Predictive Features of Persuasive Legal Texts

Karl BRANTING, [a,1] Elizabeth TIPPETT [c], Charlotte ALEXANDER [b], Sam BAYER [a], Paul MORAWSKI [a], Carlos BALHANA [a], Craig PFEIFER [a],

[a] *The MITRE Corporation, McLean VA, USA*
[b] *Georgia State University, Atlanta, GA, USA*
[c] *University of Oregon, Eugene, OR, USA*

Abstract. This paper explores the degree to which variations in the citation patterns, style, and textual content of legal briefs on motions for summary judgment are predictive of rulings on those motions. In an empirical evaluation on a corpus of briefs in support of motions for summary judgment, the most predictive features were several novel graph metrics, including the characteristics of a brief's two-hop neighborhood in a bipartite document/precedent citation graph, and the vertex degree in an Implicit Citation Graph. These results indicate that strong and weak briefs differ systematically in citation patterns. Eight stylistic features were also identified as associated with success (a ruling consistent with the brief), and nine other features were found to have no correlation. Finally, prediction based on text features, such as ngram models, was found to be only weakly predictive of success.

1. Introduction

Recent research has shown that decisions set forth in published opinions can often be predicted by machine-learning models trained on the texts of the statements of facts in those opinions [1] [2] [3], suggesting the feasibility of automating key portions of the development of systems for decision support, judicial caseload management, and many other applications of legal prediction. However, critics of this work have observed that the statements of fact in published opinions are typically highly selective summaries of the original case record, tailored for consistency with the decision. More realistic scenarios would base legal predictions on case records in the form on which adjudicators themselves actually base their judgments.

One such scenario is legal prediction based on attorneys' persuasive writings, such as legal briefs. Attorneys present fact patterns within the context of relevant case law, and then make legal arguments in favor of the outcome most favorable to their client. Decision prediction based on attorneys' briefs may therefore be a closer approximation of the actual task faced by courts and other adjudicators than prediction based just on fact statements.

[1] Corresponding Author: E-mail: lbranting@mitre.org.

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

While it is unrealistic to expect machine-learning model trained on briefs and decisions to fully evaluate the merits of legal claims, there may nevertheless be many characteristics of briefs that make them more or less persuasive to judges. Identifying features that affect the persuasiveness of briefs could be useful for attorneys (by providing feedback to improve litigation effectiveness), political scientists and jurisprudential scholars (to increase understanding of legal processes), litigants (to improve understanding of factors needed for success), and courts (to identify briefs for triage, benchmarking across judges, etc.).

This paper explores the degree to which variations in the citation patterns, style, and textual content of legal briefs are predictive of subsequent rulings. To reduce confounding factors, we focus on a single subject matter area—federal employment law—and a single procedural setting—motions for summary judgment. A corpus of summary judgment briefs and decisions is described in Section 2. Several novel applications of graph analysis to corpora of briefs are set forth in Section 3, showing that strong and weak briefs differ systematically in citation patterns. Section 4 shows that stylistic factors in briefs, while less predictive citation patterns, nevertheless have a measurable influence of rulings, and Section 5 shows naive approaches to ruling prediction based on the text of briefs are only weakly predictive. The implications and suggestions for future research are set forth in Section 6.

2. The Summary Judgment Corpus

The Summary Judgment Corpus (*Corpus*) consists of a random sample of 864 federal employment cases involving summary judgment motions in the years 2007-2018. The cases were drawn from the PACER[2] system, via Bloomberg dockets, and are limited to those having one of the following Nature of Suit codes: "Civil Rights – employment," "Labor – Fair Labor Standards Act," and "Labor – Family and Medical Leave." The experiments described here were limited to 444 cases that include at least an initial brief and an opposition brief (including reply and surreply briefs, if any) and for which the motion for summary judgment was either granted in full or denied in full (thus finessing the complexities of motions granted in part and denied in part). A team of law students downloaded the briefs and opinions, reviewed each opinion, and coded the result of the ruling. In 98% of the cases, the defendant/employer filed the motion for summary judgment.[3] There is significant class skew in the decisions, with about 76% of motions for summary judgment being granted,[4] so in our experiments we evaluate accuracy using Matthews Correlation Coefficient (MCC) [4] since it is generally a more informative measure of predictive performance on skewed data sets than alternative measures. How-

[2]https://pacer.uscourts.gov/

[3]Plaintiffs (employees) make up a somewhat higher proportion of movants overall, but they are overrepresented among cases where the judge partially granted the motion in part and denied it in part. Due to the procedural posture of motions for summary judgment, it is nearly impossible for the plaintiff in a case to win an entire case on summary judgment.

