Applying an Interactive Machine Learning Approach to Statutory Analysis

Jaromír ŠAVELKA [a,b,1], Gaurav TRIVEDI [a] and Kevin D. ASHLEY [a,b,c]
[a] *Intelligent Systems Program, University of Pittsburgh, USA*
[b] *Learning Research and Development Center, University of Pittsburgh, USA*
[c] *School of Law, University of Pittsburgh, USA*

Abstract. Statutory analysis is a significant component of research on almost any legal issue and determining if a statutory provision applies is an integral part of the analysis. In this paper we present the initial results from an attempt to support the applicability assessment in situations where the number of statutory provisions to be considered is large. We propose the use of a framework in which a single human expert cooperates with a machine learning text classification algorithm. Our experiments show that an adoption of the approach leads to a better performance during the relevance assessment. In addition, we suggest how to re-use a classification model trained during one statutory analysis for another related analysis. This points to a new way of capturing and re-using knowledge produced in the course of statutory analysis. Our experiments confirm the viability of this approach.

Keywords. Statutory Analysis, Interactive Machine Learning, Text Classification, Technology Assisted Review, Predictive Coding

Introduction

In this paper we examine an application of an interactive machine learning (ML) approach to relevance assessment in statutory analysis. Since interactive ML has been successfully applied to classification tasks in many domains, such as web image search, network data analysis, or (notably) electronic discovery (e-discovery), we are interested if and how the interactive ML approach could be helpful in determining which statutory provisions retrieved with an information retrieval (IR) system are relevant in an analysis of a given legal issue. We want to test our hypothesis that the interactive ML framework could be a useful extension to traditional legal IR systems for statutory analysis. We address this question by assessing the ability of a classification model to gradually learn from the feedback provided by a human expert and the ability to improve the suggestions

[1] Corresponding Author: Intelligent Systems Program, University of Pittsburgh, 3939 O'Hara Street, Office 520, Pittsburgh, PA 15260, USA; E-mail: jas438@pitt.edu.
This work was supported by the University of Pittsburgh's University Research Council Multidisciplinary Small Grant Program and in part by the Cooperative Agreement 5P01TP000304 from the Centers for Disease Control and Prevention. Its contents are the responsibility of the authors and do not represent the official views of the CDC. The development of the tool and the second author were supported by NIH grant R01LM010964. We also acknowledge the guidance from Harry Hochheiser towards the tool development and helpful comments provided by Matthias Grabmair. We thank anonymous reviewers for their valuable suggestions.

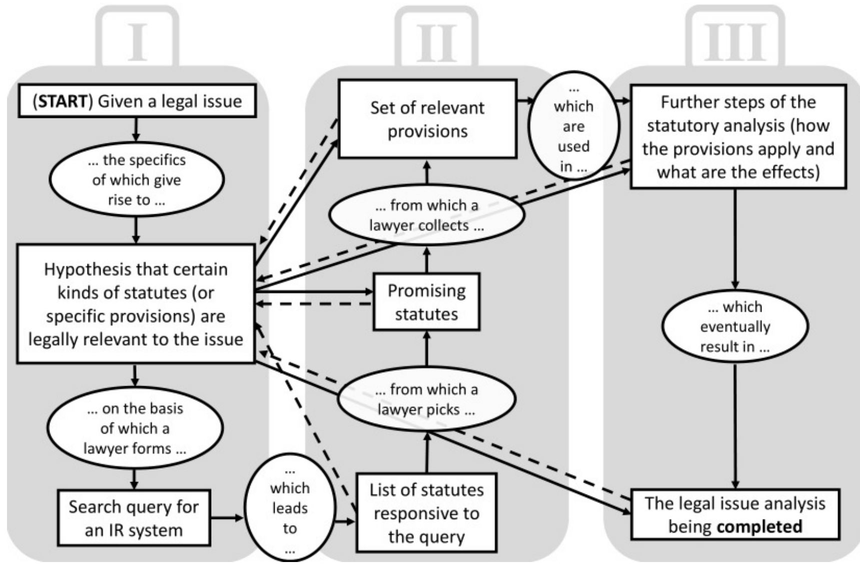

Figure 1. Schematic depiction of the process of statutory analysis.

as additional feedback becomes available. We also ask if a model resulting from one analysis could be re-used in a related statutory analysis.

