How Sensitivity Classification Effectiveness Impacts Reviewers in Technology-Assisted Sensitivity Review

Graham McDonald, Craig Macdonald, Iadh Ounis
University of Glasgow, Glasgow, Scotland, UK
firstname.lastname@glasgow.ac.uk

ABSTRACT

All government documents that are released to the public must first be manually reviewed to identify and protect any *sensitive* information, e.g. confidential information. However, the unassisted manual sensitivity review of born-digital documents is not practical due to, for example, the volume of documents that are created. Previous work has shown that sensitivity classification can be effective for *predicting* if a document contains sensitive information. However, since all of the released documents must be manually reviewed, it is important to know if sensitivity classification can *assist* sensitivity reviewers in making their sensitivity judgements. Hence, in this paper, we conduct a digital sensitivity review user study, to investigate if the accuracy of sensitivity classification effects the number of documents that a reviewer correctly judges to be sensitive or not (reviewer accuracy) and the time that it takes to sensitivity review a document (reviewing speed). Our results show that providing reviewers with sensitivity classification predictions, from a classifier that achieves 0.7 Balanced Accuracy, results in a 38% increase in mean reviewer accuracy and an increase of 72% in mean reviewing speeds, compared to when reviewers are not provided with predictions. Overall, our findings demonstrate that sensitivity classification is a viable technology for assisting with the sensitivity review of born-digital government documents.

1 INTRODUCTION

Government documents must be manually reviewed to identify any *sensitive* information, e.g. personal or confidential information, before they can be released to the public, e.g. through the UK Freedom of Information Act [6] (FOIA). In the UK, the sensitivity review of paper documents requires an exhaustive manual review, usually by expert sensitivity reviewers, of *all* of the documents that are released.

However, this human-only sensitivity review process is not practical for the sensitivity review of born-digital documents, due to the volume of digital documents that are created and the lack of structure in the collections that are to be reviewed [25]. Therefore, it has been widely recognised that a technology-assisted review (TAR) approach to the sensitivity review of digital government documents is both necessary and unavoidable [1, 12, 21].

The automatic classification of sensitive information [17, 18] is one emerging technology that has the potential to be the basis for technology-assisted sensitivity review (TASR). However, although it is generally accepted that some form of TASR is necessary [2], it is also accepted that: (1) *all* government documents will continue to be manually sensitivity reviewed for the foreseeable future [25]; (2) governments will not be able to recruit enough reviewing resources [25]. One of the roles of sensitivity classification in TASR will, therefore, be to provide the reviewers with useful information that can assist them in making accurate reviewing decisions more quickly. Moreover, this, in-turn, could potentially enable governments to increase the reviewing resources by recruiting more less-experienced (less expensive) reviewers and assisting them to conduct accurate sensitivity reviewing.

In this work, we conduct a within-subject controlled user study to investigate if providing reviewers with automatic sensitivity classification predictions helps them to perform sensitivity review. The study participants each reviewed three *batches* of government documents to identify two FOIA sensitivities [7], namely *international relations* and *personal information*. Each batch of documents had an associated *effectiveness* level of sensitivity classification predictions, either *None*, *Medium* or *Perfect* predictions. The study evaluates how the effectiveness (accuracy) of sensitivity classification predictions affects the number of documents that a reviewer correctly judges to contain, or to not contain, sensitive information (reviewer accuracy) and the length of time that it takes for a reviewer to sensitivity review a document (reviewing speed).

Our findings show that automatic sensitivity classification, with an effectiveness close to that of the sensitivity classifiers from the literature (e.g. from McDonald *et al.* [17, 18]), can result in significant improvements in the accuracy of sensitivity reviewers, compared to when no predictions are provided (+38%). Moreover, we find that providing reviewers with sensitivity classification predictions results in a statistically significant 72% increase in the mean reviewing speed of reviewers (repeated measures ANOVA ($p < 0.05$)).

The contributions of this paper are two-fold: Firstly, we present the first examination of the benefits of automatic sensitivity classification predictions for human sensitivity reviewers; Secondly, our study shows that sensitivity classification is a valuable technology for TASR, and can significantly increase the speed and accuracy of reviewers when they sensitivity review born-digital documents.

