Studying Judicial Behavior with Text Analysis
Bao Kham Chau & Michael A. Livermore

Traditional legal research methods rely on reading and interpreting legal texts. Even with an army of capable assistants, the amount of data that can be generated in this way is quite limited. Analyses of “small data” can lead to useful insights, but researchers can now move beyond conventional methods of data collection and analysis. In recent decades, the digitization of legal texts has enabled empirical legal scholars to use computational text analysis techniques to study large digital corpora of legal documents. These computer-based tools can help researchers study long-standing questions and also open new, previously inaccessible epistemic doors (Livermore and Rockmore 2019).

As with any act of translation, however, the transformation of legal documents into data capable of quantitative analysis can result in some loss of information. Legal texts, such as judicial opinions, are intrinsically high-dimensional. This means that the number of attributes that might be potentially germane to a research question—i.e., the number of dimensions that must be used to fully characterize an object like a legal opinion—are extremely high relative to the number of available observations. Reducing this dimensionality is a prerequisite for many forms of quantitative analysis. But this dimensionality reduction comes at the cost of losing potentially valuable information. The challenge for researchers who use law-as-data techniques is to strike a balance between rich, high-dimension representation of legal texts (which can be impractical to analyze) and low resolution, simplified representations that are tractable but incomplete (Frankenreiter and Livermore 2020).

For the field of empirical study of judicial behavior, managing this tradeoff is worth the effort. Techniques from the fields of machine learning, natural language processing, and text mining have already been put to productive use to study questions about judicial decision making around the world—from how the emotional valence of legal briefs predicts the outcomes of cases in the U.S. Supreme Court (Black et al. 2016) to how private litigants in the People’s Republic of China use administrative litigation to seek assistance from the state in their disputes (Liebman et al. 2020). Writing at the turn of the twentieth century, Justice Holmes famously predicted that “blackletter” study of the law would be supplanted, in his future, by “statistics” and “economics” (Holmes 1897). It does not seem outlandish, toward the beginning of the twenty-first century, to update Holmes’s forecast. While doctrinal analysis (i.e., blackletter law) and traditional statistical and economic analysis of law likely have a long future ahead of them, recent trends provide ample reason to predict that machine learning and natural language processing will take up a central place in the methodological toolkit used to study the law and legal institutions.

This chapter provides a survey of the exciting new literature that uses computational methods to extract information from legal documents to answer research questions on law and legal institutions. In particular, it discusses how legal scholars use computational tools such as term-frequency vectors and word embedding to translate legal texts into data and subject this data to quantitative analysis. This law-as-data approach supplements the traditional modes of analysis and empowers legal scholars to examine the complex interplays between the judiciary and other sociopolitical institutions.

1. From Law to Data

An important step in integrating computational text analysis techniques into legal scholarship on judicial behavior is the conversion of judicial texts into computer-readable data. Traditionally, this conversion was done manually, with dedicated research assistants diligently combing through mountains of documents, hand coding or categorizing them, and entering these categorized documents into some type of digital spreadsheet (Johnson et al. 2009, Ruger et al. 2004). Law & Zaring (2010), for example, worked with their research assistants to manually code U.S. Supreme Court opinions before inputting
them into a logistic regression model in their study of the use of legislative history by U.S. Supreme Court Justices.

Manually coding legal documents into machine-readable format, however, is both time and resource intensive. Additionally, hand coding is subjective in ways that could bias results (Burla et al. 2008, Glazier, Boydstun, and Feezell 2021). These limitations thus narrow the scope of research projects and the scale of a corpus that can be analyzed.

The advent of more sophisticated computational tools solved some of these problems while presenting new challenges. Almost all recent judicial texts, for example, do not need to be translated into machine-readable format, as they are “born digital.” Similarly, a growing portion of legacy texts has been digitized through optical character recognition (OCR) software and could be subjected to quantitative analysis. In this new digital world, the problem is not caused by the scarcity of high-fidelity data, but rather the overabundance of it.

This is because texts are considered to be unstructured data by computers. Structured data, such as Excel spreadsheets containing records of banking transactions or airline reservations, follows a particular format and is easily accessible by both humans and machines. Raw texts, by contrast, are considered unstructured because computers could not readily process the complex linguistic arrangements of the human languages. Without structures that are comprehensible to the computer, most text amounts to little more than files filled with noise.

