Automating Judicial Document Analysis

L. Karl Branting
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102, USA
lbranting@mitre.org

ABSTRACT

Collections of documents filed in courts are potentially a rich source of information for citizens, attorneys, and courts, but courts typically lack the ability to interpret them automatically. This paper presents technical approaches to three applications of judicial document interpretation: detection of document filing errors; matching orders with the motions that they resolve; and predicting the outcome of routine cases. In empirical evaluations on filings from two representative large US District Courts, the highest accuracy in identifying filing errors was achieved by combining procedural context features with high information-gain lexical features; TF/IDF similarity was found to be an effective criterion for finding motions that correspond to orders; and induction over the texts of prior simple and routine decisions was found to produce a model capable of accurately predicting outcomes from case facts without any manually engineered features or factors.

1 INTRODUCTION

The transition from paper to electronic filing in national, local, and administrative courts, which began in the late 1990s, has transformed how courts operate and how judges, court staff, attorneys, and the public create, submit, and access court filings. However, despite many advances in judicial access and administration brought about by electronic filing, courts are typically unable to interpret the contents of court filings automatically. Instead, court filings are interpreted only when they are read by an attorney, judge, or court staff member.

Machine interpretation of court filings promises a rich source of information for improving court administration and case management, access to justice, and analysis of the judiciary. The development of large-scale text analytics makes such interpretation increasingly feasible, as collections of court documents are, in effect, annotated by the metadata generated when they are submitted, by corrections when they are audited, or, for those documents that are motions or claims, by the decisions of judges or other decision makers.

However, there are numerous challenges to automating the interpretation of case filings. Courts often accept documents in the form of PDFs created from scans. Scanned PDFs require optical character recognition (OCR) for text extraction, but this process introduces many errors and does not preserve the document layout, which contains important information about the relationships among text segments in the document. Moreover, the language of court filings is complex and specialized, and the function of a court filing depends not just on its text and format, but also on its procedural context. As a result, successful automation of court filings requires overcoming a combination of technical challenges.

This paper describes the nature of court dockets and databases, sets forth three classes of representative judicial document analysis tasks–docket error detection, order/motion matching, and decision prediction–proposes technical approaches to each of the tasks, and presents preliminary empirical evaluations of the effectiveness of each approach.

2 COURT DOCKETS AND DATABASES

A court *docket* is a register of document-triggered litigation events, where a *litigation event* consists of either (1) a pleading, motion, or letter from a litigant, (2) an order, judgment, or other action by a judge, or (3) a record of an administrative action (such as notifying an attorney of a filing error) by a member of the court staff. Each docket event in a typical electronic case management system includes (1) metadata generated at the time of filing, including both case-specific data (e.g., case number, parties, judge) and event-specific data (e.g., the attorney submitting the document, the intended document type) and (2) a text document in PDF format. Each of the two court databases in which the experiments described below were performed contained filings for over 400,000 cases involving over 1,000,000 litigants, attorneys, and judges, over 10,000,000 docket entries, and more than 4,000,000 documents.

3 DOCKET ERROR DETECTION

There are many kinds of docket errors, including defects in a submitted document (e.g., missing signature, sensitive information in an unsealed document, missing case caption) and mismatches between the content of a document and the context of the case (e.g., wrong parties, case number, or judge; mismatch between the document title and the document type asserted by the user). Some errors consist of violations of a particular court's local rules and are therefore unique to that court. Other events, such as filing in a wrong case, constitute errors in any court. In either case, detection of defects at submission time could spare attorneys the embarrassment of submitting a defective document and the inconvenience and delays of refiling. For court staff, automated filing error detection could reduce the quality control (QC) auditing staff required for filing errors, a significant drain of resources in many courts.

3.1 Error Detection through Text Classification

In the court in which the first set of experiments were conducted, the QC staff review filings to detect a variety of docket errors, including the following four error types:

In: Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2017), June 16, 2017, London, UK.

Published at http://ceur-ws.org

- Event-type errors, i.e., specifying the wrong event type for a document, e.g., submitting a Motion for Summary Judgment as a Counterclaim. In experiments involving this court, there were 20 event types, such as complaint, transfer, notice, order, service, etc.
- Main-vs-attachment errors, i.e., filing a document, such as an exhibit, that should be filed as an attachment to another document, as a main document or filing a document, such as a Memorandum in Support of a Motion for Summary Judgment, that should be filed as a main document, as an attachment.
- Show-cause order errors. In some courts, only judges are permitted to file show-cause orders; it is an error if an attorney does so.
- Letter-motion errors. In some courts, certain routine motions can be filed as letters, but all other filings must have a formal caption. Recognizing these errors requires distinguishing letters from non-letters.

