Predictive Coding Proof of Capabilities Study

TYPE OF MATTER: Second Request
CLIENT INDUSTRY: Law firm working with multinational manufacturer
CONSILIO® PRIMARY CONTACT: Pete Feinberg (pete.feinberg@consilio.com, 202.822.6222 x802)
DOCUMENT VERSION: v7
DATE OF UPDATE: December 30, 2015

CONTENTS

EXECUTIVE SUMMARY

This proof of capabilities study sought to compare a predictive coding workflow to the full linear workflow on a fast moving, complex, large scale second request matter that made more complex by very cursory culling, a rolling collection, multiple languages and loose collection hygiene. By using the same processed data to execute and confirm the validity of Consilio's predictive coding solution, we were able to extrapolate practical savings of approximately 63% - about $960,000 – over what we estimate the client to have incurred to execute the full linear review, at 95% recall and 95% precision level.

Further, this case study validates Consilio's claim that our predictive coding solution is capable of handling multilingual datasets with a single responsiveness model; a material departure from the practice employed by many firms and vendors whereby the corpus is sub-divided by document principle language and sent to native-speaking, language-specific review teams for full linear review.

Lastly, this study shows that Consilio's predictive coding solution not only models emails and rich text documents well, but also does particularly well at modeling Excel and PowerPoint file types – which are of particular importance in second request matters.

This capabilities study focused on responsiveness tagging only, and did not consider a second level privilege or redaction workflow.

BACKGROUND

A year prior, the client law firm had serviced a second request that was sizable in scale (over 2.3 million documents for review), complex (6 different language documents identified in corpus), and fast moving (full production to the regulatory agency was due 60 days after the matter began).

The details of the second request and subsequent linear review performed were as follows:

- 692 datasets collected from 60 custodians across 3 countries in a rolling collection that was – as a result of the nature of the second request – relatively broad and non-specific. The data collected was not culled via keyword, date or other cleansing (of non-responsive) method.
- Processing of 95% of the collected 2.945 TB of unstructured data was completed in 34 days; including process flows to OCR of PDF images, perform eFile password cracking, and perform machine language translation.
- Full linear review for responsive docs was done in the US with contract attorneys who began 3 days after the initial receipt of documents, and persisted until full production completed 59 days after the initial release of documents, and scaled up to 216 concurrent reviewers.

ABSTRACT OF PROOF OF CAPABILITIES STUDY

The client approached Consilio to prove our predictive coding capabilities using the source data from this already processed, closed second request matter, and to report back on what the review cycle and cost are likely to have been had predictive coding been applied in lieu of doing a full linear review. And, the client requested details on how the eDiscovery workflow is likely to have gone had predictive coding been applied in lieu of a full linear review approach. Since Consilio had not yet commercialized its predictive coding capabilities at the time of the original matter, this predictive coding capabilities study is done in arrears after the original matte closed.

This study went further still by assessing the feasibility of using a single predictive coding model on a corpus of documents known to contain many different languages. And, Consilio looked to explore the accuracy of the predictive coding model on Microsoft Excel spreadsheet and PowerPoint file types specifically – as those file types have been reported by clients to be less accurately modeled when those clients used other predictive coding software vendor solutions.

WORKFLOW, SETTING UP STUDY DATABASE

Consilio reinstituted the archived review database from the original source matter, and randomly – with no further culling or bias - selected 25% of the documents to model for this study (510,362 unique documents). Consilio selected 25% because this would indicate a reasonable scale for the study but not unduly burden Consilio with uncovered storage or processing expense. For each of those 510,362 documents that would make up the source docs for this study, Consilio processed text and meta data, and imported the text and meta data for those docs into a new Global RPM review database specifically for the purpose of this predictive modeling exercise. During processing, about 5,4~~3~~00 of the 510,362 documents were discovered to be system files and no-text documents, which were not loaded into the predictive coding engine , leaving 504,969 documents as part of our model corpus. This exercise took about 7 hours to complete. Note that in a typical workflow, those system file and no-text exception documents would be foldered for review.