1. Background and Motivation

Statutory analysis is the process of determining *if a statute applies* (the focus of our work), how it applies, and the effect of this application. [16, p. 61] The process of statutory analysis (emphasizing the relevance assessment) is schematically depicted in Figure 1. The researcher starts the analysis with an *initial hypothesis* about what statutes (or specific provisions) are relevant and what are the effects of their application. To test the hypothesis he or she formulates the initial search query for a legal IR system. The list of results as well as an inspection of promising statutory provisions are likely to influence the initial hypothesis. Thus, the process of statutory analysis has an iterative nature. In the end the researcher arrives at the *set of relevant provisions* which he or she examines in order to solve the given legal issue. The aim of our work is to provide support, beyond traditional IR, for the researcher to compile the list of relevant provisions.

Figure 1 divides the statutory analysis into three parts. Region I corresponds to the retrieval of statutes from a legal IR system. Region II represents the *relevance assessment*. Region III can be understood as the application activities starting with the set of relevant provisions and finishing with the completed analysis. While the activities in region I are reasonably well supported by existing legal IR systems, the activities from regions II and III typically receive very little support if any. In our work we focus on supporting these activities in region II (i.e., relevance/applicability assessment of the individual statutory provisions). We aim to extend computer support for an attorney's activity from after the statutes are retrieved to the completion of the analysis.

In particular, the framework discussed in this paper supports the applicability assessment in situations where the issue is open-ended with many unclear aspects and the goal of the analysis is to anticipate questions that may arise. We focus on this type of analysis because it is labor intensive and could benefit from automation. Consider, for example, statutory analysis to ensure total regulatory compliance of an industrial facility, or to explore the legal landscape for a new business or an existing business entering a new jurisdiction.

Many other situations require a systematic processing of large volumes of statutory texts such as a collection of resources for a commercial organization's, legal expert's or educational institution's knowledge base. Sometimes, it may be useful to perform such a coarse-grained analysis first and more fine-grained analysis later. This may, for example, reflect a distribution of work between a junior lawyer or paralegal, who compiles a list of potentially relevant provisions, and a more senior lawyer, who uses that list as the starting point for the actual analysis.

2. Task Definition, Proposed Solution and Working Hypotheses

We define the task we aim to support in the following way: the input is a medium to large-sized set of statutes retrieved from a legal IR system in response to a query about a legal issue. For example, the legal issue of interest may concern whether a coal mining facility complies with all local, state and federal laws about worker's safety. An IR query for cases relevant to that issue might be "(coal OR mine) AND safety AND (work OR worker)". The goal of our system is to output a subset of the input set that contains the statutory provisions that are most likely to be relevant to the legal issue.

To support identifying the provisions most relevant to the legal issue, we use an interactive ML framework. The framework is based on an iterative interaction between a human expert and a ML classifier. A human expert gives feedback to the ML classifier while the ML classifier provides the user with suggestions backed with explanation. This means that the human expert may be presented with the results of the automatic classification, the model's confidence, and important features. The user may correct the model where necessary or suggest features he or she believes to be important. Thus, the relevance assessment becomes a *dialogue* between the human expert and the ML classification model.

It is our working hypothesis that, after the human expert marks a small portion of the statutory provisions as relevant or not, the system will be able to provide reasonable suggestions about the relevance of the remaining provisions. As the human expert marks an increasing number of provisions, the suggestions provided by the system will become more accurate. Moreover, we hypothesize that a model trained during a statutory analysis can be helpful in supporting future analyses if there is some relation between the two (e.g., same analysis in different time, partially overlapping subject matter, or looking for similar statutory provisions in the new jurisdiction).

