On the Effectiveness of Portable Models versus Human Expertise under Continuous Active Learning

Jeremy Pickens
OpenText
Denver, USA
jpickens@opentext.com

Thomas C. Gricks III, Esq.
OpenText
Denver, USA
tgricks@opentext.com

KEYWORDS

continuous active learning, human augmentation, human expertise, transfer learning, portable models

1 INTRODUCTION

eDiscovery is the process of identifying, preserving, collecting, reviewing, and producing to requesting parties electronically stored information that is potentially relevant to a civil litigation or regulatory inquiry. Of these activities, the review component is by far the most expensive and time consuming [8]. Modern, effective approaches to document review run the gamut from pure human-driven processes such as boolean keyword search followed by linear review, to predominantly AI-driven approaches using various forms of machine learning. A review process that involves a significant, though not exclusive, supervised machine learning component is typically referred to as technology assisted review (TAR).

One of the most efficient approaches to TAR in recent years involves a combined human-machine (IA, or intelligence amplification) approach known as Continuous Active Learning (CAL) [5]. As with any TAR review, a CAL review will benefit in some measure by overcoming the cold start problem: The machine typically cannot begin making predictions until it has been fed some number of training documents, aka seeds. In an early CAL approach, initial sets of training documents were selected via human effort, e.g., manual keyword searching. This approach to selecting seed documents relies on human knowledge and intuition.

Recently in the legal technology sector, another seeding approach that does not rely on human assessment of the review collection but is based on artificial intelligence (AI) methods and derived from documents outside the collection has been gaining momentum. For this technique, which is often referred to as "portable models", and known in the wider machine learning community as transfer learning, initial seed documents are selected not via human input, but by predictions from a machine learning model trained using documents from prior matters or related datasets. Portable models take a pure AI approach and eschew human knowledge in the cold start seeding process.

Notwithstanding the benefits asserted by the proponents of portable models as a seed-generation technique, we are aware of no formal or even informal studies addressing the overall impact of portable model seeding on the efficiency of a TAR review relative to human-driven seeding. It is an open question whether technology assisted review seeded by portable models offers a clear, sustained advantage over approaches that begin with human input. Therefore, this work constitutes an initial study into the relationship between human vs machine seeding and overall review efficiency.

2 MOTIVATION

Separate and apart from the inherent value of an assessment of the impact of portable models on TAR, there are two principles attendant to the creation of portable models that serve as a further motivation for this study: (1) the increased regulatory pressure to maintain personal privacy; and (2) the growing need for stringent cyber security measures. Consideration of both principals is generally recognized as an essential step in the development and utility of modern AI applications, given their breadth and proliferation.

Recent years have seen an increased scrutiny from EU and United States regulatory agencies. Data collection and reuse is under heavy examination as regulators seek to minimize data collection and maximize privacy and security. Portable models are a form of data reuse; the models would not exist were it not for the original data. As such, there are rights and obligations around the use of the data that goes in to training portable models, and a strong need for clearer assessments of risk when porting models. As Bacon et al noted [1]:

> The use of machine learning ("ML") models to process proprietary data is becoming increasingly common as companies recognize the potential benefits that ML can provide. Many IT vendors offer ML services that can generate valuable insights derived from their customer's proprietary data and know-how. For companies that have not yet established their own ML expertise in-house, these services can offer significant business advantages. However, there may be cases where one party owns the ML model, another party has the business expertise, and a third party owns the data. In such cases, significant intellectual property ("IP") and data protection and security risks may arise. Naturally, most companies that invest in building an ML model are looking for a return on their investment. From a financial perspective, such companies focus on using the IP laws and related IP contract terms, such as IP assignments and license grants, to maximize their control over the ML model and associated input and results. Data protection laws can run counter to these objectives by imposing an array of requirements and restrictions on the processing of various types of data, particularly to the extent they include personal

Proceedings of the Second International Workshop of AI and Intelligent Assistance for Legal Professional in the Digital Workplace (LegalAIIA 2021), held in conjunction with ICAIL 2021. June 21, 2021. São Paulo, Brazil. Published at http://ceur-ws.org
LegalAIIA 2021, June 2021, São Paulo, Brazil

> information. The interplay between these competing considerations can lead to interesting results, especially when a number of different parties have a stake in the outcome.

