

The contents of court opinions comprise among other things *fact statements* and *value statements*. The former concerns what the legal professionals know about the factual grounds of a case, given all evidence then available to them. The latter, on the other hand, concerns what legal (and ethical, whenever relevant) principles are applicable given what have been stated as facts. The classic dichotomy of "facts

✉ Elliott Ash
ashe@ethz.ch

Yu Cao
yc825@linguistics.rutgers.edu

Daniel L. Chen
daniel.chen@iast.fr

[1] Department of Linguistics, Rutgers, New Brunswick, USA

[2] Center for Law and Economics, ETH Zurich, Zurich, Switzerland

[3] Institute for Advanced Study, Toulouse School of Economics, Toulouse, France

versus law" has been a major theme of legal discourse in Common Law countries and is known to be a major component of judicial reasoning (Greenberg 2004).

This study uses computational techniques to develop a document classifier that automatically distinguishes between fact and value statements in court opinions. The resulting delineated corpora can be used for a range of empirical studies on how judges reason towards decisions. For example, do judges alter facts to fit their judgments?

Automated fact-value distinction has indeed found many applications in recent empirical legal studies. To name but a few, Shulayeva et al. (2017) highlights the immediate relevance of this distinction in judicial citations, since identifying factual grounds is the first step in drawing on legal precedents to support current decisions. Smith (2014) shows that judges are more likely to exercise policy preference in legal disputes focusing more on interpretations of facts, but less likely to do so in cases focusing more on interpretations of legal principles. In controlling the textual factors that might influence the likelihood of a case to be remanded from appellate courts to district courts, Sarel and Demirtas (2017) have considered whether a case raises more factual questions or more legal questions.

This paper presents a new corpus and approach to this problem. We start with an expert-labeled corpus (labeled by the judges themselves), where paragraphs are annotated as related to facts versus values (discussions of law). We use a new featural representation of documents based on syntactic dependencies, which captures what linguistically distinguishes fact statements from value statements. As this is exploratory research, we compare a number of models for classifying facts vs values. In a cross-model disagreement analysis, we show that our features exploit the linguistic structure of ruling actions and are less likely to be misled by lexical factors, though the classifier is prone to ignorance of the identity of agents of those actions.

To demonstrate the usefulness of the model, we use it for a downstream empirical analysis. We find that value sections of court opinions are more informative of the ideological direction of an opinion than the fact sections (although both are predictive). Future work could use fact-value labels of legal text for many relevant empirical investigations.

The organization of the paper is as follows. After a review of related works in Sects. 2 and 3 provides the motivations and details of our feature extraction. Section 4 reports two experiments, the supervised learning and the disagreement analysis that inspects model behaviors against individual texts. After showing an empirical application of our model in Sects. 5 and 6 concludes.

2 Background and related works

By fact-value distinction we mean the distinction between *linguistic statements* about facts (i.e., descriptive) and values (i.e., normative). We do not get into the broader epistemological distinction between facts per se and values per se (e.g., Mulligan and Correia 2017; Schroeder 2016).

Fact-value distinction is similar to and sometimes confused with *subjective-objective* and *fact-opinion* distinctions (Corvino 2014). From Corvino's discussion the differences between the three distinctions are repeated as follows:

- *Facts vs. values* "[fact statements] describe the world; [value statements] evaluate it."
- *Subjective vs. objective* subjective statements are "mind-dependent" (in the sense that the truth of the statement is sensitive to the choice of attitude-holders); objective statements are "mind-independent" (in the sense that the truth of such a statement can be verified independently of attitude holders).
- *Facts vs. opinions* fact statements are "objective and well supported by the available evidence"; opinion statements are "either subjective or else not well supported by the available evidence."

These should have made clear the importance of keeping the three distinctions apart. For one thing, it is controversial whether all value statements are subjective (e.g., Corvino mentions that many argue against the view the moral beliefs are subjective), and opinions can be descriptive (Corvino's own example, *God exists*) rather than normative.

That said, subjective-objective distinction, a.k.a. subjectivity classification, is by far a better studied text classification task, typically at sentence-level and in juxtaposition with sentence-level *sentiment classification*, i.e., to determine whether a subjective sentence expresses a positive or negative attitude; see Liu (2010) for a review. Works on subjectivity classification have capitalized on supervised learning methods such as the naïve Bayesian classifier, using features like unigrams, syntactic dependencies, and occurrences of the terms or syntactic patterns in a pre-determined or bootstrapping-induced dictionary (Hatzivassiloglou and Wiebe 2000; Riloff and Wiebe 2003; Wilson et al. 2004; Yu and Hatzivassiloglou 2003).

