Inducing Predictive Models for Decision Support in Administrative Adjudication

L. Karl Branting, Alexander Yeh, Brandy Weiss, Elizabeth Merkhofer, and Bradford Brown

The MITRE Corporation
7515 Colshire Dr
McLean, VA, 22102, USA
{lbranting,asy,bweiss,emerkhofer,bcbrown}@mitre.org

Abstract. Administrative adjudications are the most common form of legal decisions in many countries, so improving the efficiency, accuracy, and consistency of administrative processes could significantly benefit a large number of citizens. We explore the hypothesis that predictive models induced from previous administrative decisions can improve subsequent decision-making processes. This paper describes three data sets for exploring this hypothesis: motion-rulings, Board of Veterans Appeals (BVA) decisions; and World Intellectual Property Organization (WIPO) domain name dispute decisions. Three different approaches for prediction in these domains were tested: maximum entropy over token n-grams; SVM over token n-grams; and a Hierarchical Attention Network (HAN) applied to the full text. Each approach was capable of predicting outcomes, with the simpler WIPO cases appearing to be much more predictable than BVA or motion-ruling cases. We explore several approaches to using predictive models to identify salient phrases in the predictive texts (i.e., motion or contentions and factual background), and propose a design for displaying this information to decision makers.

1 Introduction

In many countries, the majority of legal adjudications are administrative, typified by routine licensing, permitting, immigration, and benefits decisions. The high volume of these administrative adjudications can lead to backlogs, inconsistencies, high resource loads for agencies, and uncertainty for citizens, notwithstanding the simplicity and uniformity that often characterizes such cases.

This paper presents the hypothesis that predictive models induced from previous administrative decisions can improve subsequent decision-making processes. The first step in establishing this hypothesis is to show the feasibility of creating models that predict the outcomes of routine administrative cases. The second step is to demonstrate how such predictive models can be used to improve decision processes. Our focus is on assisting individual decision makers by using predictive models to (1) identify the aspects of the instant case that are most relevant to its outcome and (2) determine the prior cases that share the most

relevant similarities to the instant case. A promising alternative approach to decision-process improvement not addressed in this paper consists of improved case routing and triage, e.g., assigning cases to specialized decision processes based on likely outcome and duration or apparent complexity. Since this approach depends heavily on the details of a given agency's decision processes, we leave it to future work.

2 Prediction in Law

Predictability is a fundamental property of legitimate legal systems. In traditional jurisprudence, prediction is based on weighing the relative strengths of arguments formulated by attorneys for and against a given proposition. In practice, however, attorneys often depend upon intuitive predictions about the outcome of cases developed from lengthy experience with judges, juries, and opposing attorneys in prior cases. Insurance companies have long used statistical models to aid in the estimation of settlement value of claims, and in the 1980's expert systems were developed to model human expertise at this task [12].

More recently, the development of corpus-based techniques for text analysis together with the increased availability of large legal text corpora has made feasible induction of models directly from the texts of case records and decisions [4]. Predictable aspects of cases include such factors as the following:

- The likelihood of success of a given motion (e.g., for dismissal or for extension of time) or claim (e.g., for Veterans' disability benefits)
- The expected award amount for a claim, e.g., the amount of a Veterans' disability award
- The expected return on civil claim, i.e., expected judgment minus expected litigation cost
- Expected litigation duration
- Recidivism probability

Predictive models of such aspects of legal cases have the potential to improve access to justice and the efficiency and consistency of case management. However, such models can be both opaque and susceptible to bias [14]. The work described in this paper is intended to mitigate these risks by focusing on predictive models as aids for improving human decision making rather than as stand-alone substitutes for human discretion.

3 Data Sets

In the United States, the agencies responsible for administrative claims, such as for veterans benefits, Social Security disability, immigration status, and Medicare appeals, all suffer from significant backlogs resulting from the inability of the agencies to handle their growing case loads with the available resources. As a first step in engagement with these agencies, whose data has privacy and sensitivity issues, we are developing prototypes on less sensitive, but representative, data sets.

