Automated Fact-Value Distinction in Court Opinions

Yu Cao, Elliott Ash, Daniel Chen

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. (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.

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 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-level classifier that automatically classifies a paragraph in court opinions as a statement of facts or a statement of values. 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

[0]Yu Cao, Department of Linguistics, Rutgers, yc825@linguistics.rutgers.edu; Elliott Ash, Center for Law and Economics, ETH Zurich, ashe@ethz.ch; Daniel Chen, Institute for Advanced Study, Toulouse School of Economics, daniel.chen@iast.fr.

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 way of creating the featural representation of a document (here a court opinion) based on syntactic dependencies, which as will be argued later, captures what *linguistically* distinguishes fact statements from value statements. On top of that, a Multilayer Perceptron (MLP) classifier is trained and tested on a circuit court opinion corpus. This corpus is expert-labeled (by the judges themselves), where paragraphs are annotated as related to facts versus values (discussions of law). We will compare the performance of our method with those of its precedents in the literature, and discuss their respective gains and losses based on text analysis of the court opinions on which the judgments given by different methods differ. It will be shown that our model has a competitive performance and it is good at recognizing facts and judicial arguments by identifying linguistic structures for ruling actions, less likely to be misled by lexical factors, but can make mistakes when it ignores the identity of agents of those actions. To show the usefulness of our classification model, we will apply it to verify a hypothesis that value sections of court opinions are more indicative of their ideological directions.

The organization of the paper is as follows. After a review of related works in Section 2, Section 3 provides the motivations and details of the current feature-based learning method. Section 4 reports two experiments: one discusses the results of supervised learning and compares the performance of the different classification models, the other zooms in on the court opinions that may shed light on their behaviors. After showing an empirical application of our model in Section 5, Section 6 concludes.

2 Background and Related Works

Two conceptual remarks are in order. First, by fact-value distinction we mean the distinction between *linguistic* statements about facts and values. The distinction between facts *per se* and values *per se* is another matter (e.g., Mulligan and Correia, 2017; Schroeder, 2016).

Second, fact-value (a.k.a. *descriptive-normative*) 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. value**s: "[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. opinion**s: 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 factual claims, i.e., descriptive (Corvino's own example, *God exists*) rather than normative.

As mentioned, fact-value distinction is similar to but different from subjective-objective distinction. That said, the latter, 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. A number of representative 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 (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 A 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 is taken to be the sum of the pre-calculated scores of bigrams in the respective dictionary, weighted by their frequencies (see B for details).

Both (Smith, 2014)'s and (Sarel and Demirtas, 2017)'s methods are reminiscent of a commonly used text representation strategy known as *bag-of-words*, except that rather than keeping track of the individual frequencies of the "words" in the bag (i.e., the dictionary), these frequencies are collapsed into a single measure. Further, in neither study has that measure been converted to a classification judgment—which can be easily done, however, as we will do in Section 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 more standard featural representation of texts (i.e., keeping different features apart) 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; accessible at `http://www.bailii.org`). 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 sen-

tence contains a citation instance. Notice that the use of the last three features makes sense only for a model that works at sentence-level, like Shulayeva et al.'s. The next section will elaborate on the dependency features used in (Shulayeva et al., 2017) as we compare those with our approach.

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. Before getting into more details, let us first make some basic observations concerning the linguistic properties of fact and values statements.

3.1 Observations

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).

Modalities and propositional attitudes are encoded with their special syntax. In English and demonstrably in many others these concepts are expressed by modals, e.g., *can, may, must, should*, etc., and propositional attitude verbs, e.g., *believe, hold, maintain, require*, etc., respectively. Crucially, both modals and propositional attitude verbs come with embedded clauses. The following value statements, taken from the *United States Circuit Court Opinion Database*, illustrate the point:

> "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, as shown in the following list of cases cited by counsel for the defendant: ..."
>
> 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, concerning only qualified immunity, not the merits. 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 to fully discharge their professional duties'."
>
> Hare v. Corinth, 135 F.3d 320 (1998)

The observation that modals are typically associated with legal principles is also made in (Shulayeva et al., 2017), and the statistical analysis in (Smith, 2014) confirms that certain propositional attitude verbs are more likely to occur in law-bound texts.

Shulayeva et al. (2017) also observe that fact statements are more likely to appear in the past tense. But this need not be the case; one can make factual claims about the present and future:

> "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)

We will not dwell on this issue for the current paper.

3.2 Features

Since the syntax of value statements employs special lexical items that occur in special syntactic structures, we expect lexical items *subcategorized* by their structural information to be the distinguishing features that tell fact and value statements apart. The syntactic dependency of a word provides exactly the syntactico-semantic information we need. Specifically, we use a word paired with the name of the dependency relation of its dependent as a feature.

