Impact of Review-Set Selection on Human Assessment for Text Classification

Adam Roegiest
University of Waterloo

Gordon V. Cormack
University of Waterloo

ABSTRACT

In a laboratory study, human assessors were significantly more likely to judge the same documents as relevant when they were presented for assessment within the context of documents selected using random or uncertainty sampling, as compared to relevance sampling. The effect is substantial and significant [0.54 vs. 0.42, p<0.0002] across a population of documents including both relevant and non-relevant documents, for several definitions of ground truth. This result is in accord with Smucker and Jethani's SIGIR 2010 finding that documents were more likely to be judged relevant when assessed within low-precision versus high-precision ranked lists. Our study supports the notion that relevance is malleable, and that one should take care in assuming any labeling to be ground truth, whether for training, tuning, or evaluating text classifiers.

1. INTRODUCTION

In supervised learning for text classification, each of a set of training documents is labeled as relevant or non-relevant by a human assessor. The training documents and their labels are used to induce a classifier to predict the relevance or non-relevance of the remaining documents in the population from which the training documents were drawn. Training documents may be selected using random sampling, or using active learning methods such as uncertainty sampling or relevance sampling [8, 7]. Simulation studies comparing these approaches typically rely on the assumption that the sampling strategy does not influence how the assessor will label a particular document. We show, in a study involving 36 paid assessors recruited from a university community, that the same documents are much more likely to be judged relevant when embedded in a set selected by random sampling or uncertainty sampling, than one selected by relevance sampling.

The influence of sampling strategy on labeling has impact beyond the selection of training sets for inducing classifiers. In technology-assisted review, where every document with a positive classification is assessed and labeled, a superior (higher precision) classifier might yield fewer labeled-relevant documents than an inferior (lower precision) classifier. In the Cranfield approach to IR evaluation (*see* [13]), a set of relevance assessments based on a pool of likely relevant documents might yield substantially different results from one based a sample of the whole. Our result calls into question the common practice of estimating measures like recall and precision from statistical samples, especially those employing non-uniform inclusion probabilities.

A number of studies [11, 12, 6, 10, 5] have shown that the proportion, as well as the order of presentation of relevant and non-relevant documents, can affect user assessment behaviour. Taken together, the results suggest that assessors are less likely to label documents relevant once they have seen a number of relevant documents, either due to presentation order or due to the overall proportion of relevant documents. This observation forms the basis of the question we addressed: Since relevance sampling produces a higher proportion of relevant documents than uncertainty sampling or random sampling, does it suppress the assessor's propensity to judge a document relevant? We further addressed the question of whether this effect was conditioned on ground truth, for several definitions of ground truth.

2. EXPERIMENTAL DESIGN

Our design specified that assessors would review batches of 100 documents for relevance to several topics, where the batches for each topic contained the same 12 known documents, and 88 documents selected by one of random sampling, uncertainty sampling, or relevance sampling.[1] There were 9 topics, and hence 36 batches in total; each batch was assessed by three different assessors.

Table 1 provides a glossary of terms specific to our experimental design.

2.1 Documents and Labels

Documents were selected from the TREC-6 Ad Hoc collection, which has been the subject of previous relevance assessment studies [14, 9, 4], and has two independent sets of relevance assessments: the official NIST binary relevance assessments ("relevant," and "not relevant") created using the pooling method, and a set of graded relevance assess-

SIGIR '16 July 17-21, 2016, Pisa, Italy

ACM ISBN 978-1-4503-4069-4/16/07.

DOI: http://dx.doi.org/10.1145/2911451.2914709

[1] Due to an error in our setup that went undetected until the assessments were complete, for some topics, the batches contained only 10 or 11 of the same known documents. The corresponding shortfall reduced the statistical power of our experiment, but does not affect its validity.