2 RELATED WORK

The need for automatic tools to identify sensitive information has been recognised by governments for a number of years [1, 12, 26, 27]. However, although there is a substantial amount of literature on masking personal information, e.g. [13, 24], it is only relatively recently that research has advanced in the field of sensitivity classification that we address in this work, i.e. classifying FOIA sensitivities.

McDonald et al. [18] was the first work to investigate automatically classifying FOIA exemptions. In [18], the authors showed that text classification [23] is a viable approach for developing sensitivity classification. The authors achieved a Balanced Accuracy (BAC) [5]

CHIIR '19, March 10–14, 2019, Glasgow, Scotland Uk

Table 1: The distribution of classification predictions for classification accuracy treatments (batches).

Classification	TP	FN	FP	TN	Sensitive	Not Sensitive	Total	BAC
None	-	-	-	-	5	15	20	-
Medium	3	2	3	12	5	15	20	0.7
Perfect	5	0	0	15	5	15	20	1.0

of 0.73 when classifying individual FOIA exemptions. When classifying two FOIA exemptions as a single category of information (on a different collection), McDonald et al. [17] achieved 0.71 BAC. The sensitivity classifiers of [17, 18] both achieved ~0.7 BAC. Therefore, in this study, we evaluate if sensitivity classification that achieves 0.7 BAC is sufficiently effective to assist sensitivity reviewers to make accurate reviewing decisions faster.

As previously mentioned in Section 1, all digital government documents will continue to be manually reviewed for the foreseeable future. Therefore, sensitivity classification must be deployed within a TAR framework. TAR is notably associated with e-discovery [22]. In e-discovery, the human reviewers typically only review the documents that are predicted to be relevant [8]. In that context, TAR has been shown to be more effective and more efficient than human only review [14]. However, in sensitivity review, *all* of the documents that are to be released must first be manually reviewed. Therefore, there is a need for studies to investigate how sensitivity classification can be of benefit to sensitivity reviewers when all of the relevant and non-relevant documents will be reviewed.

Berardi et al. [3] evaluated how sensitivity classification could be deployed within TASR to increase the overall accuracy of a human review, when there are insufficient available reviewing resources. Berardi et al. [3] built on the work of McDonald et al. [18] and showed that ranking the classified documents by the *Utility* [4], or expected increase in the overall classification effectiveness, that would be achieved if a mis-classified document was corrected, resulted in substantial improvements in overall classification effectiveness (+ 14% F_2) when only part of the collection was reviewed. Differently from the work of Berardi et al. [3], in this paper, we investigate if sensitivity classification can *assist* reviewers to make accurate sensitivity reviewing decisions more quickly.

3 EXPERIMENTAL METHOD

To investigate if sensitivity classification increases the number of correct sensitivity judgements (reviewer accuracy) and the speed at which reviewers sensitivity review digital documents (reviewing speed), we investigate four research questions, as follows:

- **RQ1**: Does providing reviewers with sensitivity classification predictions increase reviewer accuracy?
- **RQ2**: Does the reviewer accuracy increase as the sensitivity classifier's accuracy increases?
- **RQ3**: Does providing reviewers with sensitivity classification predictions increase reviewing speed?
- **RQ4**: Does the reviewing speed increase as the sensitivity classifier's accuracy increases?

We conducted a controlled user study under laboratory conditions. Participants, i.e. *reviewers*, reviewed 3 *batches* of documents to identify any documents that contained sensitive information relating to either of two UK FOIA sensitivities [7], namely: international relations; and personal information. Each batch of documents had an associated *treatment* of sensitivity classification predictions.

