The Role of Metadata in Machine Learning for Technology Assisted Review

Amanda Jones	Marzieh Bazrafshan	Fernando Delgado	Tania Lihatsh	Tamara Schuyler
ajones@h5.com	mbazrafshan@h5.com	fdelgado@h5.com	tlihatsh@h5.com	tschuyler@h5.com

Introduction

In a 2006 *Law Technology News* article, Craig Ball made the following bold statement regarding the role of metadata in eDiscovery:

> It's the electronic equivalent of DNA, ballistics and fingerprint evidence, with a comparable power to exonerate and incriminate. Metadata sheds light on the context, authenticity, reliability and dissemination of electronic evidence, as well as providing clues to human behavior.

And by June 2007, in its seminal work on best practices and principles for electronic document production, the Sedona Conference had formally recognized the importance of producing “accessible metadata that will enable the receiving party to have the same ability to access, search, and display . . . information as the producing party." Indeed, it is now widely acknowledged and accepted within the eDiscovery industry that metadata is a critical component of any electronically stored data.

Opinion remains divided, however, regarding the role of metadata in machine learning for technology-assisted review (TAR), particularly with respect to the algorithm development process. The Grossman-Cormack Glossary of Technology-Assisted Review (2013) asserts that using metadata features is “typical” in the development of machine learning algorithms. Likewise, in a 2013 blog post, Ralph Losey maintains that one virtue of predictive coding systems is their use of a “complex analytic system that looks at the entire document . . . includ[ing] metadata.”

On the other hand, Equivio and kCura have both produced documentation stating that machine learning systems typically rely upon extracted text only and that experts engaged in providing document assessments for training should, therefore, avoid considering metadata values in making responsiveness calls (Sharp 2012; kCura 2014). In 2014 in his blog “The eDiscovery Nerd,” attorney Joshua Tolles writes that “anyone suggesting that metadata is part of the [machine learning] algorithmic analysis of a given document misunderstands how the core algorithmic analysis occurs.”

Regardless of whether most machine learning algorithms currently incorporate metadata, if we accept the importance of metadata to the eDiscovery process, it is reasonable to propose that metadata can and should be incorporated into the machine learning process for TAR. Questions still remain, however, regarding the extent to which metadata fields should be utilized, which fields are likely to be most constructive, and which techniques would prove most efficacious for leveraging the contents of these fields in algorithm development.

Few active discussions of these topics exist in eDiscovery literature. Grossman and Cormack’s TAR Glossary (2013) specifically identifies subject, sender, recipient, date, and file type fields as potential sources of worthwhile metadata features. Cormack and Grossman (2014a) also discuss using cc, bcc, and time sent fields as part of their TAR protocol. Finally, in their patent, they extend the set of possible metadata feature sources to include revision history, as well (Cormack and Grossman 2014b). This set of fields is intuitively appealing given the probable correlations between responsiveness and the participants, timing, subject and format of a conversation.

Oard and Webber (2013) present a more expansive view of potentially useful metadata for TAR. They discuss several distinct classes of metadata each of which offers numerous fields of metadata that have the potential to contribute positively to machine learning-based document classification. They do note, though, that "building an actual working system also requires making many detailed choices about data cleaning, feature construction, and feature representation."

Cormack and Grossman (2014a) provide some insights into the methods they adopted for integrating select metadata field contents into their TAR process. Specifically, they describe an approach to feature engineering that involves creation of a "text representation of each document (including a text representation of the sender, recipient, cc or bcc recipients, subject, and date and time sent)." Through this process, metadata values are placed on par with content derived from document body text.

Cheng and Jones (2013) briefly discuss an alternative approach to "incorporating document metadata information . . . in ways that preserve the special status of metadata." They suggest this can be accomplished effectively by incorporating relevance scores generated from independent metadata models based on logistic regression analyses. The full details of this process are not explored in that work, however, and the questions of which and how many metadata fields to utilize remain open.

The issues surrounding the role of metadata in machine learning for TAR warrant further investigation and clarification. In this paper, we address several foundational points in a series of comparative analyses designed to provide viable answers to several of the central questions. At a minimum, we hope to establish that metadata can be incorporated into the algorithm development process with positive impacts on machine learning results for document classification. Further, while we defer treatment of the question of exactly *which* specific metadata fields are best suited for machine learning in eDiscovery, we do tackle the question of how extensive the selection of metadata should be, finding that greater inclusivity is generally better for our purposes. Finally, we show the promise of incorporating metadata into the machine learning process in ways that more effectively tap the added layer of information intrinsic to the values in these fields and that more effectively capture the complementary perspective metadata contributes to document classification endeavors.