Batch	100 documents presented to assessors for review, consisting of known documents, and context documents
Known documents	12 common documents presented for assessment with different contexts
Context	The manner in which the documents other than the known documents in a batch are selected
NIST assessments	Relevance assessments rendered by NIST for TREC 6
Waterloo assessments	Relevance assessments rendered by the University of Waterloo at TREC 6
Rel	The relevance class of a document, as determined by some combination NIST and/or Waterloo assessments
CAL	Context in which documents are selected iteratively using relevance sampling [2]
SAL	Context in which documents are selected iteratively using uncertainty sampling [2]
SPL	Context in which documents are selected at random [2]

Table 1: Glossary of terms used throughout this work.

ments ("relevant," "iffy," and "not relevant") constructed by the University of Waterloo using interactive search and judging [1]. We augmented each set of assessments with an additional category "unjudged" for documents that were excluded from the pool. The net effect is that each document has one of 12 combinations of three NIST and four Waterloo assessment categories. We chose at random one document with each combination of assessments as the 12 known documents for each of 9 topics.

Our selection of topics was predicated on the fact that, of the 50 topics in the collection, only the nine we chose had at least one document labeled with each of the 12 combinations of assessment categories. Using each sampling method, as discussed below, we selected a list of 90 documents as detailed below, and inserted the known documents at fixed positions, chosen at random. Surplus documents (at positions beyond 100) were discarded.

For random sampling, the list of 100 documents was a uniform random sample of the 560,000-document TREC corpus, in random order. Following Cormack and Grossman [2], we label this protocol, simple passive learning ("SPL"). For uncertainty sampling and relevance sampling, we employed the simple active learning ("SAL") and continuous active learning ("CAL") methods [2, 3], training a classifier to retrieve 10 documents, adding those documents to the training set, and repeating the process nine times. `Sofia-ML` was used as the base classifier, configured to minimize logistic loss, and applied to a tf-idf representation of the documents. The initial training set consisted of a positively labeled pseudo-document consisting of the topic description, plus 100 negatively labeled documents selected at random without regard to their relevance. The training set was used to train the classifier, which was used to compute the likelihood of relevance for each document in the collection. For SAL, the 10 documents with likelihood closest to 0.5 were selected and added to the training set; for CAL, the 10 documents with greatest likelihood were selected. Training labels were derived from the Waterloo assessments: "relevant" and "iffy" were labeled positive; "non-relevant" and "unjudged" were labeled negative.

Our list of 90 context documents consisted of the documents retrieved by these nine iterations, in the order retrieved. Table 2 depicts the prevalence of positive (Waterloo "relevant" or "iffy") documents in the corpus, as well as the number in each batch, including known documents.

Topic	Corpus Prevalence (%)	Context Count		
		CAL	SAL	SPL
301	0.24	40	6	5
304	0.13	26	4	4
306	0.11	41	6	6
307	0.14	65	5	5
319	0.17	63	6	4
324	0.09	74	5	5
332	0.08	41	5	5
337	0.11	83	6	6
343	0.12	33	5	6

Table 2: Corpus prevalence and number of positive documents for each context, where Waterloo "relevant" and "iffy" assessments are considered positive. The counts for each batch include known documents.

2.2 Assessment Protocol

Following approval from the ethics review panel, we recruited 36 participants at large from University of Waterloo, including undergraduate students, graduate students, and faculty. At the outset, a participant was assigned one of the methods and topics at random, and was told that they would be remunerated $20 for reviewing all 100 documents. If the participant took 2 hours or less to judge the documents and achieved at least 25% recall and 25% precision with respect to the NIST assessments for the 100 documents, they were paid a bonus of $10 and offered the opportunity to assess up to two additional batches. For each subsequent batch, a new context and topic were selected at random, such that no participant saw the same context or the same topic more than once. Participants whose assessments did not meet the criteria were paid $20 and not invited to continue. The criteria and bonus were used to encourage participants to perform to the best of their ability. The recall and precision cutoffs were chosen with the intention of ensuring quality but not forcing users to be NIST-quality. 19 participants assessed 3 batches, 7 participants completed 2 batches, and 10 participants completed 1 batch. In total, three batches representing each context and each topic were assessed. The the continuation criterion was intended to limit the impact of poor assessments, while the random context and topic assignment without repetition was intended to mitigate learning effects.

