epiq results

Technology-assisted review models and investigative features explained

People. Partnership. Performance.

Executive Summary

With the dramatic growth of digital information, corporations and legal teams need new tools to reduce cost and risk. One area that is a key focus is document review, which is often the most important part of an eDiscovery project, and the most costly. For large-volume cases, document review can account for approximately 73% of the total cost of eDiscovery, according to a 2012 RAND study[1].

Technology-assisted review (TAR) uses techniques such as predictive coding (the ranking of documents based on how potentially relevant they are to the matter), filtering and email threading to automate some parts of document review. TAR can be a cost-effective and efficient way to reduce the size of large document sets. The use of TAR is very common in the US and UK, with reported judgments endorsing the technology since 2012 and 2016 respectively. Other jurisdictions, including Australia and Ireland, have also seen TAR used in reported cases. With the growth of electronic data and the increase in cross-jurisdictional matters, European organisations may soon be using TAR as well, even if only for matters originating outside their jurisdiction.

Not only does TAR reduce the effort, time, and cost related to eDiscovery, it also improves accuracy. In many cases, the quality and consistency of the results can exceed what a team of first-level reviewers can accomplish if money and time were not a factor.

While TAR solutions have brought many benefits, early iterations of predictive coding technologies (sometimes called TAR 1.0) were not without their limitations. As technology has developed, a new iteration of predictive coding, called TAR 2.0 or continuous active learning (CAL), has addressed many of those limitations. TAR 2.0 features have reduced costs and timelines even further, and have again improved the accuracy of review. Many of the technology solutions that provide TAR 2.0 come with additional features to enhance results further, with capabilities like sentiment analysis, clustering, and visual summaries. The right iteration of TAR depends on the specifics of your case and the outcome you're trying to achieve.

1 Nicholas M. Pace and Laura Zakaras, "Where the Money Goes. Understanding Litigant Expenditures for Producing Electronic Discovery" (2012) https://www.rand.org/pubs/monographs/MG1208.html

TAR 1.0

TAR 1.0 includes two primary methods of machine learning: simple passive learning (SPL) and simple active learning (SAL), which are simply different ways of selecting the documents that will train the system. In this version of predictive coding, a reviewer knowledgeable about the case, often a senior lawyer, reviews and codes for relevancy a random set of documents to use as a control set for training. Once that is complete, the TAR engine uses these judgments to build a classification/ranking algorithm that will pull in other relevant documents. It tests the algorithm against the already-tagged control set to gauge its accuracy in identifying relevant documents.

Depending on the results, further training may be needed until the system is considered "stable". The search algorithm is judged to be stable when it no longer gets better at identifying relevant documents. Once this system is stable, the TAR engine runs a classification/ranking algorithm against the entire document population. A knowledgeable lawyer should then review another random sample of ranked documents to determine if the algorithm did its job of pulling the relevant documents to the top of the ranking. Once this is complete, the documents can be ranked from most to least likely to be relevant. The documents most highly scored for relevance can be reviewed first, and in certain circumstances, those below a previously determined score can be discarded.

The main advantage of TAR 1.0 is the clarity it gives the legal team. After training the model, the review team knows exactly how many documents they will have to review and a fairly good idea how many relevant documents the review will return. However, TAR 1.0 also has several limitations.

First, TAR 1.0 requires an authority on the matter to train the system. This is often a senior lawyer who is busy and bills at a high rate and it may not be cost-effective to deploy such an expensive resource. Second, the sample given to the document reviewer is completely random, which in a document population with few relevant documents may mean the reviewer spends more time identifying non-relevant documents than relevant ones. New platforms are addressing this challenge, but it still exists to some extent. Third, TAR 1.0 is not good at handling rolling loads. Adding new documents to the system can render the control set invalid, as they were not part of the random selection process. In the end, multiple training rounds may be required, using even more of that senior lawyer's time.

TAR 2.0

TAR 2.0, also known as continuous active learning or CAL, is a form of supervised machine learning. The computer uses a search engine and relevance ranking to select documents for review, coding, and continued training until it can no longer find relevant documents. There is generally no separation of training and review, because by the time the computer stops learning, all documents it deemed relevant have already been identified and manually reviewed.

The first step in CAL involves the computer being exposed to the initial training set, which is normally generated using a simple search via the investigative toolset of the TAR platform. These documents are then manually coded for relevance. Once the computer has a suitable number of relevant documents it will start to train itself, using relevance ranking to present the reviewer a set of likely relevant documents, which are again manually reviewed. After each review batch, the process repeats and the computer's predictions become more accurate and the volume of relevant documents being returned increases. The computer continues going through these steps until it can no longer find any more relevant documents, or until the decision is made that the cost to review outweighs the relevant information being returned.

CAL uses a statistical machine-learning algorithm to repeatedly estimate the likelihood that each yet-to-be-reviewed document will be responsive, based on examination of documents previously reviewed by a lawyer and determined to be relevant or non-relevant. CAL computer algorithms rank a document's relevance in the collection by analysing

features such as words, phrases, or metadata that indicate relevance or non-relevance as measured against the training documents. CAL resembles an internet search engine because its presentation of documents to the user ranks them from most to least likely to be relevant. As it works, CAL refines its decision-making process based on a user's feedback.

CAL algorithms should not be confused with unsupervised machine-learning algorithms used for clustering, near-duplicate detection, and latent semantic indexing, because the latter receive no input from the user and do not rank or classify documents. Clustering is an example of unsupervised learning, some of whose groupings may turn out to be useful and meaningful, or not.