These practices have understandably influenced research in automated fact-value distinction in legal contexts. In the law, these distinctions are important because they can influence decisions in high-stakes legal disputes. They can also influence policy through judges setting precedents. In Smith (2014), a list of terms highly indicative of factual statements and a list of terms highly indicative of legal statements are manually created based on a statistical analysis of 142 annotated opinions drawn from *United States Courts of Appeals Database* (Hurwitz and Kuersten 2012). For a given opinion, a function of the *standardized frequencies* (see "Appendix 1" for details) of the terms in each list is taken as a quantitative measure of the extent to which the opinion concerns the kind of statements the respective list pertains to. Similarly, applying Laver et al.'s (2003) *Wordscore* algorithm, Sarel and Demirtas (2017) use two dictionaries, *Black's Law Dictionary* as an index of legal texts and *The Oxford Thesaurus* as an index of factual texts, to calculate a score of a given text that measures its legality or factuality. The score in question sums up the pre-calculated scores of the bigrams in the respective dictionary, weighted by their frequencies (see "Appendix 2" for details).

Both Smith's and Sarel and Demirtas's methods are reminiscent of a simple text representation strategy known as *bag-of-words* (BOW), except that rather than

keeping track of the individual frequencies of "words", they are collapsed into a single measure. In neither study has that measure been converted to a classification judgment—which can be easily done, however, as in Sect. 4—since their foci are establishing the numeric correlation of that measure with another variable of interest.

To date the only study in this area that sets accurate classification as its primary goal is Shulayeva et al. (2017), where the authors adopt the standard featural representation of texts and train their naïve Bayesian classifier on 2659 annotated sentences collected from 50 common law reports at the British and Irish Legal Information Institute (BAILII). Shulayeva et al.'s model employs a wide range of features besides unigrams, including part of speech tags, dependency pairs, sentence length, sentence position, and a Boolean feature that indicates whether the sentence contains a citation instance. The use of the last three features makes sense only for a model that works at sentence-level, like Shulayeva et al.'s.

As a final note, it is worth restating that the distinction between facts and values, both in concept and in language, has special meaning in law that differs from other contexts. The results of empirical analysis of the fact-value distinction in legal language will likely not be extrapolable to broader language environments.

3 Methods

In this study we transform a document into its featural representation based on the syntactic dependencies it contains, like Shulayeva et al. (2017). But unlike the latter, here dependencies serve to subcategorize lexical items, and the lexical items so subcategorized are all that is needed. Below let us start with an observation concerning the linguistic properties of fact and values statements.

3.1 Observation

Factual propositions make claims about what the state of affair was, is, or will be like, whereas normative propositions make claims about what the state of affair *should* or *could* be like, implicitly or explicitly comparing the likelihood or desirability for different state of affairs to obtain. In other words, normative propositions are by nature factual propositions embedded under *modalities* or *propositional attitudes* (see McKay and Nelson 2014; Menzel 2017 for a review), which are encoded in English with modals, e.g., *can, may, must, should*, etc., and attitude verbs, e.g., *believe, uphold, maintain, require*, etc., respectively.

The above linguistic observation, we emphasize, applies to English language in general and judicial opinions in particular. The following value statements, taken from the *United States Circuit Court Opinion Database*, illustrate the use of modals and attitudes:

> The principle **established** has also been **affirmed** by so many decisions in the courts of New Jersey, that it **may** now be considered as the settled law of that state...

Roman Catholic Church v. Pennsylvania Railroad, 207 F.1d 897 (1913)

It bears **repeating that** this appeal is brought only by the individual officers, not the City of Corinth, ... And, it is well to **remember that** qualified immunity serves a number of quite important goals. Courts have **expressed** a concern over 'the deterrent effect that civil liability **may** have on the willingness of public officials...
Hare v. Corinth, 135 F.3d 320 (1998)

In instances in which we **uphold** the trial court's **determination that** the appeal is not taken in good faith... payment of the full appellate filing fees and costs, less what has already been collected, **must** be made within 30 days or the appeal will be dismissed for want of prosecution.
Robert L. Baugh v. Joe M. Taylor E. Nevelow P. Evans, 117 F.3d 197 (1997)

By contrast, these value indicators are typically missing in fact statements:

