Active Learning Stopping Strategies for Technology-Assisted Sensitivity Review

Graham McDonald
University of Glasgow, UK.
graham.mcdonald@glasgow.ac.uk

Craig Macdonald
University of Glasgow, UK.
craig.macdonald@glasgow.ac.uk

Iadh Ounis
University of Glasgow, UK.
iadh.ounis@glasgow.ac.uk

ABSTRACT

Active learning strategies are often deployed in technology-assisted review tasks, such as e-discovery and sensitivity review, to learn a classifier that can assist the reviewers with their task. In particular, an active learning strategy selects the documents that are expected to be the most useful for learning an effective classifier, so that these documents can be reviewed before the less useful ones. However, when reviewing for sensitivity, the order in which the documents are reviewed can impact on the reviewers' ability to perform the review. Therefore, when deploying active learning in technology-assisted sensitivity review, we want to know when a sufficiently effective classifier has been learned, such that the *active* learning can stop and the reviewing order of the documents can be selected by the reviewer instead of the classifier. In this work, we propose two active learning stopping strategies for technology-assisted sensitivity review. We evaluate the effectiveness of our proposed approaches in comparison with three state-of-the-art stopping strategies from the literature. We show that our best performing approach results in a significantly more effective sensitivity classifier (+6.6% F_2) than the best performing stopping strategy from the literature (McNemar's test, $p < 0.05$).

ACM Reference Format:
Graham McDonald, Craig Macdonald, and Iadh Ounis. 2020. Active Learning Stopping Strategies for Technology-Assisted Sensitivity Review. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401267

1 INTRODUCTION

In technology-assisted review [5] (TAR), document classifiers are often deployed to *assist* human reviewers to find relevant documents in *high-recall* tasks, where the underlying document collection is too large to be exhaustively manually reviewed and appropriate keywords are not known *a priori* to search for all of the relevant examples. For example, TAR has been widely adopted for assisting lawyers to find digital documents that are relevant to the proceedings of a legal trial, i.e., e-discovery [13], and for the systematic review of published medical findings [8]. More recently, TAR has also been shown to be effective for *sensitivity review* tasks, for example to assist human reviewers to find sensitive information in government documents [11] so that the sensitive information can be protected and the documents can be released to the public.

In TAR, the relevance judgements that a reviewer makes are often used as document *labels* to train a document classifier that can assist the reviewers to quickly find additional relevant information. Moreover, *active learning* [17] is often deployed to reduce the time and the reviewing effort (i.e. the number of labels) that is required to train the classifier. Indeed, active learning has recently been shown to be an effective approach for learning an effective sensitivity classifier in technology-assisted sensitivity review [11].

In general, active learning approaches select the documents in the collection that are expected to be the most informative for the classifier so that the reviewer can label these documents before the documents that are expected to be less useful. Therefore, when active learning is deployed in TAR, the active learning process dictates the order in which the documents are reviewed.

For technology-assisted sensitivity review, however, it is desirable to allow the reviewers to decide the order in which the documents are reviewed. For example, when sensitivity reviewing government documents, the reviewers often rely on information from other related documents or documents that were produced within a similar time frame to evaluate potential sensitivities [7]. For example, the phrase *Some money was not accounted for in the company's finances* could be judged to be not-sensitive in many situations. However, if a related document reveals that the chief executive of the company is, for example, a high-ranking politician in another country then the information could be judged to be sensitive, since releasing it could damage relations between the two countries.

With this in mind, when deploying active learning in technology-assisted sensitivity review, it is important to know when an effective sensitivity classifier has been learned, so that the *active* learning can be stopped and the order in which the documents are reviewed can then be selected by the reviewers. In this work, we propose two active learning stopping strategies for technology-assisted sensitivity review. Moreover, we compare our approaches against three state-of-the-art active learning stopping strategies from the literature. We show that our best performing proposed approach results in a significantly more effective sensitivity classifier (+6.6% F_2) than the best performing stopping strategy from the literature (McNemar's test, $p < 0.05$).

2 ACTIVE LEARNING

Active learning [17] is a family of methods for continuously learning a classifier in an *interactive* manner where the active leaning strategy selects the documents that should be *labelled* by a human reviewer at each iteration of the learning process so that the classifier can be re-trained using the additional labelled documents.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401267

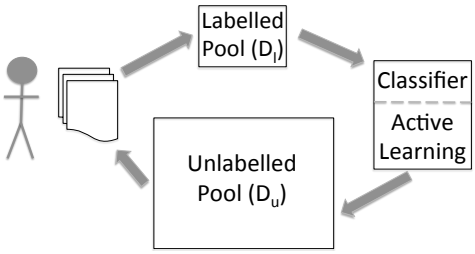

Figure 1: Pool-based active learning.