3. Related Work

The first explicit recognition of the interactive ML approach is [5] where the authors apply the techniques to image classification. Since then it has been applied in many diverse areas such as Web image search [1], making sense of large network data [4], or as an extension to text search and filtering on large text corpora [8]. In our work we apply the approach to statutory provisions.

Relevance feedback (RF) is a technique used to improve the results of an IR system. The system employs a user's feedback on the relevance of selected documents from the results list. RF can go through one or more iterations [12, p. 197]. Our work is like an extreme example of improving results with RF since the expected number of iterations goes well beyond what is usually anticipated in traditional RF frameworks.

The interactive ML approach has been applied in e-discovery in technology assisted review (TAR) or predictive coding frameworks such as a human-aided computer cognition approach [9], a TAR framework with multi-pass manual coding [3], and a highly-scalable classification framework for e-discovery [11]. TAR has been applied to heterogeneous texts associated with litigation but, apparently, not to homogeneous, structured texts like statutes. Thus, the specific challenges probably differ in e-discovery and relevance assessment of statutory provisions.

Legal IR systems return too many documents to read and assess [14]. A typical response to this challenge is enhancing IR systems with techniques used in case-based reasoning [14], dynamic document classification [13], keyword extraction [18], recommendation systems [17,25], or conceptual retrieval [7]. In our work we propose to enhance IR systems with an interactive ML component.

4. Example Application of Interactive ML to a Statutory Analysis

4.1. The Analysis

We work with two data sets that were produced during a project carried out at the University of Pittsburgh's Graduate School of Public Health. The project can be understood as a large-scale statutory analysis. The goal of the analysis was to assess and compare selected US states' regulatory frameworks concerning preparedness and response of the public health system (PHS) to public health emergencies. To be more specific, the researchers identified thirty different types of agents (e.g., doctor, emergency management, or special response team) as being part of the PHS. A starting point for the analysis was to compile (for each of the selected states) an exhaustive list of relevant statutory provisions conferring rights, obligations, or prohibitions on any of the agents in the context of the PHS' preparedness and response to public health emergencies.[2]

The task is very difficult and open-ended. Consider the following two examples:

(i) K.S.A. § 80-1921(a)(1) TOWNSHIP OFFICERS FIRE DEPARTMENT OR COMPANY
The township board of any such township shall have full direction and control over the operation of such township fire department. The board shall have the power to: Provide for the organization of volunteer members of such department and pay compensation to such members for fighting fires, responding to emergencies or attending meetings;

(ii) K.A.R. § 28-29-31(c)(2)(B) SOLID WASTE MANAGEMENT
Each person storing the tires shall meet the requirements of subsection (b) of this regulation and the following requirements: provide access to each storage area for fire-fighting equipment by either of the following means: obtaining certification from the local fire department stating that there is adequate access to each storage area for fire-fighting equipment;

[2]Additional information about the project can be found at www.phasys.pitt.edu or in [19].

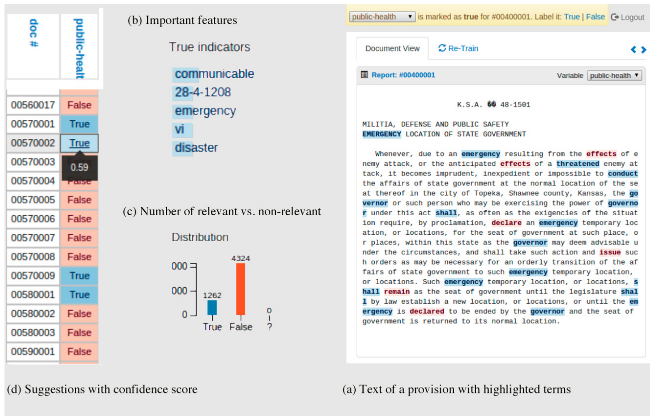

Figure 2. Some of the elements of the interactive tool that was used in our experiments. It shows a statutory provision with highlighted terms as important features (a), important features/terms with weights (b), distribution of relevant/non-relevant provisions (c), and list of labeled provisions with confidence score (d).

