White Paper

Binary Bakeoff:

Ranking Popular TAR Tools

Controlled Study

White Paper

Binary Bakeoff: Controlled Study Ranking Popular TAR Tools

ABSTRACT

Epiq evaluated the performance of three popular, commercially-available Technology Assisted Review (TAR) tools, including Relativity's Active Learning (RAL), Reveal's Brainspace (BRS), and Reveal's Reveal AI (RAI, formerly NexLP), to identify a clear winner if such a distinction existed. Our analysis includes results when using each tool's flagship queue to select the appropriate training documents, compared with the results when a human consultant selected training documents vis-à-vis the algorithms' supervised learning selections. Although the TAR tools did exhibit distinct superiority with respect to model accuracy during our analysis, we discovered something far more interesting: human intervention during the training, specifically the selection of training documents, improved model results across the board, surpassing even the best result without human intervention. We conclude that an experienced consultant can generate better results with any of the tools via custom training round selection through goal-oriented analysis of the results compared to the best "on the rails" offering. In fact, the most important criterion is not the selection of the tool itself but the consultant overseeing the workflow.

TERMINOLOGY

Terminology used in this paper references *"The Grossman-Cormack Glossary of Technology Assisted Review."* [1]

TESTING PARAMETERS

For this test, we evaluated Relativity Active Learning, Brainspace Continuous Multi-Modal Learning (CMML), and Reveal AI COSMIC in a TAR 1.0 workflow. Based on prior linear review from a population of 1,478,999 TAR-eligible documents in Relativity, all human coding decisions formed the basis of both validation and training. Epiq analyzed document file types, text quantity, text quality, and other metadata features to confirm that the documents would make good candidates for machine learning (i.e., document text represented the content of the document). All TAR-eligible documents were ingested into each tool using its specific process.[2] All coding was conducted directly in Relativity utilizing the required field types and choices for each active learning tool, and the bespoke interfaces for each tool ingested the coding from Relativity into each tool's machine learning algorithm for training.

CONTROL SET

To evaluate the quality of the active learning model, the analyst selected 3,000 random documents as a control set. The control set was explicitly excluded from any training documents.[3] Again, the review decisions from the prior linear review provided the control set coding: no additional review was performed to avoid "fitting"

[1] Maura R. Grossman and Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate* Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013)

[2] Differences in feature sets (e.g., metadata) were ignored since normalizing would beg the question of any legitimate machine-learning superiority. That is, BRS and RAI extract features from metadata and text to create entities; whereas, RAL considers extracted text only. It would be begging the question to assume that the addition of features beyond extracted text made a meaningful difference in the results. For RAL, only extracted text was analyzed. For BRS, entity extraction was enabled along with mapping of all standard email metadata and date fields. For RAI, standard entity extraction was enabled along with standard Storybook mapping.

[3] All training rounds explicitly excluding the control set documents to avoid "teaching the test." Although we understand some might consider any accidental inclusion of control set documents for training permissible, we want to avoid any potential bias in the control set.

the control set to the model. Final control set statistics yielded 287 relevant documents and 2,713 not relevant documents for an estimated richness of 9.57% (95% CL, 8.54% - 10.68%).

TRAINING

Training began with a 400-document prevalence sample selected at random from the entire corpus to avoid skewing. The prevalence sample yielded 29 relevant and 371 not relevant documents for an unbiased initial training seed set. After the prevalence sample or random seed set was applied, we leveraged each tool's training queue for future training rounds or sets.

- For Brainspace CMML, we utilized the Diverse Active selection for documents in rounds of approximately 150 documents. Diverse active training rounds are purportedly ideal for "hedging against human bias" and "favors documents that are different from each other and from previous training documents, documents that are similar to many other dataset documents, and documents that have a score near 0.5 under the current predictive model." [4]
- For Reveal AI COSMIC, we utilized the Active Learning mode which is "curated to provide documents for review that will teach the classifier the most, limiting the overall amount of time required to train the classifier." [5] Because the COSMIC interface is designed to work with smaller batches or so-called "cycles," we approximated a checkout size of 100 documents in conjunction with the Cosmic Sync functionality to push the cycle directly to Relativity for coding. Since the COSMIC selection uses a Fibonacci sequence to select documents, the actual selection approximated 144 documents; so, we used a similar target training round with Brainspace CMML.
- For Relativity Active Learning, both the document selection and training intervals are automatic (i.e., training documents cannot be batched, and training happens at preset intervals); so, we built a "click bot" to code documents in the training field as they were served by the coverage queue. The coverage queue differs from the prioritized review queue in Relativity Active Learning where the former selects documents primarily from the middle of scoring spectrum while the latter selects primarily the highest scoring documents. Both queues serve a single document at a time; so, it was not possible to select a batch of documents for bulk coding. We built gating measures to prevent the click bot from coding more than 100 documents per hour which approximates the speed of a Subject Matter Expert (SME).