The regular train crews had done and still do this work. They are employees of the railroads—called the tenant lines—which use the station's terminal facilities.
Washington Terminal Co. v. Boswell, 124 F.2d 235 (1941)

Following this injection, Amanda began to show signs of illness—fever, lethargy, and seizures. She was given a second DPT shot on April 20, 1979, after which the seizures became more frequent.
Beck v. Secretary of the Department of Health and Human Services,
924 F.2d 1029 (1991)

The Board denied the request on June 6th for failure by petitioner to justify the delay in requesting an extension of time. On August 29th, petitioner requested reconsideration of the denial.
Jacinto S. Pinat v. Office of Personnel Management, 931 F.2d 1544 (1991)

Shulayeva et al. (2017) also associate modals with legal principles. The statistical analysis of Smith (2014) confirms that certain propositional attitudes are more likely to occur in law-bound texts.

3.2 Features

Since value statements usually embed fact statements under modals and attitudes, a proper feature extraction needs to reflect this structural property. Syntactic dependencies are an easily obtainable encoding that satisfies the requirement, and have been successfully applied to subjectivity classification, (e.g., Wilson et al. 2004), stance classification (e.g., Hasan and Ng 2014), and fact-value distinction (e.g., Shulayeva et al. 2017).

As an example, in Fig. 1, the fact that *brought* depends on *repeating* as a clausal complement (CCOMP) marks embedding *this appeal is brought* ... under *repeating*,

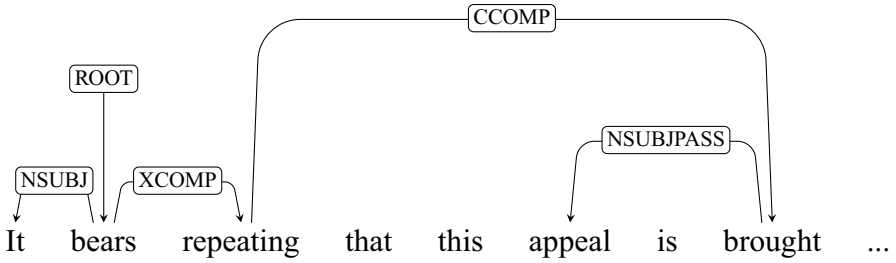

Fig. 1 Major synactic dependencies in *it bears repeating that this appeal is brought ...*. Hare v. Corinth, 135 F.3d 320 (1998)

an attitude verb. The embedded clause itself passes for a fact statement while the whole sentence conveys a ruling.

There are a number of ways to employ dependency features (see Wilson et al. 2004; Hasan and Ng 2014). We propose to use word-dependency-name pairs, e.g., `repeat-CCOMP`. This amounts to subcategorizing words with dependency labels, e.g., in place of a single feature `bear`, we have `bear-NSUBJ`, `bear-XCOMP`, etc.

Compared to Shulayeva et al. (2017) where word-dependent pairs are used (e.g., `repeat-bring`), our approach generalizes across dependents in the same syntactic environment (e.g., `repeat-bring` and `repeat-give` reduce to `repeat-CCOMP`) and thus is less prone to the problem of feature sparsity.

One could argue that the occurrence of certain tokens alone, especially in case of nouns, is sufficiently indicative of whether a statement is about facts or values (see Smith 2014 for instances). Thus following previous works, we also incorporate unigram features.

We still need to consider (engineering) the values of the features. Multiple options are there in the literature, such as raw counts, counts clipped at one, frequencies, etc. In a pilot study we find the following frequency-like measure works well:

$$\frac{\mathsf{Count}(\texttt{repeat-CCOMP})}{\max(1, \lg \mathsf{Len}(d))}.$$

That is, the count of a feature is weighted by the logarithmic length of the document, if greater than one. Obviously, the measure grows less quickly in proportion to the document length than word frequency.