Data

We drew upon three different data sets for the purposes of our experimentation:

a. Data Set 1: 4,500 documents were drawn at random from a subset of 10,586 individual documents coded as Responsive or Not Responsive to Topic 301 from the 2010 TREC Interactive Task. (Family-level assessments were not considered.) The subset from which the randomly selected 4,500 documents were drawn represents documents for which topic assessments, body text, and metadata were all readily available. Details regarding the attributes of this data set are as follows:

Data Set 1	
Review Type:	Responsive Review
Litigation Type:	Class Action Suit
Industry:	Energy
Rate of Relevance:	16.30%
Custodian Info:	149 custodians
Doc Type Info:	Email 56%; Word Processing Docs 14%; Spreadsheets 7%; PDFs 3%; Other 20%
Size Info:	Mostly small docs; > 48% 0-1,000 bytes; 94% 0-50,000 bytes
Date Info:	Date range: Jan 1998 - Sep 2002 (95% 2000-2002)

b. Data Set 2: 4,500 documents were drawn at random from a proprietary set of 20,000 business documents coded as Responsive or Not Responsive. Details regarding the attributes of this review and data set are as follows:

Data Set 2	
Review Type:	Responsive Review
Litigation Type:	Contract Dispute
Industry:	Technology
Rate of Relevance:	16.10%
Custodian Info:	23 custodians
Document Type Info:	Email 49%; HTML 15%; Word Processing Docs 5%; Spreadsheets 5%; PDFs 2%; Presentations 4%; Plain Text 4%; Other 20%
Size Info:	Mostly small-to-medium docs; 60% 1,000-10,000 bytes; 95% 0-50,000 bytes
Date Info:	Date range: Sep 2005 - Aug 2014 (90% 2009-2014)

c. Data Set 3: 4,500 documents were drawn at random from a proprietary set of 11,088 business documents coded as Responsive or Not Responsive. Details regarding the attributes of this review and data set are as follows:

Data Set 3	
Review Type:	Responsive Review
Litigation Type:	Patent Infringement
Industry:	Manufacturing
Rate of Relevance:	15.30%
Custodian Info:	5 Custodians
Document Type Info:	Email 52%; Word Processing Docs 8%; Spreadsheets 26%; Other 14%
Size Info:	Mostly small-to-medium documents; 73% 0-10,000 bytes; 88% 0- 50,000 bytes
Date Info:	Date range: Aug 1999 - Jun 2014 (69% 2011 to Jun 2014)

Data Set 1, the TREC 2010 Enron population, was impoverished in terms of metadata compared to Data Sets 2 and 3. The latter two better represent the full breadth and quality of metadata that can be gleaned from modern data processing techniques. Taken together the three data sets reflect the variability in metadata availability that is common across eDiscovery projects.

Methods

From each of the three data sets of 4,500 documents, we selected a random subset of 3,000 documents to use as a Control Set, and we tested machine learning models built from the remaining set of 1,500 documents against the Control Sets.

For all machine learning model development, we used the Support Vector Machine implementation provided in the LIBSVM library (Chang and Lin 2011). We used a polynomial kernel of degree 2, and used the cross validation option of LIBSVM to obtain probability estimates.

The metadata fields that were considered eligible for utilization in our experiments are listed in the Appendix, along with information showing which fields were used for each data set.

The following fields were defined as belonging to the Standard Metadata set: Author, Sender, Recipient, Copy, Subject, Title, File Name, Document Type, File Extension, Sent Date, Created Date, Sender Domain, and Recipient Domain. These were selected for inclusion in the Standard Metadata set because a) they are the fields most commonly cited by practitioners providing details regarding their use of metadata in machine learning for TAR, b) they are commonly available or derivable for most modern eDiscovery populations, and c) they have obvious potential for correlating directly with document relevance.