Event-type errors appear to be the most common docket errors in U.S. District courts.

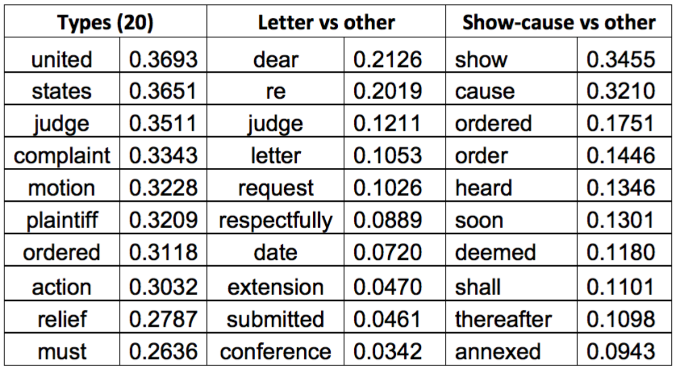
Each of these filing errors can be detected by classifying a document with respect to the corresponding set of categories (event type, main vs. attachment, show-cause order vs. non-show-cause order, or letter vs. non-letter) and evaluating whether the category is consistent with the metadata generated in the docket system by the filer's selections. Event-type document classification is particularly challenging both because document types are both numerous and skewed, having a roughly power-law frequency distribution in the test set.

The first set of experiments attempted to identify each of the four docket errors above by classifying document text and determining whether there is a conflict between the apparent text category and the document's metadata. Classification was performed with the lingpipe[1] LMClassifier, which performs joint probability-based classification of token sequences into non-overlapping categories based on language models for each category and a multivariate distribution over categories.

3.1.1 Term Selection and Document Truncation. Court filings can be thought of as comprising four distinct sets of terms:

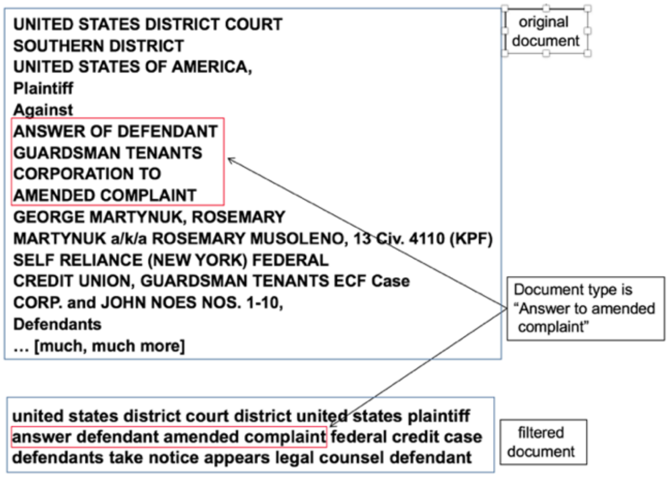
- Procedural words, which describe the intended legal function of the document (e.g., "complaint," "amended," "counsel")
- "Stop-words," which are non-content common words, such as "of" and "the"
- Words unique to the case, such as names, and words expressing the narrative events giving rise to the case; and
- Substantive (as opposed to procedural) legal terms (e.g., "reasonable care," "intent," "battery").

Terms in the first of these sets–procedural words–carry the most information about the type of the document. These words tend to be concentrated around the beginning of legal documents, often in the case caption, and at the end, where distinctive phrases like "so ordered" may occur.



Types (20)		Letter vs other		Show-cause vs other	
united	0.3693	dear	0.2126	show	0.3455
states	0.3651	re	0.2019	cause	0.3210
judge	0.3511	judge	0.1211	ordered	0.1751
complaint	0.3343	letter	0.1053	order	0.1446
motion	0.3228	request	0.1026	heard	0.1346
plaintiff	0.3209	respectfully	0.0889	soon	0.1301
ordered	0.3118	date	0.0720	deemed	0.1180
action	0.3032	extension	0.0470	shall	0.1101
relief	0.2787	submitted	0.0461	thereafter	0.1098
must	0.2636	conference	0.0342	annexed	0.0943

Figure 1: The information gain of the 10 highest-information terms for 3 legal-document classification tasks.



UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT
UNITED STATES OF AMERICA,
Plaintiff
Against
ANSWER OF DEFENDANT
GUARDSMAN TENANTS
CORPORATION TO
AMENDED COMPLAINT
GEORGE MARTYNUK, ROSEMARY
MARTYNUK a/k/a ROSEMARY MUSOLENO, 13 Civ. 4110 (KPF)
SELF RELIANCE (NEW YORK) FEDERAL
CREDIT UNION, GUARDSMAN TENANTS ECF Case
CORP. and JOHN NOES NOS. 1-10,
Defendants
… [much, much more]

united states district court district united states plaintiff answer defendant amended complaint federal credit case defendants take notice appears legal counsel defendant

Figure 2: Reduction of a full document to just high information-gain terms.

We hypothesized that only procedure terms are relevant to the type of a document, so we explored approaches to filtering non-procedure terms. Elimination of irrelevant terms can not only speed execution, but in some cases has been shown to increase accuracy [13].

Three approaches to term selection were investigated: two ad hoc and domain-specific; and one general and domain-independent. The first approach was to eliminate all terms except non-stopwords that occur in the Federal Rules of Civil Procedure [7]. A related alternative approach was to remove all terms except for non-stopwords occurring in "event" (i.e., document) descriptions typed by filers when they submit into the docket system. The third approach was to select terms based on their mutual information with each particular text categories [3]. The first lexical set, termed FRCP, contains 2658 terms; the second, termed event, consists of 513 terms. Separate mutual-information sets were created for each classification task, reflecting the fact that the information gain from a term depends on the category distribution of the documents. For example, Figure 1

[1] http://alias-i.com/lingpipe/

Table 1: Thresholds and size of large and small high information-gain term sets.

	showcause	main_attch	types	letter
ig_small	0.01 (135)	0.025 (262)	0.1 (221)	0.0005 (246)
ig_large	0.0025 (406)	0.0125 (914)	0.05 (689)	0.00001 (390)

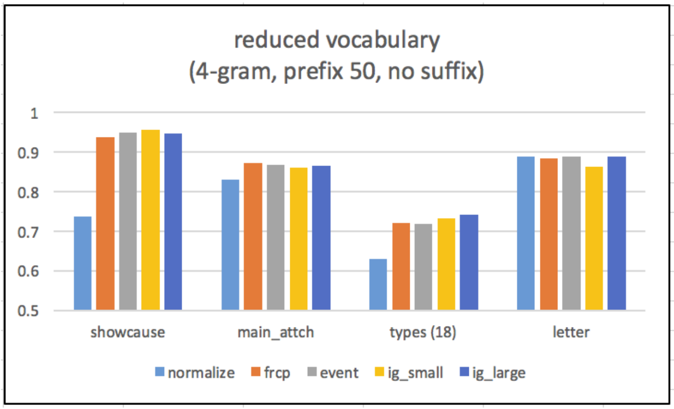


Figure 3: Classification accuracy as a function of reduced vocabulary (8-fold cross validation using a 4-gram language model, 50-token prefix length, and no suffix).

shows the 10 highest information terms for three different classification tasks: event-type classification, distinguishing letters from non letters, and show-cause order detection, illustrating that the most informative terms differ widely depending on the classification task.

Figure 2 illustrates the reduction of full document text to just high information gain terms, which typifies the vocabulary-reduction process.

Several approaches to document truncation were explored as well. The first was to limit the text to the first *l* tokens of the document (i.e., excise the remainder of the document). If *l* is sufficiently large, this is equivalent to including the entire document. A second option is to include the last *l* tokens of the suffix as well as the prefix.

3.1.2 Evaluation of Alternative Term Reduction Approaches. Two different information-gain thresholds were tested for each classification type, intended to create one small set of very-high information terms (ig_small) and a larger set created using a lower threshold (ig_large). The thresholds and sizes of the large and small high information-gain term sets are set forth in Table 1. The text of each document was obtained by OCR using the open-source program Tesseract [20]. Each text was normalized by removing non-ASCII characters and standardizing case prior to term selection, if any.

Figure 3 shows a comparison of four vocabulary alternatives on the four text classification tasks described above. These tests measured mean f-measure in 8-fold cross validation using a 4-gram language mode, 50-token prefix length, and no suffix. In the baseline vocabulary set, normalize, non-ASCII characters, numbers, and punctuation are removed and tokens were lower-cased. The results show that classification accuracy using an unreduced vocabulary was significantly lower than the best reduced vocabulary performance for show-cause order detection and type classification. Term selection had little effect on accuracy for the letter and main vs. attachment detection tasks. No reduced-vocabulary set consistently outperformed the others. This indicates that restricted term sets derived through information gain perform roughly as well as those produced using domain-specific information, suggesting that the reduced vocabulary approach is appropriate for situations in which domain-specific term information is unavailable.