4 Experiments

To evaluate the effectiveness of our dependency-based featural representation and to understand its behavior in judicial opinions, we conducted two experiments. In the first one, the supervised learning, a MLP classifier on top of our featural representation is trained and validated on the aforementioned dataset fraction. For comparison purposes, we also implemented the methods of Shulayeva et al. (2017), Smith

Table 1 Headers taken as the ground truths

Fact-indicating	Value-indicating
Background, evidence, evidence of, existence of, fact, facts, factual, findings of fact, procedural history	Abandonment, ability, acceptance of, accrual of, adequacy of, administrative, admissibility of, admission of, affidavit of, affidavits of, allegations of, analysis of, appeal of, applicability of, applicable, application of, assignments of, challenges to, claims against, common law, compliance with, competency of, conclusion, consideration of, constitutional, constitutionality of, contentions of, decision of, discussion, dismissal of, district court, federal, improper, jurisdiction, law, motion to, rule, sentencing, standards for, statutory

(2014), Sarel and Demirtas (2017) and used BOW and doc2vec (Le and Mikolov 2014) as the baselines. It is shown that our model has achieved a competitive performance. The second experiment focuses on cases where the classification predictions by different models disagree. A qualitative analysis provides some insights about pros and cons of our featural representation, how it would avail and when it might fail.

4.1 Corpus and preparation

Our corpus comprises the full set of judicial opinions from CourtListener.com, spanning over a wide range of U.S. courts and years. This corpus includes 216 courts and the corpus goes back to the 1800s for some courts.

Out of this comprehensive corpus, we start by identifying cases where we can extract labeled data for facts and law (value). There are many opinions where we can delineate the "fact" and "value" sections of the opinions using the expert labels provided by the authoring judges. In particular, we identify sections by headers using regular expressions based on preceding roman numerals.

We then categorize headers that clearly identify fact sections versus law/value sections. For example, a header beginning with *adequacy of* or *challenges to* is taken to label a value document concerning legal standards, whereas a header beginning with *factual background* or *procedural history* is taken to label a factual document. The complete lists of the headers we assume to be fact-indicating and value-indicating are given in Table 1. Following this procedure, we obtain 23,497 case sections.

Using these headers is not the only way that one could have built a labeled corpus. A disadvantage is that fact sections often contain some value statements, and value sections often contain law statements. We also lose a lot of data from sections that are not labeled according to our pattern matching algorithm. The most stringent, but also most costly, solution would be to have legal experts read opinions and extract fact and value statements at the sentence level. We attempted this at an early stage of this project and quickly realized that it would be prohibitively costly. Even

then, there were many sentences that the law students could not confidently annotate as purely factual or purely legalistic.

We therefore made a choice that the annotations of judges (in the section headers) were sufficient for our purpose. We found that with statement-level annotations, they were much higher in the fact-labeled sections. From reading the sections ourselves, we could tell that the labels were identifying paragraphs that tended to discuss facts on the one hand and law/values on the other. Future work could do more to build an effective corpus of fact vs value statements.

The next step is to consider the linguistic unit under which to do our analysis. From a linguistic perspective, paragraphs can be regarded as the smallest discourse unit where a number of congruent sentences jointly develop a single idea. We thus consider the proper granularity level on which a fact-value classifier works to be that of paragraphs. Besides the practical needs of many downstream empirical analyses, our intuition is that readers can better determine whether a paragraph is about facts or values than they can do with either a single sentence, whose interpretation is susceptible to those surrounding it, or a longer document, which may consist of paragraphs pertinent to both facts and values. In addition, whole sections vary in their length much more than paragraphs. Therefore token frequencies are more comparable by paragraph.

With this in mind, we segment each case section into a paragraph, where paragraph boundaries are annotated by HTML markup. We end up with a dataset consisting of 1,301,609 paragraph-sized documents, 36.5% of which are fact-bound and 63.5% of which are value-bound. This class imbalance is due to more section headers having our value labels, relative to our fact labels. We take 80% of dataset to be the training-development set, and hold out the remaining 20% as the test set.

Our learning task requires dependency parsing, a time-consuming procedure for any natural language processing toolkit on the market, e.g., spaCy (Honnibal and Johnson 2015) used here, in face of the sheer size of the dataset. Thus for now we use only a small fraction of the latter, containing 1000 fact instances and 1000 value instances randomly sampled from the training-development set. This fraction forms the basis of our supervised learning experiment; its scale is still larger than or comparable with those of the corpora used in the works reviewed previously.

All the texts in the dataset are cleaned by removing footnote numbers, but numbers for sections, chapters, and law references are preserved. A pilot study over a small development dataset shows that numbers of the latter category but not the former are indicative of texts on application of legal principles.