While the provision in (i) was manually classified as relevant for the analysis the provision in (ii) was classified as non-relevant. Thus the provision in (i) contains a specific obligation for one of the PHS agents concerning preparedness or response to public health emergencies while the provision in (ii) does not (the person to whom the obligation applies is not a part of the PHS).

4.2. Data Sets

Our experiments used the statutory texts from *Kansas* and *Alaska*. The raw statutory texts were processed into the trees of provisions as described in [21,20]. The resulting data set for Kansas consists of 304 statutory documents divided into 4,022 individual provisions out of which 802 are relevant. The Alaska data set contains 135 statutory documents divided into 1,564 provisions out of which 474 are relevant.

4.3. Software Environment

For the experiments we use the interactive tool described in [23,24] (some elements are depicted in Figure 2).[3] The tool was developed (at the University of Pittsburgh) for an interactive classification of clinical texts, but we easily adapted a subset of its functionality for our experiments. The tool is equipped with an interactive GUI that can be used to classify provisions. Users' decisions about each provision are recorded and, upon a user's request, the ML classifier is re-trained. The tool suggests the labels for unprocessed provisions (d in Figure 2), informs a user about its confidence (d) and prominent features (highlighted terms in a and b) leading to each suggestion. Users can suggest features that could be important by highlighting a term and clicking a respective button (a). The tool also provides summary statistics about the analysis (c).

The tool uses a support vector machine (SVM) classifier [10] with a linear kernel for classification. It works by identifying unigram features (terms) extracted from the

[3]The tool is not publicly available at the moment but the source code will be released in the near future. A demo and updates for the released source code will be available at http://vimeo.com/trivedigaurav/emr-demo.

documents. It learns these features for each classification class (e.g., here there are two classes: relevant to the legal issue and irrelevant) by looking at the training examples. During the iterative model building process, the tool considers the growing set of training examples supplied by a human expert.

5. Two Experiments on an Example Statutory Analysis

5.1. Experimental Designs

We conducted two experiments evaluating the application of interactive ML to statutory analysis. The goal of the first experiment was to examine a situation in which the analysis starts from a clean slate. We refer to this experiment as the *cold start experiment* and it was performed on the Kansas data set. The second experiment evaluates a re-use of the ML classifier in a related analysis. We refer to this experiment as the *knowledge re-use experiment* and it was performed on the Alaska data set.

The cold start experiment begins with dividing the statutory documents from Kansas into an *unprocessed data set*, i.e., the set of the documents that are considered as yet unseen, and a *test set* (50 randomly chosen statutes). After that, we randomly pick one document from the unprocessed data set, remove it from the set and assign its provisions with the labels that were manually assigned by the PHS analysis annotators. The manually classified provisions are then included in a *processed data set*. After that a classification model is trained on the processed data set and we use it to suggest labels for all the documents in the unprocessed data set and the test set. At this moment we evaluate the performance of the ML classifier. We iterate the procedure until the unprocessed data set is empty (one statute is removed from the unprocessed data set during each iteration step).

In the knowledge re-use experiment we compile a list of relevant statutory provisions for Alaska starting off from the classification model produced during the cold start experiment. We re-use the knowledge (i.e., the classification model) created during the Kansas statutory analysis for the benefit of the Alaska analysis. Note, that the goal of both analyses is the same but they are performed on different states' statutes. Otherwise, the experimental process is identical to the cold start experiment. We use 30 randomly chosen documents as a test set.