– **Motion Rulings**
Our first data set consists of 6,866 motion/order pairs drawn from the docket of a United States federal district court.[1] Motions may be granted, denied, or granted in part and denied in part, and a single order may rule on multiple motions, potentially granting some and denying others. To obviate these procedural complexities, our initial data set is restricted to orders that either rule on a single motion or that have rulings of the same type for multiple motions, i.e., all granted or all denied. Each training instance consists of the text of the motion, which may contain OCR errors (the original filings were in PDF format), together with a classification as either "granted" or "denied."

– **Board of Veterans Appeals Decisions**
Adjudicative bodies vary in the extent to which case facts and decisions are published. Many adjudicative bodies publish only decisions but not the factual record on which each decision is based. In many agencies, the original decisions are not published, but only appellate decisions. The absence of published case records can create a cart-and-horse problem in which agencies are unwilling to share sensitive data for an unproven decision-support tool, but the decision-support tool can't be demonstrated because there is no access to the data on which it must be trained.

A method of finessing this problem exploits the convention that decisions generally contain statements of the fact of the case. Decisions with clear sections can be segmented, with the statement of facts treated as a summary of the actual case record, and the decision treated as the classification of those facts in terms of legal outcome. This "bootstrapping" approach was used to demonstrate the feasibility of predicting decisions of the European Court of Human Rights in (Aletras et al. 2016) [1]. Of course, decision drafters routinely exclude facts that are irrelevant to the decision and often tailor statements of relevant facts to fit the intended conclusions. As a result, bootstrapping is merely a proxy for the actual task of predicting decisions from raw case facts. However, demonstrating that decisions can be predicted from statements of fact, even if those statements are filtered, is an essential first step in demonstrating the feasibility of prediction in more realistic settings.

Board of Veterans Appeals (BVA) cases[2] have clear sections: Issues, Introduction, Findings, Conclusions, and Reasons. The Issues and Introduction sections contain only facts and contentions, and the decision on each issue is set forth in the Conclusions section. BVA cases often involve multiple issues, but issues are consistently numbered in Issues, Findings, and Conclusions sections. We therefore split each published BVA opinion with n issues into n instances, one for each issue, in which the facts consist of an issue and the

[1] Document filings in US federal courts are "semi-public" in that they are publicly accessible through PACER (https://www.pacer.gov/login.html), but per-page charges and primitive indexing impede wholesale document mining.

[2] https://www.index.va.gov/search/va/bva_search.jsp.

entire Introduction, and the classification is extracted (using regular expressions) from the numbered paragraph of Conclusion that corresponds to the Issue (i.e., that has the same numbering). The possible decisions on each issue are (1) the requirements for benefits have been met, (2) the requirements have not been met, (3) the case must be remanded for additional hearings, and (4) the case must be reopened. Conversion of all published BVA cases in this fashion yields 3,844 4-class instances or 1605 2-class (met or unmet) instances.

Unfortunately, the Findings section of BVA cases sometimes contain conclusions about facts not discussed in the Issues and Introduction section, so these sections are an incomplete proxy for the actual case record. This incompleteness makes it impossible in principle to predict the outcome of all BVA cases from just the Issues and Introduction.

– **WIPO Domain Name Dispute Decisions**
The World Intellectual Property Organization (WIPO) publishes decisions resolving complaints brought against the holder of a domain name that "is identical or confusingly similar" to a trademark belonging to the complainant.[3] WIPO cases have only two possible outcomes: the domain name is transferred to the complainant or it is not. WIPO cases are clearly segmented into seven sections: Parties, Domain Name, History, Background, Contentions, Findings, and Decision. The facts of each instance consist of the concatenation of the first 5 sections, and the classification is "transferred" or "not transferred." The WIPO data set consists of 5587 instances with a roughly 10-to-1 class skew in favor of "transferred."

4 Prediction

The first step in confirming the hypothesis that predictive models induced from previous administrative decisions can improve subsequent decision-making processes is to demonstrate that decision outcomes can be predicted. We experimented with 3 predictive techniques: hierarchical attention networks; support vector machines (SVM); and maximum entropy classification.