The specifics of the queue selections fall outside the purview of this white paper, and we encourage the reader to inquire about the specifics of not only Relativity's document selection algorithm but also Reveal's selection design with the understanding that some selection criteria may be trade secret.

TRAINING TARGET

Anecdotally, Epiq notes that most accepted implementations of a TAR 1.0 workflow ultimately select a cutoff yielding 70% - 80% recall, further established by Maura Grossman's *In Re Broiler Chicken* order where she indicated that "...a recall estimate on the order of 70% to 80% is consistent with, but not the sole indicator of, an adequate (i.e., high-quality) review." [6] We ultimately decided to target 75% recall with a minimum of 60% F_2-measure to suspend further training. Since all cutoff measures are ultimately arbitrary, we do not see the need to explain the selection beyond a reasonable means of targeting and prioritizing a minimum level of recall while balancing the value of precision. Typical training efforts fall in the 1,500 – 2,500 total SME reviewed (inclusive of the prevalence sample) range; so, we established a 2,000-document training maximum. Thus, training would continue until the meeting the target recall and F_2-measure or exceeding 2,000 training documents, whichever occurred first.

[4] *Brainspace v6.3 Continuous Multimodal Learning*, p.52. June 8, 2020. PDF.

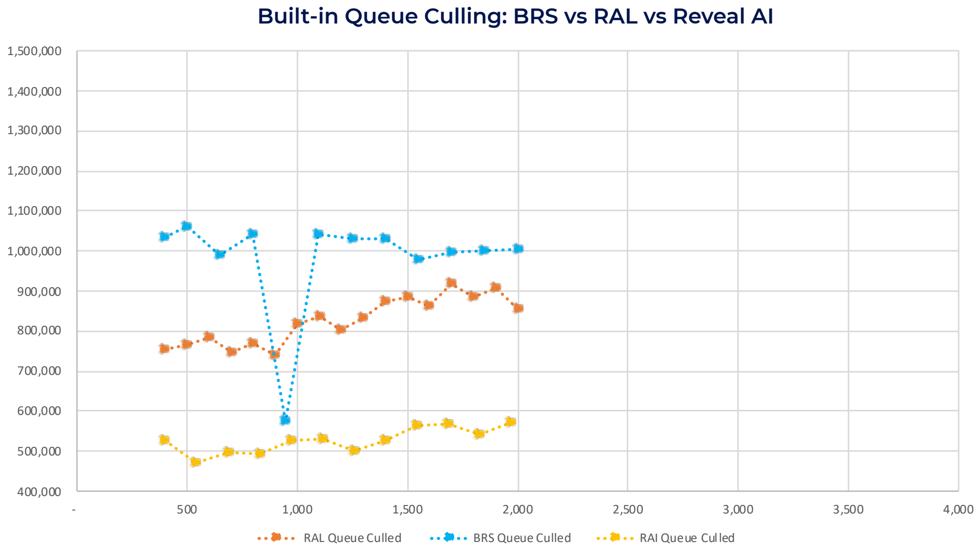
[5] https://review-help.revealdata.com/en/Create-COSMIC-Group.html

[6] Maura R. Grossman, *In Re Broiler Chicken Antitrust Litigation, Order Regarding Search Methodology for Electronically Stored Information, For the Northern District of Illinois, Eastern Division, Case No. 1:16-cv-08637,* p. 10 (January 3rd, 2018).

