An Overview of Machine Learning Applications to Technology Assisted Review for E-Discovery Part Three: Is TAR More Effective than Exhaustive Manual Review?

By Sarah Bullard, formerly of DisputeSoft

This is the third article in a four-part series on technology assisted review (TAR), a process that uses machine learning to increase efficiency and decrease the cost of document review in discovery.

The first article discussed the differences between supervised and unsupervised machine learning algorithms, and why supervised learning algorithms are more commonly used with TAR for e-discovery. The second article explored two machine learning applications pertinent to TAR: optical character recognition and natural language processing. The purpose of this article is to compare TAR with exhaustive manual review—that is, review performed solely by human beings with knowledge of the subject matter.

Is TAR More Effective than Exhaustive Manual Review

Recall, Precision, and FI Score

Three measurements are typically used to determine the success of an e-discovery process. These measurements are precision, recall, and F1 score. Koo Ping Shung, data scientist and cofounder of DataScience SG, defines these terms in his article, "Accuracy, Precision, Recall or F1?"

In Part One, we learned that a "classification problem" refers to a problem in machine learning in which the user wants the algorithm to sort uncategorized items into predefined categories. To understand precision, recall, and F1 score, it is important to first understand that in any application of machine learning to a classification problem, the algorithm will likely return results that can be sorted into four categories:

1) true positive – An item *correctly* identified as belonging to the positive class
2) false positive – An item *incorrectly* identified as belonging to the positive class
3) true negative – An item *correctly* identified as belonging to the negative class
4) false negative – An item *incorrectly* identified as belonging to the negative class

An excellent analogy clarifying these terms can be found in an article by Opex Analytics titled "Precision and Recall: Understanding the Trade-Off."

With this in mind, *recall* is calculated according to the following formula:

$$\frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Negative}}$$

Recall is the number of relevant documents identified divided by the total number of relevant documents reviewed. Therefore, recall identifies how complete the algorithm results are, in the form of the percentage of relevant documents classified correctly by the TAR algorithm.

Precision is calculated according to the following formula:

$$\frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Positive}}$$

Precision is the number of relevant documents identified divided by the total number of documents reviewed. Therefore, precision identifies how accurate the algorithm results are, in the form of the percentage of documents predicted to be relevant that are truly relevant. For example, suppose that an algorithm is used to identify the number of baseballs vs. softballs in a basket of containing 15 softballs and 12 baseballs. The algorithm identifies 9 softballs, 6 of which are correctly identified (true positive) and 3 of which are incorrectly identified (false positive). Thus, the program's recall is 6/15 = 40% and precision is 6/9 = 66.67%.

The importance of an algorithm's rate of recall vs. precision differs depending on the situation. High recall is beneficial when the cost of false negatives is high. For example, suppose that an algorithm is used to identify candidates at risk of type 2 diabetes. In this instance, a high number of false negatives (at-risk individuals not identified by the algorithm) might lead to the preventable illness of several patients. High precision is beneficial when the cost of false positives is high. For example, suppose that an algorithm is used to identify low-risk technology startup companies worthy of investment. In this instance, a high number of false positives (high-risk startups identified as safe investments) might lead to a significant loss of funds.

While precision and recall are useful metrics, remember that their purpose is to determine the effectiveness of a given classification method or function. Two additional metrics are used to analyze the precision and recall simultaneously to determine that effectiveness. These metrics are known as accuracy and F1 score. Accuracy is calculated according to the following formula:

$$\frac{True\ Positive + True\ Negatives}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

Although accuracy gives a fair "big-picture" idea of how a model is performing, it is generally not the best metric to use. This is because, in most cases, false positives or negatives will be disproportionately costly from a business perspective. Consider the examples above. It is usually much less dangerous to diagnose someone with a disease when they do not have it than to assure someone that he is healthy when in reality he is not. It is also usually much less dangerous, financially, to avoid investing in a low-risk technology startup than it is to invest in a high-risk startup under false information. Therefore, misclassifications should be given more weight than correct classifications when evaluating a model's performance. To get around this problem, Koo suggests using the F1 score instead of accuracy. The F1 score is calculated according to the following formula:

$$\frac{2PR}{(P + R)}$$

Generally, a classification model seeks a balance between precision and recall. F1 score is used to balance precision and recall while removing the largely inconsequential number of true negatives and therefore increasing the weight of false positives or negatives, which usually have more noteworthy impacts on businesses. The F1 score combines precision and recall in such a way that reveals the impact of false positives and negatives. Koo's argument in favor of the use of the F1 Score asserts that the "F1 Score might be a better measure to use if we need to seek a balance between precision and recall…"

