WINSTON & STRAWN

eDiscovery & Information Management White Paper

The Tested Effectiveness of Equivio>Relevance in Technology Assisted Review

Scott M. Cohen
Elizabeth T. Timkovich
John J. Rosenthal

February 2014

Table of Contents

Project Overview

Winston & Strawn has been working for several years with the Equivio>Relevance predictive coding technology ("Equivio"). As a continuation of that work, and at the request of a large institutional client (the "Client") seeking to evaluate the potential cost savings and effectiveness of several different "Technology Assisted Review" ("TAR") methodologies / technologies[1], Winston & Strawn participated in a study (the "Project") whereby we: (i) applied Equivio's predictive-coding technology to a set of Client documents that previously had been reviewed and coded for the Client by outside contract attorneys (the "Data Set"); (ii) estimated the approximate time and costs required to conduct a defensible TAR of those documents using Equivio predictive coding for initial review and culling; and (iii) compared the results of our predictive-coding process to those of a typical, purely human review. The overarching goal of the Project was to complete a substantially accurate predictive-coding analysis within the shortest time possible, requiring the least amount of human (expert) review possible. Based on this stated goal and the size of the Data Set, Winston & Strawn at the outset predicted an estimated completion of its predictive-coding process within 14 days—which proved accurate.

The chosen Data Set consisted of approximately 300,000 documents from a previously resolved pre-litigation matter in which contract attorneys reviewed the documents in Relativity, coding them for relevance, privilege, importance, and a few additional categories (the "Human Review"). These same documents – along with the Human Review coding results – were provided to Winston & Strawn for application of predictive coding using its licensed Equivio>Relevance software. Equivio>Relevance uses a computer-based algorithm to identify potentially responsive and privileged documents and to cull out non-responsive documents.

The following is a summary of the workflow, the results of its application to the chosen Data Set, and an estimation of the likely time and cost for conducting a full TAR of the Data Set, compared to a purely human review.

[1] The terms "TAR" and "predictive coding," as used herein, refer to the use of automated search and retrieval technology to assist in the identification and review of documents.

Breakdown of Data for Review

Winston & Strawn was provided with an overall Data Set of 301,300 documents, but only 293,185 documents were analyzed for the Project. The remaining 8,115 documents were excluded because they did not contain any extracted text[2]—a requirement for the Equivio system to import and analyze the data. The breakdown of the reviewed Data Set is reflected in Figure 1 below.

Category	Count	Percentage
Total document population	293,185	100.00%
Documents marked as responsive during Human Review	18,391	6.27%
Documents produced after completion of Human Review[3]	19,762	6.74%

Figure 1 – Data Set Breakdown

The Equivio>Relevance TAR Workflow

Winston & Strawn's TAR workflow using its licensed Equivio>Relevance technology (the "Equivio Workflow") follows the high-level steps depicted in Figure 1, below. A summary of the time required to complete the Equivio Workflow for this Project can be found at the end of this section.

1. Preparation

The first step in the workflow was the "Preparation" stage. This step includes staging the data as well as defining the issue(s) and training the designated "expert" attorney(s) to understand the process and use the software. These issues included duplicative source folder paths and the need to convert the .xml load files into flat file formats. This process took four days to complete, but with properly formatted data would have required one day.[4]

Next, we imported the data into the Equivio application and set up the application with the appropriate review issue tags and users.

Overall, the Preparation phase for this Project took five days to complete.

[2] Documents that do not contain extracted text are initially identified as exceptions by the processing engine and are subjected to a separate verification/correction workflow before inclusion in any document review. Any such documents found in predictive coding data sets are removed from data set and subjected to standard linear review.
[3] The difference between documents marked in the Human Review as "responsive" and those marked as "produced" resulted from inclusion of non-responsive family members in the final production set.
[4] Typically, if there is the need to cull any data prior to importing a data set into Equivio for review, this task also would be done in the Preparation phase. This would include any objective culling such as domain or date range filtering, as well as the application of any agreed-upon search terms to be utilized. Because the Project was based upon a prior review data set, no such culling was needed in this case.

