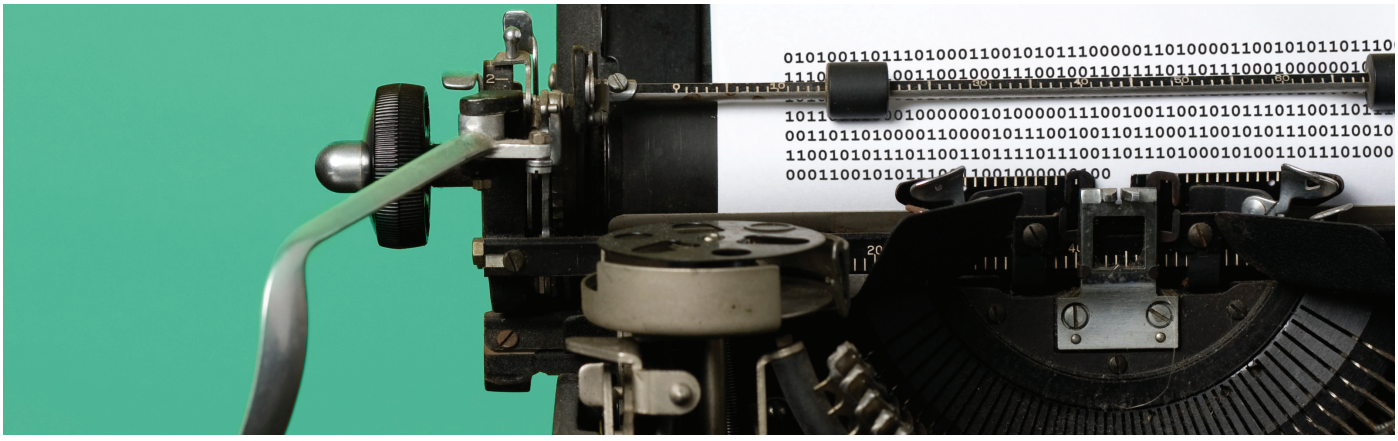


Predictive Coding: It's Here to Stay

Predictive coding programs are poised to become a standard practice in e-discovery in the near future. As more courts weigh in on predictive coding, it is increasingly clear that soon there no longer will be a question of whether predictive coding can be used. Instead, counsel should focus on how and when this technology should be applied.

WALLIS M. HAMPTON
COUNSEL
SKADDEN, ARPS, SLATE, MEAGHER & FLOM LLP

Wallis focuses his practice on business litigation. He defends companies and their directors and officers in securities and fiduciary duty cases, class actions, government and regulatory investigations and complex commercial disputes. Wallis has published extensively on best practices for e-discovery.

Traditionally, the gold standard for identifying potentially responsive electronically stored information (ESI) has been keyword search terms using Boolean logic (for example, "stock /2 option"). Under this method, search terms are electronically applied to identified ESI, with attorneys then reviewing documents that contain those search terms to determine if they are relevant or privileged. Because the legal industry is not an early adopter of technological advances, this traditional method of human review has lingered even in the face of staggering volumes of ESI.

Relatively recently, though, a number of companies have developed advanced algorithms to electronically identify and cull potentially relevant ESI. These more advanced computer-assisted methods of review, including predictive coding, initially caught on somewhat slowly. However, that hesitance has started to dissolve as courts have begun to bless the use of predictive coding and the potential advantages of predictive coding have become more publicized.

Computer analytics like predictive coding programs are poised to become a standard practice in e-discovery in the near future. Moreover, it is even possible that some courts may consider them mandatory for large cases at some future point. The real

uncertainty is not whether predictive coding will be used, but how and when it should be used. Consequently, counsel would be well-served to educate themselves on predictive coding now.

This article examines:

- The basic technology behind predictive coding.
- The ways in which predictive coding can be incorporated into case preparation.
- How courts have recently viewed the use of predictive coding.
- The advantages and disadvantages of using predictive coding.

UNDERSTANDING PREDICTIVE CODING

Broadly speaking, predictive coding refers to the use of a software program to identify documents that are relevant to a particular case or issue. Predictive coding involves a machine learning process and a combination of different algorithmic tools.

