Variability in Technology Assisted Review and Implications for Standards

Jianlin Cheng, Amanda Jones
Xerox Litigation Services

jianlin.cheng@xls.xerox.com
amanda.jones@xls.xerox.com

1. Introduction:

Xerox Litigation Services (XLS) believes that establishing standards for technology-assisted review (TAR) is a crucial endeavor. TAR's place in the e-discovery tool set is a relatively new and vulnerable one; its recent successes could easily be undermined by practitioners engaging in irresponsible implementation of the technology. Thus, if TAR is to have a lasting place in e-discovery, it is imperative for its proponents, those who have worked diligently to gain credibility and acceptance for TAR in the e-discovery industry, to agree upon some foundational best practices.

The question, though, is: which aspects of the process should be standardized, and how? We believe the answer is that validation of TAR results and generation of performance metrics should be governed by fundamental statistical principles and best practices, but that caution should be exercised when imposing rules on other aspects of the TAR process. For instance, we would support formalized guidelines asserting that TAR models should always be tested against random samples that are fully representative of the population to which the model will be applied; that documents that have been used for training a model should never be included in the set of documents used for testing it and that vendors and users of TAR should always be cognizant of the differences between generating recall measurements and generating simple estimates of the rate of occurrence of particular features in a population. We believe it is important for everyone engaged in TAR to have access to performance metrics that have been generated using statistically sound and transparent methods and to understand what those metrics mean and what they do not.

XLS also believes that there is value in maintaining flexibility in TAR processes and that the industry should support ongoing innovation and creativity with regard to TAR implementation. We maintain that there are nearly as many legitimate and useful ways to execute a TAR project as there are worthwhile goals and applications for TAR results. Reasonableness and defensibility should always play a role in shaping TAR protocols, as should the intended use of the TAR results (e.g., prioritization or QC enhancement alone, as opposed to culling and/or wholesale document coding). Additionally, though, we contend there are numerous matter-specific factors that impact TAR performance and these factors, along with observed TAR performance itself, should play roles in shaping the specific process details adopted for any given TAR implementation.

The next sections explore a few of the many sources and types of matter-specific variability impacting TAR performance. We present a series of controlled model performance comparisons[1] with the aim of illustrating how useful it can be for the TAR process to remain dynamic and creative, so that the the full potential of this approach to review can be realized for any given matter – and also to demonstrate how "full potential" can and does vary from one matter to the next.

[1] To ensure fair comparisons, all of the models discussed in this paper were generated using a basic PLSA algorithm, unless otherwise explicitly noted. None of the customizations or parameter fine-tuning that would typically be applied in a live client engagement were utilized, as this would have introduced additional variables. Similarly, all of the models compared in this paper were built using training and testing sets that were equal in size to one another, unless the comparisons themselves dictated otherwise. Finally, except in those instances where exemplar concentration was being examined as a variable, the rate of occurrence of positively and negatively coded documents in the training and testing samples was equalized across the models being compared. Therefore, while the models utilize real-world data and authentic coding from actual litigation matters, the results presented here are hypothetical and do not reflect the actual results from any XLS technology-assisted review.

2. Intrinsic Matter-Specific Variables Impacting TAR

There are a number of key factors for any TAR project that users of machine learning approaches generally cannot control, but that nonetheless have a significant impact on the quality of TAR results. These are properties inherent and unique to each matter – e.g., the particular subject matter sought, the specific composition of the corpus within which that subject matter is sought, and the complex interaction of the two.

2.1 Richness

We use the term "richness" to refer to the concentration of positive exemplars in a population. Depending on the context, it may refer to the rate of responsiveness, rate of privilege, or rate of a specific issue code of interest. Richness can be influenced by e-discovery administrators, through the use of culling strategies designed to eliminate off-topic material from a review population, but it cannot be completely controlled.

The importance of richness for TAR may seem obvious. It stands to reason that, when there is a very low concentration of positive exemplars available for algorithms to learn from and generalize, the algorithms will be less successful at comprehensively recognizing or accurately classifying the material of interest. Still, we tested this assumption to verify its legitimacy.

To do so we constructed two models for responsiveness classification, drawing the random training and testing samples for both from the same original source of coded documents. The only difference between the training and testing materials for the two models was that, for one model, we replaced a random selection of half of the responsive documents with a random selection of non-responsive documents, resulting in one training and testing population that was half as rich as the other.

As expected, results were much stronger for the model built and tested using the richer samples. The low-richness model achieved a maximum F1 of 35.81%, whereas the high-richness model achieved a maximum F1 of 50.26%[2].