Summarizing over the tests, the the highest mean f-measure based on text classification alone and the particular combination of parameters that led to this accuracy for each classification task were as follows:

(1) **Event type: 0.743** (prefix=50, 4-gram ig_large vocabulary, 20 categories)
(2) **Main-vs-attachment: 0.871** (prefix=256, 6-gram, event vocabulary)
(3) **Show-cause order: 0.957** (prefix=50, 5-gram, ig_small vocabulary)
(4) **Letter-vs-non-letter: 0.889** (prefix=50, no 4-gram, ig_large vocabulary)

3.2 Incorporating Procedural Context Features

The accuracy of event-type detection (f-measure of roughly 0.743 under the best combinations of parameters) is sufficiently low that its utility for many auditing functions may be limited. An analysis of the classification errors produced by the event-type text classification model indicated that a document's event type depends not just on the text of the document but also on its *procedural context*. For example, motions and orders are sometimes extremely similar because judges grant a motion by adding and signing an order stamp to the motion. Since stamps and signatures are seldom accurately OCR'd, the motion and order may be indistinguishable by the text alone under these circumstances. However, orders can be issued only by a judge, and judges never file motions, so the two cases can be distinguished by knowing the filer. In addition, attachments have the same event type as the main document in CM/ECF. So, for example, a memorandum of law is ordinarily a main document, but in some courts a memorandum can be filed as an attachment, in which case its event type is the same as that of the main document to which it is attached.

Contextual information potentially relevant to a document's type includes: whether it was filed as a main document or as an attachment; the filer (e.g., attorney, clerk, judge); the type of the case (e.g., criminal, civil, multi-district); and the document length (e.g., memoranda are typically long; minute orders typically short). Combining these non-lexical features with text features requires a different classifier than the language-model classifier used in the first set of experiments.

We compared the performance of SupportVector Machine (SVM) learning (WEKA's implementation of Platt's algorithm for sequential

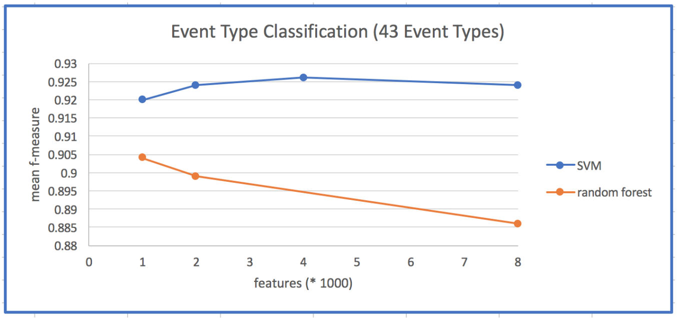


Figure 4: Event type classification accuracy as a function of reduced vocabulary (10-fold cross validation using a 2-gram language model, normalization of dates, numbers, and parties, 100-token prefix length, and minimum token frequency of 32, with 43 event types).

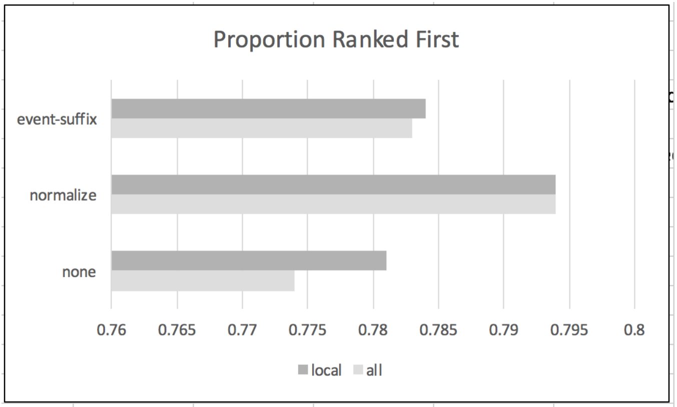


Figure 5: The proportion of groups for which the order is more similar to the triggering motion than to any other motion.

minimal optimization [10, 16]) and Random Forests [5], both in the WEKA [9] implementation, on the task of filing event classification. For each filing event, the document text was normalized by filtering stop words, normalizing dates and numbers to standard tokens, and replacing each instance of a party name with the role of that party (e.g., DFT, PTF). The result was combined with the contextual features and converted into a sparse n-gram frequency vector from which the {1,2,4,8} thousand highest information gain features were selected (unsurprisingly, the contextual features always had higher information gain than any lexical feature). The training set consisted of 28,763 main documents having 43 distinct types representing 2 month's filings in a large US District court.