THE MACHINE LEARNING PROCESS

In general, attorneys "train" the program by identifying a set of relevant documents (seed set) from a broader set of potentially relevant documents. In one common approach, experienced attorneys who are intimately familiar with the case create the seed set and code each document individually for relevance, privilege or specific issues. The program then analyzes this seed set to understand the types of documents that are relevant to the case. By applying its algorithms to the seed set documents' content and coding, the program learns to identify relevant documents and offers preliminary coding decisions.

Through an iterative process, the program is then trained with additional documents. During this process, an experienced attorney might review the program's coding decisions and accept or reject those classifications. The program then incorporates this feedback into its coding decisions. Just as human reviewers reach different decisions on the relevance of the same document, a predictive coding program may not agree with the experienced attorney's decisions in every instance. The goal is not total congruity but to have the predictive coding program agree with the attorney's coding for a predetermined percentage of the documents (for example, 95%). The iterative process normally repeats for several cycles until the program's predictive coding is sufficiently accurate when compared to the attorney's coding.

Once the iterative training process is finished, the program then looks for relevant documents by applying its coding to the entire data set. What the attorneys do with the documents identified depends on the particular workflow adopted by the attorney review team.

Search Predictive Coding: A Primer for more on how predictive coding works.

COMMON TOOLS

The actual range of potential methodologies and algorithms to perform the training process is sweeping. While every predictive coding tool has its unique algorithms and features, they tend to use similar techniques and processes. Some of the more common methodologies include:

- **Concept searching.** Instead of searching one particular word, a concept searching algorithm considers the meaning of a word to identify potentially relevant documents. It relies on different sources to provide the context in which the word appears, including dictionaries, thesauruses, taxonomies (an organization scheme that looks for similar concepts), ontologies (an organization scheme that looks for related concepts) or mathematical formulas that consider the context in which the word appears. For example, if one of the original search terms was "car," a taxonomy-based algorithm might look for automobiles, trucks and pickups, while an ontology-based algorithm might look for items related to cars, such as drivers and service stations.
- **Contextual searching.** These algorithms consider how and where specified search terms appear in the document, rather than focusing exclusively on search term matches. For instance, if two of the original search terms were "car" and "insurance," the algorithm might focus on whether these related concepts appeared repeatedly in the same discussion.
- **Metadata searching.** Some algorithms focus on certain metadata fields, such as the author, recipient and date fields, to identify relevant materials. For example, if a certain communication between John Smith and Jane Doe on January 23, 2002 is relevant, a metadata searching algorithm might assign a higher priority to other communications between those people during the same time period.

The program can also organize the data using one or more of the following tools:

- **Probability theory.** An algorithm based on probability theory makes decisions about how likely a document is to be relevant. For example, a probability algorithm using concept searching might conclude that a document containing 15 relevant search terms or phrases is more likely to be relevant than a document containing only one relevant term.
- **Relevance ranking.** Many predictive coding programs use their algorithms to rank how likely a document is to be relevant. To illustrate, a document containing 15 relevant search terms might have a ranking of "85," while a document with one search term might have a ranking of only "20."
- **Clustering.** This method groups documents with similar content (as determined by the algorithm), permitting a reviewer to view all documents that appear related to a single concept. For example, a clustering algorithm might group all the e-mails that appear to relate to the same topic, even if they came from different e-mail threads.
- **Sorting documents by issue.** Documents can be sorted and ranked by issues identified by the human reviewers during the training process. This approach can be particularly helpful in identifying the key documents on particular topics at an early stage, or before the start, of the litigation.

Even within these categories, each algorithm is unique. Vendors develop their own proprietary programs and, for

obvious reasons, do not share all the details on how their algorithms work.

Search Reducing E-Discovery Costs: Applying an Analytical Approach for information on how predictive coding, with other analytical tools, can help reduce the volume of ESI at each stage of the litigation process.

USING PREDICTIVE CODING IN LITIGATION

Some attorneys are reluctant to use predictive coding as the primary review tool until the law and practice around it have developed further. Others face resistance from opposing counsel, or even their own clients. However, predictive coding can be used in a variety of ways even if it is not part of the formal methodology used to identify responsive documents. For instance, attorneys can use it to:

- Identify key strengths and weaknesses in a client's case during early case assessments and preliminary investigations.
- Streamline aspects of document review when responding to document requests.
- Analyze a document production received from an opposing party or a third party.
- Prepare for depositions, expert discovery, summary judgment motions and trial.