[4]This figure is not representative of the success of summary judgment motions overall, as it excludes motions that were granted in part and denied in part.

ever, we include the more familiar F-measure as well. The class skew means that classification based on the majority rule (i.e., if movant, predict "win"; if respondent, predict "lose") achieves an MCC of 0.481 and a frequency-weighted F-measure of 0.740.

We distinguish two perspectives on formalizing decisions as instances for machine learning:

- *Brief-oriented.* In this perspective, each instance consists of a set of features derived from the briefs filed by a single party, together with the ruling on the motion that the briefs addressed. The ruling on the motion is labeled as a "win" or "lose" based on whether the brief supported the party who filed the summary judgment motion (the *movant*) or the party who opposes the motion (the *respondent*). For example, if the court grants the motion for summary judgment, the movant's brief would be labeled a "win" and the respondent's a "lose." Conversely, if the court denies a motion for summary judgment, the respondent's brief in opposition to the motion would be labeled a "win," and the brief by the movant a "lose."
- *Case-oriented.* In this perspective, each instance consists of the features derived from both the initial brief and the opposition brief, and the label consists of "granted" or "denied." In this approach, it is generally necessary to tag each feature to distinguish whether it came from the movant's or the respondent's brief.

In the discussion below the predicted "outcome" is either "win/lose," in brief-oriented experiments or "grant/deny" in case-oriented experiments. All results were calculated in 10-fold cross validation.

3. Prediction from Citations

In persuasive legal writing, lawyers cite cases to support particular legal propositions in their briefs. For example, a defendant/employer might assert that resigning is legally distinct from being fired, and then cite a case in which a court distinguished resigning from being fired.[5] This suggests that citations can be viewed as proxies for legal arguments and that the relative effectiveness of alternative arguments can be estimated by measuring the relative degree of association between the citations representing those arguments and outcomes. We therefore performed a series of experiments involving outcome prediction from citations. We used a modification of the the CourtListener[6] citation finder code to identify all citation spans in our corpus.

3.1. Citation Frequency Vectors

The first hypothesis that we tested was that outcomes could be predicted using a straightforward representation of briefs as citation frequency vectors, i.e., as features in which each value represents the number of times that a particular precedent was cited in a given brief. We created a brief-oriented data set in which each movant and respondent brief was

[5] See e.g. Iovanella v. Genentech, Inc., Case 2:09-cv-01024, Defendant Genentech, Inc's Memorandum of Law in Support of its Motion for Summary Judgment, Document 37-1 (April 9, 2010) (case analyzed in corpus).

[6] https://www.courtlistener.com/

Features	MCC	Mean F1
15,826 unique citations	0.152	0.563
The 100 highest-IG citations	0.401	0.611

Table 1. Performance in win/lose prediction of 675 movant and respondent briefs based on citation frequency vectors using the Weka implementation of SMO.

Features	MCC	Mean F1
Cosines of citation frequency vectors	0.127	0.683

Table 2. Performance in grant/deny prediction based on the cosine similarity of citation frequency vectors between (1) movant and respondent, (2) movant and the court, and (3) respondent and the court.

represented by a sparse vector of 15,826 unique integer citation features. As shown in the first row of Table 1, the predictive accuracy based on this representation is relatively low, although greater than chance. This weak performance may be unsurprising given that the number of features is an order of magnitude greater than the number of instances.