As shown in Figure 4, the SVM was consistently more accurate, with little variation in accuracy as a function of the number of features, although accuracy was slightly higher at 4,000 features than other feature set sizes. By contrast, the accuracy of the random forest diminished with increasing numbers of features. The highest accuracy SVM configuration, f-measure of 0.926, was much higher than the maximum observed with text-only classification (albeit, in a different court). This suggests that including procedural context features is essential for accurate document filing type identification in judicial databases, and that algorithms that can handle both textual and categorical features should be used for this task.

4 ORDER/MOTION MATCHING

In many federal courts, docket clerks are responsible for filing orders executed by judges into the docket system, a process that requires the clerk to identify all pending motions to which the order responds and to link the order to those motions. This entails reading all pending motions, a tedious task. If the motions corresponding to an order could be identified automatically, docket clerks would be relieved of this laborious task. Even ranking the motions by their likelihood of being resolved by a given order would decrease the burden on docket clerks. Moreover, order/motion matching is a subtask of a more general issue-chaining problem, which consists of identifying the sequence of preceding and subsequent documents relevant to a given document.

A straightforward approach to this task is to treat order/motion matching as an information-retrieval task, under the hypothesis that an order is likely to have a higher degree of similarity to its corresponding motions than to motions that it does not rule on. An obvious approach is to present pending motions to the clerk in rank order of their TF/IDF[2]-weighted cosine similarity to the order.

The evaluation above showing that term selection improves document classification raises the question whether term selection might be beneficial for order/motion matching as well. A second question is whether the IDF motion should be trained on an entire corpus of motions and orders or whether acceptable accuracy can be obtained by training just on the order and pending motions.

To evaluate the effectiveness of this approach to order/motion match, a subset of the document set described above was collected consisting of 3,356 groups, each comprising (1) an order, (2) a motion that the order rules on (a *triggering motion*), and (3) a non-empty set of all motions that were pending at the time of the order but not ruled on by the order (*non-triggering motions*). The mean number of motions per group was 5.87 (i.e., there were on average 4.87 non-triggering motions). For each group, all motions were ranked by similarity to the order under the given metric. The proportion of triggering motions that were ranked first and mean rank of the triggering motion were calculated from each group's ranking.

These groups were evaluated using three term selection approaches: the raw document text (which often contains many OCR errors); normalization, as described above; and event terms. The two alternative TF/IDF training models were applied to each of the three term selection approaches, for a total of 6 combinations. For each combination, the mean rank of the triggering motion among all the motions was determined.

Figure 5 shows that the highest accuracy, as measured by the proportion of triggering motions that were ranked first among all pending motions, was achieved by normalizing the text without term selection. Intuitively, reduction to procedurally relevant terms improves the ability to determine what docket event a document performs, but can reduce the ability to discern the similarity between corresponding pairs of documents. TF/IDF training on just the order

[2] Term Frequency/Inverse Document Frequency

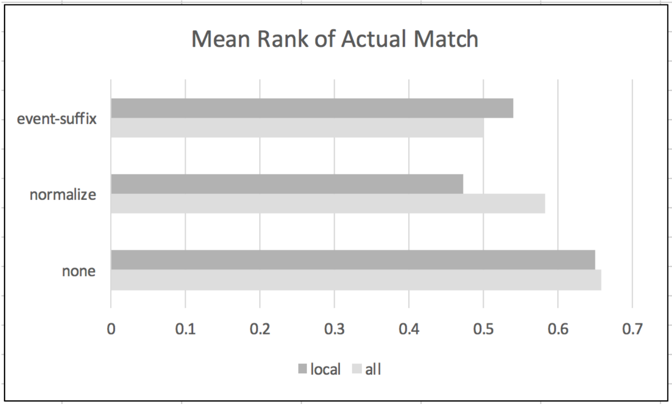


Figure 6: The mean rank of the triggering order among all pending orders, zero-indexed (lower is better, zero is perfect).

and pending motions (*local*) is at least as accurate as training over all orders and motions (*all*). Figure 6 shows the mean rank (zero indexed) of the most similar motion under each of the six conditions. The best (lowest) mean rank was achieved with normalization and local TF/IDF training.

It is not unusual for a single order to rule on multiple pending motions. A more realistic assessment of the utility of pending motion ranking is therefore to determine how many non-triggering motions a clerk would have to consider if the clerk read each motion in rank order until every motion ruled on by the order is found. One way to express this quantity is as mean precision at 100% recall. In the test set described above, using text normalization and local TF/IDF training, mean precision at 100% recall was 0.83, indicating that the number of motions that a clerk would have to be read was significantly reduced.