Active learning approaches have been shown to be effective for reducing the amount of data that is required to train a classification model in a number of scenarios where there is a large amount of data available but the *cost* of *labelling* the data is high. For example, synthesising *membership queries* [1] in scenarios where the input space is well defined (i.e., the dimensions and ranges of the classification features are known in advance) or *selective sampling* [2] in streaming scenarios where the active learning strategy decides if each individual example should be labelled as it becomes available.

In this work, we investigate a *Pool-based* active learning [10, 17] scenario. As illustrated in Figure 1, in *Pool-based* active learning there is a small pool of documents, D_l, that have associated class labels. Additionally, there is a larger pool of unlabelled documents, D_u, that the active learning component of the classifier can select documents from to have labelled by a reviewer. At each iteration, the active learning component selects the k documents from D_u that are expected to be the most informative for the classifier, so that these documents can be labelled and the classifier can be retrained.

Approaches for pool-based active learning include searching the classifier's parameter space to find the parameter values that enable the classifier to correctly predict the class of each of the labelled documents [12], and estimating the achievable reduction in classification error from labelling each of the documents [15]. However, such approaches are excessively computationally expensive [17].

One approach that has been shown to be effective for pool-based active learning is *Margin* uncertainty sampling [9, 17]. Uncertainty sampling assumes that the documents that the classifier is most uncertain about will also be the most informative for the classifier. Margin uses the difference between the classifier's prediction confidence scores (or probability estimates) of the first and the second most likely class labels (according to the classifier's predictions) as a measure of the classifier's uncertainty about the true class of a document. Margin is defined as $M(d_i, l_1, l_2) = |P(l_1|d_i) - P(l_2|d_i)|$ where for a document d_i, l_1 and l_2 are the classifier's most confident and second most confident prediction scores. Documents with a small margin score are expected to be the most informative for the classifier.

McDonald *et al.* [11] showed that Margin uncertainty sampling significantly reduced the amount of labelled data that was needed to train an effective sensitivity classifier. In this work, we also deploy Margin uncertainty sampling to learn a sensitivity classifier. However, differently from McDonald *et al.* [11], in this work we evaluate the effectiveness of stopping strategies to predict when an effective sensitivity classifier has been learned.

In active learning, defining an acceptable level of classification effectiveness is often not practical. Evaluating the classifier requires an additional set of documents with labelled examples of the classes that we aim to classify. If such an evaluation set existed then the classifier could potentially be learned using these additional documents. Moreover, it is not possible to generate a representative evaluation set from the documents that have been labelled as part of the active learning process, since the active learning strategy has specifically selected these documents to improve the classifier [17].

In practice, an active learning stopping strategy aims to stop the active learning process as close as possible to the point at which an optimal classifier is learned, i.e, the most effective classifier that could be learned on the entire document collection, given the particular classifier that has been deployed [3]. An overly *aggressive* stopping strategy stops the active learning process too soon, resulting in a sub-optimal classifier, while an overly *conservative* approach allows the active learning process to continue after an optimal classifier has been learned. In the case of sensitivity review, this results in the reviewers' ability to perform sensitivity review being impacted without any additional benefit to the classifier.

Many of the stopping strategies from the literature can only be deployed with specific classifiers, for example support vector machines [6, 16], or for specific active learning approaches, such as committee-based active learning [14]. Relying on such strategies for technology-assisted sensitivity review would limit the classification and active learning approaches that could be deployed, as new approaches are developed. Therefore, in this work, our proposed stopping strategies, and the strategies that we evaluate from the literature, are agnostic to the choice of classifier that is deployed.

3 ACTIVE-LEARNING STOPPING STRATEGIES

In this section, we present the active learning stopping strategies that we evaluate for technology-assisted sensitivity review. We firstly present our proposed approaches before, secondly, describing the stopping strategies from the literature that we evaluate.

Proposed Approaches: Our proposed active learning stopping strategies for technology-assisted sensitivity review use the classification confidence scores that are generated by the uncertainty sampling active learning strategy. The intuition behind our proposed approaches is that as the classifier learns more about the properties of sensitive information, and learns to predict sensitivity more accurately over time, then the classifier's confidence in its predictions will increase. We note that, although we deploy Margin uncertainty sampling in this work, our proposed approaches could be deployed with any uncertainty sampling approach.

