How the Accuracy and Confidence of Sensitivity Classification Affects Digital Sensitivity Review

GRAHAM MCDONALD, CRAIG MACDONALD, and IADH OUNIS,
University of Glasgow, UK

Government documents must be manually reviewed to identify any *sensitive* information, e.g., confidential information, before being publicly archived. However, human-only sensitivity review is not practical for born-digital documents due to, for example, the volume of documents that are to be reviewed. In this work, we conduct a user study to evaluate the effectiveness of sensitivity classification for *assisting* human sensitivity reviewers. We evaluate how the accuracy and confidence levels of sensitivity classification affects the number of documents that are correctly judged as being sensitive (reviewer accuracy) and the time that it takes to sensitivity review a document (reviewing speed). In our within-subject study, the participants review government documents to identify real sensitivities while being assisted by three sensitivity classification *treatments*, namely *None* (no classification predictions), *Medium* (sensitivity predictions from a *simulated* classifier with a balanced accuracy (BAC) of 0.7), and *Perfect* (sensitivity predictions from a classifier with an accuracy of 1.0). Our results show that sensitivity classification leads to significant improvements (ANOVA, $p < 0.05$) in reviewer accuracy in terms of BAC (+37.9% *Medium*, +60.0% *Perfect*) and also in terms of F_2 (+40.8% *Medium*, +44.9% *Perfect*). Moreover, we show that assisting reviewers with sensitivity classification predictions leads to significantly increased (ANOVA, $p < 0.05$) mean reviewing speeds (+72.2% *Medium*, +61.6% *Perfect*). We find that reviewers do not agree with the classifier significantly more as the classifier's confidence increases. However, reviewing speed is significantly increased when the reviewers agree with the classifier (ANOVA, $p < 0.05$). Our in-depth analysis shows that when the reviewers are not assisted with sensitivity predictions, mean reviewing speeds are 40.5% slower for sensitive judgements compared to not-sensitive judgements. However, when the reviewers *are* assisted with sensitivity predictions, the difference in reviewing speeds between sensitive and not-sensitive judgements is reduced by ~10%, from 40.5% to 30.8%. We also find that, for sensitive judgements, sensitivity classification predictions significantly increase mean reviewing speeds by 37.7% when the reviewers agree with the classifier's predictions (t-test, $p < 0.05$). Overall, our findings demonstrate that sensitivity classification is a viable technology for assisting human reviewers with the sensitivity review of digital documents.

This manuscript comprehensively extends the ACM CHIIR short paper titled "How Sensitivity Classification Effectiveness Impacts Reviewers in Technology-Assisted Sensitivity Review" by the same authors [McDonald et al. 2019]. The manuscript provides extensive additional contributions investigating how the classifier's confidence in its predictions impacts how quickly human reviewers sensitivity review documents and an in-depth analysis of the user study that investigates: (1) If there is an additional reviewing time overhead when judging sensitive documents compared to not-sensitive documents, (2) if automatic classification can reduce this reviewing time overhead, and (3) the impact on reviewing times from sensitive and not-sensitive predictions when the reviewer either agrees or disagrees with the classifier.

Authors' address: G. McDonald, C. Macdonald, and I. Ounis, School of Computing Science, University of Glasgow, Sir Alwyn Williams Building, Glasgow, Scotland, UK, G12 8RZ; emails: {Graham.McDonald, Craig.Macdonald, Iadh.Ounis}@Glasgow.ac.uk.

1046-8188/2020/10-ART4 $15.00
https://doi.org/10.1145/3417334

CCS Concepts: • **Information systems → Decision support systems**; *Users and interactive retrieval*;

Additional Key Words and Phrases: Technology-assisted review, document classification, user study, freedom of information

ACM Reference format:
Graham McDonald, Craig Macdonald, and Iadh Ounis. 2020. How the Accuracy and Confidence of Sensitivity Classification Affects Digital Sensitivity Review. *ACM Trans. Inf. Syst.* 39, 1, Article 4 (October 2020), 34 pages.
https://doi.org/10.1145/3417334

1 INTRODUCTION

More than 100 countries around the world implement *freedom of information* laws, or *acts*, that provide the public with access to information that has been produced by the government. For example, in the United Kingdom (UK), freedom of information is enacted through the Freedom of Information Act 2000 [c 36] (FOIA). Moreover, the UK Public Records Act 1958 [c 51] legislates that all documents that are of historical value (e.g., minutes from important meetings) must be transferred to a public archive, for example, The National Archives (TNA),[1] within a specified time period from the document's creation.