EARLY CASE ASSESSMENT

Using predictive coding to review client documents as part of counsel's early case assessment protocol can help counsel sort through the client's information and assess the strengths and weaknesses of the case.

As noted above, many predictive coding programs can rank and sort documents by likely relevance. Review teams can initially focus on the documents identified as most likely to be relevant, which often will contain many of the key documents that form the backbone of the case. This early case assessment can pay many dividends, from permitting early risk analysis to identifying key witnesses and allowing more efficient allocation of resources.

REVIEWING CLIENT DOCUMENTS

Predictive coding has obvious value when it comes time to search and review the client's documents for production to the other side. Before using predictive coding as part of a formal review process, counsel often will want to get the other side's consent and, if necessary, the court's approval before incurring the substantial, associated costs. As a practical matter, it probably is easier for counsel to sell predictive coding to an opposing party or its counsel who have experience with it or at least have substantial experience with e-discovery generally. Similarly, a judge with substantial e-discovery experience is likely to be more receptive to predictive coding.

However, even where the parties do not agree to use predictive coding in lieu of the traditional keyword searches, attorneys nonetheless can incorporate predictive coding tools into their internal workflows to make the review more efficient and effective. For example, counsel may use predictive coding to:

- **Prioritize pre-production review.** Attorneys may use traditional keyword searches to identify the universe of potentially relevant documents and use predictive coding to organize and prioritize the review of those documents. In this situation, predictive coding would not change the methodology used to select which documents are responsive, but simply the way that methodology is implemented. Counsel can use the rankings to better organize the review, including by staffing the most experienced or expensive reviewers on the documents that are most likely to be relevant, and the least experienced or expensive reviewers on the rest. For instance, a litigant might decide to have its primary law firm review all documents that scored between 80 and 100 on a probability ranking and assign the remaining documents (which likely represent the majority of the documents) to contract attorneys.
- **Sort documents by potential privilege.** While predictive coding has not proven particularly reliable at privilege calls, it can be used to rank the likelihood that particular documents are privileged. As with relevance calls, the potentially privileged documents can be allocated to different reviewers based on the likelihood of the document being privileged. Moreover, clustering and e-mail threading can help reviewing attorneys ensure consistency on privilege calls across similar documents.
- **Quality control a planned production.** Counsel can compare the results of a linear, human document review with the predictive coding on the same set of documents to assess whether any decisions on relevance or privilege need to be revisited.

REVIEWING OTHER PRODUCTIONS

Counsel also can use predictive coding to review document productions received from opposing parties and third parties. Because the content and organization of these productions may be completely unknown, the ability to quickly rank these documents by potential relevance is extremely valuable, particularly in a fast-moving case.

As with review of client documents, counsel can use one or more of the following tools to organize and understand documents received from opposing counsel or a third party:

- Concept and metadata searching.
- Relevance ranking.
- Clustering.
- Sorting documents by issue.

Additionally, using predictive coding to review other parties' productions can alert counsel to missing categories of information that counsel expected to receive in the production.

Search Discovery Deficiency Letter for a sample letter alerting opposing counsel to perceived deficiencies in its production and requesting additional discovery materials to remedy the deficiencies, with explanatory notes and drafting tips.

OTHER STAGES OF LITIGATION

Predictive coding has potential uses at other stages of litigation as well, including for:

- Deposition preparation, for example, to assemble deponent-specific materials with high relevance rankings.
- Expert report and deposition preparation, for example, to identify documents concerning the subject of the expert's report and testimony.
- Preparing or responding to summary judgment motions.
- Trial.

LESSONS FROM CASE LAW

For years, predictive coding was mired in a state of limbo. Most practitioners continued to use the traditional keyword searches and courts largely ignored the issue. That has started to change in the last two years, as federal and state courts issued a series of decisions addressing whether predictive coding can and should be used.