2. Assessment

The second step in the workflow was the "Assessment" stage. Assessment is the first of two stages (the other is the Training stage) in which expert review of Data Set documents occurs. Assessment begins with the creation of the "control set", a randomly selected set of 500 documents used to evaluate system performance during later stages. Equivio notes in its documentation that, in order to ensure statistical validity of the results, it is critical to separate the "control" documents from the "training" documents. For any matter, the "classifier" – the system component that calculates the relevance score for documents in the Data Set – is created by analysis of the training documents. The performance of the classifier is measured against control set documents. These control documents represent the "gold standard" against which the progress of training is monitored, the performance of the classifier is checked, and results are quantified. In terms of statistics generated, the system uses the control documents to estimate the richness of the collection, and (in later stages) the recall and precision achieved by the classifier.[5]

The Assessment phase for this Project consisted of a review by the designated experts of 540 documents that were selected randomly by the Equivio system. The experts tagged these sample documents as Relevant or Not Relevant. The system then used these 540 documents to generate an initial estimate of richness.[6] For the Project 120 documents (out of 540) were tagged by the experts in the Assessment Phase as Relevant, which equated to a richness estimate of 22.2%.

The Assessment phase took place over the course of two days.

3. Training

The experts next proceeded to a third, iterative "Training" stage. Training began with the system selection of an initial training sample of 40 documents that were then presented to the experts for coding. Once the first sample was coded, the system began an iterative process of selecting sample sets of 40 documents for expert review. While the first training sample was a random sample, all subsequent sample sets were selected using an Active Learning approach. Under Active Learning, each training sample is selected based on what has been learned from the experts' coding of previous samples. In selecting sample documents, the system's objective is to maximize the sample's contribution to the training process—in other words, to choose a sample that will teach the system as much as possible about the population of documents. Based on this criterion, the system selected samples that provide comprehensive coverage of the population (reducing under-inclusiveness), while also fine-tuning and nuancing the concept of relevance that the classifier has developed to date (reducing over-inclusiveness). This multi-layered approach ensures that the classifier is exposed to a cross-section of relevant documents that is as broad as possible, and in so doing, broadens its concept of relevance to capture more of the relevant documents, while, on the other hand, refining the classifier to eliminate false positives.

---

[5] "Recall" is, in a nutshell, the measure of the ability of a system to present all desired (*i.e.*, relevant) documents; and the recall percentage is the number of relevant documents retrieved by the predictive-coding system, divided by the number of relevant documents in the entire data set. A lower recall percentage indicates a larger number of false negatives. "Precision" measures the ability of a system to present only relevant items and determines the accuracy of the review process. The precision percentage is calculated by dividing (i) the number of documents marked as relevant by the system that truly are relevant by (ii) the total number of documents marked as relevant by the system. A higher precision percentage is better, because a lower precision percentage indicates a larger number of false positives.

[6] "Richness" can be defined loosely as the prevalence of relevant documents among the entire document collection during the system training phase.

The experts tagged the sample documents as Relevant or Not Relevant. If the experts could not decide, they could mark the document as Skipped. During the training process, the system issues alerts if an expert provides inconsistent input. For example, if an expert marks one training document as Not Relevant and a near-duplicate as Relevant, the system will ask the expert to verify these tags.

Once the experts completed all 40 documents in the sample, the system calculated training status. Three states are possible: (i) not stable, (ii) nearly stable or (iii) stable. Until stability is reached, the experts continue on to the next sample. For this Project, stabilization was reached after 26 iterations and took between three and four days.

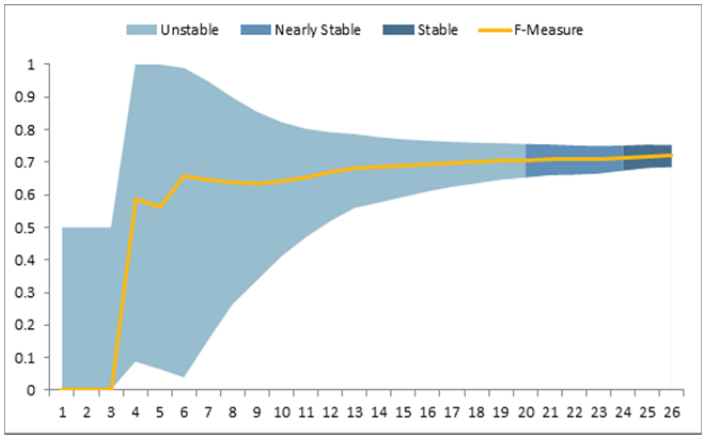

Figure 2 – Training status

In the graph in Figure 2, the yellow line represents the F-Measure (the harmonic mean of precision and recall), and the shaded area represents the margin of error. The system reaches the stabilization point when the marginal contribution of additional samples to the enhancement of the classifier approaches zero.

