Using Metrics Through the TAR Process

Traditional human review of potentially responsive documents has proven unable to keep up with exploding amounts of data in a cost-effective, reliable and consistent way. As technology assisted review (TAR) becomes a prevalent method for review, legal teams must understand how to incorporate metrics as part of a defensible approach.

Since TAR software and processes are still relatively new, legal teams must develop defensible, quality-tested approaches. In this area, metrics play a critical role. "Metrics should not be viewed as simply a testing mechanism performed once in a while throughout the discovery lifecycle. Rather, metrics are an integral part of the process itself," says Nancy Woods, Legal Technology Services Director at Kramer Levin Naftalis & Frankel LLP. In order to develop useful, relevant metrics, legal teams need to understand the different types of measurements available and how they should be used in each stage of the discovery process.

UNDERSTANDING THE DATA SET

When considering a TAR-based approach, the legal team must start with an understanding of the overall data set and how, or if, TAR can be used effectively for a specific matter.

Understanding the data set will allow the team to begin to develop strategies and identify what must be measured, tracked and captured early on, which should minimize efforts later in the process.

There are two basic types of metrics. There is the documentation of inputs into the systems, some of which are dictated by software and are unchangeable. The second type consists of outputs from the system.

> "Metrics should not be viewed as simply a testing mechanism performed once in a while throughout the discovery lifecycle. Rather, metrics are an integral part of the process itself. "
>
> *- Nancy Woods, Legal Technology Services Director, Kramer Levin Naftalis & Frankel LLP*

Most TAR projects begin with the creation of a target view, which is the data set that remains after the team has performed initial filtering, such as by date and keyword. Understanding the data set will help the team to develop metrics around inputs into the system. Not every type of document lends itself effectively to TAR, which is why it is so important for the team to understand the data set overall.

The types of documents prevalent in a given investigation or lawsuit may limit the appropriateness of a technology-assisted approach. For example, many spreadsheets contain information that may not be easy for TAR to decode thoroughly. So if many of the potentially responsive documents are spreadsheets, or if custodians do a great deal of work using spreadsheets, TAR may not be useful for this particular matter.

Different types of custodians might affect the results, so along with the different file types that custodians use, custodians' job functions should be considered. The type of potentially responsive information produced by those in sales roles may be very different than for those in IT or the C-suite. While TAR may represent a sensible approach for a subset of the data produced by these different groups, it may not be as effective across the entire data set. (This is especially true when looking to apply the decision model used on one set of custodians to another set of custodians.)

The team should also make sure to understand the nuances that files might contain within the data set. If custodians have deliberately tried to be secretive and use vague or common terms or code words, TAR might not be able to identify or distinguish them. When a great deal of subtext is involved—or if the criteria for a responsive or non-responsive document hinges on a particularly nuanced legal analysis or a complicated decision-tree—a computer-based approach may not be as effective. DECISION POINT: At this point, the legal team needs to identify metrics that will help to decide whether TAR is sensible and viable for this particular matter.

DETERMINE THE RATE OF RELEVANCE

With TAR, determining and proving the rate of relevance of the search is critical to the defensibility of the project. The right metrics, properly applied, can go a long way towards proving defensibility.

In order to determine the rate of relevance, the legal team should take a completely random sample, of typically at least 2,000 documents. This provides a representative baseline so that the legal team can compare enough results in the sample in order to test the effectiveness of the planned approach. Typically, a single reviewer or a small group that is most knowledgeable about the case should oversee this aspect and make categorizations, in order to create the highest quality set with fewer chances of contradictory findings.

DECISION POINT: Using metrics from the sample, the legal team can determine the expected volume of relevant documents within the larger set of documents to categorize.

CREATE TRAINING MATERIALS AND PROTOCOLS

Metrics should also be used to create a training set. The set should include several key components:

- A subset of the target view, the seed set, which will be used to "train" the system. This is typically somewhere between 3,000-9,000 documents in order to achieve statistical validity. The larger the number of documents, the better the margin of error. This decision should be based upon risk tolerance and the goals of the review.
- The rate of relevancy that is established from the sample view.
- The margin of error, or how precise the team wants to be. For example, it may be desirable to have a privilege review be very precise, while the first pass might be somewhat broader. This rate can be adjusted with an acceptable margin of error between 0.5% and 5%. The margin of error can be configured, but whatever decision the team makes should be recorded and reported.
- A test of the accuracy of the predictive model. This can be done by applying the model back to the sample set in order to do a side-by-side comparison.
- An agreed-upon confidence level (95% or 99% are both common) should be set and documented.

DECISION POINT: At this point, the team can use metrics to determine if the system has been "trained" well enough to deliver results within acceptable ranges of confidence and margin of error.