Steps to Create The Study Database		
Step	Description	Compressed Time
1	Reinstitute original source matter from nearline storage	14 hours
2	Randomly grab 25% of original source DB as source documents	2 hours
3	Generate text files of source docs	16 hours
4	Ingest text and meta data of "model corpus" into predictive coding engine	7 hours
	Total	41 hours

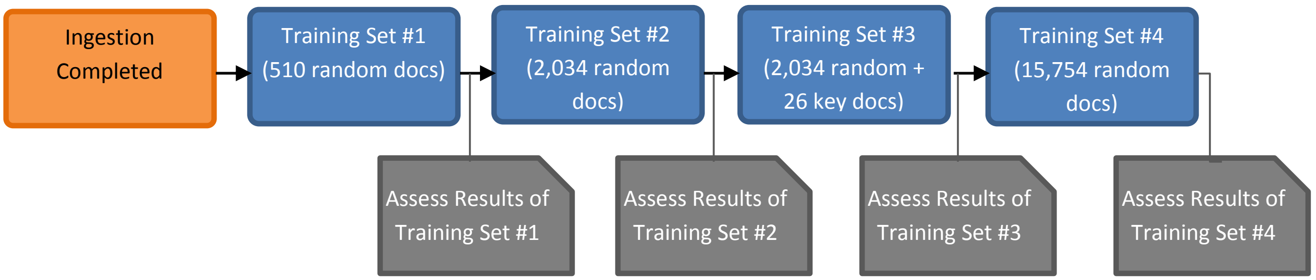
Accounting of Documents Through Ingestion Into Predictive Coding Model
510,362 docs selected from original matter ("source documents")
5,393 system file and no-text docs identified ("exception documents")
504,969 remaining docs ingested into predictive coding engine without exceptions ("the model corpus")

File type analysis of the model corpus concluded that nearly 30% of all documents in the model corpus were PowerPoint (.ppt, .pptx) or Excel (.xls, .xlsx) file types. Consilio had heard in the market that PowerPoint and Excel files are particularly troublesome at accurately modeling by other vendors' predictive coding engines – which made this dataset a more challenging test.

This approach assured that the predictive coding model and dataset for the proof of capabilities study was fully representative of the original matter, and would in no way influence the outcomes of the study. This model corpus created is representative the original matter and affords us a real case scenario, not a "washed" laboratory experience, to adjudge the capabilities of the Consilio predictive coding engine in real-life practical application.

WORKFLOW, TRAINING THE PREDICTIVE CODING MODEL

Once the text and meta data ingestion of the model corpus was complete, Consilio began to train the computer model. There were four (4) discrete steps taken to train the computer model, with assessment of the results after each step as exemplified in this flow diagram below.



The steps of training and each subsequent assessment step are discussed below.

DESIRED GOALS OF THE PREDICTIVE CODING MODEL

Before Consilio began the iterative training stages, goals needed to be established so that Consilio could definitively determine when enough training was sufficient. At the time of this study, Consilio determined that following goals would suffice – as these goals are commonly set by our clients in current matters where predictive coding is utilized:

1. 90% recall (ie. The computer model would find 9 out of 10 responsive documents in the corpus) with
2. 70% precision (ie. For every 10 documents the computer modeled as responsive, the computer would sweep in 3 or fewer non-responsive documents), and
3. No (zero) documents deemed by the computer model to be "non-classifiable" (ie. Documents that the computer cannot determine to be either responsive or non-responsive). *Note: Any document the computer deemed to be "non-classifiable" would require manual eyes-on review.*

TRAINING SET #1 AND #2: GETTING THE COMPUTER MODEL REFINED TO MEET DESIRED GOALS

For Training Set #1, Consilio randomly selected 510 (about 0.1% of the model corpus) documents from the model corpus of 504,969 documents. For each of those randomly selected docs, Consilio found the responsiveness tagging (1 or 0) from the original source matter as coded by the human reviewers during the full linear review. Consilio loaded those training document tag decisions into the predictive coding engine and trained the computer model with an "optimized run" for those exemplar documents – a process that took 12 minutes to complete.

The predictive results were then assessed by Consilio's Project Management team, and a decision was made that further computer model training was necessary to achieve the desired results. With a first training set of only 510 random exemplar documents, the computer's predicted results were 90% recall and 55% precision with no unclassifiable documents. So while our recall and unclassifiable document goals were met, the computer model was not yet refined enough to achieve our desired precision target of 70%.