Our first proposed approach, *TotalConf*, measures the classifiers' overall confidence in correctly classifying the remaining unlabelled documents. The intuition for this approach is as follows: at each iteration of the active learning process, uncertainty sampling provides us with a single score for each document that tells us how confident the classifier is about classifying the document. Moreover, these confidence scores are comparable across different iterations of the active learning process, as long as the number of potential classes is fixed. With this in mind, this approach assumes that as the classifier improves, its overall confidence will also increase until it has learned how to effectively classify sensitivity.

TotalConf assumes that when the classifier's mean confidence level for the remaining documents stabilises, then the classifier's effectiveness is no longer improving. Therefore, this approach stops the active learning process if the classifier's *TotalConf* score does

not increase for ϵ iterations. The *TotalConf* score is calculated as follows:

$$TotalConf = \frac{\sum_{d_u} |l_1 - l_2|}{|D_u|} \quad (1)$$

where d_u is an unlabelled document in D_u and $|l_1 - l_2|$ is the Margin score for d_u.

The second stopping strategy that we propose, *LeastConf*, measures the classifiers' confidence for the documents that are selected to be reviewed. The intuition for this approach is as follows: at each iteration of the active learning process, the active learning strategy selects the documents that the classifier is least confident about. Similarly to *TotalConf*, this approach also assumes that as the classifier improves, its confidence will increase until it has learned how to classify sensitivity. However, differently from *TotalConf*, this approach monitors the documents that are selected to be reviewed, i.e. the documents that the classifier is least confident about, and assumes that when the classifier's confidence stops increasing for these documents it has reached maximal confidence. Therefore, this approach stops the active learning process if the *LeastConf* score does not increase for ϵ iterations. *LeastConf* is calculated as follows:

$$LeastConf = \frac{\sum_{d_s} |l_1 - l_2|}{|D_s|} \quad (2)$$

where d_s is a document in the set of documents, D_s, that have been selected to be reviewed and $|l_1 - l_2|$ is the Margin score for d_s.

Approaches from the Literature: The first active learning stopping strategy that we evaluate from the literature stops the active learning process when the classifier's predictions stabilise [3], denoted as *StablePred* in Section 5. The approach monitors the classifier's predictions made on the unlabelled set, D_u, and stops active learning when the predictions have stabilised for ϵ iterations. To measure how stable the classifiers' predictions are, the approach calculates the Cohen's κ [4] agreement between the predictions from the current iteration and from the previous iteration. This approach stops the active learning process if the κ score is greater than a threshold, θ, for a predefined number of iterations.

The second stopping strategy from the literature that we evaluate, *Classification Change* [19], monitors the number of documents in the unlabelled set that the classifier changes its predictions for at each iteration. If the classifier does not change its predictions for the remaining unlabelled documents for ϵ iterations, this approach, denoted as *ClassChange* in Section 5, assumes that the classifier has reached its maximum effectiveness and stops the active learning.

The third, and final, stopping strategy from the literature is *Min-Error* [18], which measures the accuracy of the classifier's predictions on the documents that it is least certain about. For each iteration, the classifier predicts the class of each of the documents that are selected to be reviewed, before the documents are presented to the reviewer. If the classifier gets all of these predictions correct for ϵ iterations, then *Min-Error* assumes that the classifier has achieved a *good enough* level of effectiveness to be able to classify the remaining unlabelled documents, and stops the active learning.

4 EXPERIMENTAL SETUP

In this section, we present our experimental setup for evaluating active learning stopping strategies for technology-assisted sensitivity review. We wish to answer the following research question: "*Is the amount of uncertainty that the classifier has in its predictions a good heuristic to know when to stop active learning in technology-assisted sensitivity review?*"

We evaluate our research question using a test collection of 3801 government documents that have been reviewed for sensitivity by experienced government reviewers. The reviewers assessed the collection for *international relations* and *personal information* sensitivities, as defined by the UK Freedom of Information Act 2000.[1] The collection contains 502 sensitive documents (13%) and 3299 non-sensitive (87%). We use 500 of these documents (435 non-sensitive, 65 sensitive) as a fixed held-out set to evaluate the effectiveness of the classifier at each iteration of the active learning process.[2] To ensure the generalisability of our findings, we run our experiments over 25 stratified samples of 2500 documents from the remaining documents in the test collection (2175 non-sensitive, 325 sensitive).