5.2. Evaluation

In both experiments, we evaluate the performance of the system after each feedback-iteration from two perspectives. The *first perspective* (ML model-oriented) assesses how well the ML classifier works. As is typical in ML experiments, the model is trained on the processed data set and evaluated on the test set. As the indicators of performance, we use a receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). [6] An ROC curve is a plot of the true positive rate (y axis) against the false positive rate (x axis) for the different possible decision thresholds. AUC can be interpreted as the probability that a classifier will rank a randomly chosen positive data point (relevant provision) higher than a randomly chosen negative one (non-relevant provision). We also report the standard IR performance measures, i.e., precision (P), recall (R) and F_1-measure (F_1).

Measure\# docs	10	50	100	150	200	254	304
AUC	.78	.81	.80	.83	.81	.81	—
P/R/F_1	.63/.16/.25	.59/.39/.47	.38/.48/.42	.41/.51/.45	.39/.56/.46	.43/.62/.51	—
P/R/F_1 (manual P+)	1/.02/.04	1/.12/.21	1/.26/.41	1/.43/.60	1/.63/.77	1/.85/.92	1/1/1
P/R/F_1 (manual R+)	.21/1/.34	.24/1/.38	.27/1/.43	.31/1/.47	.37/1/.54	.51/1/.68	1/1/1
P/R/F_1 (semi-auto)	.75/.17/.28	.75/.48/.58	.69/.6/.64	.79/.70/.74	.81/.84/.83	.89/.94/.91	1/1/1

Table 1. Cold start experiment from the ML model (top) and interaction (bottom) perspective.

Measure\# docs	10	30	50	70	90	105	135
AUC	.79	.79	.81	.82	.84	.83	—
P/R/F_1	.46/.62/.53	.55/.6/.58	.58/.66/.61	.53/.72/.61	.58/.69/.63	.58/.64/.61	—
P/R/F_1 (manual P+)	1/.05/.09	1/.14/.25	1/.28/.44	1/.47/.64	1/.65/.79	1/.76/.86	1/1/1
P/R/F_1 (manual R+)	.32/1/.48	.35/1/.51	.41/1/.59	.47/1/.64	.56/1/.72	.66/1/.79	1/1/1
P/R/F_1 (semi-auto)	.53/.44/.48	.52/.60/.56	.74/.63/.68	.72/.87/.79	.91/.85/.88	.92/.90/.91	1/1/1

Table 2. Knowledge re-use experiment from the ML model (top) and interaction (bottom) perspective.

The *second perspective* (interaction-oriented) assesses how well the system of a human expert interacting with an ML classifier works (semi-auto) as compared to the human expert classifying manually without ML support (manual). This manual human classification is used as two baselines: (i) a precision-focused baseline (manual P+) of a human classifier, who assumes that all of the still unprocessed documents are not relevant; and (ii) a recall-focused baseline (manual R+), of a human classifier who assumes that all of the still unprocessed documents are relevant. In this evaluation perspective we monitor P, R and F_1 on the whole data set (including the provisions that have been manually labeled) of each of the classifiers, semi-auto, manual P+ and manual R+. All of the classifiers use the same labels for the processed data set (i.e., the manually created labels). The difference is how they deal with the provisions from the processed data set. The manual P+ considers all the provisions from the unprocessed data set not relevant (that is why precision is always 1) while the manual R+ considers all the provisions from the unprocessed data set relevant (that is why recall is always 1). The semi-auto classifier uses the predictions of the current ML model for the unprocessed data set.

5.3. Results

The results of the cold start (Kansas) experiment are summarized in Table 1 and Figure 3. The results from the first (ML model-oriented) evaluation perspective (top left parts of the Figure and top part of the Table) show a reliable improvement of the ML classifier on the test set with the increasing size of the processed data set. The ability of the classifier to distinguish between relevant and non-relevant provisions has grown from 0.73 (close to a border line between a poor and a fair classifier) to 0.82 (usually considered a good classifier) in terms of AUC (plot KS 1p in Figure 3). This improvement also shows in the shapes of the ROC curves reported in Figure 3 (KS 1p ROC 10, 150 and 254). Similarly, the improvement in R (from around 0.2 to 0.62) with P stable around 0.4 is quite promising showing that with enough training examples the classifier can perform *reasonably well* (KS 1p).