Let W be the set of tokens in any document d. Let function Dep map a token w to the set of all its syntactic dependents, and function Rel_w map a syntactic dependent v of w to the name of the dependency relation that v bears to w. Then d would be represented by the following set of features:

$$\bigcup_{w \in W} \{(w, \mathsf{Rel}_w(v)) \,|\, v \in \mathsf{Dep}(w)\}.$$

Each feature in the set is of the form $(w, \mathsf{Rel}_w(v))$, a token w paired the dependency relation name of its dependent v. It is in this sense that we say $(w, \mathsf{Rel}_w(v))$ is w subcategorized by $\mathsf{Rel}_w(v)$.

One could reasonably argue that the incidence of certain terms alone is sufficiently indicative of whether a statement is about facts or values, even though those terms, especially in the case of nouns (see Smith, 2014 for instances), are not accompanied by any syntactic dependents. This is the motivation for using a bag of unigrams in the works reviewed previously. To incorporate the effect of bags of unigrams into the current picture, we may simply assume that a word trivially depends on itself, i.e., $w \in \mathsf{Dep}(w)$ for any w, and choose whatever name that has not been used by any nontrivial dependency relation as the value of $\mathsf{Rel}_w(w)$. Obviously, the net effect of taking these steps is inserting unigrams into the featural representation of a document.

Finally, we may consider what value to assign to the features constructed above. Multiple options are there in the literature, such as raw counts, counts clipped at one, frequencies, etc.

The trials in the development stage favor a frequency-like measure as such (omitting parentheses for pairs):

$$f\left(w, \mathsf{Rel}_w(v)\right) = \frac{\mathsf{Count}\left(w, \mathsf{Rel}_w(v)\right)}{\max\left(1, \log_{10} \mathsf{Len}\left(d\right)\right)}.$$

Thus the value of feature $(w, \mathsf{Rel}_w(v))$ in a document d is still sensitive to the length of d, but not as readily as its frequency, i.e.,

$$\frac{\mathsf{Count}\left(w, \mathsf{Rel}_w(v)\right)}{\mathsf{Len}\left(d\right)}.$$

3.3 Variants of Dependency Features

Syntactic dependency features have been successfully used in studies on 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). Wilson et al. and Hasan and Ng mention some of the dimensions along which the usage of a dependency pair may vary. For any triple $(w, \mathsf{Rel}_w(v), v)$, where the word w bears the relation $\mathsf{Rel}_w(v)$ with its dependent v, one can make at least these choices: (i) to use w or the part of speech tag of w, (ii) to include or drop $\mathsf{Rel}_w(v)$, and (iii) to use v or the part of speech tag of v. We end up with at least eight variants.

The feature construction introduced above uses w and $\mathsf{Rel}_w(v)$, and drops v altogether. (Shulayeva et al., 2017) on the other hand uses w and v, and drops $\mathsf{Rel}_w(v)$. So that for a document with a token set W, the following features are collected:

$$\bigcup_{w \in W} \left\{ (w, v)) \,|\, v \in \mathsf{Dep}(w) \right\}.$$

This approach, as Shulayeva et al. argue, may outdo simple bigrams in that what ties two words together is dependency, presumably more informative than mere adjacency. But a word pair (w, v) as such will not explicitly reveal the syntactico-semantic information that subcategorizes w and thus helps to distinguish between fact and value statements.

4 Experiments

We have conducted two comparative experiments to evaluate the effectiveness of our dependency-based featural representation as well as to understand the behavior of a classifier built on that basis. The first experiment is one of supervised learning, where a MLP classifier is trained and tested on a fraction of the case corpus. Its performance is compared with three other classifiers that implement (with necessary adaptions) the methods in (Shulayeva et al., 2017), (Smith, 2014), and (Sarel and Demirtas, 2017), respectively. We also add two baseline classification strategies, i.e., bags-of-words and doc2vec (Le and Mikolov, 2014). All these latter classifiers are trained and tested on the same corpus as ours is, about which more details are given below.

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.*

Table 1: Headers taken as gold standards.

4.1 Corpus and Preparation

The labeled corpus is the full set of judicial opinions from CourtListener (courtlistener.com) that included annotated section headers clearly demarcating the "Facts" section of the opinion. The opinions are from a wide range of U.S. courts and years. Using the headers, the opinions were split into sections, and then the sections were split into paragraphs.

The resulting labeled corpus contains 23,497 documents (case sections) of various lengths with automatically extracted headers. We choose those documents with a header that clearly indicates the nature, fact-stating or value-stating, of the paragraphs contained in those documents. 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 *facts* 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.

Longer documents are decomposed into individual paragraphs (or a shorter series of paragraphs), ending up with a corpus consisting of 1,301,609 paragraphs (or short series of paragraphs), 36.5% of which are fact-bound and 63.5% of which are value-bound. It is unsurprising that this corpus should contain much more value statements than fact statements, as fact-indicating headers are largely outnumbered by value-indicating ones. We take 80% of this corpus to be the training-development set, and hold out the remaining 20% for testing purpose.