TRAINING QUEUE RESULTS

Unfortunately, none of the tools' bespoke queues could achieve the target metrics with 2,000 training documents. Brainspace performed the best across the three tools with 50.71% F_2-measure after 2,000 training documents. Relativity Active Learning finished second with 45.32% F_2-measure at 2,000 training documents. Reveal AI finished a distant third with 35.92% F_2-measure at 1,971 training documents. Figure 1 illustrates the culling effect that each tool had based on its full training population. In fact, we performed further testing after evaluating the initial results to test the asymptotic effect of Relativity Active Learning, and we nearly met the Brainspace results after extending training to 3,400 documents and realizing diminishing returns with three additional training rounds.

Figure 1 – Documents culled as a function of model training (built-in queues)



CONSULTANT-DRIVEN TRAINING SELECTION

Given the failure to achieve target metrics, we next evaluated whether the active learning tools training-selection algorithms represented the limiting factor. It is not uncommon to see models struggle with relatively low richness. In fact, we have found that we typically need very few not relevant documents to train machine learning on negative features but require more relevant documents to identify sufficient positive features to score documents adequately. The challenge then becomes how to identify a sufficient diversity of relevant documents without overwhelming the active learning algorithm with negative examples.

The human consultant utilized the exact same starting parameters for repeating these training exercises: identical population, identical control set, and identical prevalence sample or seed set. Training diverged after the random seed set to allow the consultant to evaluate the control set with the current cutoff to achieve the target recall and select samples that would prioritize diverse responsive documents with the greatest impact on the model. The same limits were placed on the consultant-curated training: 75% target recall with 60% F_2-measure not to exceed 2,000 training documents. In this test, Reveal AI finished first with 64.25% F_2-measure; Brainspace finished a close second with 63.38% F_2-measure; and Relativity Active Learning finished third with 61.54% F_2-measure. Figure 2 illustrates the culling effect that each tool had based on its full training population.

Although Reveal AI achieved the highest combination of statistics, Brainspace achieved the target metrics earliest. Relativity Active Learning was the only tool to fail the target recall and F_2-measure within the prescribed limit of 2,000 training documents, but it exceeded the metrics with the same number of queue training documents and ultimately hit the target after 3,712 training documents. For perspective, the worst performer (RAL) from the consultant-curated analysis surpassed the best performer (BRS) from the bespoke training queues (see Figure 3). This significant performance increase from worst consultant-driven performer (63.38% F_2-measure) vis-à-vis the best bespoke queue performer (50.71% F_2-measure) underscores the superiority and, at times, practical necessity of leveraging human analysis of the review and retrieval goals to facilitate an efficient and effective training set.

Figure 2 – Documents culled as a function of model training (consultant-selected training)