DEFENSIBILITY OF PREDICTIVE CODING

The landmark decision in *Moore v. Publicis Groupe* represented the first time that a court affirmatively approved a party's use of predictive coding, though both sides had agreed to use it and simply disagreed on the details. The court concluded that predictive coding "now can be considered judicially-approved for use in appropriate cases," but cautioned that it was not holding that predictive coding was required in all cases, or that the protocol used in that case would be appropriate for other cases. (*287 F.R.D. 182, 193 (S.D.N.Y. 2012), aff'd, 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012)*.)

Likewise, in *National Day Laborer Organizing Network v. US Immigration & Customs Enforcement Agency*, the court examined the reasonableness of various government agencies' search efforts in response to a Freedom of Information Act request. There, the court questioned the general effectiveness of keyword searches, noting that "there is increasingly strong evidence that '[k]eyword search[ing] is not nearly as effective at identifying relevant information as many lawyers would like to believe.'" Although the court did not order the agencies to use predictive coding, it did encourage them to do so. (*877 F. Supp. 2d 87, 109, 111-12 (S.D.N.Y. 2012)*.)

In the wake of *Moore*, courts faced with predictive coding issues generally have approved of, or encouraged, its use (see, for example, *Hinterberger v. Catholic Health Sys., Inc., No. 08-380, 2013 WL 2250603, at *3 (W.D.N.Y. May 21, 2013)*; *Global AeroSpace Inc. v. Landow Aviation, L.P., No. CL-61040, 2012 WL 1431215, at *1 (Va. Cir. Ct. Apr. 23, 2012)* (overruling objection to defendant's use of predictive coding without prejudice)).

Search Predictive Coding in Action: How It Compares to Human Review for a case analysis of the predictive coding results from the *Global Aerospace* litigation.

There is, however, some uncertainty over the status of a hybrid method, where a party applies traditional search techniques like keyword searches and de-duplication to limit the full data set and later applies a predictive coding program to the filtered data. There is no judicial consensus on how and when this method may be employed. (See, for example, *In re Biomet M2a Magnum Hip Implant Prods. Liab. Litig, No. 12-2391, 2013 WL 1729682, at *3 (N.D. Ind. Apr. 18, 2013)* (permitting keyword filtering before predictive coding); *Fed. Housing Fin. Auth. v. JP Morgan Chase & Co., No. 11-5201, hr'g tr. at *111 (S.D.N.Y. July 31, 2013)* (party agreed to forgo keyword filtering)).

Additionally, at least one court has suggested that predictive coding may not make sense financially if the universe of potentially responsive documents is fairly small (see *EORHB, Inc. v. HOA Holdings LLC, No. 7409, 2013 WL 1960621, at *1 (Del. Ch. May 6, 2013)* (withdrawing order requiring predictive coding based in part on parties' concerns about a limited document volume)).

Overall, although the law of predictive coding is still in its infancy, the number of courts addressing the issue is clearly on the rise. Courts seem to be moving towards permitting, but not requiring, this technology. Litigants that take reasonable positions and strive to work through their disputes with their opponents will typically be much better positioned to prevail in a predictive coding dispute.

COOPERATION AND TRANSPARENCY

The courts have required varying levels of transparency and cooperation in the predictive coding process. For example, *Moore* required opposing counsel to provide full access to the seed set and to meet and confer about a search methodology (*Moore, 287 F.R.D. at 186-88, 200-203*).

In contrast, another court restricted the discoverability of a party's seed set. In that case, a party applied a predictive coding program to identify relevant documents for production over the objection of the opposing party. When that opposing party sought to discover the irrelevant documents included in the seed set, the court held that production of privileged or nonresponsive materials was outside the scope of discovery and beyond its power to compel. (*In re Biomet M2a Magnum Hip Implant Prods. Liab. Litig., No. 12-2391, 2013 WL 6405156, at *1 (N.D. Ind. Aug. 21, 2013)*.)

At the same time, the *Biomet* court questioned a party's refusal to turn over the training documents and urged the producing party to make its process more transparent (*In re Biomet, 2013 WL 6405156, at *2*; see also *Fed. Housing Fin. Auth. v. JP Morgan Chase & Co., No. 11-5201, hr'g tr. at *8-9, *14-15 (S.D.N.Y. July 24, 2013)*).