Many government documents contain *sensitive* information that is exempt from public release through FOIA, such as personal information, issues of national security, or information that would be likely to damage the international relations of the country. Therefore, all government documents that are to be considered for public release must first be manually *sensitivity reviewed* to identify and protect any sensitive information. However, the volume of *born-digital* documents, such as e-mails, that are to be reviewed is much greater than that of paper documents. Therefore, the need for new technologies to *assist* with digital sensitivity review has long been widely recognised [Allan 2015; Lain 2013; The National Archives 2016]. Moreover, the need for assistive technologies to help to identify and prevent *sensitive information leakage* from information retrieval (IR) systems has recently been identified as an increasingly important and emerging area of research to enable a greater level of access to collections that potentially contain sensitive information [Roegiest et al. 2019].

Automatic sensitivity classification [McDonald et al. 2014] can potentially play an important role in assisting sensitivity reviewers. Within other technology-assisted review tasks, such as e-discovery [Oard et al. 2010], document classification is typically deployed to identify relevant documents and reduce the number of documents that have to be reviewed. However, *all* government documents will continue to be manually reviewed for the foreseeable future [The National Archives 2016]. This raises an interesting question: "How can sensitivity classification be deployed to *assist* human sensitivity reviewers?"

A central role of sensitivity classification will likely be to provide the reviewers with useful information, e.g., classification predictions, to reduce the amount of time that is required to accurately sensitivity review documents. Accordingly, it is important to know if sensitivity classification predictions are beneficial for reviewers and if they do actually reduce reviewing times.

This work builds on the work of McDonald et al. [2019] that reported initial findings, from the study reported in this manuscript, on how the accuracy of sensitivity classification impacts human reviewers. In this work, we conduct a within-subject user study under laboratory conditions (using real government documents with real sensitivities) to investigate how two properties of sensitivity classification, namely the accuracy of the classifier (building on McDonald et al. [2019]) and, ad-

[1] http://www.nationalarchives.gov.uk/.

ditionally, the classifier's confidence in its predictions, affect two aspects of the sensitivity review, namely 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). Moreover, in this manuscript, we investigate whether the classifier's confidence impacts the reviewer's agreement with the classifier and how this in turn affects reviewing times. Furthermore, this manuscript provides an additional in-depth analysis of the user study to investigate the following: (1) If there is an additional reviewing time overhead when judging sensitive documents compared to not-sensitive documents, (2) if automatic classification can reduce this reviewing time overhead, and (3) the impact on reviewing times from sensitive and not-sensitive predictions when the reviewer either agrees or disagrees with the classifier.

A repeated measures ANOVA finds that, in our study, automatic sensitivity classification with an effectiveness in line with sensitivity classifiers from the literature (e.g., McDonald et al. [2017b]) results in a statistically significant difference in the reviewers' accuracy, in terms of balanced accuracy (BAC), over the different sensitivity classification treatments, *None*, *Medium*, and *Perfect*, $F(1.176, 12.33) = 24.892, p < 0.0005, \eta^2 = 0.781$. Providing the reviewers with sensitivity classification predictions results in a 37.9% increase in reviewer accuracy, in terms of BAC, for the *Medium* classification treatment level, and a 60.0% increase in terms of BAC for the *Perfect* treatment levelcompared to the *None* classification treatment. Moreover, we find that assisting reviewers with sensitivity classification predictions results in a statistically significant difference in reviewing speeds, as measured by normalised processing speed (NPS) [Damessie et al. 2016] in words per minute (wpm), over the classification treatment levels (repeated measures ANOVA, $F(1.131, 7.915) = 78.89, p < 0.0005, \eta^2 = 0.919$). In our study, reviewing speeds increase by 72.2% in the *Medium* treatment level, and 61.6% in the *Perfect* treatment levelcompared to the *None* treatment level.