Our learning task requires dependency parsing, a time-consuming step for any natural language processing model on the market, e.g., spaCy (Honnibal and Johnson, 2015) used here, in face of the sheer size of our corpus. Thus for now we use only 1000 fact statements and 1000 value statements, randomly chosen from the training-development set, to form the basis of our supervised learning experiment. The scale of this fraction is still larger than or at least

	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

Table 2: 5-fold cross validation.

comparable with those of the corpora used in the studies 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 experiment 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 by the machine learning library scikit-learn (Buitinck et al., 2013). We compare our method with 5 other methods, three from the literature and two baselines commonly used in text classification, as detailed below.

- **DEPN(AME)**: our method. Representing a document as the set of lexical items subcategorized by the names of the dependency relations they enter. The feature vocabulary is limited to the top 4000 word-dependency pairs occurring most frequently in the training set. The feature value is a frequency-like measure (see Section 3.2). The representation is fed to a MLP classifier with two hidden layers, both with a dimension of 500 components.[1] Other settings of the MLP are as scikit-learn's default.

- **DEPW(ORD)**: a variant of DEPN, Shulayeva et al.'s (2017) way of using dependency features (see Section 3.3). The construction of feature vocabulary, feature value assignment, and the MLP classifier set-up are the same as above.

- **SMITH**: a bags-of-words like strategy, an implementation of Smith's (2014) method with adaptions. Representing a document as a vector of the standardized frequencies of words that are statistically shown to be indicative of either fact statements or value statements (see Appendix A for details). The representation is fed to a Logistic Regression classifier.

- **W(ORD)S(CORE)**: a bags-of-words like strategy, an implementation of Laver et al.'s (2003) Wordscore algorithm (cf. Sarel and Demirtas, 2017). Each word occurring

[1] We choose MLP because it supports incremental learning (i.e., training in batch), which is very helpful when we move on to training our model over the entire corpus in the future.

	DEPN	DEPW	SMITH	WS
DEPN	-	92	83.5	78.7
DEPW	-	-	80	81.8
SMITH	-	-	-	84.8
WS	-	-	-	-

Table 3: Pairwise coincidence ratio.

in the training set is assigned a score based on statistics, as a measure of its fact- or value-inclination. A document is represented as a (re-scaled) score that sums up the pre-calculated scores of the words it contains, weighted by their frequencies (see Appendix B for details).

- **BOW**: a baseline strategy. Representing a document as a bag of words (unigrams). The construction of feature vocabulary are feature value assignment are the same as DEPN. The representation is fed to a Logistic Regression classifier.

- **D2V**: a baseline strategy. Representing a document as a 500-dimension vector based on a neural model, doc2vec (Le and Mikolov, 2014; implemented by the Gensim toolkit; Řehůřek and Sojka, 2010). The representation is fed to a MLP classifier, whose set-up is the same as DEPN.

The results of 5-fold cross validation are reported in Table 2. The WS model achieves the highest F1 scores in detecting both fact and value statements, but another bags-of-words like model, SMITH, does not perform as well. The F1 scores of the DEPN 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. report better F1 scores (≥ 81) of their model when trained and tested on their manually annotated corpus. This confirms once again the old caveat that the construction of a corpus, in particular how its gold standard labels are created, has a far-reaching implication for subsequent modeling. We are not in a position to assess the noise of our corpus annotations, but rather, in what follows, we will inspect some of the opinion paragraphs on which the judgments of the models evaluated here differ. By doing so, we may hopefully gain some insights into the behaviors of these models.

4.3 Disagreement Analysis

The second experiment leaves aside the two baseline models and focuses on the behaviors of the first four models evaluated above. As there is no need for cross validation here, we re-train

the four models on a larger fraction of the training-development set, comprising 10,000 fact statements and 10,000 value statements. The re-trained models are tested against 100 examples randomly chosen from the test set, half of which are labeled as facts and the other half values.

Generally speaking, 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 bags-of-words like models are the second most similar model pair. Interestingly, though DEPN turns out to be the least similar model to WS, it fairs 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 settle 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*, *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

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

Table 4: 5-fold cross validation, liberal-conservative distinction.

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 Section 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 Subsection 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 C 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 5-fold cross-validation to compare the predictive force of fact-weighted n-gram representations and value-weighted n-gram representations. Table 4 gives the results.

Despite the absolute performance, a classifier using value-weighted n-gram representations performs better in identifying both liberal-inclined decisions and conservative-inclined decisions, confirming 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 meth-

ods 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.

A 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 statisti-

cally indicative of fact statements.

B 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.

C 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:

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

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., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 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. and 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)*, pages 751–762.

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

Honnibal, M. and 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*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

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

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

McKay, T. and Nelson, M. (2014). Propositional attitude reports. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Menzel, C. (2017). Possible worlds. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Mulligan, K. and Correia, F. (2017). Facts. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

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

Sarel, R. and Demirtas, M. (2017). Delegation at the us federal appellate courts: The power to remand as a double-edged sword.

Schroeder, M. (2016). Value theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Shulayeva, O., Siddharthan, A., and 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., and Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. In *AAAI, 2004*, pages 761–779.

Yu, H. and 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*, pages 129–136. Association for Computational Linguistics.