2.3 User Interface

Participants conducted their assessments using a full-screen HTML interface, that displayed panels containing:

- The topic title and description.
- The document, with topic title words highlighted.

- Progress information, including the number of documents reviewed, elapsed time for the current document, cumulative time per document, and target time per document.

- Single-click action buttons to render an assessment and proceed to the next document.

2.4 Evaluation

The outcome of general interest is the probability $\Pr[\text{User}^+]$ that an assessor will render a positive judgement. The primary predictor variable is the context within which the document is assessed. Accordingly, we wish to measure the conditional probability $\Pr[\text{User}^+|\text{Context}]$ in order to test the hypothesis that $\Pr[\text{User}^+|\text{CAL}] < \Pr[\text{User}^+|\text{SAL}\vee\text{SPL}]$. Assuming this hypothesis to be supported, we wish to test whether, individually, $\Pr[\text{User}^+|\text{CAL}] < \Pr[\text{User}^+|\text{SAL}]$ and $\Pr[\text{User}^+|\text{CAL}] < \Pr[\text{User}^+|\text{SPL}]$.

A second predictor variable is the relevance class *Rel* of the document, as determined by some combination of Waterloo and NIST assessments. To preserve the statistical power of our experiment, we restrict our consideration to the class W-RI, and its complement W-NU, where W-RI denotes any combination of assessments for which the Waterloo assessment is either "relevant" or "iffy." We assumed that the hypothesis $\Pr[\text{User}^+|\text{W-RI}] > \Pr[\text{User}^+|\text{W-NU}]$ was extremely unlikely to be rejected, and concerned ourselves instead with the two hypotheses:

$$(1)\Pr[\text{User}^+|\text{CAL}\wedge\text{W-RI}] < \Pr[\text{User}^+|(\text{SAL}\vee\text{SPL})\wedge\text{W-RI}]$$

$$(2)\Pr[\text{User}^+|\text{CAL}\wedge\text{W-NU}] < \Pr[\text{User}^+|(\text{SAL}\vee\text{SPL})\wedge\text{W-NU}]$$

To estimate the conditional probabilities, we computed the fraction of positive assessments for documents satisfying the specified predictors. To evaluate the significance of hypothesized differences, we applied a paired binomial test, where possible, to corresponding batches. There are an equal number of batches for each context, and for each of the relevance classes W-RI and W-NU. Across these sets, batches may be matched by topic and by the specific combination of relevance assessments, leaving us to match within corresponding triples of batches, each reviewed by a different assessor. We matched the members of these triples by the assessor's experience: As far as possible, the first batch reviewed by one assessor was matched to the first batch reviewed by another assessor; the second batch reviewed by one assessor was matched to the second match reviewed by another assessor; and so on.

Predictor	$\Pr[\text{User}^+|\text{Predictor}]$	p-value
Context: CAL	0.42 (0.36,0.48)	-
Context: SAL	0.54 (0.48,0.59)	0.0002
Context: SPL	0.54 (0.48,0.60)	0.0002
Rel: W-RI	0.63 (0.58,0.68)	< 0.0001
Rel: W-NU	0.39 (0.34,0.43)	

Table 3: Probability of a study participant making a positive assessment, with 95% confidence intervals, for the primary predictors. For context, p-values were computed using a two-tailed paired binomial test; for relevance, p-values were computed using a z-test for difference in proportions. the CAL context.

Predictor	$\Pr[\text{User}^+|\text{Predictor}]$	p-value
CAL and W-RI	0.52 (0.43,0.61)	-
SAL and W-RI	0.67 (0.58,0.75)	0.0037
SPL and W-RI	0.70 (0.62,0.78)	0.0005
CAL and W-NU	0.33 (0.26,0.41)	-
SAL and W-NU	0.42 (0.34,0.50)	0.0288
SPL and W-NU	0.40 (0.32,0.48)	0.1214

Table 4: Probability of a study participant making a positive assessment, with 95% confidence intervals, for combined predictors. p-values were computed relative to CAL, using a two-tailed paired binomial test.

