New York Law Journal
A NEW YORK LAW JOURNAL SPECIAL SECTION
An ALM Publication
E-Discovery
WWW.NYLJ.COM
MONDAY, OCTOBER 5, 2015

Practical Considerations In Using **Predictive Coding**

BY GARETH EVANS AND JENNIFER REARDEN

BIGSTOCK

Predictive coding has tremendous appeal, at least in theory. As a practical matter, however, many have been deterred from using it because various hurdles can arise. Nevertheless, with some forethought and preparation, and by involving those with the right expertise, many of the hurdles can be overcome, or at least minimized, and parties may more often realize the potential benefits of predictive coding.

What Is Predictive Coding?

Predictive coding—often referred to as "technology assisted review" or "TAR"—uses mathematical and statistical algorithms to determine whether documents are likely to be relevant. To do so, it utilizes machine learning, in which reviewers code sample documents drawn from the overall document population.

Essentially, the predictive coding tool identifies other documents in the population that share similar features with the sample documents coded as "positive" (i.e., relevant or responsive) or "negative" (i.e., irrelevant or non-responsive).

How Does It Work?

To understand how to make predictive coding practical, you first need to have a general understanding of how it works.

The traditional workflow for predictive coding has involved commencing machine learning with a "seed set" of pre-coded documents. The seed set can consist of a sample selected at random, through the use of initial search terms, documents already determined to be relevant documents, or through other means.

After processing the seed set, machine learning is then refined through itera-

GARETH EVANS *and* JENNIFER REARDEN *are partners at Gibson, Dunn & Crutcher's Orange County and New York offices, respectively.*

tive review of "training sets." These are batches of documents that the tool selects for reviewers to code until the predictive coding model is "stabilized," i.e., when additional training does not result in any meaningful improvement in results.

Some predictive coding tools select training documents strategically instead of just randomly, e.g., documents that appear to be close to the boundary between "positive" and "negative," or samples from clusters of similar documents. Using these techniques, the model may achieve stabilization more quickly.

The tool then applies the learning from the seed and training sets to the entire document population. It identifies the likelihood that the remaining documents are either "positive" or "negative," often with relevance scores. A higher score does not necessarily mean that a document is more relevant, but rather that the tool has determined that it has a greater likelihood of being relevant.

Predictive coding can also be effective on foreign language documents, including Asian languages.

Quality Control

An additional step that is frequently taken, although not always deemed necessary, is to "validate" the effectiveness of predictive coding through a quality control check. Reviewers code a random sample drawn from the overall document population, excluding documents from the seed and training sets. This sample is known as the "control sample" or "validation sample."

The coding of the control sample is then compared to the tool's decisions on the same documents. If the number of "false positives" and "false negatives" in the predictive coding results—as compared to the control sample—is acceptable, the training is complete. If not, you may seek to improve the results with further training.

Review Before Production

A few years ago, when predictive coding first gained some notoriety as a technology for document review, some envisioned documents being blindly produced after only the "computer" reviewed them.

The typical workflow that has emerged in practice, by contrast, is to review, prior to any production, documents that the predictive coding tool has identified as likely relevant. This allows for false positives—i.e., irrelevant documents—and privileged documents to be removed before production.

Benefits of predictive coding can also extend to the party requesting documents. A more efficient process—with reviewers having to review fewer irrelevant documents—can result in faster productions.

Continuous Training

Predictive coding technology has been evolving. One noteworthy development has been the appearance of tools utilizing a training methodology known as "continuous active learning" or "CAL." CAL, in effect, combines the training and final review phases described above.

After initially training the predictive model with a seed set, a CAL tool will present reviewers with documents that it has identified as likely relevant and others it has strategically selected for training. The review continues—and the model is continuously trained—until all the relevant documents have been found at the desired rate of recall.

Vendors of CAL tools claim that they train the predictive model faster and that reviewers end up reviewing fewer irrelevant documents than with other tools.

What's in It for the Producing Party?

For the producing party, significantly increased speed, substantial cost savings and improved accuracy are among the potential benefits of an effectively implemented predictive coding protocol.

These benefits are becoming increasingly important as volumes of electronically stored information skyrocket. Meeting court-ordered deadlines, and the often short deadlines that governmental investigators require, has become increasingly challenging. Indeed, we have seen a rise in sanctions for missed deadlines.

