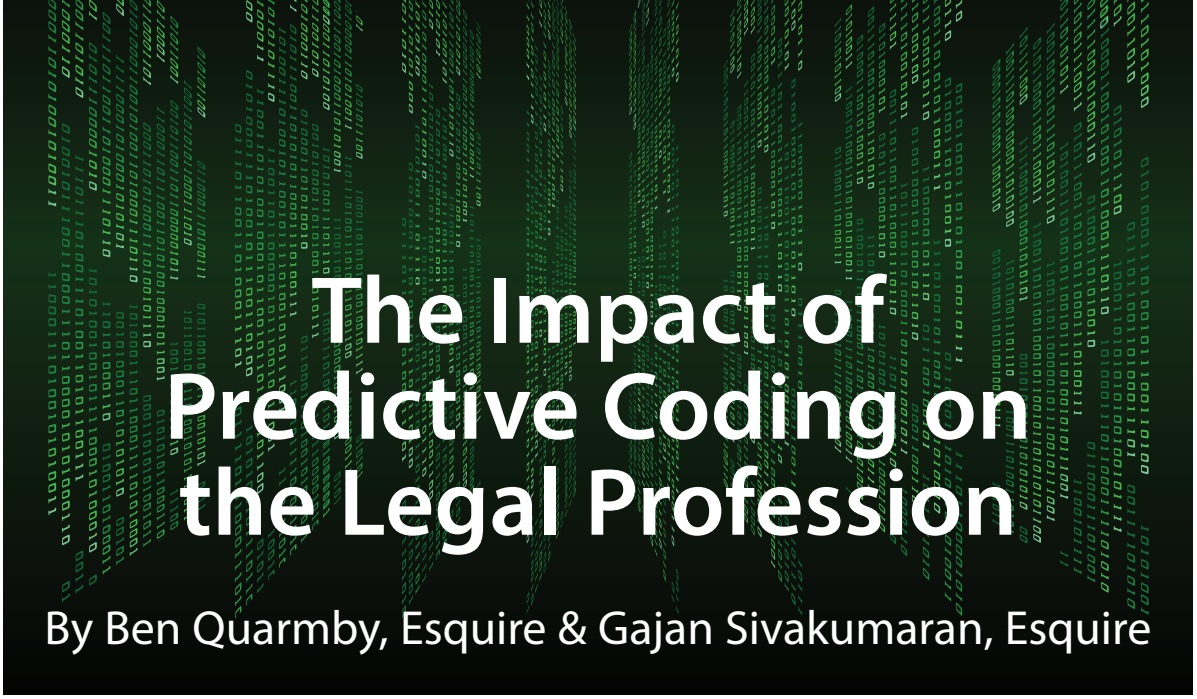

The Impact of Predictive Coding on the Legal Profession

By Ben Quarmby, Esquire & Gajan Sivakumaran, Esquire

PHOTO CREDIT: © iStockphoto.com/fran245

The ubiquity of e-mail in our business life has triggered a well-documented explosion in the volume of electronic discovery in civil litigation. It has also led to a tectonic shift in the practice of law, and to a new norm in which the review of documents relevant to a case has to be performed by armies of attorneys—often third-party contractors—sifting through terabytes of electronic data. But technological innovation may cause the ground to shift again, with the gradual introduction of predictive coding.

A High-Performing Time-Saver

Predictive coding is a tool that allows litigants to sort high volumes of data at great speed. It relies on software to analyze human coding decisions made on a relatively small number of documents (e.g. the decision to code a document as responsive, or privileged, or relevant to a particular legal issue); to decipher the relationships between the words in those coded documents; and to automatically apply the lessons learned from those documents to much larger document sets. Attorneys can thus prioritize documents for review, or even categorically exclude documents from review, based on the preliminary conclusions of the predictive software tool.

Predictive coding allows lawyers and their clients to maintain significant control over electronic discovery costs. Much of the preliminary sorting work that might have been performed by junior attorneys or paralegals can now be handled by predictive coding software, saving both time and money for the client. Electronic discovery vendors offer anecdotal evidence of 50 to 75 percent reductions in the number of documents to be reviewed, and 70 to 85 percent reductions in review costs.

Interestingly, the reliance on software rather than human review also leads to a significant increase in accuracy. Human reviewers, particularly on large-scale projects, are surprisingly inaccurate by any measure: precision (percentage of documents identified by the search method that meet the search criteria), recall (percentage of all responsive documents identified by the search method), and F-measure (a combination of both recall and precision). Large groups of reviewers are also much more likely to apply different coding standards to the same review project, leading to inconsistent coding results. Predictive coding, by contrast, allows for a far more accurate and consistent process.