4.1 Hierarchical Attention Networks

In our first approach, we extended the hierarchical neural network model presented in Yang et al. [15] to predict the outcome of legal cases from free-text sections of their case records. The original model takes as input a sequence of sentences. A sentence representation is built for each sentence with a bidirectional gated recurrent unit (GRU) layer over word embeddings. An attention mechanism determines the weight of each time-step's contribution to a sentence vector. Then, a second GRU layer operates over the sentence vectors, an attention mechanism is applied, and the weighted sentence representations are

[3] http://www.wipo.int/amc/en/domains/decisionsx/index-gtld.html.

summed to form a hidden document representation. In prior work the document representation was used to predict the ratings of Yelp and movie reviews.

The hierarchical model was extended to account for the deeper structure of legal case documents. With the intuition that human decisions are informed by some combination the text in each section, we altered the model architecture for each dataset. The WIPO cases take as input three sections: history, background and contentions. We feed each section separately into Yang et al.'s document model, sharing weights. The resulting section representations are combined to create the case representation. The architecture used for BVA cases, shown in Figure 1, considers two sections: the issue and the introduction. The issue is nearly always only one sentence, so was treated as a single sequence of words. The introduction may be tens of sentences long and is passed through the hierarchical architecture described in the paper. The case representation is a learned transformation of the issue and introduction sections.

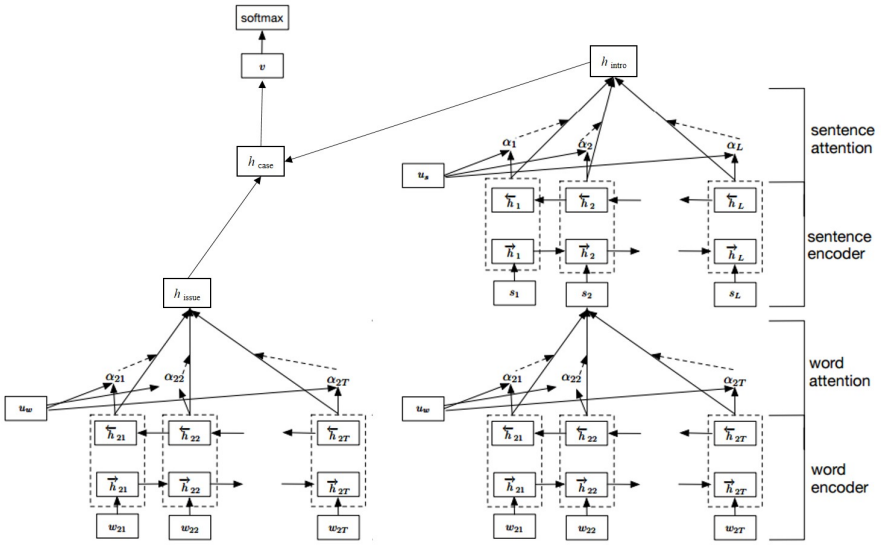

Fig. 1. Hierarchical neural model architecture for BVA cases. h_case is a learned function of h_issue, built from the words in the issue section, and h_intro, built from a hierarchical combination of the words-in-sentences and sentences in the case's introduction section.

In our experiments, a fully-connected layer appeared to better combine sections' hidden representations than a recurrent layer. We therefore used a hidden layer size of 50 for the WIPO cases and 64 for the BVA cases. We pre-trained word embeddings using the word2vec algorithm of [9]. For the WIPO cases, we pre-train on only the WIPO dataset; for the BVA cases, we use a separate dataset of approximately 50,000 appeals. We apply 30% dropout to delay over-

fitting these small datasets and use the Adam optimizer. Our models are trained on 80% of data, developed with an additional 10%, and the remaining 10% is reserved for testing.

The BVA model achieved a mean F1 of .738 and overall accuracy of 74.7%. The architecture reached a mean F1 of .944 on the WIPO cases, with an F1 of .64 for the 10-times-less frequent negative class. That model has 94.4% accuracy.

4.2 Support Vector Machine

The second approach to decision prediction was Support Vector Machine (SVM) learning. For the WIPO and BVA data sets, 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,[4] loaded into WEKA [7], and evaluated in 10-fold cross-validation using WEKA's implementation of Platt's algorithm for sequential minimal optimization[13, 8]. Because of memory issues, the WEKA SVM was run against only a subset of the entire WIPO data set consisting of 649 instances from each category.

In 10-fold cross validation the SVM approach achieved a mean F1 of 0.731 on the BVA data set, with an overall accuracy of 73%. A mean F1 of 0.950 was achieved on the WIPO data set, yielding an overall accuracy of 90.5%.