The second evaluation perspective (interaction-oriented) shows that, except at the very beginning when the processed data set is small (less than 25 documents), the use of the interactive tool consistently outperforms the baseline (manual assessment) in both (recall and precision oriented) scenarios (KS 2p P, R and F_1).

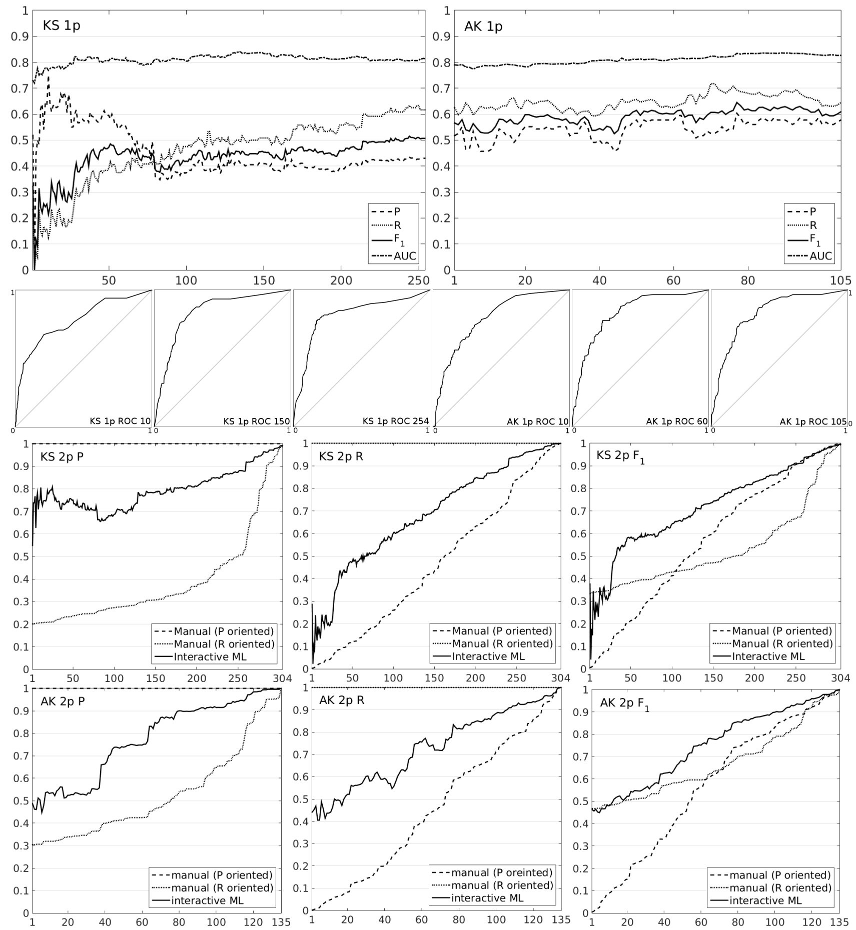

Figure 3. The results of the cold start (Kansas) and the knowledge re-use (Alaska) experiment. In the Figure KS stands for Kansas, AK for Alaska, 1p and 2p for the first (ML model-oriented) and second (interaction-oriented) evaluation perspectives, P for precision, R for recall, F_1 for F_1 measure, and ROC with a number for an ROC curve of the ML classifier trained on the specified number of documents.

The results of the knowledge re-use (Alaska) experiment are reported in Table 2 and Figure 3. The results from the first evaluation perspective (ML model-oriented) show a better starting position in terms of all the performance measures when compared to the Kansas experiment (AK 1p in Figure 3). AUC starts at 0.79 as opposed to 0.73 while the F_1 measure starts above 0.5 as opposed to approximately 0.2 (compare KS 1p and AK 1p in Figure 3). Despite the formidable starting position there still is a slight increase of the performance over the course of the experiment. The AUC rises from 0.79 to 0.83 and the F_1 measure increases from 0.55 to 0.61 (AK 1p).