Once the system is stable, the user can initiate the calculation of relevance scores for the remainder of the Data Set population. Each document receives a relevance score in the range of 0 through 100, with higher scores indicating a greater degree of relevance.[7] The graph in Figure 3 shows the distribution of Data Set documents based on the score that each document received.

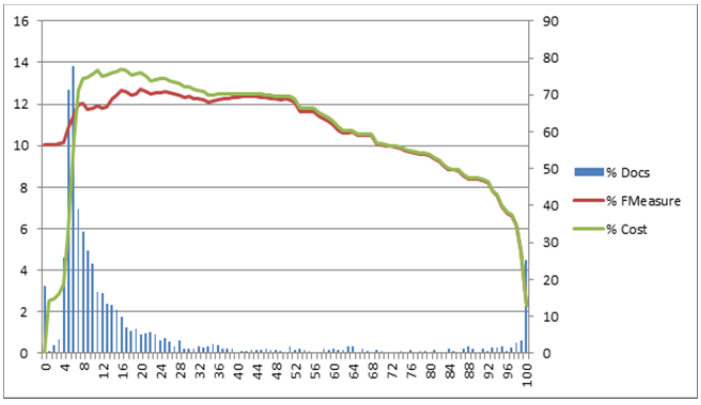

Figure 3 – Relevance score distribution

[7] A relevance score of 0 indicates a document was designated by the experts during assessment or training as Not Relevant. A score of 100 indicates that the experts have designated the document as Relevant. All other scores between these two extremes are system generated.

4. Decision

This brings us to the next stage in the workflow: the "Decision" stage. The Decision stage is the point at which a determination is made as to the minimum relevance score used to designate inclusion or exclusion of documents in the set of documents selected for review. Because every matter has differing requirements regarding a delicate balance between the cost and risk associated with the document review, the business decisions to be made in each case will be different. The decision environment allows the user to easily assess the impact of the relevance score cutoff on a variety of metrics affecting the review. Users can select documents for inclusion in the review set and see what the associated recall will be within that review population. The user sets a cutoff relevance score that corresponds to a specific percentage of the population. Documents with relevance scores below the cutoff mark are culled and will not be reviewed.[8] The Equivio decision-support environment provides information to help the user select the cutoff point. Based on the distribution of documents by relevance scores, the system generates review-to-relevance ratios as seen in Figure 4.

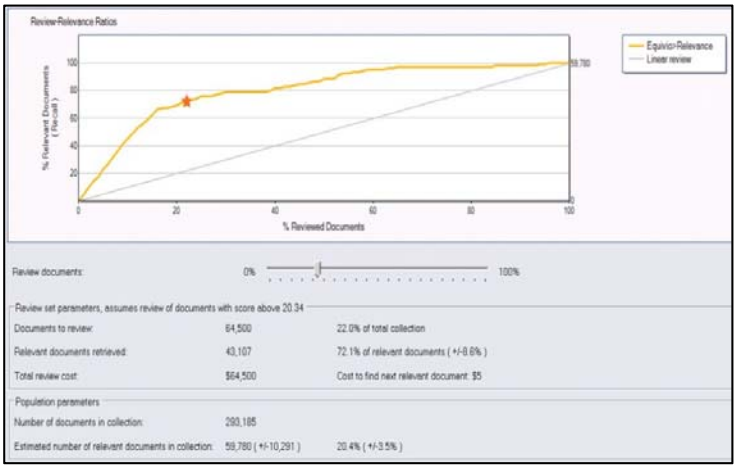

Figure 4 – Decision environment

For this Project, the cutoff point selected resulted in a calculated recall level of 88.6%.[9]

5. Verification / Quality Control

The final phase in the workflow is the "Verification" stage. The Equivio system supports *Test the Rest* (testing prevalence of relevant documents in the area below the cutoff score) and *Test the Method* (a.k.a., Discrepancy Analysis for refining recall and precision estimates of final results and measuring the relative performance of Equivio compared to an alternative culling or review method). For this Project, *Test the Method* was used. [10]

Test the Method was based on comparing the original Human Review to the Equivio predictive coding results.