4.3 Maximum Entropy Classification

The third approach to decision prediction that we explored was Maximum Entropy (Maxent) classification [3] (often termed *logistic regression*). We used the jCarafe[5] implementation of Maxent, which adds regularization to mitigate overfitting, to build a model to predict whether a motion will be granted. Our features consisted of the party filing the motion, the judge ruling on the motion, the sub-type of motion, and the sequences of 1 to 4 tokens (alphanumeric character sequences having non-alphanumeric characters on both the left and right sequence borders) that occur in the text of the motion.

We observed that the motions contain many tokens that appear only in one motion and seem to be the result of OCR errors (as noted above, the documents were filed in PDF format, and some were created by scanning images to PDF). To remove these artifacts, any token that only appeared in only one motion in the collection was removed.

There are many different sub-types of motions, e.g., for extension of time to file, for summary judgment, etc. We found that better accuracy was obtained by training separate models for motion subtypes rather than training a single model for all subtypes. Accordingly, we split the motions into the following 3 large classes of sub-types and build a separate prediction model for each class:

- Extension-type motions, such as a motion to extend a filing due date, which tend to have higher grant rates than motions in general

[4] http://www.cs.waikato.ac.nz/ml/weka/arff.html
[5] https://github.com/wellner/jcarafe

– Motions of the letter sub-type, which tend to have a slightly lower grant rate than motions in general
– Motions not included in either of the 2 classes above

We used 10-fold cross validation to build and test separate models for each of the 3 large classes above and then combined the results. The combined results were an accuracy of 75%, and the recall, precision and balanced F-score for "granted" were 54%, 66% and 59% respectively.

The predictive results for the three experiments are summarized in Table 4.3 below:

	maximum entropy	SVM	hierarchical attn. network
motion-rulings	0.742	0.757	
BVA		0.731	0.738
WIPO		0.950	0.944

Table 1. Frequency-weighted mean F1 for predictive algorithms applied to three decision data sets. Note that the SVM result on the WIPO data set is on a balanced subset, rather than the entire WIPO collection, whereas the hierarchical attention network was applied to the entire skewed set.

5 Decision Support

The results of the prediction experiments indicate that routine adjudications and orders are predictable to some extent from models trained from text representing the facts of the case (in the WIPO and BVA data sets) or the motion text (for the order-prediction data set). Since this approach does not perform argumentation mining and has no explicit model of the applicable legal issues and rules, there is a limit to the predictive accuracy that this approach can achieve except in highly routine and predictable domains, such as WIPO decisions. However, our objective isn't replacement of human discretion, but rather support for human decision making. Our hypothesis is that predictive models can assist human decision makers by identifying the portions of the predictive text, e.g., statements of case facts or motion texts, that are most predictive of the outcome. We hypothesize that a decision maker may benefit from having the predictive text identified even when the decision disagrees with the models prediction. This hypothesis is based on the observation that one of the challenges of decision making is sifting through irrelevant portions of the case record to locate the most important facts.

We distinguish two uses of predictive text:

– Highlighting the parts of a document most relevant outcome, e.g., granting or denying a motion, or accepting or rejecting a claim for benefits, so that the decision maker can quickly identify the facts determinative of the outcome.

– Highlighting the parts of one document most relevant to assessing the similarity or difference between the cases.The Common Law doctrine of *stare decisis*, under which a decision in one case is binding on subsequent similar cases, is generally inapplicable to administrative adjudications, even in countries with Common Law legal systems. Nevertheless, we hypothesize that enabling decision makers to compare the current case to the most similar prior cases could make decision making faster and more consistent.

We therefore turn to the issue of how predictive texts can be identified. In the context of algorithms for prediction based on text, identification of the most predictive text is a special case of the more general problem of feature selection [6]. We first discuss methods that are independent of the predictive model, then turn to those that are derived from the predictive model.