The second, perhaps more important challenge with respect to portable models is the possibility of data leakage. In recent years, computer security and machine learning researchers have increased the sophistication of membership inference attacks [9]. These attacks are a way of probing black box, non-transparent models to "discover or reconstruct the examples used to train the machine learning model" [6]. The basic process is that:

> An attacker creates random records for a target machine learning model served on a [portable model] service. The attacker feeds each record into the model. Based on the confidence score the model returns, the attacker tunes the record's features and reruns it by the model. The process continues until the model reaches a very high confidence score. At this point, the record is identical or very similar to one of the examples used to train the model. After gathering enough high confidence records, the attacker uses the dataset to train a set of "shadow models" to predict whether a data record was part of the target model's training data. This creates an ensemble of models that can train a membership inference attack model. The final model can then predict whether a data record was included in the training dataset of the target machine learning model. The researchers found that this attack was successful on many different machine learning services and architectures. [6]

Carlini et al [3, 4] further elaborate on the potential for portable models to reveal private or sensitive information:

> One such risk is the potential for models to leak details from the data on which they're trained. While this may be a concern for all large language models, additional issues may arise if a model trained on private data were to be made publicly available. Because these datasets can be large (hundreds of gigabytes) and pull from a range of sources, they can sometimes contain sensitive data, including personally identifiable information (PII): names, phone numbers, addresses, etc., even if trained on public data. This raises the possibility that a model trained using such data could reflect some of these private details in its output.

Tramer et al [10] note that entire models may even be stolen via such techniques, even when the adversary only has black box (observations of outputs only, rather than internal workings) access to the model: "The tension between model confidentiality and public access motivates our investigation of model extraction attacks. In such attacks, an adversary with black-box access, but no prior knowledge of an ML model's parameters or training data, aims to duplicate the functionality of (i.e., "steal") the model…We show simple, efficient attacks that extract target ML models with near-perfect fidelity for popular model classes including logistic regression, neural networks, and decision trees."

Given the potential dangers associated with modern AI applications such as portable models, we therefore ask: Do portable models provided a cognizable sustained advantage over human augmented IA processes sufficient to warrant their use in the face of privacy and cybersecurity concerns? If not, perhaps the safer and more appropriate approach is to continue using traditional human-driven techniques.

3 RELATED WORK

The key foundation in our investigation is the observation that the current state-of-the-art document review TAR process is based on continuous active learning (CAL) [5]. Given seed documents, the basic CAL process induces a supervised machine learning model which then predicts the most likely responsive, unreviewed documents. After some (relatively small) number of those top-ranked predictions are reviewed and coded, another model is induced and the next most likely documents are queued for review. The process continues until a high recall target is hit.

Review workflows that are based on CAL have what might be called a "just in time" approach to prediction. Rather than attempting to induce a perfect model up front, CAL workflows dynamically adjust as the review continues. Often this means that early disadvantages, and even early advantages, wash out in the process. For example [citation anonymized for review] found that four searchers each working independently to find seed documents found different and different numbers of seeds. But after separately using each seed set to initialize a CAL review, approximately the same number of documents needed to be reviewed to achieve high recall. This study asks similar questions in the context of portable models—whether there is a significant improvement in review efficiency when using portable models relative to traditional, non-AI techniques.

Another common portable model theme is the claim that the more historical data they are trained on, the better their predictions will be. While that may be true in some instances, it may not be in others. What constitutes privileged documents in one matter might have a different set of characteristics as privileged documents in others matter. What constitutes evidence of fraud, or sexual harassment in one matter might be different than in other matters. No amount of "big data" gathered from dozens (hundreds? thousands?) of prior matters and composed into a monolithic portable model may be relevant to the current problem if the patterns in the current problem don't match the historical ones. Therefore a question that every eDiscovery practitioner should be asking herself is where the best source of evidence for seeding the current task lies. As [2] notes: "The real goal should not be big data but to ask ourselves, for a given problem, what is the right data and how much of it is needed. For some problems this would imply big data, but for the majority of the problems much less data is necessary."

4 RESEARCH QUESTIONS

We engage three primary research questions. The first question level-sets the value of the pure AI (portable model) approach. The second two questions compare the portable model approach to a human-initiated process.

- **RQ1** Does a CAL review seeded by a portable model outperform (at high recall) one seeded randomly

- **RQ2** Do portable models initially find more relevant documents than does human effort
- **RQ3** Does a CAL review seeded by a portable model outperform (at high recall) one seeded by human effort

When attempting to consider these questions in general, issues naturally arise: What portable models are we talking about? Trained on what data? And how close was that data to the target distribution? And what humans seeded the comparison approach? And what was their prior knowledge of the subject matter?

These questions matter, and while we cannot answer them for every possible training set and human searcher, we have structured the experiments in such a way as to give the most possible "benefit of the doubt" to the portable model, and the least possible benefit to the human searcher. Thus if there are significant advantages of portable models over human effort, these should be most readily apparent when portable models are given the most affordances and humans the least.