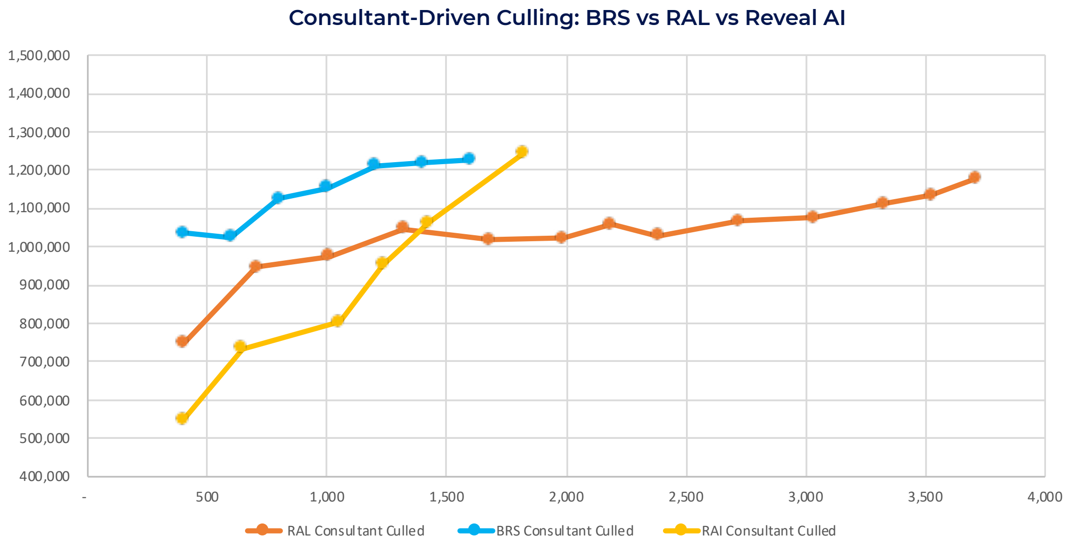
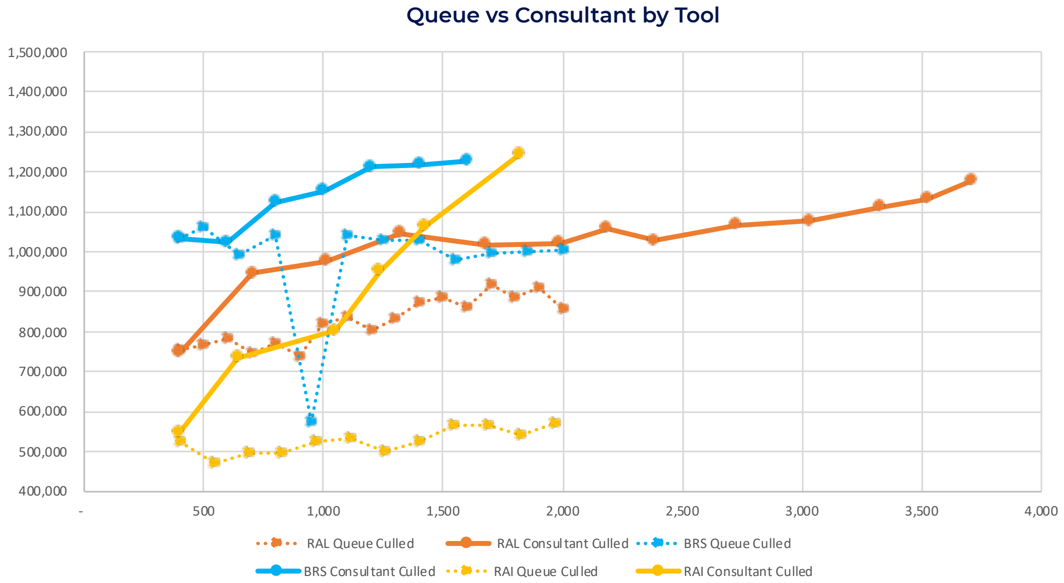


Figure 3 – Documents culled as a function of queue vs consultant-curated training



Although we do not consider any of our training methodologies trade secrets, the specifics of how we selected training documents fall outside the purview of this white paper. In practice, there are many ways to achieve similar results, and the results from this analysis indicate that human-curated training documents are superior to existing algorithms, despite their complexity. Certainly, care should be taken to avoid "teaching the test," and consultants in this exercise did not view or consider existing coding or document-level scoring to select training documents. Further, use of the control set was limited to document scores and the cutoff without consulting the actual content of the relevant or not relevant documents from same.

CONSULTANT VS MODEL ALGORITHM

Clearly, the results of the testing demonstrated that the consultant intervention in the training of the model yielded the most significant improvement in the results (i.e., F_2-measure for target recall). If this were observed in a single tool only, we might excuse it as an aberration, but the consistent improvement across the board suggests that the human understands the true goal of the project which is to maximize the culling effort while returning the most reasonable number of responsive documents. Where most bespoke training algorithms appear to be focused on training the model qua model, the consultant understands that the model scores serve a greater purpose simply for stack-ranking documents. The scores themselves have no special value short of a spectrum of distinction from other documents either for prioritized review (TAR 2.0) or cutoff selection (TAR 1.0). Understanding the goal of the review and the unique capabilities of each tool along with their relative technical strengths and weaknesses, the consultant is best-suited to recommend not only the workflow but also the tool for the job.

CONCLUSION

In each instance, the consultant or human-selected training documents surpassed the quality of the model-selected training documents. The human can synthesize the importance of training to a target result; whereas, the tools select documents based on pre-built goals that may not be effective on every review. Bespoke queues can deliver excellent training results in many instances, and this paper should not be misconstrued to indicate that queues should be ignored or avoided in their entirety. Instead, we encourage the reader to consider that machine learning should not be relegated entirely to the machine. After all, active learning utilizes supervised learning with SMEs reviewing documents to train the binary classifier. Why should we ignore the performance gain from human analysis of the active learning output and resulting human input from same to identify more effective training documents for machine learning? Machine learning represents an opportunity to augment and accelerate human analysis, not replace it.