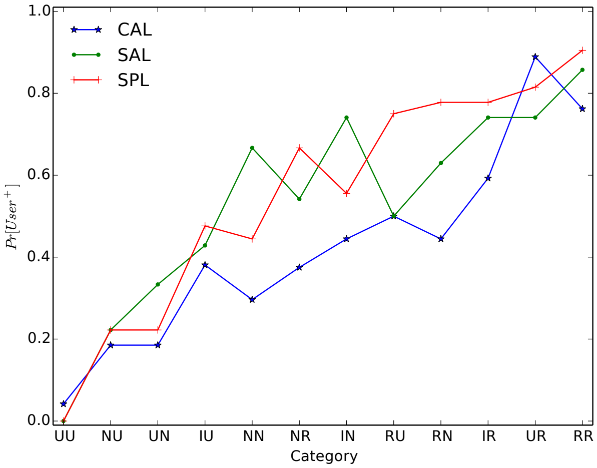

Figure 1: Probability of positive assessment given a context and elementary relevance class. Relevance classes are denoted xy where $x \in R, I, N, U$ denotes Waterloo relevant, iffy, non-relevant and unjudged, and $x \in R, N, U$ denotes NIST relevant, non-relevant, and unjudged.

3. RESULTS

Table 3 shows the results of our primary hypotheses, that context and relevance class separately influence the probability of a positive assessment. Separately, SAL and SPL both yield a substantially and significantly higher probability of positive assessment than CAL; W-RI yields a substantially and significantly higher probability of positive assessment than W-NU.

Table 4 shows the combined effect of context and relevance class. With respect to the W-NU relevance class, SAL and SPL separately yield a substantially and significantly higher probability of positive assessment than CAL. With respect to W-RI, the difference appears to be substantive, but only the difference between CAL and SAL appears to be significant, and even so would not be significant under Bonferroni correction for multiple hypothesis testing. The difference $\Pr[\text{User}^+|\text{CAL}\wedge\text{W-RI}] - \Pr[\text{User}^+|(\text{SAL}\vee\text{SPL})\wedge W-R)]$ is significant ($p \ll 0.0288$), by the following argument: for the null hypothesis to be true, it would be necessary that $\Pr[\text{User}^+|\text{CAL}\wedge\text{W-RI}] \geq \Pr[\text{User}^+|\text{SAL}\wedge\text{W-RI})]$ *and* $\Pr[\text{User}^+|\text{CAL}\wedge\text{W-RI}] \geq \Pr[\text{User}^+|\text{SPL}\wedge\text{W-RI})]$. The probability of both of these occurring by chance cannot exceed the probability of either one, which implies

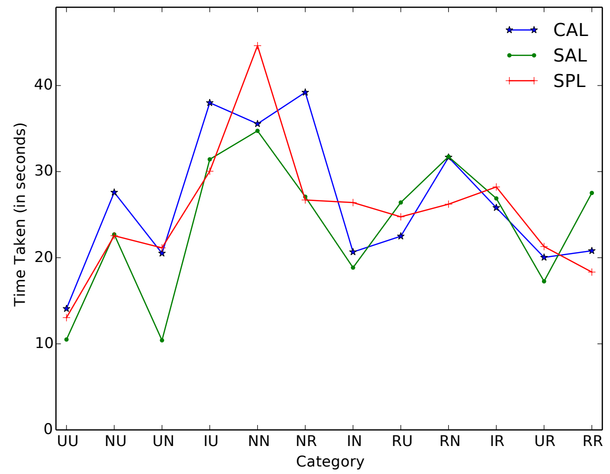

Figure 2: Average time for assessment given a context and elementary relevance class. Relevance classes are denoted xy where $x \in R, I, N, U$ denotes Waterloo relevant, iffy, non-relevant and unjudged, and $x \in R, N, U$ denotes NIST relevant, non-relevant, and unjudged

$p < \min(0.0288, 0.1214) = 0.0288$. For each combination of Waterloo and NIST assessments, Figure 1 plots the probability of a positive user assessment. Consistent with our statistical findings, the curves for SAL and SPL are generally superior to the curve for CAL. It appears that for cases where one of the Waterloo or NIST assessments is "relevant" and is not discordant with the other (*i.e.*, the other is "relevant" or "unjudged") there may be an insubstantial difference between CAL and the other contexts. Whether this observation reflects chance or an effect is a subject for future research.