4.2 Supervised learning

The classifiers used in this experiment are implemented with the machine learning library scikit-learn (Buitinck et al. 2013). As this is exploratory research using a new corpus, we feel it is important to compare many supervised learning approaches. We

Table 2 Fivefold cross validation

	F1 for facts	F1 for values
depn	73.38	74.66
depw	72.77	73.79
smith	71.85	71.4
ws	**77.11**	**77.67**
bow	72.57	73.29
d2v	67.18	65.36

Bold indicates it is the highest-F1 specification (best at predicting facts, in the left column, and best at predicting values, in the right column)

compare an array of methods from the literature and two baselines commonly used in text classification, as detailed below.

(i) DEPN: our method, using words subcategorized by dependency names as features. The feature vocabulary is clipped to the top 4000 items occurring most frequently in the training set. The feature value is as introduced in Sect. 3.2. The representation is fed to a MLP classifier with two 500-dimensional hidden layers.[1] Other settings of the MLP are as scikit-learn's default.

(ii) DEPW: using word-dependent as features, following Shulayeva et al. (2017). The construction of the feature vocabulary, feature value assignment, and the MLP classifier set-up are the same as DEPN.

(iii) SMITH: a BOW-like method implementing Smith (2014) with adaptions. It uses predetermined words indicative of facts or values as features and standardized frequencies as feature values. Vectorization outputs are fed to a Logistic Regression classifier.

(iv) WS: a BOW-like method implementing WordScore (Laver et al. 2003; Sarel and Demirtas 2017). Each word in the training set is assigned a fact- or value-inclination score. The score of a document is the re-scaled sum of the frequency-weighted scores of the words it contains (see "Appendix 2").

(v) BOW: a baseline using word, i.e., unigram features. The construction of feature vocabulary and feature value assignment are the same as DEPN. Vectorization outputs are fed to a Logistic Regression classifier.

(vi) D2V: a baseline using Gensim toolkit's (Řehůřek and Sojka 2010) implementation of doc2vec (Le and Mikolov 2014) for document vectorization. The outputs are fed to a MLP classifier, with the same set-up as DEPN.

The results of fivefold cross validation are reported in Table 2. The ws model achieves the highest F1 scores in detecting both fact and value statements, but another BOW-like model, smith, does not perform as well. The F1 scores of the depn

[1] We choose MLP because it supports training in batches, which is very helpful when we move on to training our model over the entire corpus in the future.

Table 3 Pairwise coincidence ratio

	DEPN	DEPW	SMITH	WS
depn	–	92	83.5	78.7
depw	–	–	80	81.8
smith	–	–	–	84.8
ws	–	–	–	–

model are slightly better than those of the depw model, which does not obviously outdo the the baseline bow model. The only neural model d2v has the lowest performance, suggesting that it is not as sensitive to fact-value distinctions as it might be in other topic-identifying domains. All other models have similar performance in identifying fact and value statements, suggesting that training on a balanced corpus like ours will not introduce identification bias to a feature-based classifier.

It is worth mentioning that Shulayeva et al. (2017) report F1 scores ≥ 81 for their sentence-level classifier trained and tested on a manually annotated dataset (2659 sentences; 60% neutral, 30% values, and 10% facts). The old caveat remains that no meaningful model comparison can be made when the training configuration or the test base differs. For the next step, it would be interesting to have our depn and depw models' performance evaluated on Shulayeva et al.'s dataset.

4.3 Disagreement analysis

In the second experiment, we inspect on the judicial documents on which the classification predictions by the previous models vary. By doing so we may gain some insights into the behaviors of these models. We leave aside the two baselines and focus on the first four models evaluated above, as there is no need for cross validation here. We re-trained the four models on a larger fraction of the training-development set, comprising 10,000 fact instances and 10,000 value instances and had them tested against 100 examples (50% facts, 50%values) randomly sampled from the test set.

The judgments given by the four models largely coincide: out of the 100 examples there are 74 on which the models agree. A pairwise comparison illustrates more details: we take the judgments given by one model as pseudo-gold standards and take the F1 score of the other model under comparison as the measure of coincidence ratio of the two models. The results are given by Table 3, where it is shown that depn and depw are closer to each other than either of them is to smith or ws, and the latter two BOW-like models are the second most similar model pair. Interestingly, though depn turns out to be the least similar model to ws, it fares better in the cross validation test than the other two models that are closer to ws. The interpretation could be either that the pairwise coincidence measure done on the current small test set is not representative enough, or that how the performance of the four models compare to each other might be shifted by a larger training set. We will not pursue the issue here but simply take Table 3 for what it is.