In order to meaningfully apply computational text analysis techniques to judicial texts, it is thus necessary to transform them into structured data. Translating these unstructured texts to their structural counterparts involves a range of choices and trade-offs particular to the research question at issue. If we merely transform raw texts into ordered sequences of words, for example, the dimensionality of such a representation would be prohibitively high. The average U.S. Supreme Court opinion length on abortion, for example, is about 8,000 words.¹ Even assuming a simple vocabulary size of 10,000 words,² we would still need a space of $10^{78.062}$ dimensions to represent these documents. Such high dimensionality requires an equally inordinate number of datapoints in order to derive meaningful and accurate insights (Feinerer & Hornik 2008). Because it is impossible to obtain such a large number of datapoints, it is crucial to find alternative low-dimensional representations that still preserve all information germane to the research question at hand.

2. Representing Legal Texts

This section discusses several of the most common tools used by scholars to translate legal texts into data that can be subject to statistical analysis. We provide a brief introduction to the technique and discuss some of the work that applies the tool to study judicial behavior. This review is meant to be illustrative rather than exhaustive.

2.1. Length

Perhaps the most straightforward way to represent a text is simply the length of the document. This single figure obviously eliminates all of the semantic content in a document, but nevertheless captures an important characteristic that distinguishes documents based on their basic information content. Large variations in document length denote differences in format: differences in length are what functionally distinguish tweets from books. Smaller variations in document length may be associated with individual style or the depth of analysis in a document.

---


² See Brysbaert, Marc et al. 2016. How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant’s age. *Frontiers in psychology* 7: 1116.
There is considerable variation in the length of judicial opinions, and some interesting trends between and within institutions. The most extensive study of this feature in the U.S. Supreme Court is Black and Spriggs (2008). That study examines trends in opinion length over time and maps those temporal changes onto the changing role of law clerks over the history of the Court. That paper concludes that increasing reliance on judicial clerks was not responsible for longer opinions or opinions with more footnotes. Leonard & Ross (2016) analyze opinion length at the level of state supreme courts in the United States to examine features at the case and institutional level associated with longer opinions. At the case level, characteristics such as amicus participation are associated with longer opinions. Institutionally, the authors find some systematic differences in the factors associated with opinion length between elected, merit selected, and appointed judges. In terms of the consequences of opinion length, Green and Yoon (2017) found correlation between the opinion lengths of the Supreme Court of India and future citations.

The length of opinions is not the only way to derive useful data from word counts. In an interesting use of length features of documents, Patton & Smith (2017) parse transcripts of oral arguments before the U.S. Supreme Court into separate utterances, the length of which is used to measure the impact of attorney gender on the frequency with which they are interrupted by the Justice.

2.2 Readability Scores

Another simple but commonly used metric to represent legal documents is a “readability score.” Generally, readability scores are intended to capture linguistic sophistication, accessibility to a broader audience, or degree of clarity. The Flesch Kincaid Grade level is a readability score that is broadly familiar. This metric is calculated based on a weighted sum of sentence length (total words divided by total sentences) and word length (total syllables divided by total words). Larger numbers are associated with higher “grades,” and are intended to denote documents that are more difficult to read and understand.

Black et al. (2016a) calculate readability scores of U.S. Supreme Court majority opinions to examine how the Court uses the rhetorical opportunity of written opinions to affect public discourse. The authors hypothesize that the Justices will attempt to influence the public by writing more clearly when they anticipate negative reactions to their opinions. The authors calculate a range of different readability scores—which are primarily based on sentence length, word length, and the sophistication of the vocabulary used—and take the first principal component, which explains roughly three quarters of the variation between the different metrics. They compare this readability score with polling data on preferences for more or less government. Overall, the authors find that Justices write more clearly when the ideological direction of an opinion conflicts with the sentiment of the public. The same group expanded their study in Black et al. (2016b) to estimate the relationship between readability and several other factors, including the parties to litigation and ideological distance with lower courts. Somewhat in contrast to the findings in Black et al. (2016a), Owens et al. (2013) finds that opinions are less readable when there is greater ideological distance between the Court and Congress.

Whalen (2015) examines similar data, focusing on broad temporal trends and associations with Justice ideology. That paper finds that the Court is trending toward less readable opinions, and also finds a mild correlation between reliability and ideology, with conservative Justices issuing opinions that are more difficult to read. Expanding the pool of data to briefs before the Court, Coleman and Phung (2010) find a contrary trend in readability, with briefs becoming somewhat easier to read over time, especially in those sections in which lawyers have more stylistic flexibility.

Outside of the Supreme Court, Budziak, Hitt, and Lempert (2019) examine readability among other stylistic features of opinions in the U.S. appellate courts. In that study, the authors were interested in the degree to which variation in readability and other stylistic features were associated with judge-level and case-level characteristics, rather than strategic factors such as fear of reversal. Generally, the authors find that variation in readability is better explained by judge and case characteristics rather than strategic
factors. In the U.K., Scheinert & Tonkin (2021) examine the readability of asylum decisions as a metric of their accessibility.