The primary manner in which portable models are given an advantage is that we train them on a set of documents that is drawn from the exact same distribution as the target collection to which they will be applied. In practice, portable models are never given this advantage. Prior cases in eDiscovery are not always exactly the same. Different collections, even from the same corporate entity, exhibit different distributions, especially as employees and business activities change and evolve over time. Naturally, the more different the source distribution, the less effective portable models will be when applied to a new target collection. However, by holding the distribution the same, this gives us an upper bound on portable model effectiveness and establishes a strong baseline against which the human effort can be compared.

At the same time, the human effort is minimized. As will be described in more detail below, a small team of human searchers worked for a collective total of approximately half an hour per topic. None of the humans were experts in any of the topics, nor did anyone have recent prior knowledge on the topics, as the events in this Jeb Bush TREC collection [7] took place a decade or more prior to when the searchers worked and most of the issues were local to Florida and did not make national news. In practice, humans are rarely given this disadvantage. They often work for more than thirty minutes on a problem and can have broad domain expertise that comes from having worked on similar cases in the past.

Thus, our experiments consist of a comparison between portable models trained in the best possible light vs human effort that is kept at a minimum. We do this because the core concept of portable models is that they will be sufficiently broad in scope so as to be able to identify relevant documents in a collection that contains documents of a similar content and context to those on which they were trained. ("Relevance" here refers to the notion of "what is desired", be it some sort of topical similarity such as age discrimination or fraud cases, or something like privilege.) That distributional similarity is not always guaranteed, and in fact it can be difficult a priori to know whether you a portable model has been trained on data similar enough to be useful. By using documents intentionally drawn from the exact same distribution, we are able to show an upper bound on portable model effectiveness. In practice, portable model effectiveness is likely to be lower, though how much lower remains to be studied.

	Collection Stats		Manual Seeding Stats		
Topic	Total Rel	Richness	Queries	Minutes	Total Docs
403	1090	0.38%	9	52	60
422	31	0.01%	11	26	63
424	497	0.17%	13	48	81
426	120	0.04%	7	26	32
420	737	0.25%	7	11	54
407	1586	0.55%	6	23	59
414	839	0.29%	12	39	108
410	1346	0.46%	10	24	69
401	229	0.08%	9	29	56
406	127	0.04%	9	25	42
433	112	0.04%	12	43	65
415	12106	4.17%	11	43	77
430	991	0.34%	9	41	72
417	5931	2.04%	10	25	85
413	546	0.19%	9	23	88
432	140	0.05%	13	29	45
402	638	0.22%	6	8	75
427	241	0.08%	9	25	82
419	1989	0.69%	7	27	50
404	545	0.19%	10	23	54
408	116	0.04%	9	23	91
418	187	0.06%	8	27	53
411	89	0.03%	9	29	64
412	1410	0.49%	8	29	64
416	1446	0.50%	8	34	51
423	286	0.10%	8	22	41
429	827	0.29%	12	35	62
409	202	0.07%	15	44	67
405	122	0.04%	9	36	66
428	464	0.16%	13	36	67
431	144	0.05%	10	28	57
434	38	0.01%	14	43	50
425	714	0.25%	12	47	76
421	21	0.01%	14	45	60
	Averages		9.9	31.4	64.3

Table 1: Collection and Manual Effort Statistics

5 EXPERIMENTS

5.1 Data

We test these research questions using the TREC 2016 total recall track document collection, topics, and relevance judgments [7]. This dataset contains 34 topics each with a varying number of relevant documents. Nonetheless, the richness of the majority of topics is under 1%, i.e. relatively low richness topics where portable models allege to be most effective. Table 1 contains statistics on each topic. The first column is the topic ID from 401 to 434, sorted in a manner that will be described in Section 6.3. The next two columns contain the number of total relevant documents and the richness for each topic. There are 290,099 total documents in the collection.

Human effort, aka manual seeding, was done with a small team of four searchers. For each topic, two of the searchers were instructed to run a single query and code the first 25 documents that resulted from that query. The other two searchers were given more interactive leeway and were instructed to utilize as many searches and whatever other analytic tools (clustering, timeline views, etc.) as they wanted, with a goal of working for about 15-30 minutes and stopping once they had tagged 25 documents. This was not strictly controlled, and some reviewers worked a few minutes longer, some a few minutes shorter. And some marked a few more than 25 documents, and some a few less, as is to be expected in normal, "in the moment" flow of knowledge work. Table 1 contains the manual effort statistics — with the total number of queries, total number of minutes, and total unique documents tagged as either relevant or non-relevant — for each topic. On average, the human reviewers worked for 31.4 minutes, issued 9.9 queries, and coded 64.3 documents so the overall effort was done at a fairly high pace and was relatively minimal in comparison to the size of the collection.