5.1 Salient Fact Detection

Information-Theoretic Relevance Measures A particularly straightforward approach to predicting relevance is *mutual information* (sometimes termed "average mutual information"), which is a measure of how much knowledge of the value of one variable reduces uncertainty about the value of another variable [5]:

$$I(X,Y) = \sum_x \sum_y P(X=x, Y=y) \log \frac{P(X=x,Y=y)}{P(X=x)P(Y=y)}$$

For example, in the WIPO domain, the phrases "Complainant has failed to" and "did not reply" are among the highest information n-grams, that is, occurrence counts of those phrases are more predictive of the case decision than occurrence counts of most other variables (i.e., phrases). Mutual information in itself doesn't distinguish phrases like "Complainant has failed to", which is associated with successful complainant from phrases like "did not reply", which are associated with unsuccessful complaints.

Point-wise mutual information (PMI) (sometimes termed "mutual information") [5] measures the extent to which each particular value of a variable is predictive of a particular value of another variable:

$$PMI(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

PMI can be either positive or negative, depending on whether the presence of one value makes the other more or less likely. Thus, the PMI between "Complainant has failed to" = true and "transferred" = true is positive, whereas the PMI between "did not reply" = true and "transferred" = true is negative.

Neural Network Attention Attention mechanisms for neural networks allow the network to learn to weights as part of its representation. Bahdanau et al. (2014) [2] introduced neural attention for natural language processing, learning a soft alignment for machine translation such that certain input words contribute

most to output words. In the context of text classification, the attention mechanism determines relative contributions of words in the input sequence to the prediction. This hierarchical model has attention over the words in each sentences and over the sentences that make up each section. The attention operates on output from a bidirectional recurrent layer, meaning that each time-step folds in context from surrounding words or sentences but is most responsive to the word or sentence at that time-step.

Extracting the attention weights enables analysis of the model's predictions. For the BVA cases, we find that medical ailments in the issue representation weight and percentage disabilities were weighted most heavily. This is illustrated in Figure 2. The attention output for the introduction section was less interpretable, perhaps because that section tends to be dedicated to legal procedural details. However, the extracted attention showed that the model was strongly focused on certain areas of the input, often showing more than ten percent attention on a few words in an introduction section of several hundred words. Further analysis of this output could determine if, for example, particular procedural steps are highly correlated with certain outcomes.

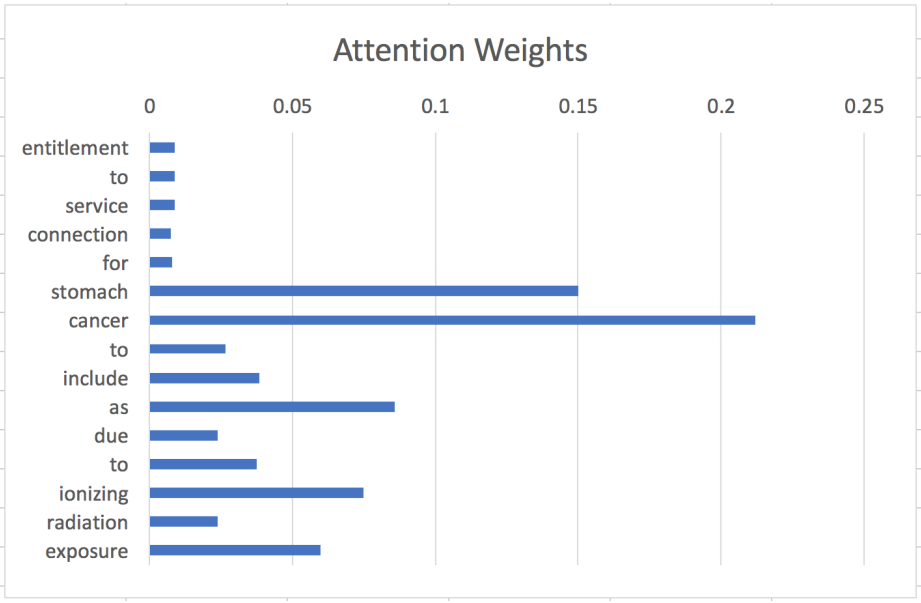

Fig. 2. Attention weights for words in a representative sentence in the introduction of a BVA case. These weights reflect the relative importance of the words in determining the label assigned by the hierarchical attention network to the case.