5 DECISION PREDICTION

Predictive models of decision making could be useful to *pro se* litigants (for help in understanding the strength of a case), to attorneys (for help in making strategic litigation decisions), and for training and decision support for judges and other decision makers. Even if the accuracy of predictive models were only approximate, they could nevertheless be valuable for decision support by helping to identify the most relevant words, phrases, or other features of a case record and the most relevant previous decisions.

Highly-accurate predictive models would require very detailed linguistic analysis of the text of case records and decisions, including argument structure, narrative analysis, etc. [8]. However, predictive models induced from simpler lexical features may be sufficiently accurate to be useful for the tasks listed above. Inducing such models can be cast as supervised concept learning over corpora of case records and decisions, where each decision is treated as a category label for the corresponding case record. This approach is feasible only for simple and routine cases for which it is possible to enumerate a small set of category labels, such as granting or denying a specific benefit or form of relief. However, such simple and routine cases are characteristic of many forms of administrative adjudication, such as immigration status and benefits entitlement.

Unfortunately, in many simple and routine administrative domains, only the decisions themselves, but not the underlying case records, are available. However, in such cases, the statement of facts in the decision can be used as a proxy for the contents of the corresponding case record. This approach was applied to decisions of the European Court of Human Rights in [1], which found that case outcomes could be predicted to some degree from statements of fact. The predictability of case outcomes from the statement of facts in the decision document doesn't conclusively demonstrate that the outcome would be equally predictable from the raw case record; the decision maker's description of the facts may have been tailored to fit the outcome. However, a demonstration that case outcomes can be predicted to some extent by models trained from fact statements alone may encourage courts and agencies to experiment with this approach to creating decision-support tools for *pro se* litigants and decision makers.

Accordingly, an experiment was performed to evaluate the feasibility of predicting decisions from the fact statements of cases in representative domain: World Intellectual Property Organization (WIPO) domain name decisions.[3] Domain name decisions resolve disputes between a domain name registrant and a third party under the Uniform Domain Name Dispute Resolution Policy (UDRP).[4] The UDRP Administrative Procedure applies to disputes concerning an alleged abusive registration of a domain name under the following criteria:

- The domain name registered by the domain name registrant is identical or confusingly similar to a trademark or service mark in which the complainant (the person or entity bringing the complaint) has rights; and
- The domain name registrant has no rights or legitimate interests in respect of the domain name in question; and
- The domain name has been registered and is being used in bad faith

WIPO decisions have a very consistent structure, including sections for History, Background, Contentions, Findings, and Decision. Just two distinct decisions are possible: transferring the domain name or denying the complaint. As a result, the decisions are well suited to the supervised concept-learning approach described above.

5.1 Experimental Design

Six thousand six hundred WIPO decisions were downloaded and parsed into the five sections described above. Each decision was labeled TRUE or FALSE based on whether the decision transferred the domain name (TRUE) or denied the claim (FALSE). The resulting set of cases had significant class skew, with 6,000 instances of TRUE but only 500 instances of FALSE. For a preliminary study, the 500 random TRUE instances were subsampled to create a balanced test set with 500 instances of each category.

This balanced set of cases was converted into a series of test sets differing in which sections were included as the text of each instance. The sections tested were as follows:

- History
- Background
- Contentions

[3] http://www.wipo.int/amc/en/domains/decisions.html

[4] https://www.icann.org/resources/pages/policy-2012-02-25-en

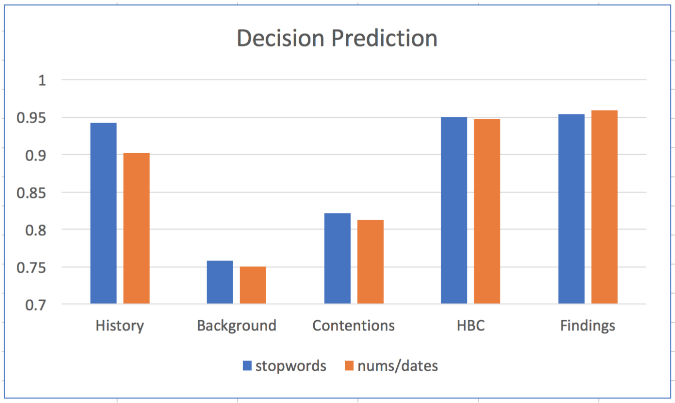


Figure 7: Mean f-measure in ten-fold cross-validation with Support Vector Machine prediction of WIPO case outcomes.

- The concatenation of History, Background, and Contentions
- Findings

The text of each instance was normalized by standardizing case and removing punctuation and, in addition, either (1) removing stop words, or (2) retaining stop words but replacing dates and numbers with standard tokens (“NUMBER" or “DATE"). The test condition in which the text consists of Findings is included for completeness, although it is not a good proxy for the case record as it contains conclusions about the facts.