Let us now focus on comparing the behavior of depn with those of others. While depn and depw are quite similar, when their judgments disagree, it appears that depw is more likely to be misled by lexical factors. For example, the following factual statement is correctly identified by depn but missed by depw, probably because the paragraph contains a lengthy reference to a legal case, here in boldface.

> Next, the reticle is blown up 200 times–the resulting enlarged reproduction being called a 'low back' or 'overlay.' Once the reticle is confirmed as containing the correct design, it is placed in a repeat camera which reduces the design to actual size and repeats it over and over again on a chrome piece or 'mask' which then becomes the actual production tool. **(People v. Superior Court (Moore) (1980) 104 Cal . App. 3d 1001, 1005 [163 Cal. Rptr. 906], italics added; see also 1984 U.S. Code Cong. Admin. News, at pp. 5760–5763.**
> Label: F; depn: F; depw: V; smith: F; ws: V.

In another value statement, though the paragraph is largely made up of factual descriptions, but the first sentence, in boldface, makes clear that those descriptions are cited as arguments in support of a judicial judgment in the background. Here, depn alone correctly understands the inter-sentential relationship:

> **The superior court provided several reasons for its finding.** First, although Bruce had some income, "it was minimal, and much was taken to support his other children." Second, "the [Eberts] neither needed nor asked for any support from [Bruce]" and "[Bruce's] testimony indicates that he would have been willing to pay something had the [Eberts] asked him to do so." Finally, Bruce "testified credibly" that he was unaware he had a legal obligation to pay support to the Eberts.
> Label: V; depn: V; depw: F; smith: F; ws: F.

Similar observations can be made when comparing depn with smith and ws. Since the latter two do not take structural information into consideration, they might be misled in description of procedural history, especially when it comes to factual description of previous court decisions. depn, this time along with depw, is immune to this disguise:

> (7) Trial counsel rendered ineffective assistance by failing to file a motion.
> (8) He was denied due process as a result of the state court 's failure to hold an evidentiary hearing on substantial, controverted and unresolved issues.
> (9) The trial court erred by refusing to grant a challenge for cause to juror.
> (10) The trial court abused its discretion by denying Barbee 's motion to suppress alleged statements made to Detective Carroll;
> Label: F; depn: F; depw: F; smith: V; ws: V.

However, sometimes depn might over-evaluate the predictive force of structural information and be misled by the latter. In the following factual statement, all other three models, including depw, give the correct judgment. But depn fails as if it is confused by structures for ruling actions like *denied that*, *asserted that*,

Table 4 Fivefold cross validation, liberal-conservative distinction

	F1 for liberal	F1 for conservative
fact-weighted	46.91	66.69
value-weighted	**49.45**	**67.29**

Bold indicates it is the highest-F1 specification (best at predicting facts, in the left column, and best at predicting values, in the right column)

averred that, but overlooks the fact that the agents of these actions are subjects (e.g., *Chris*, *Bs*) involved in the case, not the judicial authority:

> **Chris** answered and **denied that grounds existed** to terminate his parental rights; in a counter-petition, he **asserted that he was entitled to** custody of Landon under the "superior rights doctrine." The **Bs** answered the counter-petition and **averred that** "[Chirs'] personal drug use and his engagement in the drug trade" **constituted** "substantial harm that allows a court to deprive a natural parent of custody of a child" and **that** "it is contrary to the best interest of the child to permit [Chris] to exercise regular overnight visitation" with Landon.
> Label: F; depn: V; depw: F; smith: F; ws: F.

Here is another example of the same kind, where both depn and depw fail:

> The confessions given to law enforcement officers in July 1992 conflict with several other versions of the crimes Shafer gave to mental health professionals and with the co-defendant's version.[3] **Shafer**, however, **confirmed** during the change of plea hearing **that** the July 1992 confessions were the true and correct versions of the crimes.
> Label: F; depn: V; depw: V; smith: F; ws: F.

The above observations are by no means comprehensive, but they do tell us something about the behaviors of the fact-value distinguishing models under investigation, which we might reasonably conjecture given the constructs of those models. In sum, the more importance a model attaches to structural information, the more likely it would rely on presence or absence of linguistic structures for ruling actions to identify value statements, and the less likely it would be misled by lexical factors. But the cost of this gain is that such a model is also more likely to ignore important lexical information that reveals the identity of the ruling (or any other) actions.