Our analysis shows that there is an additional reviewing time overhead when our study participants review sensitive documents and that providing the reviewers with sensitivity predictions can reduce this additional overhead by ~10%. Furthermore, we evaluate the impact on reviewing speeds from *sensitive* and *not-sensitive* predictions when the reviewer either agrees or disagrees with the prediction. We find that for sensitive judgements, sensitivity classification predictions increase mean reviewing speeds by a statistically significant 37.7% when the reviewers agree with the classifier's predictions (paired samples t-test ($t(7) = 2.564, p = 0.037, d = 0.91$)). However, in our study, mean reviewing speeds for sensitive judgements decrease by 3.4% when reviewers disagree with the classifier. This decrease is not statistically significant (paired samples t-test, $t(7) = 0.723, p = 0.493, d = 2.83$).

The remainder of this article is structured as follows: Related work is presented in Section 2. In Section 3, we provide an overview of the digital sensitivity review task and assistive technologies for sensitivity review before providing details of the evaluation framework that we use in our user study. In Section 4, we present the hypotheses that we evaluate before providing details of our user study experimental setup in Section 5. We present the results of our user study in Section 6 before presenting our reviewing time analysis in Section 7. Finally, we present our conclusions in Section 8.

2 RELATED WORK

We first present work relating to sensitivity classification for identifying FOIA sensitivities in government documents in Section 2.1 before discussing work relating to technology-assisted review and how classification technologies can be deployed to assist with the sensitivity review of digital government documents in Section 2.2.

2.1 Automatic Sensitivity Classification

The need for automatic tools to identify sensitive information has been recognised by governments for a number of years [Allan 2014; DARPA 2010; Thompson and Kaarst-Brown 2005; Tough 2018]. However, although there has been a substantial amount of research looking at identifying and masking personal information, e.g., Fung et al. [2010], Sweeney [2002], and Sánchez and Batet [2016], it is only relatively recently that research has advanced in the field of sensitivity classification that we address in this work, i.e., toward automatically identifying information that is exempt from public release through Freedom of Information (FOI)[2] laws.

McDonald et al. [2014] was the first work to present a prototype system for automatically classifying FOI exemptions. In that work, the authors investigated classifying two FOI exemptions and showed that text classification [Sebastiani 2002] could be a viable approach for developing sensitivity classification. Additionally, McDonald et al. [2014] extended text classification with additional features based on the frequency of subjective sentences in a document and the expected risk associated to mentions of specific countries. The approach of McDonald et al. [2014] achieved a balanced accuracy[3] (BAC) [Brodersen et al. 2010] of 0.73.

Souza et al. [2016] investigated classifying the original security categorisation of U.S. State Department cables, i.e., *unclassified* (U), *limited official use* (L), *confidential* (C), and *secret* (S). In that work, as features, Souza et al. used metadata, such as who sent/received the document, what the document was about, and keywords that the author used to categorise the document, along with the document's text to evaluate 12 different classification models. The authors selected the best performing seven models to deploy an ensemble classifier to predict security categorisations. Souza et al. [2016] evaluated their approach through a set of binary classifications (U vs. $L \cup C \cup S$, $U \cup L$ vs. $C \cup S$, $U \cup L \cup C$ vs. S, and U vs. $C \cup S$) and found that their approach worked best when classifying U vs. $C \cup S$, achieving 0.92 F_1.

McDonald et al. [2017b] investigated the effectiveness of semantic, syntactic, and textual feature sets for classifying FOI exemptions. They derived semantic document representations from word embeddings [Mikolov et al. 2013] to extend text classification. McDonald et al. evaluated their approach against text classification extended with additional parts-of-speech or textual *n*-grams, and combinations of all the feature sets, and found that extending text classification with additional semantic and textual features was most effective (0.71 BAC, 0.54 F_2).

Recently, there has been some interest in integrating sensitivity classification into retrieval models to protect sensitive information that has been indexed by a search engine. Sayed and Oard [2019] proposed a technique for integrating sensitivity classification into a learning to rank approach to define a loss function that penalises information that should not be displayed to the user. However, the approach of Sayed and Oard [2019] did not focus on evaluating sensitivity classification.

