Quality Control for predictive coding in eDiscovery
Advances in technology are changing the way organizations perform eDiscovery.

Most notably, predictive coding, or “technology assisted review,” is becoming more widely accepted as part of the document review process. While it promises to be a powerful tool to reduce eDiscovery costs, the strategies, implications, and leading practices for predictive coding are still evolving.

More and more courts are taking up the question of acceptability of the use of predictive coding under the rules of civil procedure. The issue is whether the use of technology to replace human review is sufficient to discharge the parties’ discovery obligations. Predictive coding has been the subject of recent court decisions, but there has not been a definitive endorsement by a court in a case in which one party objected to the use of predictive coding by the other. It is clear, however, that high standards of quality control during predictive coding will help lower the risk of a dispute over an eDiscovery tool.

While the courts’ view will continue to evolve over time, predictive coding appears likely to become a standard tool in eDiscovery, and litigants should change their approach to quality control in eDiscovery as a result. This paper will discuss several strategies that companies can use now for improving quality control while using predictive coding for document review.

---

How technology has changed the eDiscovery process

In addition to increased efficiency and reduced costs, technology has made the eDiscovery process more complex, and this complexity increases risk. Statistical sampling has become a vital tool for quality control under predictive coding, and the complexity of the predictive coding workflow has highlighted the need for project management strategies to mitigate risk.

Under the traditional model of document review, documents were examined, coded, and organized by hand. The first electronic tools for document review were designed to recreate this process, as electronic documents were divided into assignments in the order in which they were loaded onto the software, and each document was reviewed linearly.

Using an electronic document depository with the ability to extract metadata from each document, however, has enabled workflow improvements and allowed for stratification, prioritization, and contextualization. Stratification means that file types—like spreadsheets or image files—can be handled separately, perhaps even by specialized review teams. Prioritization based on search terms, date ranges, or other metadata can be used to similar effect. Finally, technology permits documents to be contextualized by extracting concepts and presenting the reviewer with clusters of similar documents. These strategies have improved review efficiency and quality.
While technical tools have led to improvements, the increased complexity only emphasizes the importance of quality control during document review, and statistical sampling is an important tool in this regard. Used early in the process, the insight gained from statistical sampling can be used to refine the training of reviewers, adjust the work flow of second-level review, or improve search-term lists and other strategies used to guide reviewers. When predictive coding technology is used, statistical sampling is indispensable.

Predictive coding projects proceed in two distinct phases: the training phase and the application phase. The training phase involves the review of documents in order to train the classifier. The application phase involves the classifier making decisions about the documents not reviewed during the training phase. Each phase requires a different approach to quality control. In addition, underlying both the training and application phases, is the issue of work flow complexity, which deals with incremental data loads and with document types that either cannot be classified or that require special treatment.

**Training:** Predictive coding software must be trained to recognize the same distinctions between “responsive” and “nonresponsive” documents as a person reviewing the documents would recognize. The benefit of predictive coding lies in the fact that human decisions are limited to the training set, and are then leveraged across the entire body of documents. Thus, each decision on a training document potentially disposes of many documents. The primary quality-control goal during the training phase is to achieve consistent and accurate coding of the training documents, since consistency and accuracy in the training documents will determine the success of the predictive coding project. Human reviewers tend to be inconsistent, including during the training phase. Their views on the classification of documents evolve, and mistakes happen.

With a large set of training documents, the predictive-coding software has the capacity, within limits, to correct mistakes made to a small portion of the documents during training, and the software will still learn the correct coding for the type of document. The impact of inconsistent coding during the training phase thus depends on the absolute number of training documents that are relevant to the specific issue. For this reason, it is important to monitor the prevalence, also known as “richness,” of responsive documents among nonresponsive documents during the training phase.

Particular care should be taken to conduct effective quality control during the training phase. One approach is a double-blind review of the training set in which the training documents are reviewed by two independent reviewers or review teams. Documents for which there is disagreement between the reviewers are then reviewed by a subject-matter authority to resolve disagreements about the training documents before they become the basis for training the classifier.

The same effect can be achieved by using the predictive-coding classifier to review the training documents. Predictive-coding classifiers abstract from specific training documents to discover patterns and similarities, and the classifier will often suggest that training documents be coded differently. Those documents should be reviewed again, ideally by a subject-matter authority, to resolve the issue. The classifier can then be retrained to avoid the problem in the future. Notably, some predictive coding solutions feature built-in consistency checks designed to eliminate disagreements between the authority and the software.

**Application:** There are several options for how predictive coding can be applied. In a traditional coding work flow, where each document is reviewed, predictive coding can be used to reveal inconsistencies and function as a powerful quality-control tool. Predictive coding can also be used as the basis of production decisions by separating documents that were not reviewed but were classified as responsive by the predictive coding technology.

In the most popular approach, predictive coding is used to eliminate from further review documents that were classified as nonresponsive, while documents classified as responsive are then reviewed. Because the majority of documents are typically classified as nonresponsive, this last approach improves efficiency while eliminating the risks of producing documents that were not reviewed by an attorney.
Regardless of the work flow used, quality control over a large number of documents classified through predictive coding remains a challenge. It is similar to quality control for large document reviews performed by humans. There are three basic options for quality control in the application phase: a second review of a subset of the documents, “judgmental” sampling by nonstatistical methods, or statistical sampling. Both a second-level review and judgmental sampling are important parts of a well-rounded quality control program. Statistical sampling, however, is a much more powerful way to provide insight into the overall population of documents and the quality of coding. Statistical sampling solutions should be built into any eDiscovery software platform. This requirement is especially important for predictive coding applications. KPMG’s proprietary enterprise-level eDiscovery software, DiscoveryRadar™, provides one example.