For Training Set #2, Consilio selected an additional 1,524 (about 0.3% of the model corpus) randomly selected documents from the model corpus of 504,969 documents. Again, Consilio found the responsiveness coding from the original source matter for each randomly drawn document, and again trained the computer model with a second optimized run on the total set of 2,034 (510 plus 1,524) exemplar documents – which took the computer 2 hours to complete.

The predictive coding results were then assessed by Consilio's Project Management team, and a decision was made that the computer model was now successfully trained to our target (1) recall of 90%, (2) precision of 70% without (3) having any documents that the computer deemed "non-classifiable".

In practice, Consilio would have recommended to a client that we cease further model training at this point.

Had these 2,034 documents needed to be reviewed in real-time, this full cycle would have taken about 1 day to complete with 4 review resources in parallel drawing from an assignment pool.

TRAINING SET #3: PROVING SUPERIOR MULTILINGUAL MODELING CAPABILITY

Because the model corpus contained a sizeable count of Portuguese documents, Consilio wanted to prove that corpuses with a composition of multilingual documents could be modeled with a single predictive coding model, and that these multilingual documents did not need to be excluded from a predictive coding workflow. In order to prove this capability, Consilio did a random selection of 26 Portuguese documents from the model corpus of 504,969 documents. Because Consilio used a known characteristic of the documents – in this case, that they were principally Portuguese language – these documents were not randomly drawn; they were considered "biased" documents.

Consilio found the responsiveness coding of each document from the original source matter and ran a third optimized run on the total set of 2,034 randomly drawn documents (from Training Set #2) plus the 26 biased Portuguese documents.

In this scenario, Consilio used its predictive coding solution's functionality to identify the biased documents to the computer model as "non-random". In Consilio's predictive coding solution, such biased exemplar documents are used to refine the computer algorithm, but are excluded from the predicted results of the software – which must only, always be based on randomly selected documents. The run time of this training set was approximately the same as the run time for Training Set #2; about 2 hours. The assessment of this hybrid Training Set #3 were identical to those of Training Set #2 – 90% recall, 70% precision and zero (0) non-classifiable documents. In our QC step (below), we see the real benefit of this approach to modeling mixed-language databases with a single predictive coding model.

TRAINING SET #4: GOING FOR A NEAR PERFECT COMPUTER MODEL

Lastly, Consilio did a final iterative training set using 15,754 randomly selected documents from the model corpus – or about 3% of the overall model corpus. In this Training Set #4, Consilio sought to determine whether training the computer model on 3% of the entire corpus could vastly improve the accuracy of the predictive coding model. This optimized run completed in 2 hours and 15 minutes. The results showed that the computer model was able to be refined to 95% recall with 95% precision and zero (0) non-classifiable documents; as near perfect accuracy as is measurable by Consilio's predictive coding engine.

Summary of Training Set Iteration Results							
Training Set Information				Results of Training Set Round			
Training Set #	# Random Seed Docs (% of corpus)	# Biased Seed Docs	Run Time of Model Training	Predicted Recall	Predicted Precision	# of Non-Classifiable Docs	Goal Achieved?
1	510 (0.1%)	0	12 minutes	90% +/-3%	55%	0	No
2	2,034 (0.4%)	0	2 hours	90% +/-2%	70%	0	Yes
3	2,034 (0.4%)	26	2 hours	90% +/-2%	70%	0	Yes
4	15,734 (3%)	0	2.25 hours	95% +/-1%	95%	0	Yes

HANDLING DISAGREEMENT DOCUMENTS

Note that for none of the Training Sets did Consilio reverse any document coding decision that the computer believed to have been miscoded by the human reviewers –either those that the computer thought were responsive and the human reviewers believed to be non-responsive, or vice versa. Such disagreement documents would normally help refine the computer model further, serving to minimize the number of randomly drawn training documents and lower the cost of training the computer model. However, for this study, Consilio had no resources on staff familiar with the definition of responsiveness, and thus we were unable to perform this accuracy-enhancing task. Thus, in this proof of capabilities study, our results are deemed to be more conservative (more training documents) than would actually have been realized by the client (fewer training documents) to achieve the same results.