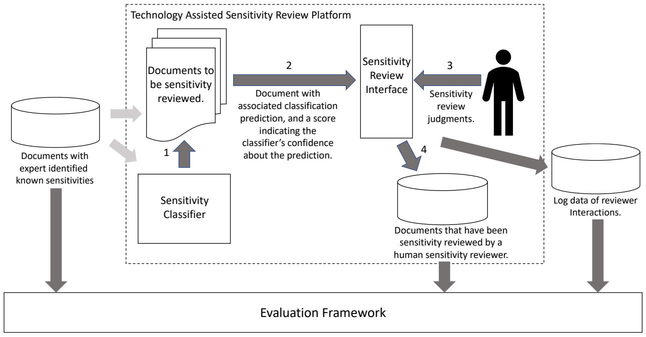

Technology Assisted Sensitivity Review Platform
Documents to be sensitivity reviewed.
2
Document with associated classification prediction, and a score indicating the classifier's confidence about the prediction.
Sensitivity Review Interface
3
Sensitivity review judgments.
Documents with expert identified known sensitivities
1
Sensitivity Classifier
4
Documents that have been sensitivity reviewed by a human sensitivity reviewer.
Log data of reviewer Interactions.
Evaluation Framework

Figure 1: A platform for evaluating the benefits of providing reviewers sensitivity classification predictions in TASR.

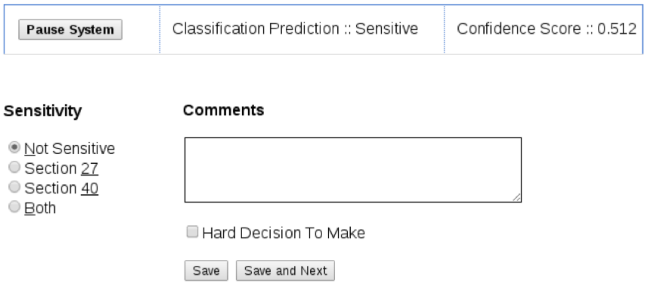
Test Collection and Expert judgements: The sixty documents used in this study are from a collection of 4000 government documents that contain real sensitivities. We obtained *gold standard* ground truth sensitivity judgements for the collection prior to the start of the study by having *expert* sensitivity reviewers from UK government departments sensitivity review the documents.

Experimental Design: The study was a within-subject design, where each participant was exposed to all of the conditions being evaluated. Participants were asked to review 3 batches of 20 documents. Each batch of documents had an associated treatment of classifier effectiveness, either *None* (i.e. no classification predictions were provided), *Medium* (i.e. the accuracy of the classification predictions was 0.7 BAC) or *Perfect* (i.e. the classification predictions agreed with the expert reviewers gold standard and, therefore, had an accuracy of 1.0 BAC).

Using the expert sensitivity reviewers' gold standard judgements, we randomly sampled documents from the collection to fit the distributions of sensitive and not-sensitive documents presented in Table 1 (25% sensitive, 75% not sensitive). For batches with *Medium* classification effectiveness (0.7 BAC), 3 documents had True Positive (TP) predictions, 2 documents had False Negative (FN) predictions, 3 documents had False Positive (FP) predictions and 12 documents had True Negative (TN) predictions (*sensitive* is the positive class).

Participants reviewed 3 batches of documents each (1**None*, 1**Medium*, 1**Perfect*), i.e. 60 documents each. To control for learning effects and fatigue, we counterbalanced the allocation of batches, i.e. we permuted the order in which batches were reviewed by different reviewers. Documents within a batch were presented in random order, consistently between reviewers. Participants were advised to proceed linearly through the batch, however, they were able to re-visit documents and change any previously made judgements. For each document, a participant recorded whether the document was "*Not Sensitive*" or contained either "*Section 27*" (international relations), "*Section 40*" (personal information) or "*Both*" sensitivities.

Reviewing Interface and Logging: Figure 1 presents the four steps of the TASR process (within the dashed box). In Step 1, a sensitivity classifier predicts which of the documents that are to be reviewed contain sensitive information. In Step 2, the documents and associated sensitivity predictions are displayed to a reviewer via a reviewing interface. In Step 3, the reviewer reads, and reviews, each document in-turn to judge whether the document is sensitive or not. In our study, we evaluate how the effectiveness of sensitivity classification predictions affect Step 3 of the TASR process. In step 4, the sensitivity reviewed documents, their sensitivity judgements and the reviewers' actions are logged for analysis.