4. ASSESSMENT TIME

We collected timing information in the course of implementing our participant retention criteria. Table 5 indicates that neither the context nor the relevance class has a substantial or significant effect on the time taken by participants to review documents. Figure 2, which plots assessment time taken against elementary relevance classes, suggests that non-relevant documents that were included in both the Waterloo and NIST judging pools may take longer to review. An interesting avenue of research would be to investigate whether this observation is a manifestation of the observation by Smucker and Jethani [12], who found that assessors took longer to make incorrect assessments. Namely, we are interested in answering the following question: Are the long prediction times observed for the NN relevance class the result of false positives?

5. CONCLUSIONS

Our results show clearly that assessors are less likely to judge documents relevant when they are presented within the context of documents selected using relevance sampling, than when they are presented within the context of documents selected using uncertainty sampling or random sampling. The effect holds for both relevant and non-relevant documents, as determined by archived assessments rendered by the University of Waterloo. Whether this effect applies to all relevant documents, or only to marginally relevant documents, remains an open question.

Predictor	Time Taken Per Doc	p-value
Rel: W-RI	26.56 (23.86,29.27)	0.1775
Rel: W-NU	23.92 (21.19,26.64)	
Context: CAL	26.44 (22.69, 30.19)	-
Context: SAL	23.53 (20.58, 26.48)	0.1984
Context: SPL	25.46 (22.21, 28.72)	0.6781

Table 5: Average time, in seconds, taken to assess documents under each condition with 95% confidence intervals for both all documents and known documents only. For context, p-values are with respect to a paired two-tailed t-test against the CAL predictor. For relevance, p-values are from Welch's t-test.

Our results call into question the practice of deeming the assessments of one individual to be authoritative [15], or of assuming that validation based on sampling is equivalent to validation using the pooling method.

6. REFERENCES

[1] G. V. Cormack, C. L. A. Clarke, C. R. Palmer, and S. S. L. To. Passage-Based Refinement (MultiText Experiments for TREC-6). In *Proc. TREC-6*, 1997.

[2] G. V. Cormack and M. R. Grossman. Evaluation of Machine-learning Protocols for Technology-assisted Review in Electronic Discovery. In *Proc. SIGIR 2014*, 2014.

[3] G. V. Cormack and M. R. Grossman. Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv Preprint*, 2015.

[4] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proc. SIGIR 1998*, 1998.

[5] M. Eisenberg and C. Barry. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *J. Amer. Soc. Info. Sci.*, 1988.

[6] M.-h. Huang and H.-y. Wang. The influence of document presentation order and number of documents judged on users' judgments of relevance. *J. Amer. Soc. Info. Sci.*, 55(11), 2004.

[7] D. D. Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. *SIGIR Forum*, 29(2), Sept. 1995.

[8] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proc. SIGIR 1994*, 1994.

[9] A. Roegiest, G. V. Cormack, C. L. A. Clarke, and M. R. Grossman. Impact of surrogate assessments on high-recall retrieval. In *Proc. SIGIR 2015*, 2015.

[10] F. Scholer, D. Kelly, W.-C. Wu, H. S. Lee, and W. Webber. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proc. SIGIR 2013*, 2013.

[11] M. D. Smucker and C. P. Jethani. Human performance and retrieval precision revisited. In *Proc. SIGIR 2010*, 2010.

[12] M. D. Smucker and C. P. Jethani. Time to judge relevance as an indicator of assessor error. In *Proc. SIGIR 2012*, 2012.

[13] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Proc. CLEF 2001*.

[14] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *J. Info. Proc. & Man.*, 36(5), 2000.

[15] W. Webber, D. W. Oard, F. Scholer, and B. Hedin. Assessor error in stratified evaluation. In *Proc. CIKM 2010*.