VERIFYING THE ACCURACY OF THE COMPUTER MODEL (QC'ING COMPUTER MODEL)

At the completion of the computer model training, Consilio sought to ensure that the computer model was in fact accurate in its predicted results. To do this, Consilio used a "blind test" technique as described below.

Consilio scored each document in the model corpus after Training Set #2– the least amount of computer training that would arrive at the desired 90% recall/70% precision/0 unclassifiable document goals. Then, after scoring each of the model corpus' 504,969 documents, Consilio randomly selected 150 documents that were above the "responsive threshold" (ie. The computer-generated score must be higher than 0.29 on a 1.00 scale to be considered responsive), and were not part of the training set. For each of those 150 "blind test documents", we compared each document's tagging from the original source matter to determine if the document was in fact tagged by the human reviewer as "responsive". [1]

At 70% estimated precision by the computer model, we would allow a maximum of 30% of the 150 "blind test' documents (all of which are above the responsiveness threshold) to in fact be coded non-responsive, and still determine the computer model's predicted results to be "certified accurate". As a result of this blind test QC

[1] *Note that while Consilio recognizes that that the human coding decisions may have been flawed (many empirical studies provide evidence of the inadequacies of human coding decisions), it provided the only possible measurement of the computer's model accuracy for the purposes of this study.*

technique, we determined that of the 150 randomly selected, 107 of them were coded by the human reviewers as responsive and only 43 were coded as non-responsive (28.6% of the 150 blind test documents). Given that 28.6% is lower than the 30% allowance, we conclude that the predictive coding model's predicted results are in fact accurate and are verified to be trustworthy.

HANDLING MULTILINGUAL DOCUMENTS WITH A SINGLE COMPUTER MODEL

Consilio has heard from clients who have used other vendors' predictive coding solutions that they typically sub-divide the corpus by language, sending any non-English language documents through a full linear review workflow, while applying predictive modeling techniques only to the English documents. Consilio believes this to be a sub-optimal and costly approach, and recommends that clients build a single predictive model for responsiveness for all documents in the corpus regardless of language. To prove that this preferred approach works, Consilio executed this multilingual path as part of this proof of capabilities study.

In this model corpus of 504,969 documents, we determined there to be a significant count of Portuguese documents among other languages as determined by Language Vector Analysis (LVA). So in Training Set #3, we augmented the 2,034 randomly drawn seed documents with a biased (non-random) selection of 26 Portuguese documents. These biased documents would be used to further refine the computer model's ability to score Portuguese documents as responsive or non-responsive, but would not in any way bias the statistics that rely on a random sample only. This is a capability that Consilio has not witnessed in other service providers' predictive coding software. The results of Training Set #3 again met our 90% recall/70% precision/0 non-classifiable document goal.

To QC the predictive coding model, we again used a "blind test" technique. First, we scored all 504,969 documents in the model corpus using the computer model created by Training Set #3. Then, we randomly selected 50 Portuguese documents that were above the responsiveness threshold of 0.29, and were not one of the 2,034 random or 26 biased Training Set #3 seed documents – our "blind test documents". For each of those 50 blind test documents, we looked at the human reviewer responsiveness coding decision and found that 42 of the 50 blind test documents were in fact coded responsive by the human reviewer. Since 70% precision would allow for 15 of the 50 to be coded non-responsive, and our measured result was that only 8 documents were human-coded as non-responsive, the computer model is in fact doing a very good job scoring responsiveness for Portuguese documents without having to create a separate computer model.

ACCURACY IN MODELING EXCEL AND POWERPOINT FILES

In the context of a second request, evidence of particular interest to regulators is often found in PPTs and Excel spreadsheets. Yet, Consilio had heard in the market that other service providers' predictive coding solutions were suboptimal at accurately scoring Microsoft Excel (xls, xlsx) and PowerPoint (ppt, pptx) files. Therefore, Consilio sought to verify that our predictive coding solution was in fact accurate at modeling these possibly troublesome file types.