Classification Prediction :: Sensitive

Confidence Score :: 0.512

Sensitivity

● Not Sensitive
○ Section 27
○ Section 40
○ Both

Comments

□ Hard Decision To Make

Figure 2: Reviewing Interface Information Panel: Reviewers can view the classification prediction (Sensitive or Not Sensitive), record sensitivity judgements and pause the system.

Figure 2 presents the information panel of the reviewing interface. When participants review a document, the panel presents the current document's classification prediction (Sensitive or Not Sensitive). The document to be reviewed is displayed below the panel in Figure 2. Therefore, when reviewing a document, the reviewers are presented with the classification prediction before they actually review the document. As can be seen in Figure 2, participants recorded a sensitivity judgement by selecting one of the four radio buttons at the left of the panel. Participants were also asked to provide a short comment about their decision. For documents that were judged to be sensitive, participants were asked to highlight any sensitive text within the document. A simple mouse-click and drag functionality facilitated the highlighting of sensitive text.

As can be seen from Figure 2, we provided the participants with a button to pause the system. Participants could use this button at any time, for example to have a comfort break or ask a question, this helped to ensure that we recorded accurate timings of when participants were focused on the reviewing task. The interface logged a timestamped record of when a participant loaded a document, saved a judgement, paused or restarted the system.

Participants, Incentives and Training: We recruited eight participants for the study from a well-known international university. Since participants were to identify information relating to personal information or international relations sensitivities, we limited participants to those whose main subject of education was politics or international law (to ensure that the participants had a good knowledge of the FOIA and were familiar with the concepts that they were being asked to review). Additionally, all of the subjects had spoken English for at least 10 years. Full ethical approval for the study was obtained from our organisation's ethics IRB.

At the beginning of the study, there was a 1 hour training session where participants were provided with training on the reviewing interface and the sensitivities that they were being asked to identify. Participants were provided with the same training manual that the expert sensitivity reviewers used. The 1 hour training session included a presentation of the manual and examples of sensitive and not-sensitive documents. Moreover, participants were given time to review a batch of 8 practice documents, and discuss their reviewing decisions with the study coordinator, before the study began.

Participants were remunerated £7.50 per hour. In total, including training times, each participant took between 10-12 hours, split over 2 sessions, to complete the study. In line with the findings of [19], participants were advised to take frequent short breaks.

Evaluation and Metrics: We compare the reviewers performance for each of the classification treatments. We report the mean performance (calculated over all reviewers) for each classification treatment, *None*, *Medium* and *Perfect*, in terms of the number of documents that a reviewer correctly judges to contain, or to not contain, sensitive information (reviewer accuracy) and the length of time that it takes to sensitivity review a document (reviewing speed).

When evaluating the reviewer accuracy, we use the expert sensitivity reviewers' judgements as a ground truth. We select BAC and F_2 as our metrics, since these measures are particularly suited to identifying sensitivity [16]. More specifically: BAC provides an easily interpretable accuracy score (0.5 indicates random) for both classes when the classes are imbalanced; and F_2 accounts for the fact that, in sensitivity review, there are greater consequences from miss-classifying a sensitive document than a non-sensitive one.

When evaluating the participants' reviewing speeds, we use Normalised Processing Speed (NPS) [11] to control for the effects of varying reading speeds and document lengths. NPS is calculated as $\frac{|d|}{exp(\log(time)+\mu-\mu_\alpha)}$, where $|d|$ is the document length and $\log(time)$ is the natural logarithm of the time taken to review d, μ_α is the mean $\log(time)$ for the reviewer who reviewed d, calculated over a particular treatment condition, and μ is the global mean $\log(time)$ calculated for all reviewers over all documents.