5 Application

As our fact-value classification model (depn) has achieved a reasonable precision and sensitivity, it would be beneficial to see how its predictions could be put to practical use. Along the fields of application mentioned in Sect. 1, here we are interested

in whether the conservative or liberal inclination of a court opinion finds a stronger correlation in the way it describes facts or the way it states values (i.e., applies legal principles). Conceivably, our hypothesis goes to the latter, since we do not expect judges' conservative or liberal policy preference to influence their accounts of facts.

To test this hypothesis, we conducted another supervised learning experiment where the predictions of our fact-value classification model obtained in Sect. 4.3 are used to create a term-frequency representation of court opinions that is relativized to either the *fact-hood* (the likelihood to be associated with fact statements) or *value-hood* (the likelihood to be associated with fact statements) of terms (see "Appendix 3" for details). A Logistic Regression classifier is then trained on a fraction (about 5%) of the U.S. Circuit Court Opinion corpus, where each opinion has been manually annotated as "conservative" or "liberal" (we ignore "neutral" cases for this application).

We did the usual fivefold cross-validation to compare the predictive force of fact-weighted *n*-gram representations and value-weighted *n*-gram representations. Table 4 gives the results. While the absolute performance is not great, a classifier using value-weighted *n*-gram representations does perform better in distinguishing liberal-inclined decisions and conservative-inclined decisions. This confirms our expectation that the value sections of a court opinion can better predict its liberalness or conservativeness.

6 Conclusion

This paper has developed a machine learning model for fact-value distinction by using lexical items subcategorized by the syntactic dependencies they enter. It has conducted two learning experiments, one to evaluate this model by comparing its performance with those of the methods proposed in the previous literature, and the other to understand how its behavior differs from its precedents by analyzing the texts on which their judgments differ. The results have established that dependency features in the way they are utilized here are useful in identifying linguistic structures that express modalities and propositional attitudes, thereby qualifying them as strong predictors for distinguishing fact and value statements. This is because value statements in the context of court opinions usually boil down to modalities and attitudes concerning judicial judgments or legal principles. Indirect support to this approach comes from yet another learning experiment, where the output of such a fact-value classifier feeds a downstream classification task that identifies a court opinion's ideological inclination.

Our results also point out a deficiency of the current approach. Value statements feature not propositional attitudes or modalities in general, but those of certain holders, i.e., judicial authorities. Thus for the future, the hope is that the techniques of a widely applied common information task, Named Entity Recognition (NER), can be incorporated into the meaning representation of court opinions, so that a fact-value classifier can be trained to concentrate on modalities, propositional attitudes, or ruling actions held by proper entities.

Appendix 1: Smith's algorithm

We describe below how the so-called standardized frequencies are calculated in Smith (2014) to determine which lexical items are statistically indicative of fact statements or value statements.

Suppose w is a word in the set W of words that appear frequently enough in the training set. Let D be the set of training documents. The frequency of w in some $d \in D$ is given by

$$f_d(w) = \frac{\mathsf{Count}(w)}{\mathsf{Len}(d)}.$$

The *standardized frequency* of w in d is defined as the ratio of the frequency of w in d to the mean frequency of w across all $d \in D$, i.e.,

$$f_d^*(w) = \frac{f_d(w)}{\mu\{f_d(w)\}_{d\in D}},$$

where μ denotes the mean.

Suppose that $D_v \subset D$ is the subset compoesd of value statements, and $D_f = D \backslash D_v$ is the subset composed of fact statements. Then we may compare the difference between the mean standardized frequency of w across all $d \in D_v$ and the mean standardized frequency of w across all $d \in D_f$, i.e.,

$$\delta_w = \mu\{f_d^*(w)\}_{d\in D_v} - \mu\{f_d^*(w)\}_{d\in D_f}.$$

If $\delta_w > 0$ and this difference is statistically significant ($p < 0.01$) then w is taken to be *statistically indicative of value statements*. A similar procedure applies to determine words that are statistically indicative of fact statements.

Appendix 2: Wordscore algorithm

Here we summarize the essentials of Laver et al.'s (2003) wordscore algorithm, couched in the terminology of supervised learning.