Test the Method involves the following three steps:

[8] In actual practice a sample of the set of documents with relevance scores below the cutoff would be subjected to review as a quality control measure.

[9] The recall value is calculated by the Equivio system as the result of a complex automated analysis of the control set assessment results, training results and the selection of the relevance-score cutoff point. A description of the exact method of calculation is beyond the scope of this paper.

[10] In typical discovery situations we recommend using a mix of both testing mechanisms.

(1) **Discrepancy matrix**: The first stage of *Test the Method* is generation of the discrepancy matrix by the system. The discrepancy matrix involves cross-matching the relevance designations for the original review documents against the designations Equivio generated for these documents. The discrepancy matrix is a simple 2x2 table, comprising four cells:

(i) Documents tagged in Human Review as R (Relevant) and also designated by Equivio as R

(ii) Documents tagged in Human Review as R (Relevant) but designated by Equivio as NR (this is one type of discrepancy document)

(iii) Documents tagged in Human Review as NR (Not-Relevant) and also designated by Equivio as NR

(iv) Documents tagged in Human Review as NR but designated by Equivio as R (this is the second type of discrepancy document)

(2) **Expert verification**: An expert arbiter reviews a random sample of the discrepancy documents. The system generates this sample, which includes exemplars from both types of discrepancy documents (Reviewer-R/Equivio-NR and Reviewer-NR/Equivio-R). The expert arbiter serves as the ultimate decision-maker or gold standard for determining the relevance or not of these sample discrepancy documents. To help ensure objectivity, the designations of the Human Review and of Equivio for the discrepancy documents are not exposed to the expert arbiter.

(3) **Results**: Recall and precision are calculated for both methods to facilitate comparison of the reviews.

Figure 5 depicts the three stages of *Test the Method* as applied to this Project:

Discrepancy Matrix

		Equivio>Relevance		
		Relevant (X)	Non-Relevant (Y)	Total
Reviewer	Relevant (A)	14902	4345	19247
	Non-Relevant (B)	40023	208305	248328
	Total	54925	212650	267575

Expert Verification

	Verification	
	Sample Size	Actual Relevant
Reviewer Relevant / E>R Non-Relevant (AY)	232	100
Reviewer Non-Relevant / E>R Relevant (BX)	285	94

Results

	Recall	Precision
Equivio>Relevance	88.6%	50.5%
Reviewer	52.4%	85.4%

Results Error Margin: 4.1%

Figure 5 – Discrepancy matrix

For this Project, 267,575 documents were used in the *Test The Method* verification phase. Of these, the Human Review and Equivio agreed on 14,902 Relevant decisions and 208,305 Not Relevant decisions. There were 4,345 documents that the Human Review tagged as Relevant but that Equivio said were Not Relevant; there were 40,023 documents that Equivio identified as Relevant but the Human Review did not. Based upon the discrepancies found during *Test the*

Method, a sample of documents was prepared for the Client's experts to evaluate and determine whether the Human Review or Equivio had made the correct call.

The experts reviewed the samples in random order, without knowing whether the Human Review or Equivio had determined the documents to be Relevant. From the 4,345 documents identified by the Human Review as Relevant, a sample of 232 was selected, and 100 of them were found by the experts as actually Relevant. From the 40,023 documents where Equivio said they were Relevant, a sample of 285 documents was selected, and the experts found 94 of them to be actually Relevant.

These samples allowed us to determine the recall and precision of both the Human Review and the Equivio system. As seen above in Figure 5, Equivio performed significantly better than the Human Review in recall (88.6% compared to 52.4%), while the Human Review had better overall precision (85.4% compared to 50.5%). Interestingly, the F-measure achieved by the human review was virtually the same as that for the Equivio system (64.9 vs. 64.3).

The *Test the Method* and Verification process took one day to perform. [11]

6. Total Equivio Workflow Time for This Project

Figure 6 charts the five Equivio Workflow steps as applied to this Project, and the time it took to accomplish each:

	Stage	Sub-step	Days*
1	Preparation	Data verification and load into Relativity	4
		Import into Zoom	1
		Set-up of Equivio>Relevance	1
2	Assessment		2
3	Training		3-4
4	Decision		N/A
5	Verification	*Test the Method*	1
*Not counting "down" days where process was available, but not used			

Figure 6 – Workflow stage duration

All told, the Equivio Workflow for this Project took **less than 13 days** – one day less than Winston & Strawn's pre-Project prediction – and required expert review (from the Assessment through Verification stages) of **2,151 total documents** (out of the 293,185-document Data Set).