When presenting our results in Section 4, we plot the mean participant score (e.g. for BAC or NPS) and 95% confidence intervals. We use the Cousineau and Morey [9, 15, 20] method to calculate confidence intervals. Importantly, this enables the reader to use the *rule of eye* to evaluate the significance of the results from the plots, i.e. we can expect $p < 0.01$ for non-overlapping intervals and $p < 0.05$ when two intervals overlap by <50%. To calculate statistical significance, we use a one-way repeated measures univariate ANOVA in pair-wise comparisons between treatment conditions, e.g. *Medium* vs. *Perfect*. We select $p < 0.05$ as our significance threshold.

4 RESULTS

To answer the research questions that we presented in Section 3, we compare the mean reviewer performance, in terms of reviewer accuracy and reviewing speed, for each of the classification levels *None*, *Medium* and *Perfect*. Firstly, we evaluate if providing reviewers with sensitivity classification predictions increases reviewer accuracy. Figure 3 presents the mean reviewer accuracy in terms of Balanced Accuracy (BAC) for each of the classification treatments, while Figure 4 presents the analogous reviewer accuracy in terms of F_2.

From Figure 3, we note that there is a clear and steady improvement in mean participant BAC scores as the effectiveness of the classifier increases, from 0.5 BAC when there are no classification predictions to 0.69 BAC (+38%) for medium classification effectiveness and 0.8 BAC when the classification predictions agree perfectly with the expert ground truth. Importantly, 0.5 BAC indicates that, on average, without classification predictions the participants' judgements were effectively random. This is indicative of the difficulty of the sensitivity reviewing task, and underlines why government departments have typically employed expert sensitivity reviewers.

Additionally from Figure 3, we note that for the *Medium* classification effectiveness treatment, the mean participant performance is almost equivalent to the level of classification effectiveness (participants = 0.69 BAC, classifier = 0.7 BAC). However, the participants

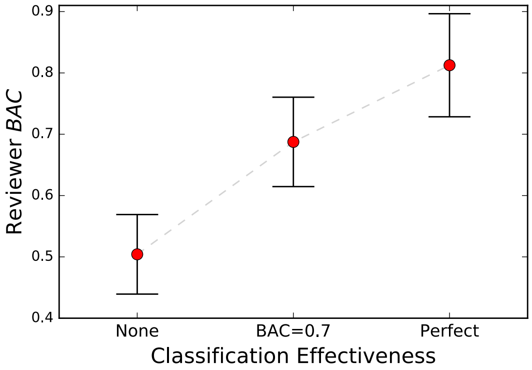


Figure 3: Mean reviewer Balanced Accuracy (BAC).

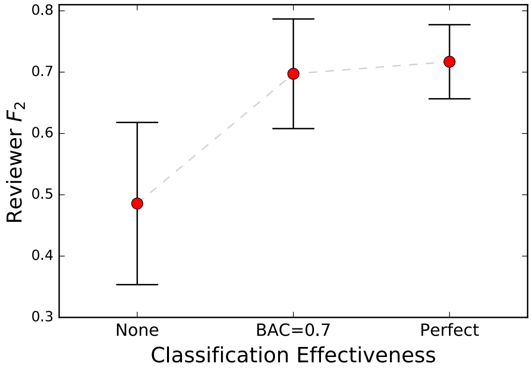


Figure 4: Mean reviewer accuracy in terms of F_2.

only achieved an accuracy of 0.8 BAC when they were provided with perfect classification predictions. This shows that even when the classifier agrees perfectly with the expert ground truth, reviewers still disagree with the classifier. Indeed, none of the participants in our study completely agreed with the *Perfect* classifier.

From Figure 4, we note that, in terms of F_2, the increase in reviewer accuracy is notably greater between the *None* and *Medium* classification than between the *Medium* and *Perfect* classification. The ANOVA test of mean reviewer accuracy between no classification and *Medium* classifier effectiveness shows significant improvements, both in terms of BAC $[F(1,7) = 23.528, p = 0.002]$ and F_2 $[F(1,7) = 7.936, p = 0.026]$. However, comparing the reviewer accuracy improvements between the *Medium* and *High* classifier effectiveness, the ANOVA test shows significant improvements in terms of BAC $[F(1,7) = 6.377, p = 0.040]$ but not in terms of F_2 $[F(1,7) = 0.560, p = 0.479]$. The main increase in reviewer accuracy between the *Medium* and *Perfect* classification are a result of more True Negative judgements since the BAC score, which accounts for True Negatives, significantly increased, while there was no significant increase in F_2, which does not consider True Negatives.