The sensitivity classification work presented thus far, i.e., McDonald et al. [2017b, 2014], Souza et al. [2016], and Sayed and Oard [2019], have addressed the development of classifiers or ranking strategies to identify or protect sensitive information. However, as previously mentioned in Section 1, all digital government documents will continue to be manually reviewed until an acceptable level of trust in sensitivity classification has developed, and, therefore, for sensitivity classification to be useful for sensitivity review it must be deployed within a technology-assisted review framework.

[2] https://www.legislation.gov.uk/ukpga/2000/36/contents,https://www.foia.gov/.

[3] Balanced accuracy (BAC) and F_2 are the most often reported metrics when evaluating the effectiveness of sensitivity classification. We provide further details of the reasons for this choice in Section 5.5.

2.2 Technology-assisted Review

Technology-assisted review is notably associated with e-discovery [Oard and Webber 2013], i.e., the task of finding all of the electronic (i.e., digital) documents that are relevant[4] to a *production request* within a legal context. In e-discovery, technology-assisted review has been shown to be more effective and more efficient than human review [Grossman and Cormack 2010]. However, differently from sensitivity review, in the context of e-discovery human reviewers typically only review the documents that have been predicted to be most likely to be relevant.

E-discovery requires an additional review to identify if a document is covered by *attorney-client privilege* and, therefore, should not be released even if it is relevant. The review for privilege is more closely aligned to reviewing for sensitivity, where the task is also to find documents that should not be released. However, there has been little research into privilege classification [Cormack et al. 2010]. Therefore, the technology-assisted review model of e-discovery cannot be directly ported to digital sensitivity review and there is a need to evaluate how technologies, such as automatic sensitivity classification, impacts the reviewers in technology-assisted sensitivity review.

Berardi et al. [2015] was the first work to directly evaluate how sensitivity classification could be deployed to assist sensitivity review. In that work, the authors evaluated the effectiveness of *utility-theoretic* semi-automated text classification [Berardi et al. 2012] for improving the cost-effectiveness of sensitivity review. Berardi et al. [2015] built on the work of McDonald et al. [2014] and found that their approach resulted in substantial improvements in classification effectiveness (+14% F_2). However, the authors did not investigate how sensitivity classification impacts sensitivity reviewers.

In this work, we also investigate deploying sensitivity classification within the context of technology-assisted sensitivity review. However, differently from the work of Berardi et al. [2015], in this work, we conduct a controlled user study under laboratory conditions to evaluate how sensitivity classification affects the reviewers' judgements and reviewing speed.

3 DIGITAL SENSITIVITY REVIEW

There is an assumption of *openness* in freedom-of-information (FOI) laws, i.e., FOI assumes that all of the information within documents that are produced by public bodies, such as the government, will be made available to the public. Digital Sensitivity Review is the process of reviewing a collection of digital documents, that are to be opened to the public, to identify any sensitive information, so that the sensitive information can be redacted, or *closed*, and the documents can be released to the public. Figure 1 illustrates the digital sensitivity review process. The input to the process is a collection of digital documents, D, that are to be transferred to a public archive. A sensitivity reviewer reads each document, $d_i \in D$, in turn and records a *sensitivity judgement*, j_i, stating if the document is *sensitive* or *not-sensitive*. For sensitive documents, j_i must also include the following: (1) a record of the passages of the document (e.g., sentences or paragraphs) that are sensitive and (2) the type of sensitivity that is present in each sensitive passage. The output of the digital sensitivity review process is the collection of reviewed documents, D_j, and the set of sensitivity judgements, J, where for each document, d_i, there is a corresponding sensitivity judgement, j_i.

3.1 Technology-assisted Sensitivity Review

As mentioned in Section 1, it is generally accepted that *all* government documents that are released to the public will continue to be *manually* sensitivity reviewed for the foreseeable future [The National Archives 2016]. However, there are a number of ways that sensitivity classification can

[4]The relevant documents are usually referred to as being *responsive* in e-discovery.