Researchers have also used readability metrics to make cross-institutional comparisons. Feldman (2019) examines readability in the U.S. state and federal courts. Goelzhauser & Cann (2014) examine readability in the state courts to examine whether there is a relationship between institutional features at the state level (specifically whether judges are appointed or elected) and opinion readability. And Madden (2021), compared the readability scores between Australian, Canadian, South African, British, and American courts and concluded that factors such as average panel size could explain some variances in readability scores.

2.3 Curated Dictionaries

Both document length and readability metrics are far removed from the meaning of the words in a text. Moving to richer textual representations that are sensitive to content, many researchers have taken advantage of tools that are based on curated dictionaries of terms. The instances of these terms in a text can then be counted to derive supplemental measures of stylistic or content features of texts.

Sentiment analysis is one such approach. In a standard approach to sentiment analysis, a dictionary of positively and negatively valenced words is collected. Words such as “bad” “terrible” “disgusting” and “annoying” would be categorized as negative, while words such as “good” “useful” “delicious” and “delightful” would be categorized as positive. These dictionaries can be bespoke for a given task or corpus, but more general sentiment analysis dictionaries are also available. Texts are processed by counting the positive and negative words to determine the overall balance of sentiment. Sentiment is typically expressed as a single number, either expressing the percentage of a document’s words that have affective content (i.e., are on the list of positive or negative words) or over a one-dimensional space that might range over -1 (all negative words) to 1 (all positive words).

Several papers have applied sentiment analysis to judicial opinions. Examining U.S. Supreme Court cases, Carlson, Livermore, and Rockmore (2016) find that there has been an increase in the usage of negative language over time. Bryan & Ringsmuth (2016) measure the sentiment of U.S. Supreme Court dissents and find that cases with dissents that have more negative language tend to attract more media coverage. Rice & Zorn (2021) develop and examine a “semi-supervised” alternative method for the construction of specialized sentiment dictionaries, and test their approach using U.S. Supreme Court cases. Busch & Pelc (2019) apply sentiment analysis to the rulings of World Trade Organization panels and Appellate Body, finding that certain types of cases tend to elicit greater amounts of affect (either positively or negatively valenced words, compared to neutral words) and that effect tends to decline as an issue area is litigated over time.

Sentiment is one of many potential dimensions that can be extracted using a dictionary-based approach. The Linguistic Inquiry and Word Count (LIWC) tool has dictionaries for dozens of psychologically meaningful categories (Tausczik and Pennebaker (2010). The ease of use, flexibility, and relatively straightforward interpretability of LIWC has led many scholars of judicial behavior to take advantage of this tool. For example, Cross & Pennebaker (2014) examine opinions issued during the early years of the Roberts Court for several LIWC expert-generated categories of words, including sentiment, but also certainty and cognitive style. Black et al (2016c) uses LIWC to measure “emotional language” in briefs to the U.S. Supreme Court. Owens & Wedeking (2011) examine Supreme Court opinions using LIWC dictionaries that they associate with “cognitive complexity.” They find substantial inter-Justice variability, but also find intra-Justice variation based on opinion type (with dissents having less complexity than majority opinions) and subject area (with opinions in criminal procedure cases having less complexity than in other areas). Corely and Wedeking (2014) use the LIWC dictionary for “certainty” to examine the association between that variable in U.S. Supreme Court opinions and the treatment of those opinions by lower courts. In their study of the relationship between opinion language and judge-level, case-level, and strategic factors, Budziak, Hitt, and Lempert (2019) also use the LIWC
tool to characterize opinions in terms of the entire suite of available categories. Boston (2020) focuses on LIWC categories of impartiality, personality, and behavioral inclinations in a separate study of the strategic use of language in the appellate courts.

It is also possible to construct bespoke dictionaries for specific research projects. For example, Hinkle et al. (2012) use an automated program running keyword searches to detect hedging and intensifying language in a study of the effect of ideological distance between U.S. District Court judges and their appellate counterparts on lower court opinions. Varsava (2018) uses bespoke dictionaries to estimate informality, forms of referring to litigants, and use of intensifiers and hedge in the opinions of the 10th Circuit to compare the writing style of the then newly minted Justice Gorsuch to his peers. Smith (2014) develops a dictionary of terms associated with “legal review” and terms associated with “factual review” in a study of the relationship between strategic factors and the emphasis that judges place on legal or factual questions in their opinions. Narenchania (2022) develops a specialized terms list to examine the justifications offered by the Supreme Court for its decision to grant certiorari.