From the viewpoint of the interaction-oriented perspective the use of the approach outperforms the baselines during the whole experiment (AK 2p P, R and F_1 in Figure 3). One difference is that in this experiment the interactive ML framework is competitive from the very beginning and does not require the processed data set to be of certain size to outperform the baseline.

6. Discussion and Future Work

The results of the experiments confirm our working hypotheses. At the same time, they clearly show that relevance assessment in statutory analysis is very difficult and there is still a long way towards a full automation. In the cold start experiment, after about 25 documents were labeled the AUC score of the classification models was above 0.8. Such a score is usually perceived as good, although not excellent. From the point of view of the standard IR measures it can be clearly seen that after about 80 documents were processed, the performance stabilized and tended to grow reliably for both P and R. The final R above 0.6 with reasonable P at 0.4 is very promising, especially if we consider that the measures would most likely still grow if additional data were available.

The direct use of the classification model trained during the cold start experiment (Kansas) was clearly beneficial in the knowledge re-use experiment (Alaska). From the ML model-oriented evaluation perspective, we can conclude that the performance of the model at the beginning of the Alaska experiment was comparable to the performance of the model in the final stage of the Kansas experiment. Thus, the re-use of the model from the previous analysis eliminated the cold start problem, i.e., the suggestions were reasonable from the beginning. The interaction-oriented evaluation confirms this.

The results show that processing of certain documents was not beneficial for the performance of the system (see sudden drops in performance in Figure 3 which were subsequently ameliorated by the additional feedback). This clearly suggests that a use of active learning [22] techniques (driving a human expert's focus on documents that are believed to improve the performance the most) could lead to a major improvement of the system. In the knowledge re-use experiment we used the model generated in the cold-start experiment with no modifications. Using techniques known from transfer learning [15] could facilitate better knowledge transfer or allow transfer between statutory analyses that are less related than the two discussed in this paper.

In this paper we focus on the performance of the ML classifier and compare it to the baseline where the task is performed manually. However, there might be other benefits of using the framework beyond reasonable suggestions about relevance of the individual provisions. For example, in [2] the authors conducted a study showing that human reviewers in e-discovery are more inconsistent in assessing documents than expected. They propose ML tools for document categorization support as a viable solution to this problem. Thus one possible additional benefit of the interaction with the system could be *more consistent and reliable results* among multiple human experts. The highlighting of important keywords could point the expert reviewer to the key parts of a provision, speeding up the reviewing process. We leave an experimental evaluation of these hypotheses for future work.

7. Conclusions

In this paper we examined if and how the interactive ML approach could help in determining which provisions retrieved with an IR system are relevant in an analysis of a given legal issue. We have shown that (i) interactively trained ML classifiers provide reasonable suggestions about the relevance of statutory provisions (ii) with increasing accuracy as more of the provisions are being processed and that (iii) it is possible to re-use the classifiers in future analyses. Importantly, the use of the interactive ML approach reliably outperforms the traditional manual assessment during the whole process of relevance assessment. As the ultimate goal of fully automating the process still seems quite distant, we have suggested multiple possible improvements of the system which we leave for future work.

References

[1] S. Amershi, J. Fogarty, A. Kapoor & D. Tan. Effective End-User Interaction with Machine Learning. *AAAI 2011*, San Francisco, CA, USA, 2011, pp. 1529-1532.

[2] T.I. Barnett & S. Godjevac. Faster, better, cheaper legal document review, pipe dream or reality? *ICAIL 2011 DESI IV Workshop*, Pittsburgh, PA, USA, 2011.

[3] J. Cheng, A. Jones, C. Privault, J.M. Renders. Soft labeling for multi-pass document review. *ICAIL 2013 DESI V Workshop*, Rome, Italy, 2013.