In response to **RQ1**, we conclude that providing reviewers with sensitivity classification predictions does increase the reviewer accuracy. In response to **RQ2**, we conclude that the reviewer accuracy does indeed increase as the classifier accuracy increases. However, in our study, there appears to be diminishing gains in reviewer accuracy as the classification accuracy increases towards perfect. We will investigate this further as future work.

Turning our attention to reviewing speed, Figure 5 presents the participants' mean NPS, in words per minute (wpm). As can be seen from Figure 5, the mean NPS of reviewers when no classification predictions are provided is 151 wpm. Providing reviewers with classification predictions (0.7 BAC) results in a mean NPS increase of 72% to 260 wpm. The one-way ANOVA between *None* and *Medium* shows that this is a significant result, $[F(1,7) = 79.549, p = 0.0001]$.

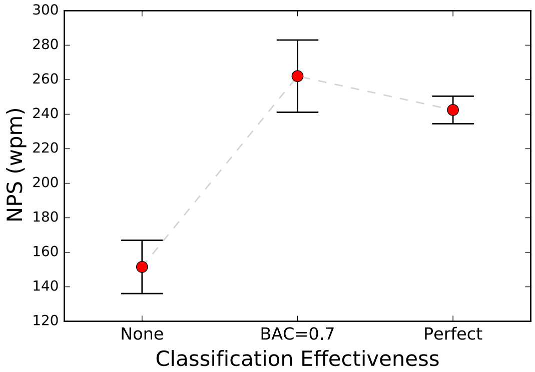


Figure 5: Mean reviewer Normalised Processing Speed (NPS).

Interestingly, we note from Figure 5 that the mean reviewing speed is slightly less when the classifier predictions are *Perfect*, 260 wpm (0.7 BAC) vs 244 wpm (perfect). The one-way ANOVA between *Medium* and *Perfect* classification shows that this decrease is not significant ($[F(1,7) = 4.210, p = 0.079]$). Therefore, the significant gains in reviewing speeds compared with no classification predictions are sustained over both levels of classification accuracy. In response to **RQ3**, we conclude that providing reviewers with sensitivity classification predictions does increase the average reviewing speed. In response to **RQ4**, we conclude that the average reviewing speed does increase between no classification and *Medium* classification. However, in our study, reviewing speeds did not increase when the classifier's accuracy increased from *Medium* to *Perfect*.

In summary, the results of our study show that providing reviewers with sensitivity classification predictions can increase the accuracy and speed of the reviewers. We argue that, our findings show that sensitivity classification with an accuracy of 0.7 BAC is sufficiently effective to assist reviewers in making sensitivity reviewing decisions. Importantly, our findings suggest that governments may be able to increase the volume of digital documents that can be reviewed, while maintaining high levels of reviewing accuracy, if they increase the number of reviewers by recruiting less experienced reviewers (at less expense than expert reviewers) and assisting them with automatic sensitivity classification predictions. This, in turn, would enable the expert reviewers to focus on reviewing the more *high risk* documents or disputed reviews.

5 CONCLUSIONS

In this work, we conducted a within-subject digital sensitivity review user study under laboratory conditions, to evaluate the benefits of automatic sensitivity classification predictions for sensitivity reviewers. We found that providing reviewers with sensitivity classification predictions resulted in significant improvements in the number of correct sensitivity judgements made by the reviewers in our study (+38% BAC) and their reviewing speed (+72% NPS), according to a repeated measures ANOVA, $p < 0.05$. Our findings provide evidence that a sensitivity classifier that achieves 0.7 BAC is sufficiently effective to assist reviews in making accurate sensitivity judgements faster when reviewing born-digital documents.