A Tool Embraced by the Courts

These advantages have caught the attention of the courts. In *Da Silva Moore v. Publicis Groupe*, No. 11-1279, 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012), Judge Peck of the Southern District of New York held that the use of predictive coding was acceptable where the parties had agreed to its use on over 3 million documents. A Virginia state court went a step further in *Global Aerospace Inc. v. Landow Aviation, L.P.*, No. CL61040 (Va. Cir. April 23, 2012), approving the defendants' request to use predictive coding over plaintiffs' objections and allowing defendants to narrow the universe of documents to be reviewed from 1.3 million to 173,000. Judges in other jurisdictions have since approved the use of predictive coding in cases: *Nat'l Day Laborer Org.*

Network v. United States Immigration & Customs Enforcement Agency, 877 F. Supp. 2d 87, 109 (S.D.N.Y. 2012); *EORHB, Inc. v. HOA Holdings, Inc.*, No. 7409-VCL (Del. Ch. Oct. 15, 2012); *In Re: Biomet M2a Magnum Hip Implant Products Liability Litigation* (MDL 2391), No. 312-MD-2391 (N.D. Ind. May 18, 2013); *In re Actos Products Liability Litigation*, No. 11-2299, 2012 WL 7861249 (W.D. La. July 27, 2012).

But these courts remain the exception rather than the rule. The stage is set for predictive coding to become the norm for practitioners, but it will only spread as fast as the courts will allow.

A Multi-Step Process

Attorneys and clients looking to apply predictive coding to a given data set often partner with a specialized electronic discovery vendor, who walks the attorneys through the various steps of the process. While each vendor may have a different approach, the core steps remain the same: Once documents have been collected and uploaded to a review platform, an attorney with deep knowledge of the case will review and code a representative sample of the documents—a "seed set"—of around 500 documents.

The software will then analyze the seed set coding and apply its conclusions to an unreviewed set of about 500 documents. The coding of that second set of documents is proofed by the original attorney, who will correct any erroneous coding and provide feedback to the discovery vendor. The corrections and feedback are then used to optimize the software's performance. The collaborative process between the attorney and the vendor is repeated until the software's error rate is reduced to 5 percent or less—an error rate that lies well below those commonly seen in human-reviewed projects.

Predictive coding is not a perfect solution and it does have its limitations. As a threshold matter, it is highly dependent on the person training it. It is also not necessarily the right choice for every case: It is far better suited to cases involving data sets in excess of 100,000 documents (and ideally in excess of 250,000 documents) than it is for smaller cases. Finally, it is a tool that will only work effectively with documents containing extractable and indexable text—thus, unlike a human reviewer, predictive coding software is unable to process pictures, databases, or spreadsheets.

An Unexpected Boon for Foreign Litigants

Predictive coding offers an interesting opportunity to keep control over e-discovery costs, a feature that is obviously of interest to U.S. litigants. But it is also a feature that is proving attractive to foreign entities looking to assert their rights in U.S. courts, and that might previously have been reluctant to take on the potentially crushing discovery costs associated with complex U.S. civil litigation.

Many entities hoping to take advantage of predictive coding have been inquiring whether the software is capable of processing non-English language documents. The short answer is yes. Indeed, because most predictive coding algorithms only analyze the word structure of a document and do not rely on dictionary definitions at all, they can be used with documents in most languages, so long as all of the documents in a given data set are in the same language. To allow their systems to perform adequately, vendors therefore use software tools to separate documents in a data set into single-language batches of documents, before running independent predictive coding analyses on each batch.

While those approaches are satisfactory with document sets in most languages, they do not necessarily work well with Chinese, Japanese, or Korean document sets. Sentences written in these languages generally lack spaces between words, making it more difficult for machine algorithms to parse whole sentences into individual units of meaning. Vendors have therefore had to devise work-arounds to allow their predictive coding tools to process these documents. Some vendors process the entire document set using a basic machine translation to translate the documents into English before applying the predictive coding analysis. Others use "tokenizers," software tools designed to process the text of Chinese, Japanese, and Korean documents and separate the sentences into individual words. Improvements in this area remain necessary and the full benefits of predictive coding may not yet be available to all entities contemplating foreign-language document discovery in the United States.

A Tool Primed for Widespread Acceptance

In conclusion, predictive coding is a tool that promises to lower the barrier to entry into court for many litigants by reducing the cost of electronic discovery. It is also one that has been shown to improve the overall quality of the legal service offered by attorneys to their clients by increasing the speed and accuracy of document processing.

And yet the use of predictive coding is not as widespread as one would expect it to be. Because the tool is being met with increasing acceptance in courts around the country, it is now the responsibility of outside counsel and the rest of the legal community to embrace the technology and routinely consider it as a discovery option in appropriate litigations. ◆

Ben Quarmby, Esquire, is a partner at MoloLamken LLP in New York, NY and a barrister member of the Hon. William C. Conner AIC. Gajan Sivakumaran. Esq. is discovery counsel at MoloLamken LLP.