5.2 Experiment Structure

In order to compare portable models against both random and human-seeded techniques in RQ1 through RQ3, there needs to be a collection on which the portable model can be trained, separate from the collection on which it and the comparative approaches are deployed. We will refer to these two collections as "source" and "target', respectively. For the reasons enumerated in Section 4, we carve out the portable model training source collection from the same distribution as the target collection, and do so by selecting documents at random. For a given topic, we:

1. Shuffle the collection randomly
2. Split the collection into k groups
3. For each group:
   - (a) Use that group as the portable model training "source" collection $\mathcal{S}$
   - (b) Train a model $\mathcal{M}$ using every document in $\mathcal{S}$
   - (c) Use the remaining groups as the "target" collection $\mathcal{T}$
   - (d) Select manual (human) seeds $\mathcal{H}$ by intersecting all found docs (see Table 1) with $\mathcal{T}$
   - (e) Select random seeds $\mathcal{R}$ from $\mathcal{T}$ until five positives examples are found
   - (f) Selected portable seeds $\mathcal{P}$ from the top of the $\mathcal{M}$-induced ranking on $\mathcal{T}$ in an amount equal to $|\mathcal{H}|$
   - (g) Use the appropriate seeds to run each experiment RQ1 through RQ3
4. Average results across all k groups for the topic, but do not average across topics

The specifics of step (3g) depends on the research question being tested. For example, for RQ1, $\mathcal{R}$ and $\mathcal{P}$ are each (separately) used to seed a continuous active learning (CAL) process. For RQ3, $\mathcal{H}$ and $\mathcal{P}$ are used. Other than the different seedings, these CAL processes are run exactly as in [5] except that updates are done every 30 documents rather than every 1000 documents. And unlike some approaches, the learning is not relevance feedback for a limited number of steps. It is truely continuous in that it does not stop until the desired recall level is achieved, which in these experiments are set to 80%.

While selecting a source collection for portable model training from the same distribution as the target collection already offers great advantage to the predictive capabilities of a portable model, i.e. puts it above where it would likely perform in more realistic scenarios, we extend this advantage even further by giving the model larger and larger source collections on which to train. We compare three primary source/target partitions: 20/80, 50/50, and 80/20, with k=5, k=2, and k=5, respectively. (In the 80/20 case, steps (3a) and (3c) are reversed, with the current group used as the target collection and the other groups used as the source collection.) The reason for the 20/80 partition is that the eDiscovery problem is a recall-oriented task. The larger the review population, aka the target collection, the more realistic the CAL process is likely to be. However, the disadvantage is that only 20% of the TREC collection is be used for training the portable model. The 80/20 partition reverses the balance: 80% of the collection is used to train the portable model, but only 20% of the collection is available to simulate the CAL review, which can be problematic for especially sparse topics. The 50/50 partition splits the difference.

Comment: An astute observer may find slight fault with the structure of this experimental setup, in that there is a small amount of knowledge overlap between the source and target partitions when doing human seeding. Specifically, the human searchers originally searched across the entire collection rather than across split collections. It is possible that a document found by a human searcher that ended up in a source partition has led the human to issue a query that found more or better documents that ended up in the target partition. Thus even though the human-found documents in only the target partition are used to seed a CAL process (Step 3d, above), the existence of some of those seeds could have been influenced by knowledge of documents in the source partition. We note this issue and make it explicit, but do not think that it affects the overall conclusions of the experiment. One reason is that even if humans had some knowledge of documents in the source partition when finding the documents in the target partition, the portable model $\mathcal{M}$ is given knowledge of *every* document, positive and negative, in the source partition. Table 3 shows the raw number of positive documents used for training in the source partition, and it swamps the documents that the humans would have looked at in their short search sessions. Thus, one can think of any overlap during human seed selection as the background knowledge that human would likely already be expected to possess when working in a real scenario. E.g. an investigator working on detecting fraud or sexual harassment likely has some implicit background knowledge of fraud or sexual harassment.