REFERENCES

[1] Sir Alex Allan. 2014. Records Review. Cabinet Office, UK Government. https://www.gov.uk/government/publications/records-review-by-sir-alex-allan

[2] Sir Alex Allan. 2015. Government Digital Records and Archives Review. Cabinet Office, UK Government. https://www.gov.uk/government/publications/government-digital-records-and-archives-review-by-sir-alex-allan

[3] Giacomo Berardi, Andrea Esuli, Craig Macdonald, Iadh Ounis, and Fabrizio Sebastiani. 2015. Semi-Automated Text Classification for Sensitivity Identification. In *Proc. CIKM*.

[4] Giacomo Berardi, Andrea Esuli, and Fabrizio Sebastiani. 2012. A Utility-theoretic Ranking Method for Semi-automated Text Classification. In *Proc. SIGIR*.

[5] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *Proc. ICPR*.

[6] c. 36. 2000. Freedom of Information Act 2000. https://www.legislation.gov.uk/ukpga/2000/36/contents. (2000).

[7] c. 36. 2000. Freedom of Information Act 2000, Part 2, Exempt Information. https://www.legislation.gov.uk/ukpga/2000/36/part/II. (2000).

[8] Gordon V Cormack and Maura R Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[9] Denis Cousineau. 2005. Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in quantitative methods for psychology* 1, 1 (2005), 42–45.

[10] Geoff Cumming and Robert Maillardet. 2006. Confidence intervals and replication: where will the next mean fall? *Psychological methods* 11, 3 (2006), 217.

[11] Tadele T Damessie, Falk Scholer, and J Shane Culpepper. 2016. The influence of topic difficulty, relevance level, and document ordering on relevance judging. In *Proc. ADCS*.

[12] DARPA. 2010. New technologies to support declassification. (2010). http://fas.org/sgp/news/2010/09/darpa-declass.pdf.

[13] Benjamin Fung, Ke Wang, Rui Chen, and Philip S Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)* 42, 4 (2010), 14.

[14] Maura R Grossman and Gordon V Cormack. 2010. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Rich. JL & Tech.* 17 (2010), 1.

[15] Geoffrey R Loftus and Michael EJ Masson. 1994. Using confidence intervals in within-subject designs. *Psychonomic bulletin & review* 1, 4 (1994), 476–490.

[16] Graham McDonald, Nicolás García-Pedrajas, Craig Macdonald, and Iadh Ounis. 2017. A Study of SVM Kernel Functions for Sensitivity Classification Ensembles with POS Sequences. In *Proc. SIGIR*.

[17] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2017. Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings. In *Proc. ECIR*.

[18] Graham McDonald, Craig Macdonald, Iadh Ounis, and Timothy Gollins. 2014. Towards a Classifier for Digital Sensitivity Review. In *Proc. ECIR*.

[19] Linda McLean, Maureen Tingley, Robert N Scott, and Jeremy Rickards. 2001. Computer terminal work and the benefit of microbreaks. *Applied ergonomics* 32, 3 (2001), 225237.

[20] Richard D Morey. 2008. Confidence intervals from normalized data: A correction to Cousineau (2005). *reason* 4, 2 (2008), 61–64.

[21] Michael S Moss and Tim J Gollins. 2017. Our Digital Legacy: an Archival Perspective. *Journal of Contemporary Archival Studies* 4, 2 (2017), 3.

[22] Douglas W Oard, William Webber, et al. 2013. Information retrieval for e-discovery. *Foundations and Trends® in Information Retrieval* 7, 2–3 (2013), 99–237.

[23] Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* 34, 1 (2002), 147.

[24] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[25] The National Archives. 2016. The Application of Technology-Assisted Review to Born-Digital Records Transfer, Inquiries and Beyond. (2016). http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf

[26] E Dale Thompson and Michelle L Kaarst-Brown. 2005. Sensitive information: A review and research agenda. *Journal of the Association for Information Science and Technology* 56, 3 (2005), 245–257.

[27] Alistair G Tough. 2018. The Scope and Appetite for Technology-Assisted Sensitivity Reviewing of Born-Digital Records in a Resource Poor Environment. *Handbook of Research on Heritage Management and Preservation* (2018).