2.4 Term-Frequency Vectors

Curated dictionaries generally rely on expert knowledge to construct relevant word lists. This approach has the advantage of leveraging the considerable linguistic knowledge of the expert community and may also facilitate the interpretation of results. However, dictionary-based representations ignore the many terms that are not represented in their word lists and could mistakenly mischaracterize neutral legal phrases such as “burden of proof” as having positive or negative valance. In addition, curated dictionaries are limited to a fairly circumspect and predetermined set of dimensions, such as positive or negative sentiment or complexity.

An alternative to the dictionary approach of counting the appearance of certain selected groups of words is simply to summarize documents over their terms. This is referred to as a “bag-of-words” or term-frequency vector representation. The bag-of-words approach is one of the most common structural representations in law-as-data research. Under this model, documents are represented as vectors that describe the occurrence of words within a document (DeAngelo & McCannon 2017, Sag 2019). The order of words in a document is ignored, and the dimensions of the bag-of-words representation are dependent on the size of the vocabulary in a corpus.

Term frequency vectors are a specific type of n-gram model, where n is the length of the sequences of words. Typically, n is “1”, meaning that only unigrams – i.e. one word “sequences” – are considered. It is possible to have bigrams (n of “2”, i.e. two word sequences) or higher n-grams. Under such an approach, documents would be represented as frequencies over all of their unique 2-word-sequences. The dimensionality would be all of the two word frequencies in the corpus—a huge expansion over a simple unigram term-frequency vector. Due to this dimensionality explosion, unigrams are typically used, although some bigrams may be included.

The general lack of word order means that term frequency vectors are far from a lossless representation of a text. The meaning of certain words depends on their positions in the sentence and what their neighboring terms are: for example, whether a word is preceded by the term “not” can have a large effect on semantic content. A term-frequency vector does not provide sufficient information to determine the meaning of a text, but can provide a general impression of subject matter, given the types of words that are used.
Figure 1. Term Frequency of the top 15 words in the U.S. Constitution

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>shall</td>
<td>180</td>
</tr>
<tr>
<td>state</td>
<td>148</td>
</tr>
<tr>
<td>united</td>
<td>70</td>
</tr>
<tr>
<td>may</td>
<td>31</td>
</tr>
<tr>
<td>section</td>
<td>22</td>
</tr>
<tr>
<td>president</td>
<td>20</td>
</tr>
<tr>
<td>constitution</td>
<td>19</td>
</tr>
<tr>
<td>house</td>
<td>19</td>
</tr>
<tr>
<td>congress</td>
<td>18</td>
</tr>
<tr>
<td>one</td>
<td>17</td>
</tr>
<tr>
<td>two</td>
<td>17</td>
</tr>
<tr>
<td>person</td>
<td>15</td>
</tr>
<tr>
<td>law</td>
<td>14</td>
</tr>
<tr>
<td>time</td>
<td>14</td>
</tr>
<tr>
<td>office</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1 illustrates bag-of-words representations using the U.S. Constitution. Table 1 contains the term-frequency table of the top 15 words in the U.S. Constitution, where function words (i.e. those that describe grammatical or structural relations rather than content; e.g., “of”, “the”, “is”, and “a”) have been removed and the letters have been converted to lowercase. From this information, it is possible to understand that the document concerns governmental institutions and perhaps even constitutional matters. But it would be impossible to read off the actual process used to adopt legislation, or the content of the Bill of Rights.

A term-frequency vector is a different way of representing documents compared to the returns from LIWC or another curated dictionary. A sentiment analysis returns the count of all affectively loaded words, perhaps distinguishing between words with positive or negative valence. Documents are represented as a single number representing either the total degree of affective language, or the balance of positive, neutral, and negative language. By contrast, a term frequency vector is a list of many numbers, each associated with a term in the document.

Term frequency vectors need not be over the entire vocabulary of a corpus. For example, many analyses of writing style examine only function words. Several studies of judicial texts use term frequency vectors over function words. Carlson, Livermore, and Rockmore (2016) use the representation—sometimes referred to as a “stylistic fingerprint”—in a study of U.S. Supreme Court decisions. That analysis finds that there is a “style of the time” that holds together the writing of contemporaneous Justices; they also find that Justices’ individual writing styles have become less consistent over time (i.e. more intra-Justice variability) while the style of the Court has become more consistent (i.e. less inter-Justice variability). They associate this with the increasing role played by clerks in drafting opinions. Similarly, Frankenreiter (2019) also measured the use of function words to conduct a stylometric analysis of the European Court of Justice and Rosenthal & Yoon (2011) use a similar measure to predict authorship of judicial opinions.

