Is it CAR, TAR, RAR...?
(Deep dive into Predictive Coding)

In our previous article "Cross your Techs and Dot your A.Is", we explored how to undertake an internal audit and/or an investigation following a dawn raid. In particular, we discussed how to narrow the scope of the exercise and how to collect the relevant data in an efficient manner while reducing legal spend. But an important question still remains unanswered: How does one navigate through this sea of data collected with a limited budget and limited time without compromising the integrity and quality of the review? The answer is simple: Predictive Coding. In this article, we explain what type of technologies are out there, and how these tools can not only help you reduce legal costs and time but also ensure that the review is conducted in an efficient and consistent manner.

The Myriad of Predictive Coding Tools

There are many acronyms used to refer to 'Predictive Coding' and the subtly different flavours thereof – e.g., '**CAR**' (Computer Assisted Review), '**TAR**' (Technology Assisted Review), '**RAR**' (Relativity Assisted Review) – all of which have different iterations (e.g., TAR 1.0, TAR 2.0, etc). But for the purpose of this article, we will keep it simple and use 'TAR' to refer generally to all of these models.

One of the earliest iterations of TAR is TAR 0.1 which includes two main methods of predictive coding: a simple passive learning and simple active learning which are simply different ways of selecting documents used to 'train the machine'

Lawyers and compliance officers are familiar with TAR 1.0 as it is now widely available in Europe. Yet, not many are aware of the existence of the subsequent iterations/models and what they can do. To many, TAR is reliant on the input of a subject matter expert (generally, a lawyer) through the review of a sample of documents which is then used to 'train the machine' which in turn will apply the same categorisation.

For example, if a business has 200,000 documents to review, it may choose to engage two experts (e.g., lawyers) to review a sample set of 2,000 documents. The technology will then analyse the pattern of the experts' coding and will then identify out of the remaining document pool which documents the experts deem relevant.

The challenge with TAR 1.0 is that it may not be the right tool for a document review where the facts are unknown and what is considered relevant at one point during the review/investigation may become less relevant. In this case, the chances are that the first round of selection may still produce too many false positive. As we all know, this type of document review is iterative and as such, it requires the experts to repeat the training round again and again as they progress the review and discover additional relevant facts. This in turn will cause more delay and incur more costs as not all relevant documents may be selected in the first round. However, it has many pros and can be used in simple disclosure exercise in follow-on damages claims for example where the facts are known

and undisputed. On the other hand, such tool may not be the right one for cartel investigations where known complex facts are limited.

TAR and the New Generations

The new generations of TAR (also referred to as "Continuous Active Learning) is less binary and determines the similarity of the textual concepts contained within them. For example, documents about football might be considered conceptually similar to each other, they might also be considered to be similar to other documents about American Football (but not quite as similar as they are to each other). When the expert reviewer then tags a document about football as being of interest, this predictive coding technology will then identify any other documents which might be relevant to both football and American football. Typically, documents are 'Ranked' as to their similarity to other documents that have already been flagged as being of interest, on a scale of 1-100 with 100 being most similar and 1 being least similar. Other documents about football might then be assigned a rank around 70, whereas documents about American Football might be assigned a slightly lower rank, say 60. Conversely, documents about something completely unrelated would be assigned a much lower rank – maybe 20 or 30.

There are various approaches and workflows one can build around these document ranks, probably the most obvious being a simple prioritisation approach where the documents are sorted by Rank, and presented to the reviewers in descending order. This results in the most likely to be relevant documents being looked at first and the least likely to be relevant looked at last.

Statistical validation or is it all an *elusion*?

So how is this validation carried out? Statistical Sampling is the short answer. The slightly longer answer is this - one common use case of Assisted Review is to cut short a review, once "most" of the relevant documents have been found, and thus saving the time and cost of reviewing all the remaining un-reviewed documents. In order to test the theory that "most" of the relevant documents have already been located, a random statistically valid sample of the as-yet un-reviewed documents is taken, and reviewed by the team. The percentage of relevant documents found within the sample, can be extrapolated to the larger un-reviewed population – if 1% of the sample was relevant, then 1% of the larger population will (statistically speaking) "probably" be relevant. We can then work out the number of Relevant documents that are 'Eluding' us (this statistical sampling process is often referred to as an 'Elusion Test'), and we already know the number of documents that we coded as Relevant during the review thus far – adding the 2 numbers together gives us a total count of all Relevant documents within the document universe, and calculating the % of these Relevant documents that we have already found gives us our current recall % if we were to stop. This allows for a very simple proportionality argument to call a halt to the review (assuming the recall % is sufficiently high of course) along the lines of : We have located 87% of all the relevant documents, to find the remaining 13% we would need to review XXX,XXX documents which would cost £YYY,YYY and take a further ZZ weeks (figures to be filled in as appropriate) – given the overall value of the case, this would be

disproportionate and so we propose to halt the review at this stage, saving the time and costs outlined above. Opposing council, or the judge may of course disagree and request or make an order for further review to find a higher % recall if the case warrants it.