[11] Note: For this Project, the Client also took the added step of subjecting Winston & Strawn's use of Equivio to a final "Golden Set" QC, conducted by an independent expert. The Golden Set methodology began with a random selection and designation of 100,000 Data Set documents by the independent expert, which designations were then compared against the designations made for those documents by Equivio, allowing the Client to verify Winston & Strawn's results.

7. Comparison: Equivio>Relevance Compared to Purely Human Review

The Equivio-based TAR process described above provided the project team with a rare opportunity to compare the use of TAR to traditional human review. Our analysis yielded the following comparison:

- The estimated time required for a typical "expert" to review the 2,151 documents reviewed by the Client's designated experts during the Assessment through post-Decision Verification phases is 43 hours.
- If the Client then proceeded, as we recommend, to additional human review of the 64,500 "potentially relevant" documents, we estimate the time required to be 1,075 hours.
- Further, to perform an additional "Test the Rest" review of 22,869 documents (*i.e.*, 10% of the 228,685 documents scored by the Equivio system as below the chosen relevance cutoff) the time required would be 382 hours.[12]
- The grand total of the time required for these three subsets of review is 1,457 hours.

By comparison, the time required for a traditional, purely human first-level review of the entire 293,185-document Data Set would be 4,886 hours, not counting any QC sampling.

Accordingly, had the Equivio TAR method been used during the actual litigation, even with the additional recommended steps of confirmatory human review and QC sampling, the total time required would have been **significantly less than** the time required for a purely human review.

Conclusion

Equivio>Relevance can be used to reduce the time, cost, and effort associated with electronic document review. As demonstrated in the *Test the Method* analysis described above, review of just 22% of the overall Data Set in this instance would yield approximately 88% of all relevant documents. (This was independently verified by the Client's "QC" analysis of the predictive-coding results.) The user can adjust the selected cutoff to yield higher recall or, alternatively, to lower review costs. The ability of the system to support graduated culling is conditional on the ability of Equivio to generate ordinal relevance scores for the documents in the population (as opposed to binary designations – Relevant or Not-Relevant – as determined "internally" by the predictive coding system). Similarly, the graduated relevance scores enable prioritized review, stratified review, and targeted quality assurance.

[12] The analysis set forth herein is predicated upon the agreed-upon procedures provided by the Client for this exercise. We note that our experience in predictive coding demonstrates significant client savings, but usually not on the magnitude reflected in the above numbers. For example, because the scope of this Project was so limited, the estimated numbers above do not reflect costs associated in reviewing all of the documents to be produced for confidentiality and/or privilege, assuming that would be required, or for "second-level" review of certain documents. It should also be noted that not all cases and document types are amenable to predictive coding. The above analysis does not include the additional cost that is necessary to review documents not amenable to predictive coding—such as image files, voice files, and number-intensive files.

Author Contact Information

Scott M. Cohen
Director, e-Discovery Support Services
scohen@winston.com
+1 212 294 3558

Elizabeth T. Timkovich
Partner; Member, e-Discovery Practice
etimkovich@winston.com
+1 704 350 7780

John J. Rosenthal
Partner; Chair, e-Discovery Practice
jrosenthal@winston.com
+1 202 282 5785

About Winston & Strawn

Winston & Strawn LLP is an international law firm with more than 900 attorneys among 18 offices in Beijing, Brussels, Charlotte, Chicago, Geneva, Hong Kong, Houston, London, Los Angeles, Moscow, New York, Newark, Paris, San Francisco, Shanghai, Silicon Valley, Taipei*, and Washington, D.C. The exceptional depth and geographic reach of our resources enable Winston & Strawn to manage virtually every type of business-related legal issue. We serve the needs of enterprises of all types and sizes, in both the private and the public sector. We understand that clients are looking for value beyond just legal expertise. With this in mind, we work hard to understand the level of involvement our clients want from us. We take time to learn about our clients' organizations and their business objectives. And, we place significant emphasis on technology and teamwork in an effort to respond quickly and effectively to our clients' needs.

Visit winston.com if you would like more information about our legal services, our experience, or the industries we serve.

*Opening Spring 2014