Top 10 Best Practices in Predictive Coding

Emerging Best Practice Guidelines for the Conduct of a Predictive Coding Project

"...design an appropriate process, including use of available technology, with appropriate quality control testing, to review and produce relevant ESI."

Magistrate Judge Andrew Peck of the Southern District of New York in Monique Da Silva Moore, et al., v. Publicis Groupe and MSL Group
Predictive coding is new to e-discovery. Predictive coding, sometimes termed computer-assisted review or technology-assisted review, is software that can be trained by a human being to distinguish between relevant and non-relevant documents. Over the past three years, the predictive coding market has transitioned from a preliminary embryonic state, a test bed for experimentation by technology geeks and early adopters, to the point today where the technology has been approved for use by courts around the US and internationally, and has become the single most talked about topic in e-discovery worldwide.

As is well known, the classification technologies that underlie predictive coding applications in the e-discovery arena are widely used in a very broad range of industrial and scientific settings, and have done so since beginning in the 1960’s. Some of the best practices that have developed in these settings are analogous to the e-discovery setting. However, e-discovery is a unique arena, especially in terms of the stringent defensibility requirements that apply. As such, it has been necessary to develop and define best practices that will address the unique needs of predictive coding applications in the e-discovery environment. The objective of this white paper is to outline the best practices that are emerging in the application of predictive coding technology to the e-discovery arena.

1. CHOOSE THE “EXPERT” WITH DUE CONSIDERATION

Don’t believe the marketing messages! Predictive coding is not magic and it’s not even a panacea for all the world’s ills. It is, however, a smart piece of software that can be trained to “imitate” the criteria used by a human being to evaluate a document’s relevance. In a sense, the software is encoding the intelligence and knowledge of an experienced attorney. But here is what they don’t tell you: Predictive coding is a garbage-in, garbage-out application, otherwise known as GIGO. With quality input, predictive coding applications can generate outstanding results. But train it poorly, you won’t be taking home any medals... This technology is dependent on the quality and consistency of the input it receives from the human being training it. As such, it is important to give due consideration to the choice of that person. In the terminology that Equivio has developed for predictive coding projects, we refer to this person as the “expert”. This term is not used lightly. The person training the software needs to be a knowledgeable attorney, with the experience and authority to make review decisions that are likely to have a significant impact on the subsequent conduct, and potentially outcome, of the case. As such, our first best practice is that due consideration be given to the selection of the expert.
2. COLLABORATIVE TRAINING

To the best of my knowledge, the collaborative training approach was first developed by Kit Goetz, Vice President of Litigation at Qualcomm, and her team. Collaborative training involves using a team of two or three “experts” to train the system together, rather than a solo expert. The collaborative approach is typically used for the first 500 or 1,000 documents. The attorneys sit together, in a room, and start training the system together. The rule is that they are required to reach consensus on the relevance designation for each document. Kit’s team has run statistics on this. The findings are instructive: in the first half hour, the attorneys disagree on 77% of the documents. After four hours, the disagreement rate has dropped to 2%. The process that is occurring in the room is that the group is progressively refining the concept of relevance that underlies the case. The distillation of a clear, well-defined, well-bounded concept document relevance helps ensure the quality of the subsequent training process. Qualcomm’s experience shows that this is something that needs to be worked on, and cannot be assumed or taken for granted. The impact on the quality of final output has been shown to be very significant.

3. TAG BY “APPLICATION-ACCESSIBLE” DATA

Most of the predictive coding systems in the e-discovery market focus their analysis on document content, rather than external metadata, such as date or custodian. As such, it is important to instruct the expert to tag documents based on the data that can be accessed by the system. For example, if the predictive coding system captures only document content, the expert should be instructed to tag documents based on document content only, regardless of metadata. It’s possible to imagine a document whose content is apparently responsive, but which falls outside the relevant date range of the case. Were this document to be tagged as not-relevant, the predictive coding application would be misled into “thinking” that the document content itself is not relevant. The best approach is to cull by metadata prior to training the predictive coding application, and in so doing avoid any potential tagging errors.

4. SUPER-ISSUE AND INDIVIDUAL ISSUES

It’s important to distinguish between the super-issue (aka master issue) and the individual issues (aka sub-issues). The super-issue relates to whether the document is relevant or not to the case, and is used to construct the review set. Documents with scores above the relevance cut-off score for the super-issue are passed on to review, while documents with scores below the cut-off are culled. The individual issues have a very different role. Let’s say that the super-issue is “Navigation”. Our individual issues might be North, South, East and West. For the super-issue – that is, Navigation -- we
will set a cut-off score in order to select documents for review. But for the individual issues, cut-off scores are not used. The individual issues are used not to construct the review set, but to organize it. For example, we may have one review team specializing in North documents, another specializing in South, a third in West and a fourth in East. Within the set of review documents, we use the individual issue scores to assign, for example, South documents to the South review team.