Linear Model Weights An alternative approach to using a predictive model to identify the most salient case facts makes use of the feature weights learned during model training. In linear models, such as maximum entropy and SVM with linear kernels, feature weights are indicative of relevance of features to the model's predictions [10], which in our application consists of predicted case decisions. These feature weights will have a similarity to point-wise mutual information (PMI) above, with features increasing the chances of a positive prediction tending to have positive weights and features decreasing the chances tending to have negative weights. One difference between PMI and the feature weights for logistic regression with regularization is that a regression feature weight differs when a feature either only occurs infrequently (weight magnitude is diminished) or is correlated with other features in the model (weight is adjusted for the effects of correlated features on the model).

For example, when maximum entropy/logistic regression (with regularization) is applied to the domain of motions for an extension of time, phrases having a relatively large positive feature weight for "granted = true" include "to dismiss" (feature weight of 0.319), "dismiss" (0.310) and "with the consent" (0.285), whereas phrases with relatively large negative feature weights include "a stipulation" (-0.385) and "stipulation" (-0.735).

5.2 User Interface for Decision Support

The data derived from the predictive models is intended to be displayed to administrative claimants and decision makers using a Graphical User Interface (GUI). The complexity of the displayed content along with a need to prevent user error necessitates a usable interface. We are therefore exploring interface designs to present this information in a manner that best facilitates its use, that is, to present data in a manner that improves a decider's speed and accuracy [11] . The preliminary design concept shown in Figure 3 contains several features that we hypothesize will assist users with deciding on cases efficiently and accurately. One feature of this design concept allows a user to view the most relevant cases in multiple ways (i.e., multi-case comparison, high-level comparison, in depth comparison). Providing multiple formats for case comparison allows a user the flexibility to decide how in-depth they would like to view the current case and previous case information. Another design feature provides the user with convenient access to relevant information (e.g., the rules) during the review process. This design concept leverages the pattern of open/close panels, which allow the user the ability to customize their view as they go through the evaluation process. In order to support efficient comparison, this design also provides a highlighting feature that is intended to allow a user to compare the similarities between current and previous cases. A future evaluation of this design concept will provide the information needed to iterate on the design patterns and features. The overall goal is to provide a satisfactory user experience while also assisting the decision maker to make quick and accurate assessments of cases. We plan on conducting an initial experimental evaluation to assess the overall ability of the combined predictive model and user interface to facilitate improved