The F1 score is a generically useful metric by which to evaluate a machine learning model, at least initially. However, Bill Dimm, the developer of algorithms from predictive coding, conceptual clustering, and near-dupe detection used in his company's Clustify software, finds the F1 score to be a "virtually worthless" measure of the success of a predictive coding process, since it weights precision and recall in an arbitrary fashion rather than weighting them in a manner that reflects the economics of a given case. For this reason, after running a preliminary evaluation of a model using the F1 score, it is wise to consider the actual impact of false positives and false negatives on the specific problem at hand, and then to create a more customized evaluation metric to use.

TAR vs. Exhaustive Manual Review

In their paper, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, computer science professors Maura Grossman and Gordon Cormack describe several situations in which TAR either matched or exceeded human review in terms of recall, precision, and F1 score, despite several factors that lent an advantage to the manual reviewers (discussed in the "Limitations" section of the paper). Throughout the paper, the human propensity for fatigue is suggested to be the reason for this discrepancy in performance, although it is not proven at any point in the paper to be the definitive cause.

Another metric in which TAR outperforms exhaustive manual review is in terms of monetary cost. One study shows that TAR can reduce time dedicated to review by 74%, and can reduce money spent on review by 30%. Another use case describes a situation in which a client saved almost $4 million in review costs by using TAR instead of manual review.

However, at least for now, it seems that active human participation is still a necessary aspect of TAR. Attorney and e-discovery academic Casey Sullivan describes a relevant situation in his 2018 article *AI-Driven Discovery Process Produces Millions of Unresponsive Docs*. Sullivan relates the disastrous result of TAR use by United Airlines for a large domestic airline travel antitrust litigation, in which the system classified far too many documents as responsive.

This resulted in recall being extremely high. United Airlines had produced 1 million more documents than all the other involved airlines *combined*. Often, there is a tradeoff between precision and recall: adjusting your model to achieve higher precision is likely to lower its recall, and vice versa. However, as discussed in the Opex Analytics blog mentioned earlier, sometimes choosing a better model can improve both precision and recall simultaneously. In the case of United Airlines, the high volume of returned documents also resulted in precision so low that the results could not be used, and the cost in both time and money of the initial failed TAR process was devastating.

As we don't have access to the specific TAR process that United Airlines followed, it's not possible to identify exactly where human input was underutilized. However, it is possible to speculate. One possible explanation is that United Airlines failed to follow standard testing procedures with the involvement of subject matter experts. This requires that the subject matter experts classify both a training set (to train the system) and a testing set (to check whether the training has been successful). Additionally, should the testing indicate that the training was unsuccessful, the subject matter experts will need to continue classifying training and testing sets until the system can accurately classify a testing set within a certain predetermined threshold. With a large data set (such as that of United Airlines), this can be a time-consuming, and therefore expensive, process.

Another possible explanation is that the classification function used was not well-suited to United Airlines' document corpus. Without qualified human beings supervising the training and testing process, it's possible and even likely that a suboptimal machine learning algorithm will be selected to classify the documents. Of course, with proper testing, this should be quickly detected, and the algorithm adjusted accordingly. However, a combination of a poorly-selected algorithm and an anemic approach to testing will yield inaccurate results.