Let A be a function that assigns an a priori score to documents in the training set D. In our case, let $A(d) = -1$ if d is a fact statement, $A(d) = 1$ if d is a value statement. It can be shown that if a priori we have even chance to come across any document in D, then the probability for a document to be d upon observing the occurrence of w in that document is given by

$$P(d|w) = \frac{f_d(w)}{\sum_{d' \in D} f_{d'}(w)}.$$

The score of a word w is calculated by

$$S(w) = \sum_{d \in D} A(d) P(d|w).$$

Thus for a given document t in the test set T, we may calculate its score as

$$S(t) = \sum_{w \in t} S(w) f_t(w).$$

To ensure that $\{S(t)\}_{t \in T}$ has the same dispersion metric as $\{A(d)\}_{d \in D}$, $S(t)$ is further re-scaled as

$$S^*(t) = \left(S(t) - \mu\{S(t)\}_{t \in T}\right) \frac{\sigma\{A(d)\}_{d \in D}}{\sigma\{S(t)\}_{t \in T}} + \mu\{S(t)\}_{t \in T},$$

where σ denotes the standard deviation.

Given our set-up of A, $S^*(t)$ is converted into a categorical judgment simply as follows: t is a fact statement if $S^*(t) \leq 0$, a value statement otherwise.

Appendix 3: Fact- and value-weighted N-gram frequencies

For a paragraph p in a case d, our fact-value classification model provides a predicted probability $f_p \in [0, 1]$ that p is about facts, with numbers near one indicating fact patterns and numbers near zero indicating law or value patterns.

We compute the counts of terms for each paragraph, including unigrams and bigrams after removing stopwords, capitalization, and punctuation, and stemming word endings. Let $\mathsf{Count}_p(w)$ be the count of a term $w \in W$ in the paragraph p, where W gives the vocabulary of n-grams in the case d.

For each paragraph $p \in d$ and each term $w \in W$, we compute the *fact-weighted count*, $f_p \mathsf{Count}_p(w)$, and *value-weighted count* $(1 - f_p)\mathsf{Count}_p(w)$. Then, the *fact frequency* of the term w in the case d is the summation over the fact-weighted counts over paragraphs in the case, divided by the mean fact-weighted counts over all n-grams:

$$F^f_d(w) = \frac{\sum_{p \in d} f_p \mathsf{Count}_p(w)}{\mu\{\sum_{p \in d} f_p \mathsf{Count}_p(v)\}_{v \in W}},$$

and correspondingly its *value frequency* is

$$F_d^v(w) = \frac{\sum_{p\in d}(1-f_p)\mathsf{Count}_p(w)}{\mu\{\sum_{p\in d}(1-f_p)\mathsf{Count}_p(v)\}_{v\in W}}.$$

References

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., et al. (2013). API design for machine learning software: Experiences from the scikit-learn project. In *ECML PKDD workshop: Languages for data mining and machine learning*, pp. 108–122.

Corvino, J. (2014). The fact/opinion distinction. *The Philosophers' Magazine*, *65*(2), 57–61.

Greenberg, M. (2004). How facts make law. *Legal Theory*, *10*(3), 157–198.

Hasan, K. S., & Ng, V. (2014). Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 751–762.

Hatzivassiloglou, V., & Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on computational linguistics, Vol. 1*, pp. 299–305. Association for Computational Linguistics.

Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1373–1378, Lisbon, Portugal: Association for Computational Linguistics.

Hurwitz, M., & Kuersten, A. (2012). Changes in the circuits: Exploring the courts of appeals databases and the federal appellate courts. *Judicature*, *96*(1), 23–34.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(2), 311–331.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Latural Language Processing*, *2*, 627–666.

McKay, T., & Nelson, M. (2014). Propositional attitude reports. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford, CA: The Metaphysics Research Lab, Stanford University.

Menzel, C. (2017). Possible worlds. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford, CA: The Metaphysics Research Lab, Stanford University.

Mulligan, K., & Correia, F. (2017). Facts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford, CA: The Metaphysics Research Lab, Stanford University.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, pp. 45–50, Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 105–112. Association for Computational Linguistics.

Sarel, R., & Demirtas, M. (2017). Delegation at the us federal appellate courts: The power to remand as a double-edged sword. Working Paper. Frankfurt School of Finance & Management. https://ssrn.com/abstract=3094634.

Schroeder, M. (2016). Value theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Stanford, CA: The Metaphysics Research Lab, Stanford University.

Shulayeva, O., Siddharthan, A., & Wyner, A. (2017). Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, *25*(1), 107–126.

Smith, J. L. (2014). Law, fact, and the threat of reversal from above. *American Politics Research*, *42*(2), 226–256.

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. *AAAI*, *2004*, 761–779.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on*

empirical methods in natural language processing, pp. 129–136. Association for Computational Linguistics.