The hypothesis that data sparsity contributed to low prediction performance suggested an experiment in which the feature set was reduced to the 100 highest information gain (IG) citations. As set forth in the second row of Table 1, this feature reduction boosted MCC to 0.401. Inspection of the citations with the highest IG indicated that they were generally proxies for fatal factual defects in a plaintiff's case. Each one essentially justified dismissing cases that fall into a certain fact pattern. Citation frequency vectors are more predictive for the simpler tasks of predicting whether a brief is by the plaintiff or defendant (MCC = 0.420, F1 = 0.690) and of predicting whether the brief is by the movant or the respondent (MCC = 0.450, F1 = 0.694).

3.2. Citation Comparisons

A second hypothesis is that it is the relationships among the citations by the parties and the court, rather than the individual citations themselves, that are predictive of outcomes. Perhaps overlap in citations between a court's ruling and a party's brief might be a proxy for the relative strength of the arguments in that brief. For example, a judge who is persuaded by the arguments in a brief might choose to incorporate the citations in support of those arguments.

To test this hypothesis we calculated the cosine between the citation frequency vectors of three pairs: (1) movant and respondent; (2) movant and the court; and (3) respondent and the court. We evaluated grant/deny accuracy on the resulting case-oriented data set in which each case is represented by these three cosine values. As shown in Table 2, predictive performance was weak, suggesting that simple citation vector similarity is not a good proxy for argument strength.

3.3. Graph Analysis

There has been extensive research on graphs derived from collections of homogeneous documents linked by embedded citations, such as Supreme Court decisions [5] and codes of statutes or regulations [6]. This approach to graph modeling is not directly applicable

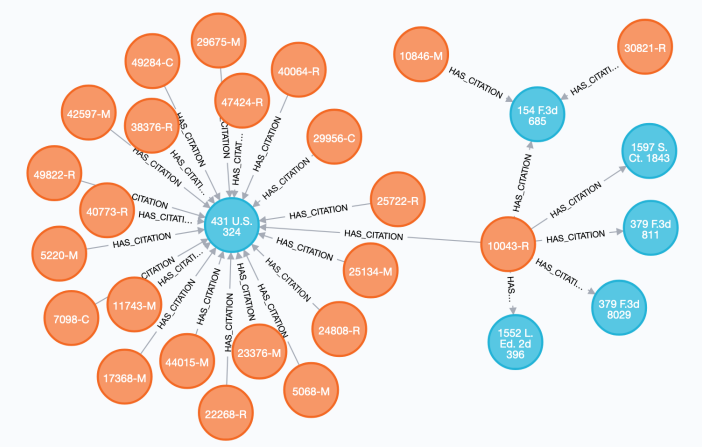

Figure 1. A bipartite graph representation of the Corpus in which nodes are case documents and precedents, and each edge connects a document to a precedent that it cites.

to the Corpus, in which the documents that contain citations (briefs) differ from those that are cited (precedents). We therefore experimented with two novel graph representations: a bipartite citation graph, consisting of both document (brief or decision) and precedent nodes linked by "has-citation" edges; and an implicit citation graph, in consisting of briefs that are connected if they share a common citation. We implemented both graphs in Neo4j.[7]

3.3.1. Bipartite Graph Experiments

Figure 1 shows a fragment of a bipartite graph representation of the Corpus in which each node represents a document (brief or decision) or a precedent and each edge connects a document to a precedent that it cites. The label on each brief node is an internal index followed by "-M," "-R," or "-C," denoting movant documents, respondent documents, or court documents, respectively.

The bipartite graph representation is conducive to analytics that exploit locality in citation space, that is, that compare briefs in terms of similar citation behavior. For each brief in our graph, we derived the following features:

1. prob2HopWin. For a given brief of type "R" or "M", the percentage of all two-hop brief nodes of the same type whose win/lose value is "win."
2. numLevel0Cites. The number of a brief's citations (i.e., vertex degree in the bipartite graph).
3. avgLevel1Cites. The mean number of citations to each precedent that the brief cites (i.e., mean vertex degree of a brief's 1-hop neighbors).
4. numSharedCites: The number of citations shared with the brief of the opposing party (i.e., the number of 2-hop paths between a brief and the opposing side's brief).