The costs of document review can also be extraordinarily high. Traditional search terms often yield high numbers of irrelevant documents. By substantially reducing the number of irrelevant documents, fewer documents require review, which can significantly reduce costs. A study by the RAND Institute in 2012 found that savings from predictive coding ranged from 20-30 percent at the low end to 77 percent at the high end.

Producing parties can use predictive coding in a variety of ways. For example, it can speed up considerably the review of large numbers of documents set aside as potentially privileged.

What's in It for the Requesting Party?

Benefits of predictive coding can also extend to the party requesting documents. A more efficient process—with reviewers having to review fewer irrelevant documents—can result in faster productions.

Additionally, requesting parties in large cases often complain that productions can amount to "document dumps,"

with large numbers of irrelevant documents produced. This can be a product of document requests that are not narrowly tailored. But to the extent it results from reviewers erring on the side of caution when reviewing large volumes of irrelevant documents, it can yield a production that is more narrowly focused on relevant documents.

Unfortunately, the reality is that some requesting parties—particularly in asymmetrical litigation (e.g., an individual or a class against a large corporation)—often seek to use burdensome document requests for leverage. Consequently, such litigants may oppose a producing party's efforts to make document search and review more efficient and less costly.

It is in precisely these types of cases that producing parties are more likely to want to use predictive coding. In negotiating predictive coding protocols, requesting parties have been known to seek to impose hurdles to deter the producing party from using or realizing the benefits of predictive coding. Examples include demanding unrealistically high recall and confidence levels (which can substantially increase the number of documents that must be reviewed in training) and access to the documents (including irrelevant documents) in the seed, training and control sets.

What Issues Will You Face?

Common issues when considering predictive coding include (1) whether your e-discovery vendor has predictive coding capabilities; (2) whether predictive coding will actually yield cost savings; (3) whether you should disclose to the opposing party your intention to use predictive coding; (4) whether you must share with the opposing party the documents used in the seed, training and validation sets; (5) whether predictive coding may be used in combination with keywords or other search methodologies; and (6) whether court approval is necessary.

Importance of Vendor Selection

A party's ability to use predictive coding often depends upon the capabilities of its e-discovery services provider (aka vendor). Parties often do not consider predictive coding until they are well downstream in a case and find themselves faced with the burden and expense of a massive document review. At that point, if they are already committed to a vendor that lacks predictive coding capabilities or they are locked into a contract in which predictive coding pricing is prohibitive, it is often too late.

Many vendors offer predictive coding software that is not their own, but rather is licensed from a software vendor. A potential issue with this arrangement is that the e-discovery vendor passes through to the end user the software vendor's pricing. That pricing is usually fixed (i.e., there is little or no flexibility), based on the volume of documents to which predictive coding is applied, and the rates are often relatively high. Because of this relatively high, volume-based pricing, predictive coding can become impractical in large document volume cases—the very cases where it may be needed most.

Using predictive coding in such situations may only be practical if one reduces the size of the document population to which predictive coding is applied, e.g., by first using search terms before applying predictive coding. In this scenario, the overall recall of the predictive coding output will be reduced to the extent that the search terms miss relevant documents.

Vendors that have developed their own predictive coding software generally will have a greater ability to be flexible and creative with pricing. In addition to lower pricing generally, we are increasingly seeing such vendors offering either a flat fee for predictive coding or bundling it with other technology-based charges (e.g., processing, use of a review platform, and hosting of data). Such vendors also are more likely to have skilled personnel experienced in successfully developing and implementing a predictive coding protocol.

Carefully negotiating predictive coding pricing when first engaging a vendor can be very important. Even with vendors that have their own predictive coding software, pricing can be prohibitive. Parties often do not focus on the pricing for predictive coding when they negotiate the vendor contract. When they are later confronted with the challenges of a large document review, they may face a Hobson's choice between high traditional document review and high predictive coding pricing. The good news is that pricing can be renegotiated sometimes.

Will Your Costs Actually Be Lower?

With cost savings being one of the principle professed benefits of predictive coding, parties can be surprised when it does not actually yield substantial savings. It is therefore important to analyze the likely overall costs in advance.

