TEXAS LAWYER
September 3, 2012
An ALM Publication

# Understand Predictive Coding Options


BY JOSHUA L. FUCHS AND BENJAMIN J. WOLINSKY

In today's world, document review in any case with moderate-to-large amounts of e-data consumes a substantial portion of the litigation budget. Although the gold standard for such review long has been an exhaustive manual process in which human eyes review every document, increasing amounts of raw data frequently make this strategy a non-option.

It's simple math. Imagine that a litigant collects 1 million potentially relevant e-documents — not a large amount by today's standards. The litigation team trains a room full of contract attorneys and gives them chunks of data to review. Assuming the contract attorneys review these documents at the rate of one per minute, even at $50 per hour, the initial review stage alone costs more than $833,000. Under that scenario, the gold standard earns its name, and not in a good way.

To the rescue comes predictive coding, also referred to as computer-assisted review. Simply put, predictive coding is using computer software to conduct some portion of the document review.

**SIMPLY PUT, PREDICTIVE CODING IS USING COMPUTER SOFTWARE TO CONDUCT SOME PORTION OF THE DOCUMENT REVIEW.**

Recent cases have thrust predictive coding into the spotlight. Reading the recent case law alone or listening to software vendors, however, will not teach in-house counsel how to evaluate whether and how to use predictive coding.

This article explains the two most common forms of predictive coding — concept-based and active learning-based — in straightforward terms and describes factors the legal department should consider when deciding whether to use predictive coding.

Both kinds of predictive coding use a series of complex linguistic and/or mathematical algorithms to identify patterns within documents. However, they differ in what happens next.

1. *Concept-based predictive coding.* In concept-based predictive coding, the litigation team conducts a series of sample manual document reviews (typically called "seed sets"). The results of those reviews are used to train the tool; usually, the litigation team must review thousands (if not tens of thousands) of documents during this process.

After the tool is trained by the seed sets, it codes the remaining document population. Then, lawyers use a manual review to statistically sample the responsive and nonresponsive results to see if the results meet a predetermined acceptable confidence level. Once the team is satisfied with the statistical sample, all of the responsive documents identified by the tool are produced to opposing counsel *without further human review.*

The cost advantages to concept-based predictive coding can be significant. Although the price to use a concept tool is greater than a traditional review platform, removing much of the human review dramatically reduces the overall costs. Further, the review should be faster and more consistent than a traditional review.

On the downside, most of the data will be produced without human review, thereby losing important litigation team knowledge. Additionally, many lawyers still find it difficult to produce documents to the opposing party without the safety net of one final round of attorney review, no matter how compelling the arguments to do so may be.

2. *Active learning-based predictive coding.* This is concept-based coding with arguably more advanced technology, yet still retaining some human safeguards. A room of contract reviewers still may exist in this model.

The process begins like concept-based predictive coding: The litigation team conducts the initial few rounds of seed sets; once trained, the tool codes the entire document set. The more likely responsive documents then generally go through a more traditional document review. As the reviewers go through those documents, marking some nonresponsive, the tool takes that input and improves its own selection criteria. It continually moves increasingly responsive documents to the top of the review pile, winnowing the number of documents it deems responsive and that humans must review.

This is still a radical departure from current practice, in that humans are not looking at the majority of *nonresponsive* documents. But lawyers generally are more comfortable with this option, because there are still real people reviewing documents before they go to the opposing party.

As with concept-based predictive coding, there is statistical sampling of the nonresponsive documents to ensure accuracy at a predetermined confidence level.

The price of an active-learning tool is generally greater than solely concept-based tools. Additionally, there will be a higher review cost because a team will review all of the responsive documents (unlike in pure concept-based predictive coding, in which humans only review the seed set and statistical sample).

TEXAS LAWYER

September 3, 2012

**BOTH KINDS OF PREDICTIVE CODING USE A SERIES OF COMPLEX LINGUISTIC AND/OR MATHEMATICAL ALGORITHMS TO IDENTIFY PATTERNS WITHIN DOCUMENTS.**

But these costs generally are lower than a traditional manual review.

Also, this option educates the litigation team about the contents of the responsive documents produced, and it minimizes the risk of inadvertent production of privileged or confidential materials. The process also may be easier to justify with a court.

## CONVINCING THE JUDGE

Judicial approval still remains a consideration with any method of predictive coding. The first step in seeking approval is to point to the support found in the rules. Federal Rule of Civil Procedure 26(b)(2)(C)(iii) requires the court to limit the frequency and extent of discovery where the burden or expense outweighs the likely benefit. Texas Rule of Civil Procedure 192.4(b) imposes a nearly identical mandate. This is important because the cost of exhaustive manual review often approaches or exceeds the amount in controversy.

Parties can suggest predictive coding as an alternative that ensures discovery is proportional to the value of the case. Moreover, an attorney can argue that predictive coding is equal to or more accurate than human review.

For example, the National Institute of Standards and Technology and the U.S. Department of Defense co-sponsor the Text REtrieval Conference (TREC). TREC's website notes it was created to "support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies." To this end, TREC's Legal Track evaluates information retrieval technology for use in the legal system. Every year, the Legal Track performs a study evaluating text retrieval technologies, and it recently concluded that predictive coding efforts can achieve results on par with, and maybe superior to, traditional human review.

Finally, an additional necessary component to predictive coding is active project management by the litigation team and software vendor. "Garbage in, garbage out" is especially true with predictive coding. Lawyers must ensure proper front-end planning, establishment of a work flow that's adequate to the intricacies of the case, and allocation of sufficient resources to train the tool and sample the results.

Predictive coding can be a viable tool to control e-discovery costs, but lawyers need to understand how it works so that they can justify its accuracy to a court and its cost to their clients' management teams. Attorneys also must be able to evaluate which form of predictive coding best fits the case, determine what priorities distinguish one case from another and choose the tool accordingly. By understanding the available tools and making informed choices, it might be possible to establish a new and true "gold standard." IHT

*Joshua L. Fuchs of Houston is a partner in Jones Day's business and tort litigation practice, a member of the firm's e-discovery committee and a member of the Sedona Conference. Benjamin J. Wolinsky is an associate with the firm's securities litigation and SEC enforcement practice.*