TAR 2.0 saves time and money

Compared to reviewing all of the documents relating to a matter, TAR 2.0 can reduce the time and number of resources required to get to the relevant information.

TAR 2.0 allows the system to continuously learn from the data being reviewed and saves costs by getting to the most relevant material faster, potentially reducing the number of documents that need review and requiring fewer document reviewers. Review teams only need to identify a few relevant documents to start to use the statistical model in CAL. The system can then score documents by likelihood of relevance and can push them to the beginning of the queue for the review team to tag.

TAR 2.0 is more tolerant of coding inconsistencies as the model will correct itself over the following iterations, possibly avoiding large quantities of 'similar' irrelevant documents from being reviewed in subsequent batches. It can also handle multiple tranches of data, adding the new documents to the model, scoring them accordingly, and learning over time any new relevant document types. Using TAR 1.0, you would generally need to re-run the full training process over the new data, adding time and money.

Is TAR 1.0 still relevant?

While, TAR 2.0 offers many improvements on the TAR 1.0 model, there are instances where TAR 1.0 still offers a better, more manageable solution. For example, where you have very large fixed datasets (several million documents), being able to train the system upfront and know exactly how many documents will then require review means that the review can be budgeted and appropriate resources gathered to reach specific deadlines. When using TAR 2.0 it is difficult to make these predictions.

Also, because of the pre-training and the use of control sets, TAR 1.0 provides more accurate statistics such as:

- the richness of the data (the percentage of expected relevant documents in the dataset);
- expected levels of recall (the percentage of potentially relevant documents you are likely to see when reviewing only a proportion of the overall dataset) and;
- the precision of the model (the accuracy of identifying those relevant documents).

Armed with this information, legal teams are able to make informed decisions about the population to review and what degree of sampling is required to achieve confidence over the remaining unreviewed data.

Investigative features of TAR

TAR workflows offer a wide range of investigative functionality that can help identify evidence quickly during many types of investigations. One of the features, sentiment analysis, allows searching for negative sentiment or fraud signals. It can find and rank emails and documents for pressure, rationalisation or opportunity, as well as negative sentiment. We have seen this put to good use on HR disputes.

Another feature, communication mapping, allows the user to visually identify links and trends between individuals which may not be immediately apparent by looking at the individual documents. This can then open up new lines of investigation or help to reduce the data pool to a more manageable number.

By providing a summary of a particular dataset, often called a baseball card, the user can then quickly understand a dataset as a whole, such as drilling into topics of discussion or excluding distribution list emails. This will often be the first step in truly understanding the data.

Advanced document clustering provides links between documents based on their content, automatically grouping similar documents together, which can allow the documents to be viewed as a set, to be compared, or to identify common threads of information.

Traditional document review platforms will allow you to search the metadata of a document; the properties associated with the file, such as created date, hash value, etc. A more advanced form of this is entity extraction, which uses natural language processing (NLP) to find and extract metadata from the content of the document. For example, it will identify all the referenced names, dates, monetary values, etc. from the document body and allow you to search and filter on those values. Using these extracted dates will assist in contract/report identification where the sign-off date could be different to the created/modified date of the document.

A single user may have many references within the data, from various email addresses to versions of their name that they use in the content. Entity normalisation automatically groups this information into a single entity so that if you are searching for 'custodian A', you will be retrieving all variations of their name and email address. This also aids visualising the data as more information is grouped to every person, giving a more accurate view of communications, e.g. where two individuals switch to secondary email addresses to discuss certain topics.

The analytics models can also be used for anomaly detection, where patterns are identified in the dataset, such as 'person A' normally sends emails during business hours, and any key differences in the pattern, such as a sudden group of emails sent on a weekend, are flagged for review. This is often a key step in fraud detection and insider trading.

Is TAR right for your case?

In most cases, we recommend the use of technology-assisted review for projects with at least 25,000 documents to review after standard culling strategies are exhausted. However, there are cases which can benefit from using CAL with as few as 5,000 documents when you need to find the most relevant material quickly. There are experts with the knowledge and expertise who can provide the most suitable solution for your case.

Summary

The smart use of technology and workflows significantly reduces cost, risk, effort, and time. This is especially pertinent at a time when controlling and reducing cost is high on the objective list for many partners and clients in corporate legal and compliance teams. TAR algorithms provide a significant advantage to teams tasked with identifying the relevant content in a mountain of unstructured business communications and documents. Based on TAR features and upgrades, there is no question that lawyers should explore its solutions. As the supervised machine learning involved in CAL becomes more autonomous in training and execution, predictive coding will continue to improve, alongside other capabilities which will provide more benefits for clients.

worldwide resourcefulness

Epiq, a global leader in the legal services industry, takes on large-scale, increasingly complex tasks for corporate counsel, law firms, and business professionals with efficiency, clarity, and confidence. Clients rely on Epiq to streamline the administration of business operations, class action and mass tort, court reporting, eDiscovery, regulatory, compliance, restructuring, and bankruptcy matters. Epiq subject-matter experts and technologies create efficiency through expertise and deliver confidence to high-performing clients around the world. Learn more at www.epiqglobal.com.

80+ offices 14 data centres 5,500+ people

epiq

People. Partnership. Performance. epiqglobal.com

Class Action & Mass Tort | Court Reporting | eDiscovery | Business Process Solutions | Regulatory & Compliance | Restructuring & Bankruptcy