5. SEPARATE CONTROL DOCUMENTS FROM TRAINING DOCUMENTS

In order to ensure the statistical validity and defensibility of the predictive coding process, it is critical that the “control” documents be kept separate from the “training” documents. The control set comprises a random, representative sample of documents from the collection. The expert tags the control documents as relevant or not; the control set then serves as the gold-standard, against which the ability of the predictive coding system to assess document relevance is tested. To state this in control-training terms, the control documents measure how well the system has been trained. Were the control documents used for training as well, we would know something about the ability of the system to assess the relevance of the control documents, but we would not know anything about the system’s ability to assess relevance across the universe of documents in the collection.

6. BUILD CONTROL BEFORE TRAINING

The control set serves as an independent yardstick for measuring the performance of the predictive coding system. It is important to create the control set prior to training. This approach is in contrast with earlier methodologies, where the control set was created after training. The “control-first” approach facilitates use of the control set as an independent measure for monitoring of training. Rather than relying on “intuition” or arbitrary measures, the “control-first” strategy equips the user with an objective, concrete measure of the training process, and with a clear, objective indication of when training can be terminated.

7. USE MANUAL TRAINING MODE

Automatic training refers to a training technique in which document relevance tags, generated by a team of reviewers in a standard review process, are fed into the predictive coding system in order to train the application. For example, for a review of one million documents, review may have been completed for 10,000 documents.
Automatic training would use the relevance tags from these 10,000 documents, to train the system to assess relevance in the remaining documents. Automatic training contrasts with manual training, in which a reviewer is assigned to train the predictive coding system as part of an intensive, dedicated training effort. Wherever possible, manual training is the preferred approach. Manual training tends to yield better quality input, as opposed to the output from a standard review process, because it is focused effort, and the “expert” is very aware of the significance of the training input on each and every document. In addition, training input from a single senior reviewer often tends to be more consistent than input from a large review team. Nonetheless, automatic mode can be a useful option in certain situations, such as plaintiff scenarios or internal investigations, where there is no onus to produce documents to opposing counsel. However, for document productions, where the defensibility consideration is paramount, the best practice, wherever possible, is to prefer manual training.

8. TRACK TRAINING CONSISTENCY

The need to track the consistency of the expert’s training input derives from the garbage-in garbage-out risk discussed earlier. Ideally, the predictive coding application monitors the expert’s input across various different dimensions, to verify that input is consistent. For example, if two very similar documents are encountered in training, and the expert tags one as relevant and one as not, best practice would dictate that this potential inconsistency be flagged for verification.

9. USE GRADUATED RELEVANCE SCORES

Predictive coding systems typically generate, for each document, either a graduated relevance score, for example, on a scale of 0 through 100, or a binary designation, relevant or non-relevant. The graduated relevance score has a number of advantages. First and foremost, the graduated approach enables the user to participate in one of the key business decisions in the e-discovery process; that is the volume of documents to be reviewed. This is a business decision, based on criteria of reasonableness and proportionality. Reasonableness and proportionality, however, vary from case to case, reflecting the mix of risk and cost that the customer is willing to bear in the case, and which, in turn, are a function of the monetary and strategic values at stake. In addition, the graduated scores enable the implementation of emerging new models in e-discovery, such as prioritized review (starting with the most relevant documents and working back) and stratified review (partitioning the review based on relevance scores, with high scoring, high potential documents assigned for in-house review, and low scoring documents assigned for low-cost contract review).
10. VALIDATE RESULTS

As emphasized in Judge Andrew Peck’s opinion in the Da Silva Moore case, quality assurance is a key component of the predictive coding process. The objective of quality assurance in a predictive coding application is to provide transparent validation of the results generated by the application. One of key tests is to verify culling decisions. For example, using the distribution of relevance scores, the user may decide that documents with scores above 24 will be submitted for review. Documents with scores of 24 and below will be culled. An emerging best practice is to “test the rest”, that is, test documents below the cut-off line, to double-check that this area of the collection does in fact contain a very low prevalence of relevant documents. The expectation is that a representative random sample, of statistically appropriate size, will be reviewed and tagged by the expert. Using these tags, it’s possible to assess the extent of relevant documents in the cull zone, and to confirm or modify the cut-off point accordingly.

In conclusion, it should be noted that predictive coding is a dynamic, rapidly developing arena. The best practices described in this white paper will undoubtedly continue to evolve. In documenting these best practices, as they have taken shape over the past year or so, the intention is not to define a universal textbook template for the generic predictive coding project, but to provide a platform from which it will be possible to develop, refine and create new and better practices as the e-discovery industry continues to assimilate the game-changing technology of predictive coding.