[4] D.H. Chau, A. Kittur, J.I. Hong & C. Faloutsos. Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning. *CHI 2011*, Vancouver, BC, Canada, 2011.

[5] J.A. Fails & D.R. Olsen, Jr. Interactive Machine Learning. *IUI'03*, Miami, FL, USA, 2003, pp. 39–45.

[6] J. Fogarty, R.S. Baker & S.E. Hudson. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *GI 2005*, Waterloo, Ontario, Canada, 2005.

[7] M. Grabmair, K.D. Ashley, R. Chen, P. Sureshkumar, C. Wang, E. Nyberg & V.R. Walker. Introducing LUIMA: An experiment in legal conceptual retrieval of vaccine injury decisions using a UIMA type system and tools. *ICAIL 2015*, San Diego, CA, USA, 2015, pp. 69–78.

[8] F. Heimerl, S. Koch, H. Bosch & T. Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), 2012, pp. 2839–2848.

[9] C. Hogan, R. Bauer & D. Brassil. Human-aided computer cognition for e-discovery. *ICAIL 2009*, Barcelona, Spain, 2009, pp. 194–201.

[10] T. Joachims. Making large-scale [SVM] learning practical. *Advances in kernel methods – Support vector learning*. MIT Press, Cambridge, MA, 1999.

[11] D. Krasner & I. Langmore. Flexible processing and classification for eDiscovery. *JURIX 2013*, Bologna, Italy, 2013, pp. 87–96.

[12] C.D. Manning, P. Raghavan & H. Schütze. *Introduction to information retrieval*. CUP, 2008.

[13] D. Merkl & E. Schweighofer. The exploration of legal text corpora with hierarchical neural networks: A guided tour in public international law. *ICAIL 1997*, Melbourne, Australia, 1997, pp. 98–105.

[14] M.F. Moens. What IR can learn from case-based reasoning. *JURIX 2002*, London, UK, 2002, pp. 83–91.

[15] S.J. Pan & Q. Yang. A survey on transfer learning. *TKDE 2010*, 22(10):1345–1359, 2010.

[16] W.H. Putnam. *Legal Analysis and Writing for Paralegals*. Delmar Cengage Learning, 3rd ed., 2008.

[17] P. Quaresma & I.P. Rodrigues. A Collaborative Legal Information Retrieval System Using Dynamic Logic Programming. *ICAIL 1999*, Oslo, Norway, 1999, pp. 190–191.

[18] E. Schweighofer & D. Merkl. A Learning Technique for Legal Document Analysis. *ICAIL 1999*, Oslo, Norway, 1999, pp. 156–163.

[19] P.M. Sweeney, E.F. Bjerke, M.A. Potter, H. Guclu, C.R. Keane, K.D. Ashley, M. Grabmair & R. Hwa. Network analysis of manually-encoded state laws and prospects for automation. *Network analysis in law*, 2014.

[20] J. Šavelka & K.D. Ashley. Transfer of predictive models for classification of statutory texts in multi-jurisdictional settings. *ICAIL 2015*, San Diego, CA, USA, 2015, pp. 216–220.

[21] J. Šavelka, M. Grabmair & K.D. Ashley. Mining information from statutory texts in multi-jurisdictional settings. *JURIX 2014*, Cracow, Poland, 2014, pp. 133–142.

[22] B. Settles. *Active learning literature survey*. University of Wisconsin, Madison, 2010.

[23] G. Trivedi, P. Pham, W. Chapman, R. Hwa, J. Wiebe & H. Hochheiser. An Interactive Tool for Natural Language Processing on Clinical Text. *IUI TextVis 2015*, Atlanta, GA, USA, 2015.

[24] G. Trivedi. Clinical Text Analysis Using Interactive NLP. *IUI Companion 2015*, Sonoma, CA, 2015.

[25] R. Winkels, A. Boer, B. Vredebregt & A. van Someren. Towards a legal recommender system. *JURIX 2014*, Cracow, Poland, 2014.