Technology Assisted Review in Disclosure

July 2024

Predictive Coding - Historic Court Approval

Predictive coding (also known as technology-assisted review or TAR) generally reduces the time and cost involved in the disclosure process.

The US decision of *Da Silva Moore v. Publicis Groupe* endorsed the use of predictive coding in the US in 2012. The United Kingdom followed suit in the High Court decision of *Pyrrho Investments Limited & Anr v MWB Property Limited and Others* in 2016.

For context, *Phyrro* involved the review of 3.1 million documents even after deduplication and search terms were applied. Master Matthews, referencing *Da Silva Moore* and I*rish Bank Resolution Corp. v. Quinn* (a 2015 Irish High Court decision where the judge granted approval to the plaintiffs for use of technology assisted review combined with predictive coding) endorsed the use of predictive coding, factoring in its effectiveness and the potential for substantial cost savings.

Legislative Endorsement of Predictive Coding

PD57AD, which governs Disclosure in the UK's Business & Property Courts, specifically provides for the use of technology (*PD57AD, paragraph 3.2 (3), 6.9, 9.6 (3) (a)*) and defines "*Technology Assisted Review*", which "*includes all forms of document review that may be undertaken or assisted by the use of technology, including but not limited to predictive coding and computer assisted review*".

TAR Application

The use of machine learning has evolved since *Phyrro* in 2016 to such an extent that predictive coding alone is not the only TAR methodology relied upon when conducting an effective Disclosure review. For the best output, predictive coding is deployed alongside visual analytics and other unsupervised learning techniques (including clustering, outlier detection, communication analysis, topic modelling, etc); using a multitude of technology assisted review tools assists review teams to conduct a more accurate and robust disclosure exercise, thus helping find key documents early and reducing overall costs.

The main TAR uses in Disclosure are:

- **Automated Document Review**
- **Text Classification and Entity Recognition**
- **Continuous Active Learning (CAL)**
- **Unsupervised Learning**

Automated Document Review

- **Predictive coding** is one form of automated document review. It involves training a machine learning algorithm to recognise the text in relevant documents, based on a sample training set of documents that have been reviewed and tagged as (i) relevant and (ii) not relevant. The algorithm scores each relevant document, analysing certain content like keywords, phrases and metadata. The algorithm then applies the information it has gathered form the relevant documents within the training set to predict the outcome of the unreviewed documents in the dataset.

- **Supervised Learning** is the other form of automated document review. Reviewers with subject matter expertise review and tag a subset of documents, which the machine learning model then applies to the remaining dataset to assess relevance.

Technology Assisted Review in Disclosure
July 2024

Text Classification and Entity Recognition

- **Text Classification** is when keywords are used to categorise documents into various categories, such as confidential, privileged, or potentially relevant/responsive to a specific list of issues for disclosure (LOIFD) issue.
- **Entity Recognition** is when a machine learning algorithm extracts from a document's metadata key entities, such as names, dates, locations, etc., to help organise and arrange documents.

Continuous Active Learning (CAL)

- **Continuous Active Learning** is an iterative process, where a predictive model continuously learns to classify relevant and not relevant documents based on the continued review and tagging by the review team. The algorithm then improves its predictions based on the reviewer tagging application. This is explored in further detail on page 3.

Unsupervised Learning

- **Unsupervised Learning** is used to try to locate and assess novel patterns or unusual documents that may be overlooked in a standard review. It is typically used in document clustering (and outlier detection), theme identification, and fraud detection.
- To provide more context, in unsupervised learning the algorithm analyses and categorises a document population without prior knowledge of the dataset, and without the assistance of any applied tagging or labelling. Clustering, communication analysis, topic modelling, anomaly (outlier) detection and dimensionality reduction are all forms of
- **Clustering** is used to group documents into clusters based on their repeated keywords and their other associated metadata. Clustering can help to group documents by their class, themes or topics as a result. Tools like Latent Dirichlet Allocation (LDA) can uncover hidden topics within a document set, grouping documents that discuss similar subjects together.
- **Clustering outlier detection** algorithms can help to locate documents that cannot be grouped into the broad clusters, thereby indicating potentially unique or relevant information that would otherwise be overlooked. An example of this in practice might be where atypical patterns or unusual words appear in documents, which are add odds with the rest of the same document type, i.e., unusual words in emails or chats in a financial institution might signify some kind of misconduct.
- Similarly, **communication analysis** examines digital communications (including emails, chat messages, social media interactions) that can help identify key participants and understand communication patters, i.e. who is communicating with whom. Communication analysis also helps to detect anomalies within a dataset. This may be of significant assistance where communications are sent at unusual times or from private email addresses by a particular individual.
- **Topic Modelling** extracts themes using algorithms such as Latent Dirichlet Allocation (LDA) which helps identify the number of instances a term occurs, thereby identifying themes in the data that would otherwise go ungrouped. By way of example, where the word "invoice" occurs frequently in similar document types, those topics (and associated documents) would be grouped together.

Technology Assisted Review in Disclosure
July 2024

- **Dimensionality reduction** uses methodologies like Principal Component Analysis (PCA) to transform data into fewer dimensions, i.e. clusters can be viewed in 2D or 3D space, which helps to visualise similarities or differences more obviously.

CAL in Further Detail

How does CAL work?

- The CAL model is constantly updated as new documents are reviewed and tagged. This ensures accuracy that reflects the latest reviewer tagging.
- The CAL model prioritises documents that are likely to be relevant for review, thereby allowing relevant content to be assessed in a ranked priority order.
- A feedback mechanism ensures that the model's predictions can be corrected if needed.
- For a wholly effective review, CAL is used in tandem with clustering and other visual analytics - this setup allows reviewers to effectively categorise their dataset and identify key documents early on.

Precision and Recall

- **Precision** is the ratio of relevant documents retrieved in contrast to the total number of documents retrieved by the system. High precision indicates that documents identified as relevant by the CAL model are mostly accurate, with few irrelevant documents / false positives included.
- **Recall** is the ratio of relevant documents retrieved in contast to the total number of relevant documents available in the dataset. High recall indicates that the CAL system successfully identifies most of the relevant documents, with few relevant documents / false negatives missed.
- **Precision and recall** are calculated continuously to test the performance of the CAL model, i.e., after the initial training phase, during the review, and at regular evaluation intervals before the final calculation at the conclusion of the CAL process) - these metrics help identify whether the model is identifying a high proportion of relevant documents (recall) while limiting the number of irrelevant documents that are subjected to review (precision).
- **Interplay between precision and recall:** Throughout the CAL process, there is often compromise between precision and recall. Adjusting the model to increase one metric may reduce the other, so achieving an optimal balance is best for an effective disclosure exercise.

Contact

Fiona Campbell

Director, Head of Electronic Disclosure

+44 (0)330 460 6620

fiona.campbell@fieldfisher.com