2.2 Subject Matter

Equally intuitive is the idea that, for any given matter, the topics of interest themselves will influence the degree to which TAR classifications will be successful. In this respect, human review teams and statistical algorithms may be quite similar – some topics are simply more difficult to interpret and code correctly than others.

Here too we tested this intuition to verify its accuracy. We constructed two models – one for Topic A and another for Topic B – using the same random sample of documents for training and testing, controlling for richness so that it would be equal across the two topics. Both topics were coded by the same review team and underwent the same degree of quality control. Nevertheless, the model for one of the topics outperformed the other. Specifically, the model for Topic A achieved a maximum F1 of 32.26%, while the model for Topic B achieved a maximum F1 of 36.89%.

2.3 Corpus

Less intuitive, perhaps, is the idea that a particular corpus itself may be a critical variable that will dictate the performance of TAR algorithms. It has been suggested, for instance, that it would be appropriate to build a model and establish performance metrics for it using one set of data and thereafter simply reuse that model for future data sets for the same matter, without retraining or retesting. However, the machine -learning algorithms of TAR do not learn subject matter in the abstract; they learn the patterns that characterize the subject matter of interest as it is realized in the specific corpus from which they are trained. Thus, it may be a serious mistake to assume that performance metrics for one data set will hold true for another. We have observed that changes in the source and composition of a data set will often lead to a degradation in the performance of TAR models.

[2] Details regarding the exact design of all of the comparisons presented in Sections 2 and 3, along with corresponding figures, are available upon request. They have been omitted here in the interest of brevity.

To illustrate this, we generated two distinct models: Model 1 using data from Corpus 1 and Model 2 using data from Corpus 2. Both of these models were designed to classify documents for responsiveness for the same matter, using coding generated and quality controlled by the same set of attorneys. Richness, training sample size, and testing sample size were equalized across the two models. When Model 1 was applied to a randomly drawn sample from Corpus 1, it achieved a maximum F1 of 30.58%. Similarly, when Model 2 was applied to a randomly drawn sample from Corpus 2, it achieved a maximum F1 of 30.30%. However, when Model 1 was tested against the random test sample drawn from Corpus 2, it achieved a maximum F1 of only 17.01%. Thus, even when sample size, richness, review team, and relevance criteria are the same across corpora, the composition of the corpus itself can lead to a pronounced decline in a model's performance metrics.

In this instance, the two corpora were, in most ways, very similar to one another. Both were predominantly email. Both represented the same basic date range and both were drawn from the same corporate sources. The only easily discernible difference was in the custodial make-up of the data. Therefore, we conclude that even under the most promising circumstances for model reusability, it is important to train a new model with fresh representative data for each new population; or, at a minimum, pre-existing models should be rigorously tested over a new random sample from the target population to confirm that performance metrics continue to meet minimum quality requirements for the project.

3. Matter-Specific Execution Variables Impacting TAR

While intrinsic matter-specific variables play an important role in shaping TAR performance, there are, in fact, many more aspects of TAR implementation that depend upon choices users make for each project. These choices involve parameter settings for the algorithm and decisions regarding the sources of information to be utilized in the model building process.

At XLS, we have experimented with many variations for TAR execution, and we continue to do so for each new project. We believe that this is the best way to consistently optimize and tailor results to the specific use case at hand.

In particular, we have been interested in capitalizing on multiple different sources of information to enhance our classification results. We believe there is great potential in drawing upon and combining insights about the population from many different perspectives and in leveraging as many correlations between corpus attributes and relevance as possible to achieve an end classification that is more nuanced and successful. We discuss several of these approaches below.

3.1 Dictionary Composition

Many, if not most, TAR algorithms utilize dictionaries composed of tokens extracted from training data. Generally, these tokens are "unigrams," corresponding roughly to individual words from the documents in the corpus. It is possible, however, to encode not only unigrams but also "bigrams," or tokens corresponding to each two-word sequence in the corpus. This idea is appealing, as there is a sense in which it introduces the possibility of creating models with some degree of sensitivity to syntactic relationships between words. More generally, it presents an opportunity to extract and leverage more information from the lexicon of the training data.

We compared the results of utilizing a plain unigram model to results obtained utilizing a model that included both unigram and bigram tokens. We found that modest improvements could be achieved using the model that included bigrams. Our bigram model achieved a maximum F1 of 55.63%, while the unigram model trained and tested on the same data achieved a maximum F1 of 52.98%.