Consilio did a blind test of all (about 81,000 by count) Excel and PowerPoint files in the model corpus that were scored by the computer above the recall threshold for the Training Set #3 (2,034 randomly drawn training exemplars plus 26 non-randomly drawn Portuguese documents). Of those 81,000 Excel and PowerPoint files above the recall threshold, 79.6% of them were human coded as “responsive”, meaning 20.4% of them were coded by the human review team as “non-responsive”. Given that the anticipated precision of the computer model was 70%, this computer model allows for up to 30% of those documents above the recall threshold to be human coded non-responsive - but only 20% of them were. Thus, we can conclude that the Consilio predictive coding solution is in fact accurate at modeling Excel and PowerPoint file types.

EXTRAPOLATING RESULTS TO THE ORIGINAL MATTER

There are two ways to extrapolate how predictive coding could have been used in this full 2.3 million document original matter; the simple extrapolation and the practical extrapolation.

SIMPLE EXTRAPOLATION

If we assume that all 2.3 million documents for matter were received on Day 0 and processed and delivered on Day 0 (utopian case), and we assume that all reviewers were fully applied reviewing documents evenly and consistently for the entirety of the 59 day review cycle, we do a simple comparison between what the linear review would have cost and elapsed time taken versus a review guided by predictive coding. We’ll call this utopian, unrealistic comparison the “simple extrapolation”.

For the full linear review, if we estimate an average of 80 concurrent reviewers for 59 days, at an estimated 60 documents per hour, and 8 hour days, and at an estimated cost of USD$40 per reviewer per hour (a blended market rate considering on shore and near shore resources), the full linear review costs - ignoring processing or collections expenses that would have been incurred in either a full linear review or predictive coding workflow - are projected to have been about USD$1.51 million, or about USD$0.66 per document.

For the predictive coding review, there would have been some cost incurred to review the seed set. If we had used predictive coding to its most refined level as represented by Training Set #4 - which achieved 95% recall, 95% precision and zero (0) non-classifiable documents – we would have had to train the computer model on 15,750 documents. That would have taken (using the same estimates as above) 6.5 days with 5 concurrent reviewers at a cost of about USD$10,000. This would have allowed us to train the computer model and generate scores for all 2.3 million documents at a very high accuracy level. Doing so would have determined that 73% of the 2.3 million corpus set is below the responsiveness threshold (in other words, documents that were likely or very likely to have been non-responsive). The remaining 27% of the documents, about 620,000 documents, were either likely or very likely to be responsive.

If we assume that each of those 620,000 likely or very likely responsive documents must be manually reviewed by a review team, at a USD$0.66 per document cost, the client would have incurred a total cost of about USD$420,000 – for a savings of about USD$1.1 million or about 66% cost savings. Reviewing 620,000 documents would have taken a review team of 40 reviewers (estimated) about 32 days – meaning the production could have completed in fewer than 40 days elapsed time, saving 20 days of time in this simple extrapolation.

PRACTICAL EXTRAPOLATION

Unlike the Simple Extrapolation, this matter represented a rolling collection with rolling processing. In reality, it took 34 days after the first data was released to the review database for 96% of the total 2.3 million records to be released to the review database. Thus, in a predictive coding workflow, we would have needed to periodically create predictive coding models on segmented batches during the rolling collection to help prioritize likely or very likely responsive documents for review.

Round	Day	Batch Released Volume (Docs)
1	4	231,041
2	11	765,315
3	32	1,113,683
Total		2,110,039

With the benefit of hindsight, we can analyze the rolling collection and document release sequencing over the 59 day matter period, and we can surmise that it would have been ideal to execute separate predictive coding batch exercises three (3) times during the 59 day period as shown in the table to the right; after Day 4, Day 11 and Day 32.

Each predictive coding run is on the batch of released documents during that window. So for example, on Day 4, at which point 231,041 documents had been released to the database, we would have run a predictive coding exercise. To do so, we would have used 3% of that batch size – assuming our goal was 95% recall and 95% precision - to train the predictive coding model for responsiveness (6,931 docs, costing USD$4,574 at USD$0.66 per doc). That would have allowed us to classify as either likely or very likely non-responsive 168,660 docs (73% of 231,041 docs), and we would have continued with human review on the remaining 62,381 likely or very likely responsive docs at a cost of $41,172 (assuming USD$0.66 per document review cost).