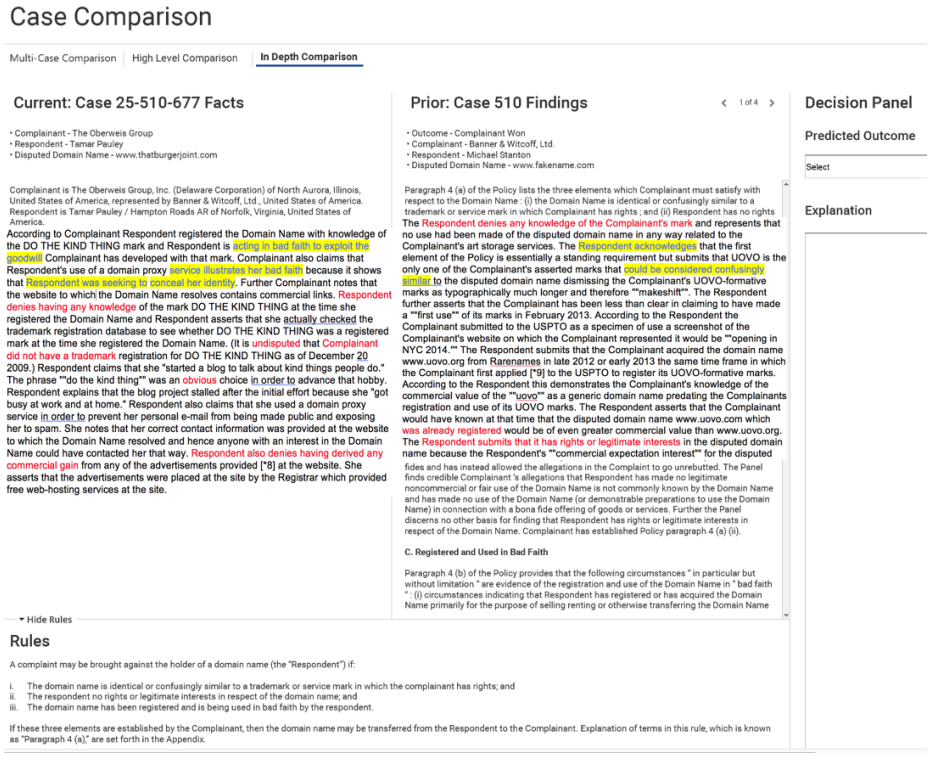

Fig. 3. A prototype decision-support interface design illustrating how the most salient case facts can by highlighted to assist with analysis of the case record and comparison between cases. Phrases highlighted in yellow are associated with rulings in favor of complainants, whereas phrases highlighted in red are associated with rulings in favor of respondents.

speed and accuracy in decision making. We hypothesize that the speed and accuracy of decision making can be improved by highlighting the phrases in cases with facts having the greatest weight under a predictive model and by retrieving prior cases with the strongest similarity to the current case in terms of the highest weight phrases. The initial evaluation will be performed using non-lawyers as subjects and WIPO case outcome detection as the predictive task, since WIPO cases have relatively simple and predictable facts and issues. After initial evaluation, we plan to conduct future evaluations that validate this concept using lawyers and other end-users of this type of decision support system.

6 Summary and Future Work

This paper explores the hypothesis that predictive models induced from previous administrative decisions can improve subsequent decision-making processes. Three data sets were developed: motion-rulings, BVA issue decisions, and WIPO domain name dispute decisions. The ability to predict outcomes in these three domains was demonstrated using three different approaches for prediction: maximum entropy over token n-grams; SVM over token n-grams; and a hierarchical attention network applied to the full text. This initial evaluation did not establish the superiority of one approach over another, but rather indicates that the outcome of routine decisions is predictable using multiple alternative models from the text of the motion or contentions and factual background alone and that predictive accuracy varies depending on the domain and the nature of the predictive texts.

We propose use of feature weights or network attention weights from these predictive models to identify salient phrases in motions or contentions and case facts. We have developed an interface design for presenting this information to improve decision making, and we propose an experimental evaluation to measure the extent to which speed and accuracy in decision making can be enhanced by enhancing the salience of phrases based on their weight in the decision model.

The ultimate objective of this work is to improve the efficiency, accuracy, and consistency of administrative decision making, the form of adjudication that has the greatest impact on most citizens, by integrating automated decision models into the human decision process. This work represents an initial step towards this objective.

Acknowledgments

The MITRE Corporation is a not-for-profit company, chartered in the public interest, that operates multiple federally funded research and development centers. This document is approved for Public Release; Distribution Unlimited. Case Number 17-1719.

References

1. Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., Lampos, V.: 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. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational Linguistics 22(1), 39–71 (Mar 1996), http://dl.acm.org/citation.cfm?id=234285.234289
4. Branting, L.K.: Data-centric and logic-based models for automated legal problem solving. Artificial Intelligence and Law 25(1), 5–27 (2017), http://dx.doi.org/10.1007/s10506-017-9193-x
5. Gallager, R.G.: Information Theory and Reliable Communication. John Wiley & Sons, Inc., New York, NY, USA (1968)
6. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (Mar 2003), http://dl.acm.org/citation.cfm?id=944919.944968
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations 11(1) (2009)
8. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K.: Improvements to platt's smo algorithm for svm classifier design. Neural Computation 13(3), 637–649 (2001)
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013), http://arxiv.org/abs/1301.3781
10. Mladenić, D., Brank, J., Grobelnik, M., Milic-Frayling, N.: Feature selection using linear classifier weights: Interaction with classification models. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 234–241. SIGIR '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/1008992.1009034
11. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
12. Peterson, M., Waterman, D.: Rule-based models of legal expertise. In: Walters, C. (ed.) Computing Power and Legal Reasoning, pp. 627–659. West Publishing Company, Minneapolis, Minnesota (1985)
13. Platt, J.C.: Advances in kernel methods. chap. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pp. 185–208. MIT Press, Cambridge, MA, USA (1999), http://dl.acm.org/citation.cfm?id=299094.299105
14. Sidhu, D.: Moneyball sentencing. Boston College Law Review 56(2), 672–731 (2015)
15. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of NAACL-HLT. pp. 1480–1489 (2016)