3.2 Metadata Utilization

TAR algorithms are generally text classifiers, meaning they are primarily concerned with the text of documents and lack any direct mechanism for exploiting documents' metadata information. Again, though, it is intuitive to think that metadata properties of documents may be correlated with relevance in unique ways that cannot be captured through statistical analysis of text alone. For this reason, we have explored methods of incorporating document metadata information into our TAR models in

ways that preserve the special status of metadata, thereby allowing us to leverage another distinct source of information to improve classifications.

Specifically, we developed techniques that involve the generation and incorporation of relevance scores from independent metadata models based on logistic regression analyses. This approach has proven quite promising. For instance, we compared a model generated with PLSA alone to a model that combined baseline PLSA scores with metadata model scores and found that the PLSA-only model achieved a maximum F1 of 51.43%, while the model that generated scores via product combination of the metadata model and PLSA scores achieved a maximum F1 of 56.11%.

Pursuing this fine-tuning further, we also experimented with the specific method of combining model scores. In the case above, we were actually able to achieve an even greater advantage by combining the PLSA and metadata logistic regression model scores via a principal component analysis. Specifically, that model achieved a maximum F1 of 60.43%.

3.3 Pre-Existing Model Inputs

While we do not believe that it is appropriate or effective to apply pre-existing models to new data sets as a sole source of classifications, we do believe that pre-existing models have the potential to enhance the quality of classifications generated for new populations when used as a supplemental source of scores. Supplementation in this way provides one more method for leveraging all available knowledge about the subject matter and documents to achieve the best possible TAR results.

We tested this idea by comparing the results of a model utilizing new PLSA scores alone to a model that incorporated both new PLSA scores and scores from past models. The model that incorporated the additional information performed better, achieving a maximum F1 of 44.00% as compared to the baseline PLSA model's maximum F1 of 36.00%.

3.4 Layering Multiple Supplementary Inputs

As a next step, we hypothesized that if these additional sources of information were independently valuable, they would be even more beneficial when used together. Specifically, we tested the benefits of using both metadata modeling inputs and previous model inputs to enhance baseline PLSA scores, beyond the benefits that either of the simple combinations would achieve.

We compared the results of several models to test the hypothesis: 1) PLSA alone, 2) PLSA with metadata logistic regression modeling, 3) PLSA with scores from past models, and 4) PLSA with both metadata logistic regression modeling and past model scores. PLSA alone achieved a maximum F1 of 28.24%. PLSA with metadata logistic regression modeling achieved a slightly higher maximum F1 of 29.80%, and PLSA with scores from past models achieved a maximum F1 of 32.73%. The PLSA model that incorporated both metadata logistic regression modeling and past model scores achieved a the highest maximum F1, though, of 35.15%. These results support our hypothesis.

Interestingly, metadata modeling alone did not improve baseline PLSA performance greatly; it was only when used in conjunction with supplementation from past models' PLSA scores that it contributed to a more noticeable enhancement of the overall results.

3.5 Counterevidence

Given the findings above, it may be tempting to conclude that bigram modeling, metadata logistic regression modeling and incorporation of scores from past models should all be hard-wired into classifiers and utilized for every project. Unfortunately, the story is not as simple as that. For each one of the tactics discussed above, we have also seen cases where the modeling variation failed to contribute to significant gains in the baseline algorithm performance. In fact, there have been instances where these approaches led to noticeable declines in performance. Thus it would be premature to include any of the above as obligatory components of TAR algorithms, especially given that each one adds a certain amount of processing overhead to the core algorithm's classification generation.

4. Conclusion

Above we considered only a few of the many parameters associated with TAR algorithm tuning. There are many more – ranging from relatively simple adjustments (e.g., varying the stop word list for dictionary filtering, varying the minimum frequency of occurrence for words to be included in dictionaries, etc.) to more complex adjustments that may involve soft/fuzzy labeling of training sets or innovative approaches to the construction of training sets. Many of these variations on the basic TAR theme are likely to show the same mixed patterns of TAR performance impact as the factors discussed above. Therefore, we are forced to conclude that there is no single best recipe for optimizing TAR performance and no guarantee of equally strong results for every project. Instead, experimentation and customization are the best avenues for achieving optimal results.

There is virtually always a way to leverage TAR results to some advantage for review workflow efficiency in large matters, if flexibility and adaptation can be embraced as key elements of the TAR process. This is why our view on standards is a performance measurement-oriented one, rather than an implementation-oriented one. We think it is crucial for users to understand what their TAR system is achieving, while retaining the latitude needed to optimize performance and, ultimately, make well-informed decisions regarding the best possible use of their TAR results.