Although term frequency vectors over function words have been used in several papers on judicial style, there are other measures that can be used to estimate stylistic characteristics of documents. For example, Wahlbeck, Spriggs, and Sigelman (2002) use a more general list of stylistic features (such as the type-token ratio, and the diversity of word length) in their study of the influence of law clerks on the opinions of U.S. Supreme Court Justices Marshall and Powell. Cheruvu (2019) calculates a measure
of lexical diversity in a study of whether the European Court of Justice’s French language mandate differentially affects the efficacy of judges from different backgrounds. Zubrod et al. (2020) utilized the Automated Integrative Complexity scoring system in an analysis of the complexity of the attorneys’ opening and closing statements of attorneys in well-known trials.

One of the benefits of the term-frequency vector representation is that it allows for useful mathematical interpretation. Geometrically, vectors with $k$ components can be understood as related to each other in a $k$-dimensional space. To represent a color based off the Red, Green, Blue (RGB) color scheme, for example, one would need a vector with three components. This vector would live in a 3-dimensional space, where the values of these components indicate how much red, green, and blue a particular color should contain. Orange, for instance, would be represented by the vector $(255, 69, 0)$, indicating that it is made up of 100% red, 64.7% green, and 0% blue.

A metric such as cosine similarity can be used to calculate the difference between vectors in geometric terms. The metric does so by measuring the cosine of the angle between the two vectors. This measurement allows researchers to focus on whether the substance of the vectors is similar. In our RGB example, the cosine similarity metric would yield a higher similarity score for orange $(255, 69, 0)$ and red $(255, 0, 0)$ than orange and blue $(0, 0, 255)$. For example, Hinkle (2016) uses cosine similarity between term frequency vector representation of opinions in the U.S. appellate courts to calculate their similarity. Oldfather, Bockhorts, and Dimmer (2012) used cosine similarity to explore the responsiveness of the U.S. Court of Appeals for the First Circuit’s opinions to briefs filed in that jurisdiction. Alternatively, term frequency vectors that have been normalized into percentages can be treated as distributions, and a measure such as Kullback-Leibler (KL) divergence can be used to calculate differences in statistical or information-theoretic terms. In their study of writing style on the U.S. Supreme Court, Carlson, Livermore, and Rockmore (2016) use KL divergence as their approach to estimating similarity.

Because term-frequency vectors over the entire vocabulary occupy a very high dimensional space, machine learning tools can be used to reduce that dimensionality in ways that allow for more useful statistical analysis. We discuss one such tool—topic models—in section 2.7 below. In using the bag-of-words model to transform judicial texts, researchers often perform some preprocessing. Questions to consider include whether 1) the capitalization of the words is important; 2) function words are needed; 3) inflected forms of the same word are different words (e.g., whether to stem or lemmatize words); 4) there should be special consideration for proper nouns and numbers; and 5) it is important to supplement the bag-of-words model with a weighting schedule such as the term frequency-inverse document frequency (tf-idf) scaling scheme. There are no globally accepted preprocessing best practices for text analysis; rather, different datasets and different research questions require different combinations of preprocessing options (Grimmer & Stewart 2013).

2.5 Parsers

Text parsers are software that separate a corpus of texts into smaller components based on a set of preprogrammed rules or logics. A text parser usually takes as inputs raw, unstructured texts and breaks them into smaller chunks (i.e., tokens) for analysis. After tokenizing the unstructured texts and transforming them into structured data, a parser can then annotate and categorize them according to some structural or grammatical constraints. A relatively straightforward example of this annotation would be part-of-speech tagging, where the tokens are assigned a grammatical category (e.g., noun, verb, adverb, adjective). “Text parsers are easy to understand” would be tagged as “Text” (adjective), “parsers” (noun), “are” (verb), “easy” (adjective), “to” (adverb), “understand” (verb). Once the texts are tokenized and

---

3 The term-frequency-inverse document frequency (tf-idf) scheme is a way to represent texts that reflect how important a word is to a document. The tf-idf value increases proportionally as the number of times a word appears in a document increases and is offset by the number of documents in a corpus that contain said word.
annotated, parsers could then use customized rules or logic to determine the syntactic relationships between the annotated tokens.