6 RESULTS

6.1 RQ1: Portable- vs Random-Seeded Recall

The results for our first question are found Table 2. Under the rubric of symmetry, the results are expressed in terms of raw percentage *point* (not percentage) differences between the precision achieved at 80% recall for the portable model $\mathcal{P}$-seeded review versus a random $\mathcal{R}$-seeded review, and averaged across all partitions for each topic. Positive values indicate better portable model performance;

Topic	20/80 Partition	50/50 Partition	80/20 Partition
	Δprecision		
403	74.4	85.2	92.2
422	9.3	28	21.8
424	75.1	77.8	65.2
426	53.8	52.2	32.2
420	76.7	71.4	78.4
407	55.8	51.7	50.4
414	8.8	6.2	6.9
410	35.6	29.3	50.1
401	23.3	26.1	17.2
406	8.7	3.8	6.7
433	54.7	48.6	39.2
415	-0.8	0.3	4.7
430	14.2	7.1	17.9
417	6.2	10	17.5
413	67.9	69.3	57.8
432	43.3	46.9	13.4
402	4.2	3.3	4
427	60.5	52.5	34.4
419	1.4	11	6.8
404	1.2	-0.4	0.3
408	0.3	0	0.3
418	0.5	0.2	0
411	0.6	0.3	0.1
412	14.4	17.8	5.9
416	1.1	2.8	0.3
423	12.5	5.8	4.9
429	76.7	80.9	70.9
409	26.8	15.2	3.1
405	66	60.4	43.1
428	15.1	12.3	15.6
431	26.8	19.2	15.4
434	36.4	51	54
425	66.7	59	59
421	1.3	9.1	3.7
	Averages		
	30.0	29.8	26.3
	p<0.000001	p<0.000001	p<0.000001

Table 2: CAL review relative precision based on portable model seeding versus random seeding

negative values the opposite. Nearly universally, $\mathcal{P}$-seeding outperforms random seeding; the p-value under a binomial test is < 0.00001.

This is a wholly expected result. Closer examination of the simulated review orderings shows that in low richness domains most of the precision loss comes not from the CAL iterations, but from the larger number of documents needed to find enough positive ones to start ranking. Note that random seeding on the target partition also outperforms fully linear review by an average of 12.2, 10.6, and 5.6 percentage points on the 20/80, 50/50, and 80/20 partitions, respectively. So even random seeding of CAL is better than no CAL at all.

Thus in answer to the question: Does $\mathcal{P}$-seeding produce an efficacious result, the answer is yes. However, the more important question is not whether portable models are useful; it is whether they are useful relative to other reasonable, simpler, or less risky alternatives. For that we turn to the remaining research questions.

6.2 RQ2: Portable vs Human Seed Initial Relevance

The results for our second questions are found in Table 3 under the Target Portable and Target Manual columns for each partition group. For example, in the 20/80 partition, where 80% of the collection is used as the target collection and on average across all 5 folds, on topic 403 human effort found 18.4 positively-coded seed documents whereas the portable model $\mathcal{M}$ found 41.6 at the same level of effort (i.e. at 60 documents, as per Table 1). On topic 421 under the 20/80 partition, humans found an average 10.4 documents and $\mathcal{M}$ found 3.2.

The average number of positive training documents in the source partition, i.e. the data on which $\mathcal{M}$ is trained, is shown. The number of negative training examples is the remainder of the fold. Averages across all 34 topics are shown at the bottom of the table, as is a binomial p-value.

These results show that when 20% of the collection is used to train $\mathcal{M}$ (20/80 parition), even though that data is literally from the same distribution as the target fold, the various $\mathcal{M}$ are able to find seed documents at a rate no better than a small amount of human effort. There is only a difference of 0.2 documents across all topics, and while there is some variation between topics the differences are not statistically significant (p=0.303). As the training partition increases, and 50% then 80% of the collection is used to train each $\mathcal{M}$, so too does the ability of the model to find more seed documents. On the 50/50 partition $\mathcal{M}$ finds on average 3.5 more documents than the human at the given effort level, and on the 80/20 partition it finds an average 1.5 more documents. Both results are statistically significant.

When the number of seeds is normalized per fold and topic by the number of total seeds found, i.e. the positive seed precision, the manual effort has an average precision of 66.0% across all folds, whereas $\mathcal{M}$ precision is 64.2%, 76.5%, and 78.1%, respectively across 20/80, 50/50, and 80/20. That is, even though the average number of documents that $\mathcal{M}$ finds on the 50/50 partition is larger (3.5) than on the 80/20 partition (1.5), the latter partition is smaller. The actual precision goes up slightly.

Thus in answer to the question: Do portable models initially find more relevant documents than does human effort, the answer is mixed. When given 20% of the collection for training, they do not. When given 50% or 80%, they do. However, the improvement is modest: a few percentage points, or a few extra documents.

