# How to Use Analytics and Predictive Coding as Securities Litigators

By Gareth T. Evans and Goutam U. Jois

As children, those of us who grew up in the 1980s and 90s pored over the pages of *Where's Waldo?* books to find the elusive Waldo. The way most of us searched for Waldo—by scanning pages, looking at characters one by one, and deciding whether each character is Waldo—is quite similar to the "traditional" method of document review in complex litigation: assigning reviewers to look at each document, one by one, and deciding whether that document is responsive to the other side's document requests.

But traditional document search and review is usually time-consuming, costly, and difficult, especially in securities litigation, where cases can involve millions of documents. There are substantial benefits to using analytics and predictive coding as tools to assist attorney review. Producing parties can identify key documents earlier and at a lower cost, and requesting parties may receive more relevant documents sooner. And massive volumes of documents can be searched and reviewed much faster and more effectively. That said, these technologies are not silver bullets. They are tools to improve the discovery process, but they still require the involvement of attorneys with expertise in their use.

**What Are Analytics?**

"Analytics" refers generally to applications that help organize documents based on their content and their similarities to other documents. These applications do not rank-order or classify documents as likely to be relevant or responsive.

Imagine the *Where's Waldo?* books described earlier. Each scene in the books depicts a large crowd, and the reader tries to locate Waldo, who wears a striped hat, striped shirt, glasses, and blue pants. In a simple, linear document review, attorneys might review every single document—analogous to scanning every face on every page to spot Waldo. The use of analytics, by contrast, can make the search and review process more efficient, helping reviewers find and follow a trail of relevant documents quickly. Some of the analytics applications are clustering, email threading, near duplication, and conceptual searching. Clustering can group documents together that are conceptually related, allowing reviewers to focus on the most important concepts. Email threading can consolidate all of the emails in a chain so that they are reviewed by a single attorney at once, rather than scattered across the review. Near-duplication and conceptual searching allow attorneys to find more documents that are similar to certain pre-identified key documents. This can be used to consolidate before one reviewer multiple versions of the same document, avoiding multiple reviewers reviewing essentially the same document and potentially making inconsistent coding calls.

To continue the analogy, analytics might sort the crowd in a *Where's Waldo?* scene into groups: people wearing glasses, people wearing hats, people in striped shirts, and so on. In the securities litigation context, you may already have identified some documents related to an important

transaction or business issue. Instead of reviewing the entire universe of documents to find other documents relating to the same transaction or issue, conceptual searching can find similar documents for you and segregate them for review. Newer "visual analytics" tools can display this information graphically, allowing attorneys to click through key concepts quickly and zero in on important documents more efficiently.

**What Is Predictive Coding?**

Analytics can be a "one-two punch" when combined with predictive coding. Where analytics helps reviewers organize and find key documents quickly and earlier, predictive coding uses mathematical and statistical techniques, informed by attorney input, to extrapolate which documents out of a large document population are likely to be relevant and responsive. In the *Where's Waldo?* analogy, a text classification and machine learning algorithm would rank the likelihood that each of the characters on the many pages of a *Where's Waldo?* book is actually Waldo, based on the examples of what Waldo looks like, provided by the reader.

Voltaire remarked that the Holy Roman Empire "was in no way holy, nor Roman, nor an empire." Similarly, predictive coding is neither "predictive" nor "coding" in the narrow senses of those words. Predictive coding is sometimes misunderstood as a *replacement* for attorney review, in which, based on minimal attorney input, a computer almost magically codes every single document as relevant or not, and the relevant documents are produced without further review. Predictive coding is usually a more iterative process, involving attorney input and quality control at various stages.

In addition, it is unusual to produce documents identified by the tool as likely to be responsive without human review. Most commonly, such documents are reviewed before actual production. The predictive coding tool's classification and ranking of documents as likely to be responsive in effect replaces first-pass review. It does not replace second-pass review, which in most cases will still be conducted by human reviewers. Rather, predictive coding eliminates the time and cost associated with human review of documents with a high likelihood of being irrelevant and nonresponsive. In massive securities litigation reviews, such irrelevant and nonresponsive documents usually make up the vast majority of the document population.