Ashley & Brüninghaus (2009) uses a text parser to extract information from the facts of already decided cases that can be used to predict the outcomes of new cases. To do so, they modified their text parser to identify special adjectives and annotate segments containing these adjectives with specific markers for later analysis. The authors then showed that the modified approach eventually yielded a significantly better prediction than the bag-of-words approach. Despite this advantage, supplementing bag-of-words output with expert knowledge might be overly burdensome. If the current selection of automated parsers does not support such annotations, scholars would have to hand code the semantic structures of the interested languages themselves. Additionally, adding semantic context to the bag-of-words model can further increase the dimensionality of the dataset and make the data unwieldy.

2.6 Anti-Plagiarism Software

Anti-plagiarism software is designed to detect textual and stylistic similarities between two or more corpus of texts and prevent the unauthorized appropriation of ideas, words, or results. Most anti-plagiarism software measure similarities through a combination of keyword, syntactic, and semantic analysis. This approach not only looks for literal textual overlaps, but also considers structural (e.g., part of speech) and conceptual (e.g., passive vs active voice) similarities. In addition to these linguistic features, some anti-plagiarism software also uses stylometric analysis to quantify writing styles (e.g., frequency of interpolation and passive voice) and compare stylistic similarities between documents. These similarity measures enable scholars to establish authorial fingerprints, which helped them investigate, for example, the extent of William Shakespeare’s collaboration with John Fletcher and Christopher Marlowe.

Textual and stylistic similarities could also yield novel insights into the judiciary. Choi & Gulati (2005), for example, used plagiarism software to analyze the authorship of judicial opinions. Similarly, Corley (2008) and Corley, Collins, and Calvin (2011) have also used various plagiarism software to track the influence of lower court opinions and parties’ briefs on U.S. Supreme Court opinions (Corley 2008, Corley, Collins, and Calvin 2011). Because these off the shelf tools are designed for different purposes, researchers using these domain-specific tools should ensure that the tools could be used to answer the research questions at issue.

2.7 Topic Models

A topic model is an unsupervised machine learning algorithm that identifies latent subject matter categories in corpora of texts (Blei 2012, Blei & Lafferty 2007). Given a textual corpus, a topic model produces topics, which in the technical topic modeling sense are probability distributions over a vocabulary, where each word in the vocabulary is assigned a non-negative weight (such that all weights sum to one). Each document is in turn summarized as a probability distribution over the topics. The highest-weighted words within a topic provide a sense of the subject matter that the distribution represents. For example, in a study of U.S. Supreme Court opinions, a topic labeled by the authors “trusts and estates” had the following highly weighted words: “estate, trust, death, property, descendent, wife, interest” (Livermore, Riddell, and Rockmore, 2017). The representation of a given document as a distribution over topics summarizes the document as a weighted mixture. These distributions – both of the topics and the words they comprise – are produced as the best fit to an underlying generative probabilistic model for the observed simple word frequencies.

A primary advantage of the topic models over simple term frequency vectors is that they can account for word co-occurrence, which helps to uncover characteristics of context. They also substantially reduce the dimensionality needed to describe documents within a corpus. A given group of documents

---

may contain tens of thousands of unique words. To represent documents as frequency distributions over those words requires vectors with an equal number of components. The number of topics generated by a topic model is much smaller, typically fewer than 500.

The most well-known topic model is a latent Dirichlet allocation (LDA) mixed-membership model (Blei, Ng and Jordan 2003). The LDA topic model served as the basis for a substantial research paradigm in natural language processing, which has sought to improve on its performance and expand its usefulness. For example, dynamic topic models are intended to detect how themes change over time (Blei & Lafferty 2006), and the structural topic model is optimized to test metadata covariates that are associated with different topics (Roberts et al. 2014).

Topic models have been used extensively in the study of courts and judges. Livermore, Riddell, and Rockmore (2017) combine a topic model with a machine learning classifier to examine how the contents of U.S. Supreme Court opinions evolve over time, rendering them less court-like as compared to the opinions issued by the U.S. appellate courts. Rice (2019) applies a topic model to the U.S. Supreme Court and argues that this more fine-grained representation could provide a better metric for empirically analyzing judicial opinions than traditional dichotomous representations of legal opinions. Rice (2017) also used topic modeling to extract two proxies — topic concentration and topic dissimilarly — and used them to investigate whether and how dissenting opinions at the U.S. Supreme Court shaped the scope of the majority’s opinions. Lauderdale & Clark (2014) supplemented the voting data of U.S. Supreme Court Justices with results from a topic model to better describe the political cleavages between Justices. In a similar vein, Macey & Mitts (2014), and Fagan (2015) use topic models to describe the inner workings of US courts.