In our experiments, for each iteration, we set the number of documents to be labelled to 20 and randomly down sample the training data when training the classifier. We deploy a SVM classifier with a linear kernel and $C = 1.0$. For each of the stopping strategies presented in Section 3, following [3], we set the threshold number of iterations to trigger the stopping strategies, $\epsilon = 3$, and our Cohen's κ threshold $\theta = 0.99$. We test for statistical significance in the achieved classification effectiveness using McNemar's non-parametric test, with $p < 0.05$. Stopping strategies that result in a classifier that is significantly more effective than the next best approach are denoted by † in Table 1.

Table 1: The achieved classifier effectiveness when each of the active learning stopping strategies is deployed.

	Precision	Recall	F_1	F_2	BAC	auROC
All_Docs	0.2819	0.7571	0.4108	0.5662	0.7215	0.7745
$Oracle_{opt}$	0.2806	0.7857	0.4135	0.5777	0.7289	0.7673
TotalConf †	**0.2631**	**0.7857**	**0.3942**	**0.5623**	**0.7137**	**0.7721**
LeastConf †	0.2575	0.7285	0.3805	0.5334	0.6933	0.7623
StablePred	0.2577	0.7142	0.3787	0.5274	0.6897	0.7470
ClassChange †	0.2525	0.7000	0.3712	0.5168	0.6813	0.7508
MinError	0.1737	0.5857	0.2679	0.3972	0.5661	0.6124

5 RESULTS

Table 1 presents the effectiveness of the sensitivity classifier, in terms of precision, recall, F_1, F_2, balance accuracy (BAC) and the area under the ROC curve (auROC), at the point when each of the evaluated stopping strategies stops the active learning process. Table 1 also shows the effectiveness of the learned classifier after all of the documents have been labelled, denoted as *All_Docs*, and an oracle stopping strategy that stops the active leaning process when the classifier achieves the highest BAC score, denoted as $Oracle_{opt}$.

Firstly, from Table 1, we note that our proposed $TotalConf$ stopping strategy results in the most effective sensitivity classifier for all of the reported metrics. Moreover, this classifier is significantly more effective (denoted as †) than that of the next best performing strategy, $LeastConf$ and +6.6% more effective, in terms of F_2, than the best performing strategy from the literature, $StablePred$. Indeed, in our experiments, both of our proposed uncertainty based stopping strategies outperform $StablePred$ in terms of recall, F_1,

[1] http://www.legislation.gov.uk/ukpga/2000/36/contents

[2] None of the active learning stopping strategies that we evaluate make use of this held-out test data.

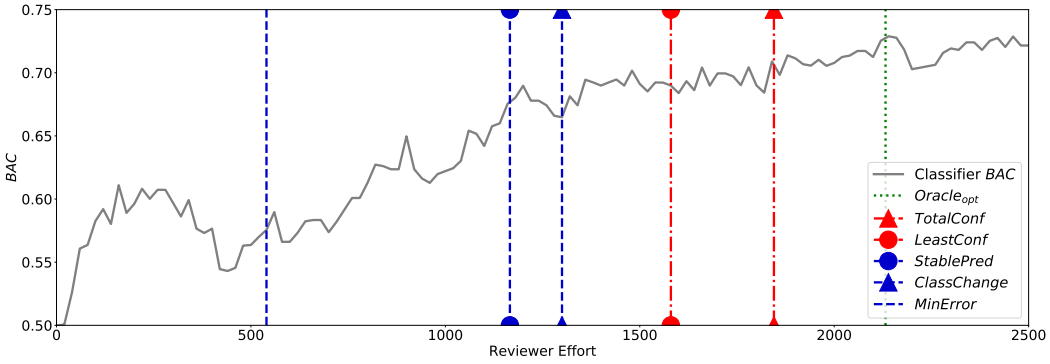

Figure 2: Active learning stopping strategies' performances in terms of balanced accuracy (BAC).

F_2, BAC and auROC. Therefore, we conclude that a sensitivity classifier's uncertainty can be a good indicator of its effectiveness.

Secondly, we note that the $Oracle_{opt}$ stopping strategy results in the most effective sensitivity classifier, in terms of recall, F_1, F_2 and BAC. This shows that it is beneficial to know when to stop the active learning process, both in terms of increasing the effectiveness of the sensitivity classifier and to avoid unnecessarily dictating the reviewing order of the documents.