The costs that a party will incur for predictive coding consist not just of technology costs—i.e., the vendor's charges for using the predictive coding tool (discussed above)—but also, importantly, professional fees, primarily attorney fees for the document review involved in training the predictive model and validating the results. Additional attorney fees may be incurred in negotiating a stipulated predictive coding protocol, if you decide to seek one, and in motion practice if the parties are unable to reach agreement and you nevertheless want advance court approval of the protocol.

The attorney fees that will be incurred in training the predictive model, validating

the results, and conducting a final pre-production review often depend on the number of documents that must be reviewed (and on the rates of those conducting the review). Many contend that "expert reviewers" should conduct training and validation.

Where the prevalence of relevant documents in the document population is very low, or the targeted recall and confidence level high, it will typically be necessary to review significantly larger samples to train the predictive model and validate the results, which likely will drive up costs. Here, the predictive coding tool used can make a difference, as vendors with active learning tools claim that training and validation is much more efficient in low prevalence situations than with random sampling based tools.

Predictive coding may not yield significant savings where the prevalence of relevant documents is very high. Under those circumstances, predictive coding is less likely to bring substantial efficiency gains over manual review. In other words, there may not be sufficient document review savings to offset the cost of using the predictive coding tool.

Disclose to the Other Side?

Whether to disclose your intention to use predictive coding to the other side, and to seek agreement on a stipulated predictive coding protocol, is another important decision point.

Courts generally encourage parties to disclose and seek agreement on a protocol (but acknowledge that such "cooperation" is not strictly required). Requesting parties may have legitimate interests in ensuring that the predictive coding protocol is sound, as the predictive model (like search terms) is designed to eliminate documents from being reviewed. Moreover, a significant risk of non-disclosure is that it may expose the predictive coding process to challenge by hindsight, similar to unilaterally using search terms without having sought and obtained advance agreement on search terms with the requesting party.

Disclosing and seeking a stipulated protocol, however, can often lead down a path of protracted negotiations and motion practice. Under some circumstances, the requesting party may actually seek to obstruct the producing party's use of predictive coding by demanding unreasonably high recall figures or demanding to participate in the training process in a manner that most producing parties will find unacceptable.

One approach that some producing parties have taken to mitigate or avoid the risks of non-disclosure, while also avoiding the downsides of disclosure, is to use predictive coding as a means of prioritizing review—for example, reviewing all the documents hitting search terms, but using predictive coding to prioritize review of documents that are most likely to be relevant.

Agree to Share Seed and Training Sets?

In the early predictive coding cases, in order to obtain both agreement with the requesting party and court approval, producing parties were often willing to share seed and training documents with the requesting party (including irrelevant documents). Since then, this issue has become one of the principal hurdles for predictive coding, as many producing parties have become much more reticent to do so.

There is some judicial support for not disclosing the seed and training sets. Magistrate Judge Andrew Peck of the Southern District of New York recently pointed out in *Rio Tinto v. Vale* that there are alternatives to producing seed sets and training documents and coding decisions—"such as statistical estimation of recall at the conclusion of the review as well as by whether there are gaps in the production, and quality control review of samples from the documents categorized as non-responsive."[1] Additionally, with CAL tools, there are no discrete training sets to share and studies have shown that seed sets have much less impact on the results.

Some parties also have agreed to a middle ground, where the producing party has provided access to samples from the seed and training sets rather than the entire sets.

Seek Advance Court Approval?

Finally, whether to seek approval from the court before using predictive coding is another decision that must be made. While advance approval is not strictly required—one court recently commented that such a request was "somewhat unusual"2—it can be helpful to submit disputes about the predictive coding to the court for resolution.

Conclusion

Predictive coding can offer substantial potential benefits, but also involves quite a few issues and potential hurdles. To successfully navigate the process, it is critical to have an e-discovery service provider with the appropriate technology, pricing and experience, as well as the assistance of counsel with the appropriate knowledge and experience of these issues.

1. *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125, 128-29 (S.D.N.Y. 2015).

2. See *Dynamo Holdings Ltd. Partnership v. Commissioner of Internal Revenue*, 143 T.C. No. 9, 2014 WL 4636526 (Sept. 17, 2014).