Outside the United States, Keydar (2020) applied topic modeling on trial transcripts from the ICTY’s Kunarac case to examine trends throughout the proceedings against Bosnian-Serb forces. In a similar fashion, Shen-Bayh (2018) used topic modeling to differentiate between theoretical versus actual jurisdiction of autocratic courts in Africa. Grajzl and Murrell (2021a,b) relied on the results generated by topic modeling techniques to describe how ideas associated with the English financial revolution shaped the sources of law used in the English courts. Dyevre and Lampach (2021) used topic modeling to determine that there is a correlation between more inclusive annulment and referral procedures and greater issue heterogeneity in the decisions of the European Court of Justice. Similarly, Carter, Brown, and Rahmani (2016) also used topic models to describe the Australian High Court’s judicial workload and analyze the relationships between the court’s legal subject matter over time. Finally, Destrooper (2018) applied topic modeling techniques to the languages of both Cambodian NGOs and the Extraordinary Chambers in the Courts of Cambodia (ECCC), analyzed the results, and showed that there is a correlation between the discursive priorities of Cambodian human rights NGOs and what the ECCC focus on in cases appearing before it.

2.8 Supervised Machine Learning

Researchers have employed machine learning algorithms to classify legal observations or predict the outcome of legal cases. There are two main families of machine learning approaches — supervised and unsupervised — both capable of handling high-dimensionality datasets. In supervised machine learning, the models are trained with datasets that have been labeled by humans. After the models are trained, they could be used to classify new data and predict outcomes. Naïve-Bayes classifier and Support Vector Machine (SVM) are two examples of supervised machine learning. Conversely, in unsupervised machine learning, there is no labeled data. Instead, the algorithms are used to analyze and discover patterns without the intervention of humans. The topic modeling approach discussed earlier is an example of unsupervised machine learning.

Many legal scholars have used supervised learning tools to categorize legal texts. Daniels & Rissland (1997) employed a supervised learning algorithm to identify relevant passages within applicable judicial opinions. Gonçalves & Quaresma (2003) explore the use of SVM to classify European
Portuguese legal texts and propose ways to reduce the computational complexity of the task. Later work has sought to identify the ideological slant of amicus curiae briefs (Evans et al. 2007) and judicial opinions (Hausladen, Schubert, and Ash 2020), to explore various types of classifiers to categorize German judicial writings (Urchs, Mitrović, and Granitzer 2020), and to apply deep learning to recognize and extract the reasonings relevant to the holdings of the Supreme Court of Taiwan (Liu & Chen 2019).

In addition to classification, legal scholars also applied machine learning techniques to predict the outcomes of legal cases (Alschner & Charlotin 2021). One of the earlier precursors to such techniques could be found in the work of Brüninghaus & Ashley (2003). This paper created a rule-based algorithm to predict the outcome of trade secret misappropriation cases. Aletras et al. (2016), Varga et al (2021), and Haidar et al. (2021) use similar techniques to predict outcomes in the European Court of Human Right, Slovakian criminal proceedings, and accident cases in Moroccan Courts (respectively). A major limitation of all of these studies is that they draw information from the fully published opinions, rather than information that was available prior to the decision being issued. Accordingly, they may be better understood as predicting the justifications offered by courts for their decisions, rather than the outcome itself.

Supervised machine learning can also be used as an aid in broader study of the courts. Rice (2014), for example, uses naive Bayes, maximum entropy, and decision tree classifiers to predict the issue areas of the published courts of appeals opinion in a study of the relationships between the U.S. Supreme Court’s attention to a particular issue and the rise of new legal arguments for that issue in the lower federal courts. Livermore, Riddell, and Rockmore (2017) use a simple classifier (logistic regressions) to determine the prediction accuracy of a model trained to distinguish the topic representations of Supreme Court and appellate court opinions.

2.9 Embeddings

To capture word order and other semantic context while avoiding the dimension explosion problem, some scholars have turned to a relatively new text representation model called word-embedding. Word-embodiments is a technique to represent the words within a vocabulary as vectors in a relatively low dimensional space. These vectors are constructed based on a predictive task in which the words in a document are used to predict other words in that document. For example, the word2vec model developed by Google is based on a skip-gram task in which the goal is to use words as inputs to predict their context (i.e. near neighbor words) as outputs. These predictive tasks are used to train supervised machine learning models, and in particular artificial neural networks, which contain hidden layers that perform matrix translations on the input layer to project them into an alternative k-dimensional space. The projection is the heart of the embedding and serves as the vector representation of each word. These vectors have been found to encode a substantial amount of semantic information and have been used in sophisticated natural language tasks such as machine translation. Word embeddings have been generalized to represent entire documents as vectors by including additional metadata (here document tags) in the model.