6.3 RQ3: Portable- vs Human-Seeded Recall

The results for our third and final question are also found in Table 3 under the Δprecision columns. Again, in the interest of symmetric magnitudes, Δprecision is the percentage *point* difference between $\mathcal{P}$-seeded versus $\mathcal{H}$-seeded CAL. While again there is some variation across topics, on average on the 20/80 partition, $\mathcal{P}$-seeding is 0.4 percentage points worse, while on the 50/50 and 80/20 partitions

$\mathcal{P}$-seeding is 0.6 and 1.2 percentage points better. However, none of these results are statistically significant (p=0.303).

Furthermore, when we look at the raw document count difference between the two conditions (not shown in the table) another story emerges. In the 80/20 partition, on those topics for which $\mathcal{P}$-seeded CAL is better, is it better on average by 186 total documents. Where $\mathcal{H}$-seeded CAL is better, it is better by 896 documents. On the 50/50 partition, $\mathcal{P}$- versus $\mathcal{H}$-seeding is 271 vs 743 documents better, and on the 20/80 partition it is 166 versus 655 documents. There is no consistent advantage of either approach over the other, but the negative consequences of $\mathcal{H}$ seeding seems to be far smaller than those of $\mathcal{P}$ seeding.

The reason for the topical sort order across all tables should now become clear: All tables in this paper are sorted by the Δprecision of the 80/20 partition. This seems to be the partition for which portable models are the strongest; they have the most training data. And sorting by Δprecision allows us to see where $\mathcal{P}$-seeding vs $\mathcal{H}$-seeding each shine. To that end, we introduce one more metric into the discussion: The WTF "ineffectiveness" metric [11, 12] encapsulates the notion of not only looking at average performance, but at outliers. A system that has good average performance but egregious outliers might want to be avoided, especially in eDiscovery where every case matters and the costs incurred by an outlier are more significant than in, say, ad hoc web search.

From this perspective, we see that where portable models perform the strongest, i.e. on the 80/20 partition where they are given 80% of the available positive documents, there are outliers in both directions. The top three $\mathcal{P}$-advantage outliers show a 60.5, 12.0, and 5.7 percentage point difference. The top three $\mathcal{H}$-advantage outliers show a 25.0, 12.6, and 11.1 percentage point difference. However, in terms of raw document counts these translate to a 670, 667, and 468 documents for the $\mathcal{P}$-advantage, and 9249, 2006, and 892 documents for $\mathcal{H}$-advantage. There appear to be fewer "WTFs" from $\mathcal{H}$-seeding.

7 CONCLUSION

We have shown that a portable model can be useful. Certainly relative to linear review, and even relative to randomly seeded CAL workflows, taking a portable approach offers a significant advantage. They are also marginally better than humans when it comes to finding initial seed documents. When it comes to sustained advantage, i.e. precision at 80% recall, the advantages fade. There is no statistically significant difference in human vs portably seeded CAL workflows, and slight evidence that the outliers for the portable approach are worse.

We note also that porting models carries with it significant risk in the form of intellectual property rights, data leakage via membership inference attacks, privacy, and security. It is every party's own subjective decision as to whether the advantages of portable models outweigh the challenges and risk. However, from the results in this study we would recommend continuing to invest in human-driven seeding (IA–intelligence augmentation) processes and not going all in on AI. At least relative to the topics studied in this paper, the modicum of effort required of the human are a fair trade relative to risk. Even when portable models are built on a corporation's own data, and models are not swapped between different owners and therefore risk is lower, we do not yet find that the portable model provides a sustained advantage.

8 FUTURE WORK

Certainly this is but one study and more studies with a wider range of models, data collections, and human effort are needed. Perhaps the humans could have done even better if given more time, were working on a domain in which they had specific expertise, or were given more powerful analytics with which to find seed documents. Conversely, portable models were given all possible advantages in this experimental structure by building them on documents drawn from the exact same distribution as the target collect, in ever increasing amounts (20%, 50%, and 80%). It is not likely that portable models will ever be trained on prior data as perfectly similar to the target distribution. Therefore, portable models might likely have performed much worse in realistic scenarios where the source and target collections are further apart. E.g. when modeling fraud or sexual harassment, does what constitute evidence of fraud or harassment in one collection express itself the same way in another collection? Future research is needed in three different areas: (a) More and larger collections from similar but not identical distributions on which to train models, or perhaps the "right" small collections on which to train models, as per [2], (b) more advanced models and better transfer learning, and (c) better and stronger baselines against which to compare.