For each selection of case sections and standardization method, the text was converted into n-gram frequency vectors for n=1–4, with only those n-grams retained that occur at least 8 times. The result was converted into sparse arff format,[5] loaded in Weka, and evaluated in 10-fold cross-validation using Weka’s implementation support vector machine (SVM) with sequential minimal optimization.

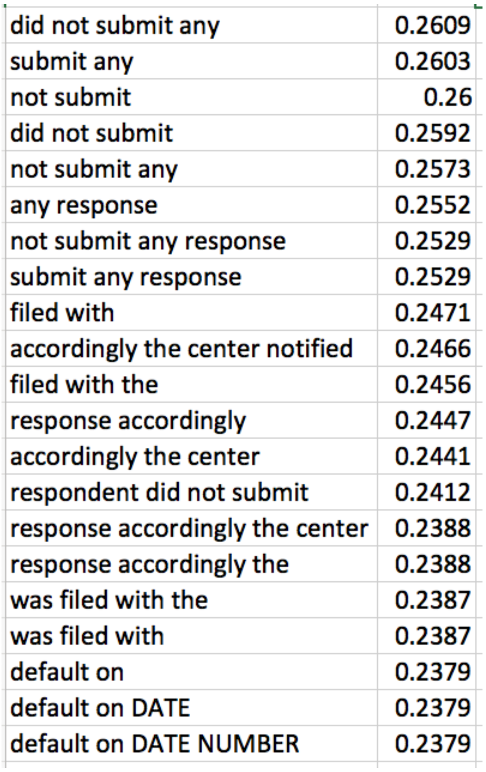
Table 2: Mean f-measure in ten-fold cross-validation with Support Vector Machine prediction of WIPO case outcomes. The text of each instance consists of the History (H), Background (B), Contentions (C), all three (HBC) or Findings (F) section.

	H	B	C	HBC	F
stopwords	0.943	0.758	0.822	0.950	0.955
nums/dates	0.902	0.750	0.813	0.948	0.960

5.2 Experimental Results

As set forth in Table 2 and Figure 7, the greatest predictive accuracy was achieved by the combination of the History, Background, and Contentions sections of each case (HBC). The predictive accuracy from these three sections, f-measure of roughly 0.95, was almost as high as the accuracy of prediction based on the text of the Findings section.

To understand why the HBC text is so predictive, it is helpful to examine the terms with the highest mutual information with the concept to be predicted, some of which are shown in Figure 8. This excerpt shows that phrases concerning filing, failure to submit a response, notification, and default are particularly strongly associated with the outcome of the case. One may view high information-gain phrases as being similar to the factors in [2] with the difference that they are induced automatically rather than being crafted manually. The SVM decision surface represents the set of tradeoffs among these factors that is most consistent with the training data, in a manner reminiscent of [6], but without the necessity of domain-specific hand-engineered factors.



did not submit any	0.2609
submit any	0.2603
not submit	0.26
did not submit	0.2592
not submit any	0.2573
any response	0.2552
not submit any response	0.2529
submit any response	0.2529
filed with	0.2471
accordingly the center notified	0.2466
filed with the	0.2456
response accordingly	0.2447
accordingly the center	0.2441
respondent did not submit	0.2412
response accordingly the center	0.2388
response accordingly the	0.2388
was filed with the	0.2387
was filed with	0.2387
default on	0.2379
default on DATE	0.2379
default on DATE NUMBER	0.2379

Figure 8: A subset of high information-gain terms in WIPO for History/Background/Contentions instances.

WIPO domain name dispute cases may be particularly conducive to predictive modeling owing to their binary outcomes and relatively stereotypical fact patterns. This experiment does not address the differences between the case record and the facts as summarized in the decision document, and the evaluation above artificially diminished the effect of class skew by subsampling to produce a balanced test set. Nevertheless, the impressive accuracy of a predictive model trained on raw text without any feature design or knowledge engineering suggests that this approach may have great promise for increasing access to justice for *pro se* litigants and improving training and decision support for decision makers in domains with many routine adjudications.

6 RELATED WORK

The history of applying text classification techniques to legal documents dates back at least to the 1970s [4]. Text classification has

[5] http://www.cs.waikato.ac.nz/ml/weka/arff.html

been recognized as of particular importance for electronic discovery [18]. Little prior work has addressed classification of docket entries other than Nallapati and Manning [14], which achieved an f-measure of 0.8967 in distinguishing Orders to Show Cause from other document types using a hand-engineered feature set.