The following fields were added to the Standard set to form the Extended Metadata set: All Custodians, Primary Custodian, Record Type, Attachment Name, Bates Start, Delivery Id, Company/Organization, Native File Size, Text Size, Normalized Date, Parent Date, Family Count, Attachment Count, Recipient Count, Copy Count, Combined Recipient Count, and Page Count. These fields were selected more opportunistically – on the basis of availability and amenability to transformation into generalizable machine learning features. This approach was adopted to allow the worth of the various fields to emerge through the modeling process itself and to mimic the reality of many eDiscovery situations where users have minimal control over the volume or quality of the metadata at their disposal.

Metadata fields that were blank or filled with a null value more than 5% of the time were omitted from the modeling process. Continuous or quasi-continuous metadata values were transformed into categorical values to enable correlations to emerge between these types of values and document relevance. For example, date values were collapsed into simple Month-Year values. Similarly, file size values were assigned to bands representing categories ranging from very small to very large.

The metric we use throughout to compare model performance is Area Under the Receiver Operating Characteristic Curve (AUROC). This metric indicates the probability that a given model will assign a higher ranking to a randomly selected responsive document than a randomly selected non-responsive document. We used the pROC package (Robin et al. 2011) to calculate the AUROC values and to test the significance of the differences we observed between models. For significance testing, we used the Delong method with a two-sided test and a significance level of 0.05 (DeLong et al. 1988).

Experiments

Metadata vs. Body Text Alone

The foundational view driving this work is that disregarding metadata when undertaking machine learning for TAR leaves a wealth of potentially valuable information untapped. The content of key metadata fields pertaining, for example, to the timeframe in which a document was authored, the participants in a conversation, or the subject matter of a file will often be correlated strongly with document relevance in eDiscovery. Thus, our first hypothesis was that incorporating metadata into the machine learning process would lead to improved results.

To test this hypothesis, we compared the results of machine learning models built from body text alone to models built from body text supplemented with the simple text contents of Standard Metadata. Results are presented in Table 1.

Table 1. AUROC for Models based on Document Body Text Alone vs. Models based on Document Body Text Supplemented with Standard Metadata Values.

	AUROC Body Text + Values from Std MD	**Conf Interval 95% Conf Level**	**AUROC Body Text Alone**	**Conf Interval 95% Conf Level**	**Statistical Significance**
Data Set 1	0.8593	0.8398-0.8788	0.8500	0.8297-0.8703	Not Significant
Data Set 2	0.9042	0.8892-0.9192	0.9018	0.8877-0.9158	Not Significant
Data Set 3	0.9341	0.9223-0.9458	0.9233	0.91-0.9366	$p < 0.00001$

While including the text from Standard Metadata with body text for the generation of features did not result in significantly superior results across the board, the improvement was highly significant for Data Set 3. There were no instances in which the addition of Standard Metadata was detrimental to model performance. Our first hypothesis was supported in at least one instance by this initial comparison.

Extended Metadata vs. Standard Metadata

Given the above and in accordance with Oard and Webber's (2013) assertion that a wide array of metadata fields "are potentially useful as sources of features for use by a classifier," we decided to push the initial hypothesis further by testing the impact of incorporating the text contents *all* of the readily available and robustly populated metadata fields for each of our data sets. For Data Set 2 and Data Set 3, this doubled the number of metadata fields being tapped for modeling. For Data Set 1, the impact was less drastic, adding only seven fields to the original eleven that were available in the Standard Metadata set. Our second hypothesis was that models incorporating Extended Metadata text alongside body text would lead to superior results when compared to results generated from models based on body text with Standard Metadata text alone. Results are presented in Table 2.

Table 2. AUROC for Models based on Document Body Text Supplemented with Extended Metadata Values vs. Models based on Document Body Text Supplemented with Standard Metadata Values.

	AUROC Body Text + Values from Ext MD	**Conf Interval 95% Conf Level**	**AUROC Body Text + Values from Std MD**	**Conf Interval 95% Conf Level**	**Statistical Significance**
Data Set 1	0.8786	0.8609-0.8963	0.8693	0.8398-0.8788	$p < 0.00001$
Data Set 2	0.9138	0.9001-0.9275	0.9042	0.8892-0.9192	Not Significant
Data Set 3	0.9728	0.9655-0.9801	0.9341	0.9223-0.9458	$p < 0.00001$

A highly significant improvement was achieved using Extended Metadata, as opposed to Standard, for both Data Set 3 and Data Set 1. In the case of Data Set 2, leveraging the Extended Metadata did not lead to a significant improvement, but results were on a par with the performance achieved by the model based on Standard Metadata. These results provided strong support for our second hypothesis.