Better and stronger baselines are not limited to more effective human effort. They also include other existing, common practices. For example, many companies dealing with sensitive information keep lexicons of search terms used to find sensitive information. While a lexicon could in some sense be thought of as an "unweighted" portable model, one difference is that it's manually constructed, transparent, and can embed human intuition and patterns never seen in prior data, i.e. lexicons do not need to be trained. Another common approach for corporations with repeat litigation is the idea of a "drop in seed". That is, instead of building large models based on huge datasets from all possible prior matters, some in the industry have developed the ad hoc practice of taking a few coded documents from previous matters, which matters are known to be similar to the current matter, and using those as the initial seeds on the target collection. This of course only works behind the firewall, as companies will not transfer documents to other companies. But given the security, privacy, and related membership inference attack risks of portable models, companies might not want to transfer their own models to a competitor, either. So in addition to comparing portable models against human seeding, they should be compared against lexicons and drop-in seeds. Perhaps these latter approaches outperform both portable models and human-seeded approaches when considering the total cost of a review and not just the document count.

The cost of the portable model (vendor charge) versus the human approach (e.g. half an hour of searcher time) needs to be considered as well, and not just the cost of the subsequent document review. In short, this is a rich space for the exploration of tradeoffs, risks, and advantages for various human-driven vs machine-driven eDiscovery processes.

Topic	20/80 Partition				50/50 Partition				80/20 Partition			
	Positive Counts				Positive Counts				Positive Counts			
	Source	Target			Source	Target			Source	Target		
		Portable ($\mathcal{P}$)	Manual ($\mathcal{H}$)	Δprec		Portable ($\mathcal{P}$)	Manual ($\mathcal{H}$)	Δprec		Portable ($\mathcal{P}$)	Manual ($\mathcal{H}$)	Δprec
403	218.0	41.6	18.4	27.8	545.0	26.0	11.5	42.8	872.0	10.4	4.6	60.5
422	6.2	6.6	12.8	-8.2	15.5	11.0	8.0	14.6	24.8	3.8	3.2	12.0
424	99.0	40.2	43.2	0.0	247.5	28.5	27.0	2.4	396.0	11.4	10.8	5.7
426	24.0	23.2	36.0	22.6	60.0	21.5	22.5	7.2	96.0	8.8	9.0	4.6
420	147.4	50.8	49.6	2.6	368.5	33.0	31.0	0.6	589.6	13.2	12.4	4.4
407	316.2	31.4	30.4	1.4	790.5	19.5	19.0	2.5	1264.8	7.8	7.6	4.2
414	167.6	36.2	6.4	2.9	419.0	24.0	4.0	0.2	670.4	9.4	1.6	3.2
410	269.0	71.6	68.0	-1.6	672.5	45.0	42.5	-2.8	1076.0	18.0	17.0	2.9
401	45.8	33.6	36.0	3.6	114.5	27.0	22.5	2.1	183.2	11.0	9.0	2.8
406	25.2	11.8	31.2	1.3	63.0	21.5	19.5	-1.7	100.8	9.4	7.8	2.0
433	22.4	28.4	30.4	-5.6	56.0	25.0	19.0	-2.2	89.6	10.8	7.6	1.5
415	2408.4	33.6	46.4	0.1	6021.0	21.5	29.0	-2.9	9633.6	8.0	11.6	1.3
430	198.0	62.6	48.0	-1.0	495.0	40.5	30.0	0.4	792.0	16.8	12.0	1.3
417	1186.2	87.2	81.6	0.5	2965.5	54.5	51.0	0.8	4744.8	21.8	20.4	0.9
413	109.2	51.2	52.8	-0.3	273.0	34.0	33.0	0.4	436.8	13.8	13.2	0.6
432	28.0	10.2	31.2	-12.1	70.0	20.5	19.5	6.1	112.0	8.6	7.8	0.4
402	127.0	49.4	43.2	0.1	317.5	34.0	27.0	0.6	508.0	14.0	10.8	0.3
427	48.2	30.8	35.2	2.1	120.5	23.5	22.0	5.2	192.8	10.2	8.8	0.2
419	396.6	28.4	40.0	-0.2	991.5	18.0	25.0	-0.6	1586.4	7.2	10.0	0.1
404	108.8	30.6	26.4	0.1	272.0	19.5	16.5	-0.1	435.2	8.2	6.6	0.0
408	22.8	20.8	16.0	0.1	57.0	20.5	10.0	0.0	91.2	8.6	4.0	0.0
418	37.4	13.4	10.4	0.1	93.5	16.0	6.5	0.0	149.6	7.6	2.6	0.0
411	17.8	9.4	9.6	-0.1	44.5	11.5	6.0	0.1	71.2	4.4	2.4	-0.1
412	280.0	55.8	47.2	1.5	700.0	37.5	29.5	0.7	1120.0	14.4	11.8	-0.1
416	279.0	36.4	29.6	0.6	697.5	20.5	18.5	0.1	1116.0	8.0	7.4	-0.2
423	57.2	16.8	16.8	0.2	143.0	13.0	10.5	1.0	228.8	5.4	4.2	-0.4
429	165.4	61.4	63.2	-1.5	413.5	39.0	39.5	-0.4	661.6	15.6	15.8	-1.2
409	39.8	25.2	28.0	11.7	99.5	23.5	17.5	-1.8	159.2	10.2	7.0	-1.9
405	23.8	31.0	32.8	-0.4	59.5	22.0	20.5	1.1	95.2	9.2	8.2	-2.4
428	92.4	55.6	48.0	2.7	231.0	34.5	30.0	-0.9	369.6	14.0	12.0	-5.4
431	28.8	24.8	30.4	-3.6	72.0	21.0	19.0	-11.6	115.2	7.6	7.6	-8.3
434	7.6	5.6	24.0	-29.8	19.0	9.0	15.0	-13.6	30.4	5.4	6.0	-11.1
425	142.8	51.2	44.0	-7.8	357.0	32.0	27.5	-10.8	571.2	12.8	11.0	-12.6
421	4.2	3.2	10.4	-23.4	10.5	5.0	6.5	-18.6	16.8	1.8	2.6	-25.0
AVG	210.3	34.4	34.6	**-0.4**	525.8	25.1	21.6	**0.6**	841.2	10.2	8.7	**1.2**
		p=0.303		p=0.303		p=0.0002		p=0.303		p=0.00004		p=0.303