There are three basic rules to remember about using statistical sampling to validate the results of predictive coding: the sampling population must be defined, the sample size must be calculated correctly, and the samples must be drawn randomly.

Using statistical sampling for quality control

Using predictive-coding technology to limit review by attorneys to the documents most likely to be relevant, can reduce time and overall eDiscovery costs significantly. In this example, document review time was reduced from 106 days to about 20 days, and cost was reduced from nearly $2 million to about $400,000.

"Validate" is used here in a meaning specific to quality control in eDiscovery; it does not refer to AICPA standards.
Define the sampling population:
Statistical sampling is used to draw an inference about the population from the sample. This process first requires the population to be defined correctly in order to interpret the results correctly. There are several approaches to defining the population. The first approach is to define the population as all documents that were not part of the training set. This comprehensive approach will yield an inference about the entire document population and will test the overall quality of the predictive-coding work flow. A second approach is to sample only the set of the documents that were coded nonresponsive in order to gauge whether or how many responsive documents were missed.

Calculate sample size correctly:
The most common calculation of sample size is a straightforward binomial formula. The most important factors determining the sample size are the desired confidence level, error rate (confidence interval), and prevalence. The confidence level reflects the likelihood that the sample is a true representation of the overall population. For example, a 95 percent confidence level means that if 100 independent samples were randomly selected, 95 of them would accurately represent the population (within the error rate). The error rate expresses the range of expected results. For example, with an error rate of +/-5 percent, if the sample shows that 10 percent of the documents were classified incorrectly in the validation sample, the actual number will be between 5 percent and 15 percent. Increasing the sample size may lower the error rate. Finally, prevalence represents the expected percentage of responsive documents in the population. As a note of caution, correctly interpreting prevalence and adjusting the sample size accordingly requires an advanced-level understanding of statistics.

Draw samples randomly:
Randomizing software can help any user draw sufficiently random results easily. Randomization becomes a challenge, however, when there are changes in the population, such as the addition of new documents. In practice, a validation sample against half the documents at the midpoint of the project cannot be updated simply by sampling the second half of the documents at the completion of the project. Both validation samples may be useful, but neither will permit a statistically-valid statement about the entire population because each sample will not have been randomly selected from the total population. Getting sampling right is an important part of making the work flow defensible and ensuring quality. The KPMG white paper The case for statistical sampling in e-discovery, provides an excellent resource on statistical sampling and statistical process control in document review.

Quality control for documents that require individual review

The training and application phases represent the core of the predictive coding work flow. By necessity, predictive coding work flows are complex, as they require the identification and tracking of different categories of documents. In addition to the bulk of the documents for which predictions are generated, there are five categories of documents that will require individual review and should be a special focus of quality control: training documents, validation samples, ambiguous documents, nontext files, and potentially privileged documents.

- **Training documents** – These documents enable the classifier to learn how a reviewer would handle a specific document. They must be reviewed to give the technology the required input and, as discussed, should be a special focus of quality control.

- **Validation samples** – These randomly selected documents must be reviewed by an attorney in order to assess the performance of the predictive-coding classifier. Validation samples are statistical samples, and the rules discussed above apply.

- **Ambiguous documents** – Given the variation in documents, case strategy used, and the complexity of the subject matter, predictive coding technology may not achieve sufficiently clear results for all documents, leaving these documents to be reviewed by an attorney. Depending on the software used, ambiguous documents may not be explicitly identified.

- **Nontext documents** – Since predictive-coding technology is based on the content of text documents, nontext documents such as image files or poor-quality scans must be reviewed.

- **Potentially privileged documents** – These documents need to be reviewed by an attorney to produce a privilege log and confirm that the information is subject to privilege.
Predictive coding technology demonstrates the general principle that more sophisticated technical tools lead to workflow complexity. While defensibility and disclosure requirements may be top of mind for outside counsel, successfully navigating the many moving parts in predictive coding technology should be the foremost project-management concern.

Predictive coding is a powerful document review tool. Nonetheless, the increased use of technology has also increased the complexity of the eDiscovery process, which can result in increased risks. Companies should consider the strategies discussed above for improving quality control during complex eDiscovery work flows, particularly for predictive coding.

An enterprise-level approach: An enterprise-level approach can reduce costs and increase efficiency by allowing the eDiscovery provider to become familiar with the data sources, share information across different cases, and to avoid an ad hoc approach for each stage of the process. The enterprise-level approach can also allow for continual improvement and consistency in workflow. For example, the protection of privileged information is one of the core concerns of eDiscovery, as privilege considerations are often subject to interpretation and different approaches. Corporations also tend to work with several law firms, and often with various teams within each firm. Consistency in maintaining claims of privilege among matters and over time is important, as any variation in how that information is handled increases risk. Documents that were produced in one matter may no longer be subject to privilege in other matters.

Checklists: The second principle of quality control in eDiscovery is the use of checklists. eDiscovery projects are extremely complex, and the use of predictive coding only adds to the complexity. Simple checklists that are followed consistently can help mitigate the risks of error. Checklists make project delivery documentable and auditable by preserving a record of the tasks performed. Checklists can also be customized for large projects and enterprise solutions. In order to unlock cost savings while minimizing risk, checklists should be living documents that are amended as optimal workflow for a particular client are developed and information is shared.

Conclusion

Effective checklists require some thought and testing. For example, checklists should use natural breaks in the workflow, be simple and logical, fit on one page, and have a clear objective. See Gawande, Atul, The Checklist Manifesto: How to Get Things Right, 2009.