[7] https://neo4j.com/

Graph Feature	Movant	Respondent	All
avgCiteWinScore	0.037	0.054	0.207
avgCiteGrantScore	0.037	0.054	
numSharedCites	0.033	0.032	0.022
prob2HopWin			0.179
prob2HopGrant			0.086
numLevel0Cites			0.035
avgLevel1Cites			

Table 3. Information gain of graph features in win prediction for briefs by movants and respondents, individually and in combination. Empty cells represent zero information gain.

Brief Type	Features	MCC	Mean F1
All	avgCiteWinScore, numSharedCites, prob2HopWin prob2HopGrant, numLevel0Cites	0.477	0.742
Respondent	avgCiteGrantScore & numSharedCites	0.189	0.664
Movant	avgCiteGrantScore & numSharedCites	0.177	0.715

Table 4. Performance in win/lose prediction based on the most predictive graph features for each set of briefs, using the Weka implementation of Multilevel Perceptron.

5. avgCiteWinScore. The mean citation win score of each cited precedent, where the citation win score is the percentage of briefs of the same type ("R" or "M") citing that precedent with value "win." The root brief is excluded from this calculation.
6. prob2HopGrant and avgCiteGrantScore: like prob2HopWin and avgCiteWinScore but based on "grant" rather than "win."

Table 3 shows the information gain from each of the graph features in win prediction for movant, respondent, and all briefs, and Table 4 shows win prediction performance on the highest information gain features for each of these 3 sets. It may be unintuitive that the most predictive features for all briefs differ from those for the movant and respondent briefs individually. However, as mentioned above, outcome prediction based on the party alone is sufficient for an MCC of 0.481. Evidently, the difference in citation behavior between movants and respondents is captured by the graph features with high information gain for the full set. However, prediction for movant and respondent briefs individually is conditioned on already knowing the value of the party, so graph features indicative of party status are less informative. The predictive performance for the movant and respondent briefs individually, shown at the bottom of Table 4 reflects the residual information from graph features after party status is factored out.

3.3.2. Implicit Citation Graph Experiments

Our second set of graph experiments involve a graph, derived from the bipartite graph, in which nodes consist of documents, and edges connect pairs of documents that cite a precedent in common. We refer to this representation as an implicit citation graph (ICG). A portion of this graph is shown in Figure 2.

Our first observation was that the ICG had 16 connected components, 15 consisting of singletons (a single brief each), and the remaining component containing all other

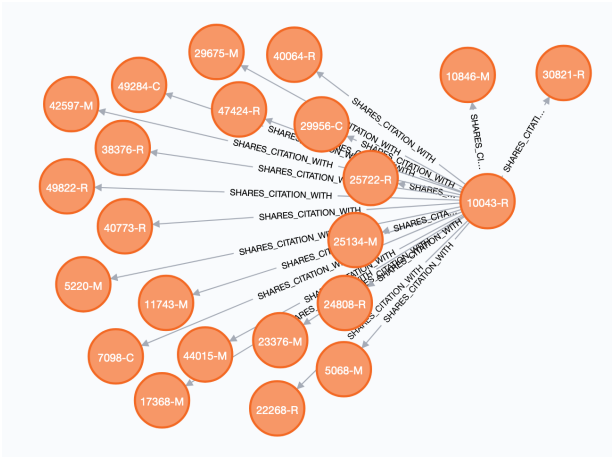

Figure 2. An implicit citation graph representation of the Corpus in which nodes are documents and precedents and edges connect pairs of documents that cite a precedent in common.

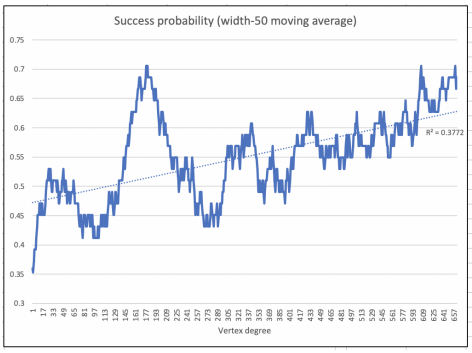

Figure 3. The mean likelihood of success over a width-50 moving window as a function of vertex degree in the Implicit Citation Graph.

briefs. The singletons were all unsuccessful, i.e., had value "lose." Intuitively, this indicates that citing only precedents that are cited by no one else is a very poor strategy. This observation suggests the hypothesis that there is a correlation between vertex degree in the ICG (i.e., the number of precedents a brief cites that are cited in some other brief) and the likelihood of success, and indeed we observed a correlation of 0.1278 between vertex degree and the win/lose decision represented as 0 or 1. This correlation can be visualized in a graph of the 50-element moving average of 0/1 values as a function of vertex degree, as shown in Figure 3.