If we repeated this exercise for batch #2 at Day 11, we would have incurred an additional USD$15,153 in cost reviewing 22,959 docs (3% of the 765,315 batch size). That would have yielded 206,635 likely or very likely responsive docs (27% of the 765,315 total batch size) that would have needed human review at a projected cost of $136,379. Lastly, for batch #3 at Day 32, we would have incurred an additional USD$22,050 to review a 33,410 document seed set (3% of 1,113,683 docs). In addition, we would have incurred USD$198,458 to review the 300,694 docs scored likely or very likely to have been responsive (27% of 1,113,683 batch set).

Batch	Day	Batch Released Volume (Docs)	Size of Training Set (3% of batch size)	Cost to Train Model (USD$0.66)	# of Docs Scored Responsive (27% of batch)	Cost to Review Responsive Docs (USD$0.66)
1	4	231,041	6,931 Docs	USD$4,575	62,381 Docs	USD$41,172
2	11	765,315	22,959 Docs	USD$15,153	206,635 Docs	USD$136,379
3	32	1,113,683	33,410 Docs	USD$22,050	300,694 Docs	USD$198,458
Total		2,110,039		USD$41,777		USD$376,009

In total, running predictive coding would have incurred $41,777 in cost to train three discrete computer models. Plus, reviewing the likely or very likely responsive docs from each batch would have incurred USD$376,009.

Lastly, because our batching would have left out any docs received after Day 32, an additional 196,459 docs were collected and processed into the review database after our last batch. Assuming all of those were manually reviewed at USD$0.66 per doc, that would have added to the project cost another $129,663. So the total projected cost of doing three discrete predictive coding batches to serve this rolling collection is estimated at $547,449 – a savings of about $962,000 over the full linear review costs, or about 63% savings. Because of the staged collections and subsequent predictive coding batching, there would have been no elapsed time savings benefit.

As a final extrapolated result, we calculate savings if we had targeted a lower recall and precision – say 90% recall and 70% precision. Those calculations are shown in the table below. Training three predictive coding models would have cost less than USD$6,000, but it would have cost just over USD$400,000 to do eyes-on review for all likely or very likely responsive documents. And as indicated above, the 196,459 docs collected and processed after batch #3 was completed would have added USD$129,663 in cost, for a total expenditure of about USD$535,000 – or about USD$1 million in savings over the full linear review.

Batch	Day	Batch Released Volume (Docs)	Size of Training Set (0.4% of batch size)	Cost to Train Model (USD$0.66)	# of Docs Scored Responsive (29% of batch)	Cost to Review Responsive Docs (USD$0.66)
1	4	231,041	924 Docs	USD$609	67,002 Docs	USD$44,221
2	11	765,315	3,061 Docs	USD$2,020	221,941 Docs	USD$146,481
3	32	1,113,683	4,455 Docs	USD$2,940	322,968 Docs	USD$213,159
Total		2,110,039		USD$5,629		USD$403,861

COMMENT ON SECOND LEVEL/PRIVILEGE REVIEW

In this study, the workflow assumption is that all documents that were identified by the predictive coding software as likely or very likely responsive would be sent to a first level review team for consideration of receiving a responsive tag or not. All documents that received a responsive tag would then be sent to second level privilege review prior to production. This is analogous to a full linear review workflow whereby all documents tagged responsive by a first level review team are then reviewed by a second level review team for privilege. Thus, to the extent that the predictive coding based classification workflow yielded the same responsive documents as the full linear review workflow, there would be no differential costs either way in second level privilege review.

Consilio's predictive coding solution is quite capable of creating concurrent models – one for responsiveness, one for privilege, etc. While this proof of capabilities was structured to be reflective of a most conservative approach, we could just as well have created a second computer model for privilege that would have allowed us to further subdivide the second level privilege review. Doing so would have undoubtedly yielded further second level review costs savings. This more advanced workflow would have scored each document in the model corpus as not only likely responsive or not, but would have also added a second score for likely privileged or not; allowing the review managers to only send to (typically more expensive) second level review resources those documents that were scored as both likely responsive as well as likely privileged. Again, for the purpose of this proof of capabilities, we chose to not model our workflows in this more advanced manner.