There is extensive current activity in predictive models trained on factors unrelated to the merits of the case such as the nature of suit, attorneys, forum, judge, and parties [19]. Recent startups marketing predictive models for litigation support based on non merits-based factors include Lex Machina [11], LexPredict [12], and Premonition [17]. The insurance industry has a long history of developing decision prediction based on the merits of a claim, but these models are typically manually constructed, e.g., [15]. Outcome prediction based the merits of the case as extracted directly from raw text is a relatively new research area, with little work outside of [1].

7 SUMMARY AND FUTURE WORK

Judicial document collections contain a rich trove of potential information, but analyzing these documents presents many challenges. This paper has demonstrated how many types of filing error detection can be formulated as text classification problems. The highest accuracy was obtained by combining lexical features that characterize the document itself with procedural context features that indicate the role that the document is intended to play. These results demonstrate the feasibility of automating portions of the process of auditing court submissions, which could significant reduce a persistentdrain on court resources.

The experiment with order/motion matching demonstrates that while term selection may improve accuracy for document classification, it can decrease accuracy for tasks that involve matching based on overall similarity rather than procedural similarity.

The demonstration of outcome prediction in WIPO decisions illustrates that for case corpora with a limited set of possible outcomes and relatively stereotypical fact patterns, decision models of impressive accuracy can be induced without hand-engineered features, simply from the fact descriptions. This approach may be particularly promising for decision support and improved access to justice in the simpler and more routine end of the judicial spectrum.

No single technology is applicable to all judicial documents, nor is any approach sufficient for all document analysis tasks. However, each addition to this suite of technologies adds to the capabilities available to the courts, government agencies, and citizens to exploit the deep well of information latent in judicial document corpora.

ACKNOWLEDGMENT

The MITRE Corporation is a not-for-profit Federally Funded Research and Development Center chartered in the public interest. This document is approved for Public Release, Distribution Unlimited, Case Number 17-0362. ©2017 The MITRE Corporation. All rights reserved.

REFERENCES

[1] N. Aletras, D. Tsarapatsanis, D. Preotiuc-Pietro, and V. Lampos. Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. *PeerJ CompSci*, October 24 2016. https://peerj.com/articles/cs-93/.

[2] V. Aleven and K. Ashley. Doing things with factors. In *Proceedings of the Third European Workshop on Case-Based Reasoning (EWCR-96)*, pages 76–90, Lausanne, Switzerland, November 1996.

[3] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, Jul 1994.

[4] J. Boreham and B. Niblett. Classification of legal texts by computer. *Information Processing & Management*, 12(2):125 – 132, 1976.

[5] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001.

[6] S. Brüninghaus and K. Ashley. Generating legal arguments and predictions from case texts. In *Proceedings of the Tenth International Conference on Artificial Intelligence & Law (ICAIL-05)*, pages 65–74, Bologna, Italy, June 6–11 2005.

[7] L. I. I. Cornell University Law School. The federal rules of civil procedure. https://www.law.cornell.edu/rules/FRCP.

[8] D. Gutfreund, Y. Katz, and N. Slonim. Automatic arguments construction–from search engine to research engine. In *2016 AAAI Fall Symposium Series*, 2016.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[10] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy. Improvements to platt's smo algorithm for svm classifier design. *Neural Computation*, 13(3):637–649, 2001.

[11] Lex machina. https://lexmachina.com/ [Accessed: 27 November 2016].

[12] Lexpredict. https://lexpredict.com/ [Accessed: 29 November 2016].

[13] R. E. Madsen, S. Sigurdsson, L. K. Hansen, and J. Larsen. Pruning the vocabulary for better context recognition. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 483–488. IEEE, 2004.

[14] R. Nallapati and C. D. Manning. Legal docket-entry classification: Where machine learning stumbles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 438–446, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[15] M. Peterson and D. Waterman. Rule-based models of legal expertise. In C. Walters, editor, *Computing Power and Legal Reasoning*, pages 627–659. West Publishing Company, Minneapolis, Minnesota, 1985.

[16] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[17] Premonition. https://premonition.ai/ [Accessed: 27 November 2016].

[18] H. L. Roitblat, A. Kershaw, and P. Oot. Document categorization in legal electronic discovery: computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1):70–80, 2010.

[19] M. Surdeanu, R. Nallapati, G. Gregory, J. Walker, and C. Manning. Risk analysis for intellectual property litigation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Law*, Pittsburgh, PA, June 6–10 2011. ACM.

[20] Tesseract. https://en.wikipedia.org/wiki/Tesseract_(software).