This correlation is intuitive; citing idiosyncratic cases may be a sign that the lawyer lacks familiarity with relevant case law, or that their legal argument is so far-fetched that they must cite to remote cases. By contrast, lawyers who can draw parallels to a large corpus of precedent, either due to their expertise or the underlying merits of the case, are likely to fare better.

4. Stylistic Features

The prose style of briefs has been shown empirically to affect judges' assessments of persuasiveness and credibility [7]. We therefore measured the predictive effect (from a brief-oriented perspective) of a number of stylistic features that have been suggested as having a possible effect on judges' rulings:

1. doc count - the total number of documents filed by the party, i.e., 0, 1, or 2, depending on whether the party files an opposition, reply, or surreply
2. citation count - the number of citations in the brief
3. hedging - words associated with trying to explain away bad facts or bad arguments, e.g., "even assuming," "albeit"
4. hyperbolic language - use of terms like "blatant," "absurd," "egregious," etc.
5. sentence count - the number of sentences in the brief
6. legal amplifiers - use of terms like "conclusory," "inadequate," "irrelevant" etc..
7. repetition - terms such as "additional," "again"
8. total string cites - the number of instances of multiple consecutive citations
9. mean string cite length - the average number of citations in each string cite
10. jurisdiction - the district in which the case was brought
11. mean sentence length in tokens
12. negative emotional state - terms indicating negative affect, e.g., "upset," "cried"
13. jury request - whether a jury was requested
14. nature of suit - as specified in the PACER system
15. court - the court in which the case was brought
16. pro se - whether the party was self-represented
17. cause of action - as specified in the PACER system

Collectively these stylistic features are modestly predictive of a party's likelihood of success: MCC = 0.389 and F1 = 0.693 are achieved with the following rules (induced by jRip[8]):

```
(docs <= 1) and (citation_count <= 26)
        => won=no (296.0/72.0)
(docs <= 1) and (legal_amplifiers <= 28)
        => won=no (106.0/37.0)
(mean_string_cite_length >= 4.666667)
        => won=no (57.0/24.0)
 otherwise
        => won=yes (435.0/121.0)
```

This rule indicates that the best predictive performance can be obtained using just features 1, 2, 6, and 9 above.

We calculated the mutual information between each of the features and the win/lose decision. Features 10 through 17 in the list had negligible mutual information with outcomes, meaning that we were unable to detect any meaningful effect from these features.

[8] https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/JRip.html

The information gain from features 1 through 8 varied from 0.117 (doc count) to 0.031 (mean string cite length). We conclude that features 1–8 are meaningful and should be considered by attorneys when drafting briefs. However, it may be that these features are most significant in boundary cases, e.g., failing to file a reply or surreply may be an indication of lack of conscientiousness; too few citations may indicate lack of effort; lengthy string cites and hyperbolic language may indicate an attempt to compensate for a weak case or poor writing.

5. Text-Based Prediction

As mentioned in the Introduction, prior research has shown that text classification techniques sometimes perform well in predicting outcomes from judges' statements of the case. In contrast, recent work suggests that the techniques perform very poorly when applied to fact statements written by laypersons [8]. We hypothesized that applying these techniques to the text of briefs might perform at an intermediate level of accuracy.

We performed two experiments to test this hypothesis. Both experiments followed a case-oriented paradigm in which the decision was predicted from text produced by both the movant and the respondent.