Field-Encoded Metadata vs. Metadata as Text

The first two experiments demonstrated that utilizing metadata content to supplement document body text in machine learning, even without distinguishing between text and metadata features, can have a significant positive impact on results. However, metadata values carry an added layer of information – field associations – that is lost when metadata content is conflated with body text content to generate features for machine learning.

Neglecting to encode *both* metadata field *and* value information may constitute an unnecessary forfeiture of valuable input. This information loss can be avoided by creating distinct features for values occurring in different fields. This would mean, for example, that NYTimes as Sender, NYTimes as Recipient, and NYTimes as a topic of conversation within the body text of an email would each function as unique features. Creating these distinctions amongst values that would otherwise be collapsed into a single feature allows each one to be independently correlated with document relevance. We hypothesized that using field-encoded metadata values would lead to improved machine learning performance when compared to models where metadata values are undifferentiated from body text features.

There are a number of options available for creating machine learning models that maintain the distinctions between features derived from the body text of a document and features derived from metadata. We first tested an approach that involved adding a simple "tag" to the metadata values to record their source, e.g. NYTimes_BODY versus NYTimes_Sender versus NYTimes_Recipient, before intermingling the values with body text for purposes of feature selection. We pursued this methodology in the first of our experiments exploring the impact of differentiating between metadata and text in modeling. Results are presented in Table 3.

Table 3. AUROC for Models based on Document Body Text Supplemented with Tagged Extended Metadata Values vs. Models based on Document Body Text Supplemented with Plain Extended Metadata Values.

	AUROC Body Text + Tagged Ext MD	Conf Interval 95% Conf Level	AUROC Body Text + Plain Ext MD	Conf Interval 95% Conf Level	Statistical Significance
Data Set 1	0.8766	0.8594-0.8938	0.8786	0.8609-0.8963	Not Significant
Data Set 2	0.9300	0.9174-0.9427	0.9138	0.9001-0.9275	p = 0.007
Data Set 3	0.9746	0.9683-0.9809	0.9728	0.9655-0.9801	Not Significant

Performance differences were not as striking in this instance. For Data Set 2, utilizing metadata tagged with field information in addition to simple body text features improved performance significantly, but the result was not highly significant and significance was not attested in the other two cases.

An alternative option for utilizing metadata values in a way that preserves their field attributes and also more strongly highlights the potential for metadata to provide a complementary profile of the document population involves modeling document metadata independently from the body text, as noted by Cheng and Jones (2013). Specifically, one model is generated using field-encoded metadata values and a separate model is generated using body text values with the results of the two being combined thereafter – a technique sometimes referred to as "late fusion" (Cheng et al. 2013). In this study, we adopted a very simple technique for combining the scores from the two models: the score from one model was multiplied by the score from the other to generate a single new score for each document. Filippova and Hall (2011) express the intuition motivating this approach in two straightforward ways: "two complementary views on the data . . . refine predictions" and "a simple model which combines the predictions made by the two classifiers outperforms each of them taken independently."

In pursuing this approach, we first compared late fusion models to models built from data in which extended metadata values were simply intermingled with body text to create features. Results are presented in Table 4.

Table 4. AUROC for Models based on Late Fusion of an Independent Document Body Text Model and an Independent Extended Metadata Model vs. Models based on Document Body Text Supplemented with Plain Extended Metadata Values.

	AUROC Body Text + Ext MD–Late Fusion	Conf Interval 95% Conf Level	AUROC Body Text + Plain Ext MD	Conf Interval 95% Conf Level	Statistical Significance
Data Set 1	0.8856	0.8704-0.9008	0.8786	0.8609-0.8963	Not Significant
Data Set 2	0.9388	0.9285-0.949	0.9138	0.9001-0.9275	$p < 0.00001$
Data Set 3	0.9777	0.9714-0.984	0.9728	0.9655-0.9801	$P = 0.01931$

These results indicate considerable promise for the late fusion approach to metadata modeling for TAR. A highly significant improvement was observed for Data Set 2 and a significant improvement for Data Set 3. A significant improvement was not observed for Data Set 1.

Comparing the late fusion technique for incorporating metadata to the method in which field-tagged metadata values are intermingled with body text values in a single model yielded an analogous, albeit less pronounced, pattern of improvement, as seen in Table 5.