Search Rule 26(f) Conference Checklist for a list of topics counsel should be prepared to discuss at the meet and confer, including predictive coding issues.

WHETHER TO USE PREDICTIVE CODING

As a legal matter, the average court today seems unlikely to order the parties to use predictive coding. The true question is whether predictive coding makes sense in the overall framework of the case. The answer will depend on a number of factors, some of which may be beyond counsel's control.

The threshold issue is whether the parties can even agree to use predictive coding. In the private litigation arena, the best scenario for predictive coding involves litigation where both sides face substantial production obligations. In those cases, the parties are more likely to have aligned interests in making discovery as efficient as possible. Conversely, a party with fewer documents to produce may be less likely to take a reasonable approach on production issues.

Beyond that threshold issue, counsel should consider the advantages and disadvantages of predictive coding in their particular case.

ADVANTAGES

The greatest advantage of predictive coding is the potential to dramatically reduce the number of documents requiring attorney review, which ultimately can save time and money (although some people have questioned how significant these savings actually are). It also can:

- Minimize or eliminate the inconsistent production and privilege calls that plague every large document review and allow for a higher level of consistency in the process.
- Identify more relevant documents than the traditional linear attorney review in which documents are reviewed one after another.
- Substantially reduce the risk of being accused of deliberately hiding relevant documents, because it is far easier to justify the nonproduction of an important document where the predictive coding program coded it as nonresponsive.

DISADVANTAGES

Predictive coding has its downsides and limitations. Most significantly, it is not yet a standard practice so there is little certainty about how a court or opposing counsel might view it. Not all predictive coding programs (or vendors) are created equal, and deciding which ones are best for a particular case can be challenging.

Further, many algorithms cannot effectively evaluate spreadsheets or documents without searchable text. Similarly, most commonly-used predictive coding programs cannot yet reliably analyze other file types, such as videos, graphics and audio files, which may be critical in certain types of cases. Thus, counsel will need a good vendor and a strong project manager to tailor the predictive coding program to meet the specific challenges in the case.

Additionally, opposing counsel may press to be actively involved in developing the search methodology for the predictive coding, including reviewing the coding for the seed set and assessing the responsiveness of particular seed set documents. Depending on the court deciding the dispute, opposing counsel may gain access to review irrelevant but still sensitive or damaging documents included in the seed set that would otherwise be shielded.

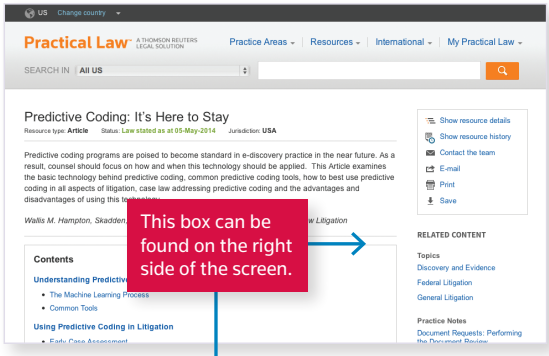
Finally, predictive coding requires significant attention from experienced attorneys during the machine learning process. A flawed seed set or training process will cascade those flaws throughout a production. To guard against this risk, counsel must commit substantial time and financial resources at the start of a case. For this reason, many have questioned whether predictive coding is more cost-effective than traditional attorney review, particularly in smaller cases.

The views expressed in this article are those of the author and not necessarily those of Skadden, Arps, Slate, Meagher & Flom LLP or its clients.

Related Content

The following is a selection of related resources that can be found on practicallaw.com

>> Simply search the resource title



RELATED CONTENT

Practice Notes

E-Discovery in the US: Overview

Document Responses: First Steps in Responding to an RFP

Document Requests: Performing the Document Review

Document Requests: What to Expect in Response to an RFP

Standard Documents

Litigation Budget Template

Budget Template: Document Production

Budget Template: Reviewing an Opposing Party's Documents

Checklists

Case Assessment Decision Tree and Costs Worksheet

E-Discovery Project Management Checklist

Document Discovery Planning Tree

Articles

Learning to Cooperate

Improving E-Discovery Outcomes with ESI Special Masters

Choosing Outside E-Discovery Service Providers