Predictive coding, therefore, is a toolthat attorneys can use to make review more efficient, principally by addressing the problems (cost, delay, risk of error, and so forth) associated with having to review large numbers of *irrelevant* documents.

**How Do You Use Analytics?**

As described above, analytics tools help reviewers find key documents quickly and earlier. Imagine, for example, a collection of documents from the underwriter in a securities case. Perhaps these include emails from the key members of the deal team. Analytics can help group the documents in various ways. For instance, if the documents are clustered by concept, the review tool can provide information about the documents that are most representative of each cluster. With that information, reviewers can then decide how to prioritize review.

Perhaps even more important, attorneys may decide that certain clusters need not be reviewed at all (or that only a sample of documents will be reviewed) because the cluster contains irrelevant documents, such as generic "blast" emails sent to all employees or other similarly unimportant documents. Because most documents in any review are usually irrelevant, analytics can help reviewers figure out which small portion of the collection is actually worth looking at closely. Visual analytics tools can provide even more granular information, including key concepts that are prevalent in certain documents and relationships between clusters of documents. When a reviewer "drills down" to view a specific document, some tools can also highlight similar documents that have been deemed relevant.

**How Do You Use Predictive Coding?**

Generally speaking, analytics tools organize documents for review but do not attempt to sort documents by their likely relevance. That is where predictive coding comes in. Most predictive coding work flows involve attorneys initially training the tool with a "seed set" of documents. A seed set may consist of documents the attorneys have already selected and coded (known as a "judgmental sample"). It may also be identified through the use of initial search terms or a random sample or selected through other means.

Once the seed set is coded, a machine learning algorithm ranks documents that are likely to be relevant. Typically, another sample, known as a "training set," is drawn from the documents that have been identified by the algorithm as likely relevant. Attorney experts review this training set and make any adjustments necessary, confirming the algorithm's output or revising it. This process is iterative and continues until the model is "stabilized"—that is, until review of additional training sets does not result in any meaningful improvement in results. Review also may include "validation" of the effectiveness of predictive coding, in which attorneys code a sample of documents (the "control sample" or "validation sample") from the overall collection and compare their results to the algorithm's decision on the same documents. If coding of the control sample is acceptably close to the algorithm's, then training is complete.

What is described above reflects the earlier (and, for now, the most prevalent) predictive coding work flow. More recently, continuous active learning (CAL) tools have been developed that, in effect, combine the training and second-pass review phases, such that reviewers review documents until the tool finds no more responsive documents. CAL tools are also more flexible, allowing rolling uploads (i.e., all documents do not need to be uploaded at the start) and continuous training (the algorithm re-ranks every document each time a new document is coded by a reviewer). Notably, because seed sets are less important in a CAL work flow, and there are no discrete "training sets," the disagreements parties may have over sharing seed and training sets may become moot as CAL work flows become more prevalent. In addition, validation of the results can be tested by reviewing samples from the documents that are "left over," i.e., those that the tool has determined likely to be nonresponsive, rather than reviewing a control set.

**Predictive Coding in Everyday Life**

Although predictive coding may seem exotic, similar technology abounds in our day-to-day experience. There are a number of websites and applications that recommend products, books, or movies based on prior selections. For example, there are music streaming services that create customized "radio stations" for the listener. The initial selection of a stream is analogous to coding a seed set. Giving songs a "thumbs up" or "thumbs down" over time helps refine the application's algorithm, similar to the refinement of responsive documents through training sets. Eventually, the algorithm stabilizes, and you have a radio station almost exclusively comprising songs you like—though, just as with predictive coding, the algorithm isn't perfect, and songs you don't like may creep in once in a while.

While these applications may be challenged in doing a perfect job with matters as nebulous and variable as musical and literary tastes, text classification is a simpler matter. For example, email management apps that sort which emails are important and which are junk, operate on a similar principle and usually do an accurate job. Moreover, when predictive coding is used, a quality control protocol is usually followed to verify how well the tool has performed.