Table 5. AUROC for Models based on Late Fusion of an Independent Document Body Text Model and an Independent Extended Metadata Model vs. Models based on Document Body Text Supplemented with Tagged Extended Metadata Values.

	AUROC Body Text + Ext MD–Late Fusion	Conf Interval 95% Conf Level	AUROC Body Text + Tagged Ext MD	Conf Interval 95% Conf Level	Statistical Significance
Data Set 1	0.8856	0.8704-0.9008	0.8766	0.8594-0.8938	Not Significant
Data Set 2	0.9388	0.9285-0.949	0.9300	0.9174-0.9427	$P = 0.049$
Data Set 3	0.9777	0.9714-0.984	0.9746	0.9683-0.9809	$P = 0.02982$

These findings suggest that generating independent metadata models combined with independent body text models via late fusion can lead to results that are significantly better than models that do not distinguish between text and metadata and models that distinguish between text and metadata via tagging alone.

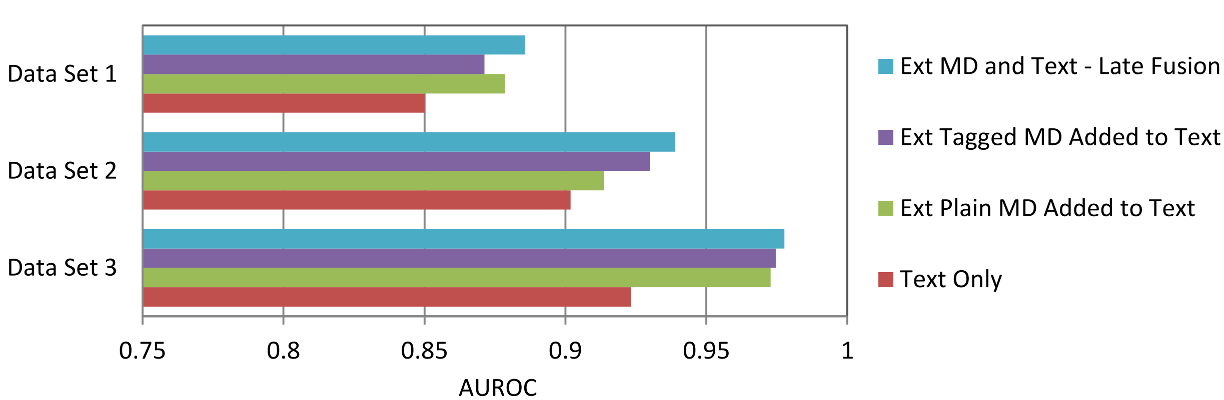
Discussion

All three of our hypotheses were supported, despite a degree of cross-topic/cross-corpus variability. We found evidence to support the idea that using metadata was preferable to omitting it, that using *all* available metadata was preferable to using a limited subset, and that exploiting both field *and* value information intrinsically associated with metadata is preferable to conflating metadata and text features.

Up to this point, our pairwise comparisons have focused on incremental increases in the complexity of the methods adopted to incorporate metadata into the machine learning process. Taking a more global view of the results, though, allows stronger trends to emerge and offers more definitive answers to the fundamental questions raised at the outset of our discussion.