Figure 2 shows how far through the active learning process each of the stopping strategies stops the active learning process (vertical dashed lines). The x-axis of the figure shows the number of documents that have been sensitivity reviewed and used to train the classifier, i.e., the reviewer effort, while the y-axis shows the classification effectiveness in terms of BAC. The classifier's effectiveness at each iteration of active learning is shown by the solid grey line.

Firstly, we observe from Figure 2 that all of the active learning stopping strategies that we evaluate tend to be overly aggressive, and hence result in a less than optimal classifier being learned. As future work, we will conduct a user study to evaluate if the benefits from enabling the reviewers to select the order in which the documents are reviewed are offset if the reviewers are assisted by less accurate classification predictions. We note, however, that although the stopping strategies stop the *active* learning process, the classifier can continue to be retrained as more documents are reviewed.

StablePred is the best performing stopping strategy that we evaluate from the literature. However, this strategy is more aggressive than either of our proposed approaches. *StablePred* is more likely to stop the active learning process in each consecutive active learning iteration, since the likelihood of agreement between the classifier's predictions increases as the number of unlabelled documents decreases. Therefore, if the effectiveness of the classifier continues to increase even when there are relatively few unlabelled documents remaining, as is often the case with learning to classify sensitivity, then it appears that this approach tends to stop the active learning too soon to be effective for sensitivity classification.

Each of the stopping strategies the we evaluate in this work stops active learning after a heuristic condition has been observed for ϵ iterations. As future work, we will investigate stopping strategies that automatically set this threshold.

6 CONCLUSIONS

In this work, we proposed two novel active learning stopping strategies for technology-assisted sensitivity review. Moreover, we evaluated the effectiveness of our proposed approaches against three state-of-the-art stopping strategies from the literature. We showed that our best performing proposed approach resulted in a significantly (McNemar's test, $p < 0.05$) more effective sensitivity classifier (+6.6% F_2) than the best performing stopping strategy that we evaluated from the literature.

REFERENCES

[1] D. Angluin. 1988. Queries and Concept Learning. *Machine Learning* 2, 4 (1988), 319–342.
[2] L. Atlas, D. Cohn, and R. Ladner. 1990. Training Connectionist Networks with Queries and Selective Sampling. In *Proc. NIPS*.
[3] M. Bloodgood and K. Vijay-Shanker. 2009. A Method for Stopping Active Learning Based on Stabilizing Predictions and the Need for User-Adjustable Stopping. In *Proc. CoNLL*.
[4] J Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.
[5] G. Cormack and M. Grossman. 2014. Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery. In *Proc. SIGIR*.
[6] S. Ertekin, J. Huang, L. Bottou, and L. Giles. 2007. Learning on the Border: Active Learning in Imbalanced Data Classification. In *Proc. CIKM*.
[7] T. Gollins, G. McDonald, C. Macdonald, and I. Ounis. 2014. On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records. In *Proc. PIR@SIGIR*.
[8] C. Lefebvre, E. Manheimer, and J. Glanville. 2008. Searching for Studies. *Cochrane Handbook for Systematic Reviews of Interventions* (2008), 95–150.
[9] D. Lewis and W. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. In *Proc. SIGIR*.
[10] A. McCallumzy and K. Nigamy. 1998. Employing EM and Pool-Based Active Learning for Text Classification. In *Proc. ICML*.
[11] G. McDonald, C. Macdonald, and I. Ounis. 2018. Active Learning Strategies for Technology-Assisted Sensitivity Review. In *Proc. ECIR*.
[12] T. Mitchell. 1982. Generalization as Search. *Artificial Intelligence* 18, 2 (1982), 203–226.
[13] D. Oard, J. Baron, B. Hedin, D. Lewis, and S. Tomlinson. 2010. Evaluation of Information Retrieval for E-Discovery. *Artificial Intelligence and Law* 18, 4 (2010), 347–386.
[14] F. Olsson and K. Tomanek. 2009. An Intrinsic Stopping Criterion for Committee-Based Active Learning. In *Proc. CoNLL*.
[15] N. Roy and A. McCallum. 2001. Toward Optimal Active Learning through Sampling Estimation of Error Reduction. In *Proc. ICML*.
[16] G. Schohn and D. Cohn. 2000. Less is More: Active Learning with Support Vector Machines. In *Proc. ICML*.
[17] B. Settles. 2012. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–114.
[18] J. Zhu and E. Hovy. 2007. Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem. In *Proc. EMNLP-CoNLL*.
[19] J. Zhu, H. Wang, and E. Hovy. 2008. Multi-Criteria-Based Strategy to Stop Active Learning for Data Annotation. In *Proc. Coling*.