**Potential Advantages in Securities Litigation**

In securities litigation, which can involve collections of millions of documents, analytics and predictive coding will likely provide increased speed, reduced cost, and improved accuracy in document search and review. Regulators (especially the Securities and Exchange Commission, Department of Justice, and Federal Trade Commission) are also increasingly receptive to—and savvy about—the use of analytics and predictive coding in investigations. By reducing the number of irrelevant documents, these tools can reduce the number of documents that are actually reviewed, reducing cost by anywhere from 20 percent to nearly 80 percent, according to a study conducted by the RAND Institute.

These tools can also benefit requesting parties. If the production process is more efficient, a requesting party might receive documents faster. The requesting party may also receive a smaller set of more relevant documents, rather than a "document dump" of documents of marginal relevance. Unfortunately, in our experience, securities cases are often asymmetrical, where one party has very few documents to produce and the other side has quite a lot. In such cases, requesting parties may try to use burdensome document requests as leverage.

Some commentators refer to the so-called "TAR tax"—concessions a requesting party will try to extract in exchange for "allowing" the producing party to use technology-assisted review (TAR). Examples are demanding unrealistically high recall and confidence levels (which can substantially increase the number of documents that must be reviewed) and access to the documents (including irrelevant documents) in the seed, training, and control sets.

Courts, however, are becoming increasingly receptive to, and knowledgeable about, TAR. Indeed, in the recent *Rio Tinto* case, Magistrate Judge Peck wrote, in an order approving the parties' TAR protocol: "One point must be stressed—it is inappropriate to hold TAR to a higher

standard than keywords or manual review. Doing so discourages parties from using TAR for fear of spending more in motion practice than the savings from using TAR for review." *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125, 129 (S.D.N.Y. 2015). Courts' increasing familiarity with TAR may make them more likely to scrutinize unreasonable requests from requesting parties.

**Some Important Considerations**

When using analytics and predictive coding, each step in the work flow may involve judgment calls. For example, clustering may reveal that certain documents are likely mass emails regularly sent to company employees but probably not relevant to the litigation. Should that set of documents be disregarded entirely? Should they all be reviewed to confirm that they are irrelevant? Should attorneys review just a sample? If so, how large should the sample be?

A predictive coding work flow raises similar questions. Most predictive coding algorithms will assign each document a relevance score (the likelihood that a document is relevant). What score should be the cutoff for relevance and, therefore, for production? Attorneys may also test the recall (the percentage of the estimated number of responsive documents that the tool has found) and precision (how well the tool did in eliminating nonresponsive documents) of their predictive coding results. How much precision and how much recall are "enough"? There are no definitive answers to these questions, at least beyond the general concepts of reasonableness and proportionality that apply to all discovery efforts.

Although prior impediments to the use of analytics and predictive coding are falling, some still exist. Pricing is coming down (though not all vendors are created equal), and judicial approval is increasing (though not all judges are familiar with these tools). Clients are increasingly interested in predictive coding.

There are also technical and logistical considerations. Although every vendor promises to offer predictive coding and analytics, quality can vary greatly. Some vendors are much more rigid in their work flows and may not be able to adapt to more complex cases, where documents are collected on a rolling basis or attorneys would like to use a hybrid work flow. Questions about what to disclose are ever-present as well. Although the specific answers to the substantive and logistical questions will vary from case to case, addressing these issues early on will generally help.

**Conclusion**

Analytics and predictive coding can offer substantial benefits in complex litigation, particularly in securities litigation and investigations, although they involve a number of issues. An attorney with expertise in their use should be involved before considering or using these potentially powerful tools. There is no "one size fits all" approach. Rather, these are useful tools that, with close supervision by experienced lawyers and knowledgeable vendors, may help make the discovery process substantially more efficient.

**Keywords:** litigation, securities, e-discovery, analytics, predictive coding, technology-assisted review, TAR

Gareth Evans is a partner in Gibson, Dunn & Crutcher's Irvine, California, office. Goutam Jois is an associate in the firm's New York City, New York, office.