Table 3: Across each of the various partitions: (1) Number of positive training documents in the portable model "source", (2) The number of positive seed documents found by the portable model $\mathcal{P}$ and the manual $\mathcal{H}$ approaches, and (3) the relative precision (Δprec) of $\mathcal{P}$-seeding over $\mathcal{H}$-seeding at 80% recall, i.e. the precision of the former minus the precision of the latter. Positive numbers indicate that $\mathcal{P}$-seeding was more effective, negative numbers that $\mathcal{H}$-seeding was more effective.

REFERENCES

[1] Brittany Bacon, Tyler Maddry, and Anna Pateraki. 2020. Training a Machine Learning Model Using Customer Proprietary Data: Navigating Key IP and Data Protection Considerations. *Pratt's Privacy and Cybersecurity Law Report* 6, 8 (Oct. 2020), 233–244.

[2] Ricardo Baeza-Yates. 2013. Big Data or Right Data?. In *Proceedings of the 7th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW 2013)*, Loreto Bravo and Maurizio Lenzerini (Eds.), Vol. 1087 (CEUR Workshop Proceedings). Puebla/Cholula, Mexico. http://ceur-ws.org/Vol-1087/

[3] Nicholas Carlini. 2020. *Privacy Considerations in Large Language Models*. Retrieved April 29, 2021 from https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html

[4] Nicholas Carlini, Florian Tramer, Eric Wallace, Mathew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting Training Data from Large Language Models. Article arXiv:2012.07805.

[5] G. V. Cormack and M. R. Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Gold Coast, Australia, 153–162.

[6] Ben Dixon. 2021. *Machine Learning: What are Membership Inference Attacks?* Retrieved April 29, 2021 from https://bdtechtalks.com/2021/04/23/machine-learning-membership-inference-attacks/

[7] Maura R. Grossman, Gordon V. Cormack, and Adam Roegiest. 2016. TREC 2016 Total Recall Track Overview. In *In NIST Special Publication 500-321: The Twenty-Fifth Text REtrieval Conference Proceedings (TREC 2016)*, Ellen M. Voorhees and Angela Ellis (Eds.). Gaithersburg, Maryland. https://trec.nist.gov/pubs/trec25/trec2016.html

[8] Nicholas M. Pace and Laura Zakaras. 2012. *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery.* Rand Corporation, Santa Monica, CA, USA.

[9] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Proceedings of Network and Distributed Systems Security (NDSS) Symposium.* San Diego, California.

[10] Florian Tramer, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs. In *Proceedings of the 25th USENIX Security Symposium.* Austin, Texas, 601–618.

[11] Daniel Tunkelang. 2012. *WTF!@k: Measuring Ineffectiveness.* Retrieved April 29, 2021 from https://thenoisychannel.com/2012/08/20/wtf-k-measuring-ineffectiveness/

[12] Ellen Voorhees. 2004. Measuring Ineffectiveness. In *Proceedings of the 27th annual ACM SIGIR conference on Research and Development in Information Retrieval.* Sheffield, UK, 562–563. https://doi.org/10.1145/1008992.1009121