In the first experiment, all text was extracted from the briefs submitted by each party and normalized by conversion removal of punctuation, numbers, and stop words. The text was converted to terms as described below, and a flag denoting the type of the party prepended to each term in a brief by that party, e.g., "M-" is prepended to each term from a movant brief, and "R-" is prepended to each term from a respondent brief. This convention permitted terms from one party to be distinguished from terms by the other. We experimented with three term representations: (1) binary (one-hot) 1–3 gram vectors, (2) 1–3 gram frequency vectors, and (3) lower-cased unigram frequency vectors. The macro-averaged MCC grant/deny prediction using these representations was 0.253, 0.123, and 0.136, respectively, indicating relatively weak predictively value.

The second experiment applied this procedure just to the parenthesized text that precedes citations, e.g.,

> (employee cannot establish pretext by asserting unsupported blanket denial) Irvin v. Airco Carbide, 837 F.2d 724 (6th Cir. 1987)

Typically, such a parenthetical succinctly expresses the proposition for which the precedent is being cited. One might surmise that, collectively, such parentheticals constitute the main arguments of a brief and that a model trained on these texts would, in effect, learn the relative effectiveness of these arguments. However, the macro-averaged for parenthetical using the 3 representations listed above were 0.032, 0.040, and 0.032, respectively, indicating predictive accuracy essentially equal to chance.

Briefs contain complex arguments and detailed references to the factual record, and a court's decision on a motion depends on both the nuances of legal and factual arguments set forth in briefs and on factors outside of the scope of the briefs themselves, such as the evolution of legal doctrine. It may be unsurprising that simple document classification techniques produce weak prediction results when applied to documents with such complex discourse structure.

6. Conclusion

This paper has shown that both the pattern of citations and various stylistic properties of briefs are associated with the likelihood of success of the underlying motion. Correlation is not causation, so it is not clear which of these factors are independent variables that could controlled by a litigant to improve the odds of success and which are the common consequence of factors that also control the outcome. Nevertheless, the results provide some insight into the influence of lawyering and legal writing on outcomes. The predictiveness of 2-hop neighborhoods in a bipartite graph is evidence that successful lawyers on each side tend to cite to a common set of cases, whereas less-successful lawyers tend to make idiosyncratic citations and to use hyperbolic language and other potential smoke screens for weaknesses in their case, their research, or their writing.

Distinguishing causal from merely correlated factors is the work of future analysis. However, the research described in this paper illustrates how machine-learning and graph analysis can be used to identify factors for further investigation, including distinguishing significant from insignificant stylistic variations and detecting latent graph characteristics that reveal hitherto unsuspected relationships between citations and decisions.

Acknowledgments

The MITRE Corporation is a not-for-profit company, chartered in the public interest. This document is approved for Public Release; Distribution Unlimited. Case Number 20-2944.

References

[1] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. CoRR. 2019; abs/1906.02059.

[2] Sulea O, Zampieri M, Vela M, van Genabith J. Predicting the Law Area and Decisions of French Supreme Court Cases. In: RANLP. INCOMA Ltd.; 2017. p. 716–722.

[3] Branting LK, Yeh A, Weiss B, Merkhofer EM, Brown B. Inducing Predictive Models for Decision Support in Administrative Adjudication. In: AI Approaches to the Complexity of Legal Systems - AICOL International Workshops 2015-2017, Revised Selected Papers. vol. 10791 of Lecture Notes in Computer Science. Springer; 2017. p. 465–477.

[4] Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. "PLoS ONE". 2017; 12(6). E0177678.

[5] Carmichael I, Wudel J, Kim M, Jushchuk J. Examining the Evolution of Legal Precedent Through Citation Network Analysis. North Carolina Law Review. 2017 December; 96(1):227–269.

[6] Whalen R. Legal Networks: The Promises and Challenges of Legal Network Analysis. Michigan State Law Review. 2016; 2:539–566.

[7] Benson RW, Kessler JB. Legalese v. plain English: an empirical study of persuasion and credibility in appellate brief writing. Loy LAL Rev. 1986; 20:301.

[8] Branting K, Balhana C, Pfeifer C, Aberdeen J, Brown B. Judges are from Mars, Pro Se Litigants are from Venus: Predicting Decisions from Lay Texts. In: Legal Knowledge and Information Systems - JURIX 2020: The Thirty-Third Annual Conference; 2020. To appear.