TRAINING THE TECHNOLOGY

Since TAR-based approaches rely on an iterative workflow, users must review every step of this phase and conduct multiple rounds. The tool models the information that has been gathered from the training set into a "dictionary" and a set of associated rules. Often, someone on the team simply presses a button to launch this, but the metrics in this phase must include a transparent process around what types of calls and what information the team is giving the system. The team must document the inputs, which can be compared to switches that define the process. When considering the inputs, the team should take into account:

- How many files do not contain text?
- What kinds of documents are involved, since the team will want to avoid some, such as email contacts?
- Whether stems of words should be considered?
- How to handle short words?
- Whether "noise words" should be excluded?
- What was the seed view ration?
- What text block exclusions were taken into account?
- What other factors unique to this matter have been taken into account?

DECISION POINT: This is where the team must use metrics to provide documentation for decisions made around iterative training.

APPLY TO ENTIRE DOCUMENT SET

The team must benchmark the results from the entire target view against the results from the training set. Comparing the predicted values to the decisions made by the original human reviewers determines how accurate the model is. The predicted results will not be 100% accurate compared to the relevancy rate, but should be within an acceptable margin.
For example, if the sample indicates a 10% (+/- 2%) relevancy rate, the predicted relevancy rate of the target view should fall within that band.

Once the team is comfortable that the decision model has been proven, then the model can be applied against the entire target view. There are two factors to consider in this area:

1. Recall—how successful the decision model has been at finding relevant documents.
2. Precision—how successful the decision model has been in finding the relevant documents compared to those that are not relevant.

Part of understanding TAR lies in becoming fluent with the terminology. Here are some common TAR terms and their definitions.

- **Confidence level**—a representation of how often a result is found, expressed as a percentage. A 95% confidence level means that 95 times out of 100, the same result would be delivered.
- **Recall**—how successful TAR was at retrieving the content it was expected to find, expressed as a percentage. For example, imagine you have a pile of iron and plastic filings in a variety of colors. If you are trying to identify all the green iron filings, a magnet would provide near 100% recall. However, it would do so with low precision.
- **Precision**—the percentage of retrieved documents that fits the criteria. While the magnet in the experiment described above would provide excellent recall of green iron filings, precision would be limited by the magnet's inability to separate by color. You would get all the green iron filings (recall), need another tool to remove the other colors (precision).
- **Statistical validity**—how often a result between different measurements can be relied upon and not attributed to random error.
- **Margin of error**—the deviation between the results of a sampling process versus actual results, typically due to the limitations due to the size of the sample, expressed as +/- X% For example, presidential elections can be predicted by polling a small sample of voters. The smaller the sample, the bigger the margin of error.
- **Token**—a string of one or more characters that become significant when grouped together. In the context of document review, tokens are typically simply words, or sometimes word stems.

The TAR engine assigns a score to each document, in essence ranking the documents in order of the likelihood it is responsive. With those scores in mind the legal team can decisions about how many documents to review (perhaps sampling those below a certain score). Alternately, the team may choose to start reviewing the highest scored documents—these are most likely to be key to the case as well—and keep reviewing until it feels it has reached the point of diminished returns. Whatever decisions the team reaches should be thoroughly documented.

At this point, the process becomes more linear and the team can compare ongoing review against performance metrics. At beginning of process, the team may want to batch out the first tranche of identified documents and then have humans review each one to see if relevancy rates are accurate. This will allow the team to determine if the TAR software has been "trained" correctly and if another training iteration should be conducted. As the review progresses and the software becomes more accurate at identifying relevant and non-relevant documents, the team can require humans to review fewer documents. By using the right metrics and applying them at appropriate intervals along the way, the team can also identify areas where problems arise and hone in on those.

Ultimately, the team may decide to apply the predictive decision model to a set of documents without full review; for example, excluding not relevant documents deemed "not relevant" from review and production, while still sampling them to provide validation. Sampling can ensure that the number of relevant documents that has been identified by TAR falls within an acceptable range of expected outcomes. For example, if the team has determined that 1% of documents should be expected, and the sample indicates 10% instead, it is time to decide how to proceed.

DECISION POINTS: At this stage, it's important to use metrics to determine if actual results match expected results and recall rates fall within an acceptable range. The team must also decide how to handle documents "left behind."

CONCLUSION

A TAR approach can represent a cost-effective, efficient and improved way of conducting document review. However, this approach requires knowledgeable users and ongoing vigilance to ensure that the software has been programmed to respond appropriately and that the team and software has adapted as new information emerges during the review process.

Legal teams should use metrics to determine if the software is returning too many non-relevant documents or failing to identify a high enough percentage of relevant files. If the review gets off track, clearly defined metrics can also bring it back on track. Along with providing peace of mind, the proper application of metrics will also improve defensibility.

Interested in learning more?

Contact us at info@elitediscovery.com

400 N. Saint Paul, Suite 1300, Dallas, TX 75201
866.896.2626 | www.elitediscovery.com