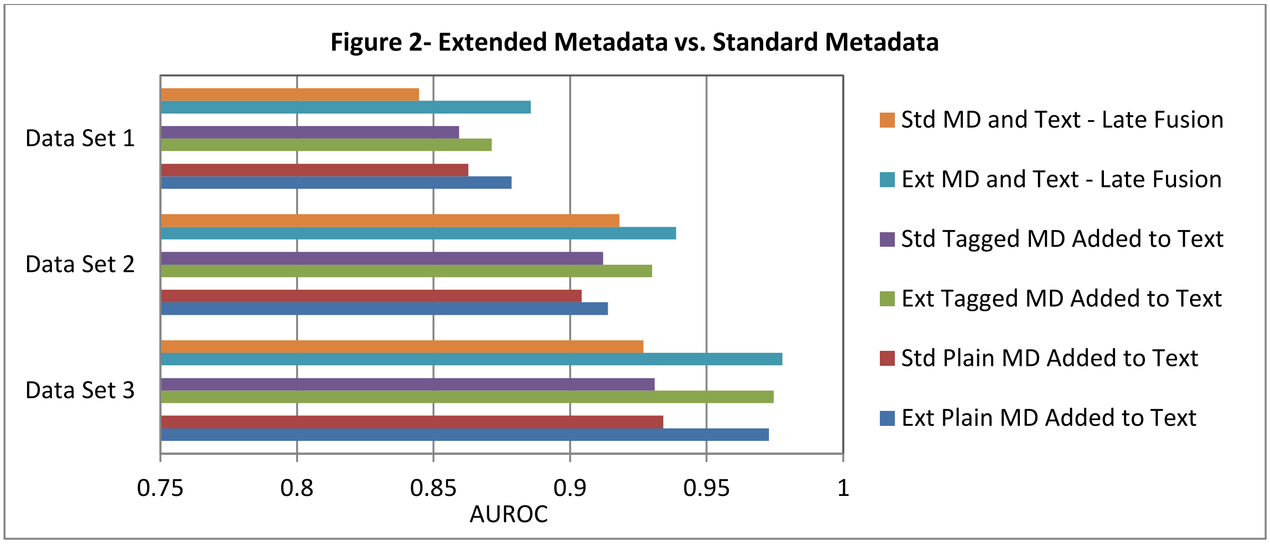
First, there is strong support for the hypothesis that metadata can be leveraged to improve the performance of machine learning models for document classification in eDiscovery. In our first experiment, we restricted our attention to the differences between utilizing the limited Standard Metadata intermingled with body text to utilizing body text alone. If we compare utilizing the *full* Extended Metadata to using body text alone, we observe more striking performance differences, as seen in Figure 1.

Figure 1- Extended Metadata vs. Body Text Alone



The models based on body text alone performed more poorly in every instance than models that incorporated Extended Metadata. These differences were highly significant across the board for Data Set 3. They were highly significant in two instances and moderately significant in the other for Data Set 2. For Data Set 1, the difference was highly significant in one instance and moderately significant for the others.

Similarly strong trends can be observed when each model created using Standard Metadata is compared to its Extended Metadata counterpart, as seen in Figure 2.



There were no cases in which a model built using Extended Metadata performed more poorly than an analogous model built using Standard Metadata only. For both Data Set 3 and Data Set 1, the improvements brought about by using Extended Metadata were highly significant ($p < 0.00001$) in each pairwise comparison. For Data Set 2, statistical significance was achieved only for the late fusion scenario, but improved performance through the use of all available metadata was observed as a general trend.

Conclusions and Future Research

The overall findings of this study support a move toward incorporating metadata as an integral component of the machine learning training process for TAR in eDiscovery. Undervaluing this data source by omitting it from model development may represent a missed opportunity and possibly even a failure to capture information material to fact discovery.

Furthermore, the results indicate that using all available metadata, rather than relying solely on the more limited set that is most often discussed in the context of machine learning, can be highly advantageous. Thus, continuing to devise new and better ways to leverage as much of the information conveyed by metadata as possible – even from fields that do not intuitively correlate with relevance – could be a worthwhile endeavor for TAR practitioners.

Finally, based on the experimental results observed for the late fusion approach to metadata and text modeling, a solid case can be made for viewing metadata as a wholly independent perspective on the data. Our preliminary exploration in this area suggests that this may be a key factor in fully capitalizing on the complementary contributions metadata can make to document classification for TAR.

Still, there are many directions remaining for future research into the role of metadata in machine learning for eDiscovery. Our study involved data sets where the rates of relevance were all relatively high, ranging from ~15%-~16%; it would be valuable to determine whether rate of relevance itself influences the impact of metadata in the modeling process. Similarly, we examined a variety of topics and corpora in our research to establish a number of broad generalizations about the role of metadata in machine learning for TAR. The impacts of employing metadata likely depend to some extent on both the specific topics and the unique corpus attributes at play in the given matter. Identifying reliable correlations between these variables would enable TAR practitioners to make informed decisions about which projects would likely benefit from investment in more complex metadata-inclusive modeling tactics.

Finally, we tested the impact of including *all* available metadata for algorithm development, as an alternative to foregoing metadata entirely or limiting its use to a small standard set. We did not examine the contributions of specific metadata fields at a more granular level. It is possible that certain classes of metadata consistently play bigger parts in the machine learning process than others. It would be valuable to know if there is a static set of key metadata fields that consistently leads to improved results without proliferating features and computational expense unnecessarily.