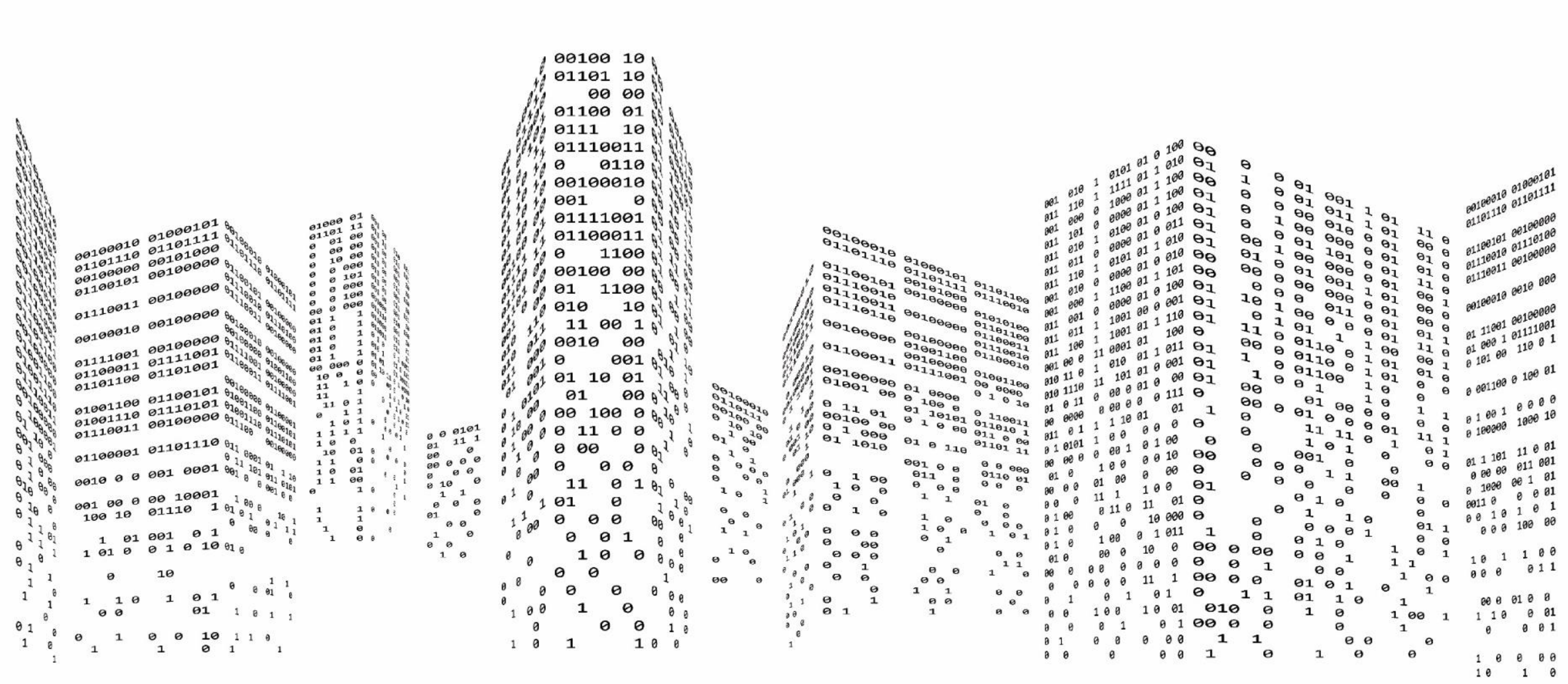
While machines can be useful tools for automating and expediting the document classification process, humans are integral to guide and correct those machines.

While machines can be useful tools for automating and expediting the document classification process, humans are integral to guide and correct those machines. The situation faced by United Airlines is just one example of how the quality of human input to a TAR algorithm is crucial to its effectiveness, and how its absence can cause major issues in e-discovery.

Conclusion

Well-managed TAR outperforms exhaustive manual review according to some studies and metrics. However, without proper supervision by experts, TAR may produce unusable results, wasting a client's time and money. To learn more about the role of machine learning applications in TAR for e-discovery, look for the fourth and final article in this series, which considers ethical concerns about TAR, and whether TAR is right for you.

To learn more about DisputeSoft's e-discovery services including identification, recovery, preservation, and analysis of systems, databases, and other non-custodial evidence, visit our electronic discovery services page and explore a representative e-discovery case: General Electric v. Mitsubishi Heavy Industries.



If you are in need of an e-discovery expert, we invite you to consider DisputeSoft.

Contact Information

Jeff Parmet, Managing Partner
301.251.6182
jparmet@disputesoft.com

12505 Park Potomac Ave. | Suite 475 | Potomac, MD | 20854