Highly defensible but only when properly validated

Predictive coding is a strong model and defensible when it is properly used, and the results are correctly validated. With any sort of technology-based solution, there is perhaps an inherent caution or even distrust when it comes to relying on the results – after all these are often technically very complex 'black-box' solutions that have been designed by highly qualified and skilled data scientists, statisticians, mathematicians and experts in AI – so it would be unreasonable to expect laypersons in those fields to be able to understand the more arcane workings of the 'Black Box'. However – it is not necessary to understand how the technology does what it does, testing the results, and validating the outcome is what matters most. One should not shy away from using this tool simply because it is too complex to understand. The results will speak for themselves.

Far-reaching potential but no universal panacea

The benefits to utilising TAR in document review are potentially far-reaching. It is generally accepted that document review is the single largest cost in most e-disclosure projects, and anything that can help to reduce that cost in a robust defensible process is obviously going to have a potentially huge impact on both time and cost. In a recent document review exercise carried out for one of our clients encompassing a document population of over 330,000 documents, Inventus implemented a CAL workflow and protocol, managing to achieve a recall of between 93-97% after having reviewed just less than 10% of the total document population (~31,000 docs reviewed), saving the time and cost associated with having to review the remaining ~ 299,000 documents.

However - Assisted Review is not right for every case – The concepts within the document universe are determined based on the text within the documents – If Relevance of a document turns on something other than the concepts contained within it, then Assisted review will be of little benefit. For example, if a specific date range is a factor in determining relevance, because this does not equate to a 'concept' a TAR model will be unable to differentiate between 2 conceptually similar documents, one of which is within a date range and one without. Similarly, for a DSAR (Data Subject Access Request) where relevance may depend on presence or absence of a specific individual's name, TAR would perhaps not be of much use. As with all technology, it comes down to picking the right tool for the job. It is important to have a basic understanding of TAR in in order to best determine when and how it should be used, and to ensure you are designing your process and workflow appropriately but ultimately, the choice of technology will largely depends on a number of factors, e.g., the nature and context of the review, the number of documents, the type of data collected, etc. The key is to always consult with your eDiscovery vendor or local expert to ensure you are using them as part of a robust and defensible workflow.

More about the Authors

Alex Woodrow is Director of Advanced Programmes at Inventus, based in London. Alex joined Inventus 10 years ago as an already seasoned e-disclosure Project Manager, with 9 years prior experience working within some of the largest law firms in the world.

Now a Relativity Expert, Alex has worked with numerous review platforms over his 19-year experience in the e-discovery industry and has developed a passionate interest in AI and its use in the analytics tools that are becoming more and more predominant in the industry.

Alex regularly consults with legal teams to develop robust and defensible workflows and processes utilising and maximising the potential benefits of Analytics solutions.

Marie Leppard is a partner at Euclid Law, The Competition Law Firm. Before joining Euclid Law, Marie was a senior associate at Clifford Chance's antitrust practice (London and Paris).

Marie assists clients on French, UK and EU antitrust investigations, complex multi-jurisdictional mergers and abuse of dominance cases. Marie's practice focuses mainly on cartel investigations. Marie has worked on numerous high-profile cross-border cartel investigations before the European Commission, the CMA, the FCA and the US Department of Justice.

Marie has also vast experience in providing clients with compliance framework and training as well as advising on the use of the latest technologies and artificial intelligence in internal audits and cartel investigations.

About Euclid

Euclid Law was created by experienced competition lawyers with a common desire to build a new competition law firm that is agile, collaborative, highly commercial in its thinking, innovative in its approach to delivering results and free from the constraints of larger law firms.

Our core expertise covers all aspects of competition law, including cartels and anti-competitive agreements, merger control, abuse of dominance, state aid, competition litigation, market investigations as well as audit and compliance. With offices in both London and Brussels, in-depth experience and a network of contacts in key jurisdictions around the world built up over many years of practice, we have the ability to advise clients across Europe and worldwide. We represent clients before EU, UK, German and Belgian authorities and courts.

More information on: https://euclid-law.eu/

About Inventus

A leading legal support services provider focused on reducing the costs and risks associated with the discovery and compliance processes through the effective use of technology solutions. Inventus has been providing services to corporate legal departments, law firms, and government agencies since 1991.

More information on: https://www.uk.inventus.com/