There is clearly an important role for metadata in the machine learning process, and this paper presents viable starting-point heuristics for deciding which metadata fields to use and possible techniques for incorporating them into a machine learning model. The challenges now are to refine our understanding of the ideal scenarios in which to utilize metadata and to identify optimal methods for incorporating the most effective set of metadata into the most efficient modeling process.

Bibliography

Ball, Craig. (2006). Understanding metadata. *Law Technology News*, 13(1), 36, 74-75.

Cormack, G. & Grossman, M. (2014a). Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. *SIGIR '14 Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 153-162.

Cormack, G. & Grossman, M. (2014b). *U.S. Patent No. 8,838,606*. Washington, DC: U.S. Patent and Trademark Office.

Chang, C.& Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1-27:27. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm

Cheng, J. et al. (2013). Soft Labeling for Multi-Pass Document Review. Paper presented at the DESI V Workshop. Rome, Italy.

Cheng, J. & Jones, A. (2013). Variability in technology assisted review and implications for standards. *Paper presented at the DESI V Workshop*. Rome, Italy.

DeLong, E. et al. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44, 837-845.

Eddelbuettel, D. & Francois, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8), 1-18.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with rcpp*. New York, NY: Springer. ISBN 978-1-4614-6867-7.

Filippova, K. & Hall, K. B. (2011). Improved video categorization from text metadata and user comments. *SIGIR 11 Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*.

Grossman, M. & Cormack, G. (2013). *The Grossman-Cormack Glossary of Technology-Assisted Review*, Fed. Courts L. Rev. 7.

kCura. (2014). Technology-assisted Review Training Workbook – v8 Fourth Edition.

Losey, R. (2013). Relevancy ranking is the key feature of predictive coding software. [Blog post] E-Discovery Team Commentary for the eLeet. Retrieved from http://e-discoveryteam.com/2013/08/25/relevancy-ranking-is-the-key-feature-of-predictive-coding-software/

Oard, D. & Webber, W. (2013). Information retrieval for e-discovery. *Foundations and Trends in Information Retrieval*, Vol. 7, Nos. 2–3, 99–237.

Robin, X. et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, DOI: 10.1186/1471-2105,12-77.

Sedona Conference Working Group. (2007). The Sedona Principles, Second Edition: Best Practices, Recommendations & Principles for Addressing Electronic Document Production.

Sharp, W. (2012). Top 10 best practices in predictive coding. Duke Law. Retrieved from https://law.duke.edu/sites/default/files/images/centers/judicialstudies/Panel_1-Background_Paper_4.pdf

Tolles, J. (2014). Less sales, more guidance in predictive coding and analytics. [Blog post] The eDiscovery Nerd. Retrieved from http://theediscoverynerd.com/2014/06/23/less-sales-more-guidance-in-predictivecoding-and-analytics/

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1), 1-29.

Appendix – Metadata Fields – Standard and Extended

The table below indicates utilization/non-utilization of each metadata field for each data set. Metadata fields that were eligible for use but not used in any model, due to insufficient attestation across all projects, are not listed.

	Metadata Field	Data Set 1	Data Set 2	Data Set 3
Standard Metadata Fields	Sender	Yes	Yes	Yes
	Document Subject	No	No	Yes
	File Name	Yes	Yes	Yes
	Email Subject	Yes	Yes	Yes
	Title	No	Yes	Yes
	Author	Yes	Yes	Yes
	Copy	Yes	Yes	Yes
	Recipient	Yes	Yes	Yes
	Created Date	No	Yes	Yes
	Sent Date	Yes	Yes	Yes
	Document Type	Yes	Yes	Yes
	File Extension	Yes	Yes	Yes
	Sender Domain	Yes	Yes	Yes
	Recipient Domain	Yes	Yes	Yes
Extended Metadata Fields	All Custodians	No	Yes	Yes
	Attachment Name	Yes	No	Yes
	Attachment Count	Yes	No	Yes
	Bates Start	No	No	Yes
	Combined Recipient Count	No	Yes	Yes
	Company/Organization	No	Yes	No
	Copy Count	Yes	Yes	Yes
	Primary Custodian	Yes	Yes	Yes
	Delivery Id	No	Yes	Yes
	Native File Size	No	Yes	No
	Family Count	No	Yes	Yes
	Normalized Date	No	Yes	Yes
	Parent Date	Yes	Yes	Yes
	Recipient Count	Yes	Yes	Yes
	Record Type	No	No	Yes
	Text Size	Yes	Yes	Yes
	Page Count	No	Yes	Yes