While embeddings are only just beginning to be used in legal scholarship, several projects help demonstrate its promise. Ash & Chen (2019) describes initial results from an embeddings analysis of U.S. appellate courts to describe relationships between embeddings and judge features, such as political party affiliation and law school. Nyarko & Sanga (2022) use a word-embedding model to compare how legal terms of art (the terms “reasonable” and “consent”) are used differently by judges and laypersons. Additional projects include Rice, Rhodes, and Nteta (2019), which used word embedding to find racial bias against African American litigants in over one million U.S. appellate court opinions. Choi, Harris, and Shen-Bayh (2022) study the effects of judge and litigant ethnicity on case outcomes and then use word embedding to analyze written judgments to assess the mechanisms through which judge and litigant ethnicity influence judicial decision making. Ash & Chen (2018) use word embeddings to extract features of the U.S. Supreme Court Justice Kavanaugh’s past judicial rulings to determine his ideological
leanings. And Ash, Chen, and Naidu (2022) use word embeddings of judicial opinions in a study of the influence of economics training on the content of judicial opinions.

As with most new computational techniques, scholars should be aware that the current word embedding models produce more accurate results for English and other Western European languages. Even for those languages, however, word embedding techniques could not handle unknown or out-of-vocabulary words. Additionally, certain word embedding programs do not recognize shared representations at sub-word levels. For example, word2vec, a word embedding algorithm, will treat a word that ends in “less” as dissimilar to the same word without it (e.g., “flawless” being distinct from “flaw”).

3. Challenges and Future Research Avenues

Computational text analysis techniques empower legal scholars to derive novel insights into the legal system. All the advantages of computational tools, however, come with their own limitations. These limitations are either the consequences of the software design process or data availability or a combination thereof.

The majority of these tools were developed by English-speaking technologists for the analysis of English texts. This leaves the tools woefully inadequate to evaluate non-English documents. In the field of sentiment analysis, for example, the number of English lexicons dwarfs those for other languages (Lo et al. 2017). Without a readily available labeled sentiment dataset for a specific language, it is difficult to use sentiment analysis techniques for that language. Translation of a non-English language into English could only take you so far, as the act of translation could reduce the accuracy of the results (Franky and Veselovska 2019, Bakliwal, Arora, and Varma 2012, Perez-Rosas, Banea, and Mihalcea 2012). In addition to being language-specific, these tools are also domain-specific (Perry & Benoit 2017). This means that sentiment analysis tools developed for one area (e.g., consumer analysis) might not produce accurate results for another (e.g., legal evaluation).

Another limitation of computational techniques involves bias in underlying data. Computer scientists often describe this phenomenon as “garbage in, garbage out,” which expresses the idea that flawed input would inevitably produce flawed output. In tools involving machine learning, for example, biased inputs could cause the machine learning models to make erroneous assumptions and produce undesirable outputs (Chau 2022). Carlson, Livermore, and Rockmore (2020) have found strong evidence that judicial attributes influence publication decisions in the U.S. courts of appeals, raising serious questions about the credibility of a long line of research that uses published opinions to draw inferences concerning the causal effect of judicial attributes and case outcomes.

The continually evolving computational landscape warrants some caution in using state of the art textual processing techniques. Because the outputs of these algorithms could be difficult to interpret, legal scholars should familiarize themselves with these cutting-edge algorithms. For example, researchers have cautioned that the difficulty of interpreting topic models creates important limitations on their appropriate use (Shadrova 2021; Caspi & Stiglitz 2020).

An important step for the field will be the expanded use of text analysis tools for languages other than English. Although important research has been done on legal systems around the world, the bulk of research using computational techniques to study courts has been in the United States. Comparative computational analysis of law research would allow researchers to study factors such as geography, national boundaries, social milieu, and language differences that are held constant in uni-cultural studies. In doing so, they would gain a better understanding of a particular regional phenomenon and avoid the pitfall of over-generalizing results that are only applicable to a particular region or set of circumstances (Epstein, Šadl, and Weinshall 2021).

Work Cited


Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55, no. 4: 77-84


Bryan, Amanda C., and Eve M. Ringsmuth. 2016. "Jeremiad or weapon of words?: the power of emotive language in Supreme Court dissents." *Journal of Law and Courts* 4, no. 1: 159-185


Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21, no. 3: 267-297


Lo, Siaw Ling, Erik Cambria, Raymond Chiong, and David Cornforth. 2017. "Multilingual sentiment analysis: from formal to informal and scarce resource languages." *Artificial Intelligence Review* 48, no. 4: 499-527