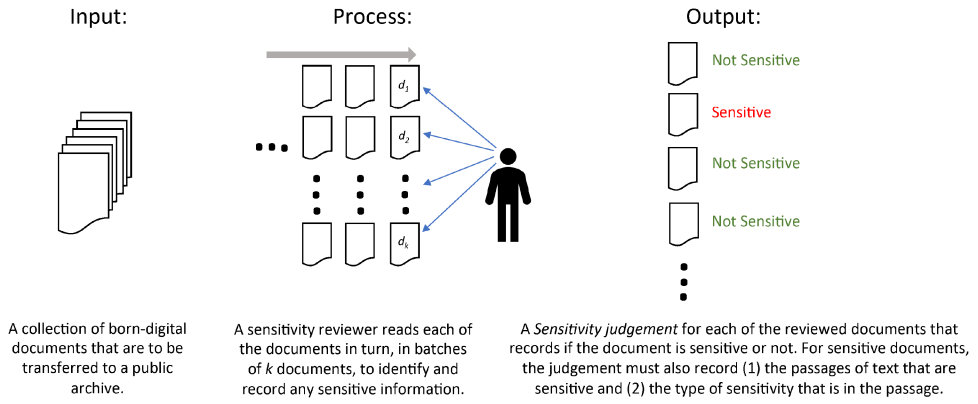

Fig. 1. Digital sensitivity review is an iterative process. The input is the documents that are to be reviewed. A reviewer reviews batches of k documents and records a sensitivity judgement for each document. The output is the reviewed documents and judgements.

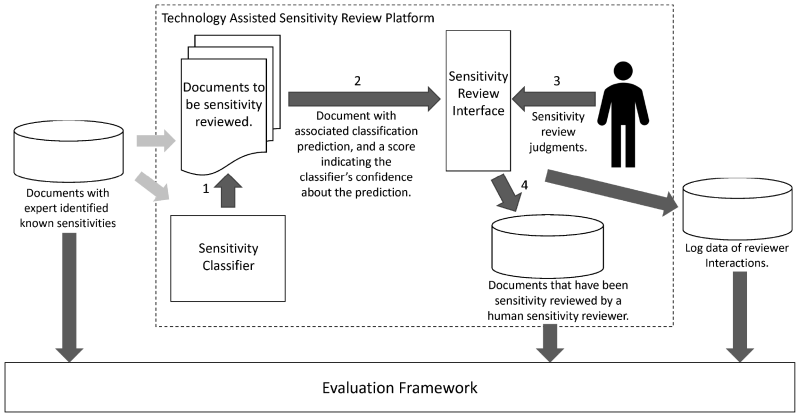

Fig. 2. A platform for technology-assisted sensitivity review (within the dashed box), along with a framework for evaluating the benefits of providing sensitivity reviewers with sensitivity classification predictions.

be deployed to assist with the digital sensitivity review process. For example, by prioritising *not-sensitive* documents to increase the number of non-sensitive documents that can be released to the public with limited reviewing resources [McDonald et al. 2018]. In this work, we focus on evaluating whether providing the reviewers with sensitivity classification predictions can reduce the time that it takes for a reviewer to review a collection of documents, while maintaining (or increasing) the reviewing accuracy.

Figure 2 presents, within a dashed box, the technology-assisted sensitivity review platform that we use in our study to assist human reviewers by providing them with sensitivity classification predictions. Our reviewing platform supports four sequential actions, labelled 1–4 in Figure 2: (1) For each document, $d_i \in D$, the sensitivity classifier, Φ, makes a prediction, $\hat{y} \in \{\text{sensitive}, \text{not-sensitive}\}$, as to whether d_i is sensitive or not; (2) the collection, D, with associated sensitivity predictions, L, is passed to a reviewing interface; (3) the reviewer then reviews each document in-turn and records a sensitivity judgement for each; and (4) the sensitivity reviewed documents, with the associated human judgements, are then written to persistent storage.

In this work, we are concerned with evaluating the benefits of providing sensitivity reviewers with automatic sensitivity classification predictions. Next, we describe the framework for assessing these benefits through our user study.

3.2 The Evaluation Framework

We deploy the technology-assisted sensitivity review platform with human *study participants* while varying the effectiveness of the sensitivity classifier within the platform. After each participant has completed reviewing their allocated documents, we can then evaluate how the sensitivity classification predictions affect how the reviewer completes the task.

In addition to the technology-assisted sensitivity review platform, Figure 2 also presents the framework that we use to evaluate the effect of sensitivity classification predictions on the reviewer's accuracy and reviewing speeds. For our evaluation, we need a reliable ("gold standard") ground truth detailing the sensitivities within the documents that the study participants review. We use the ground truth of a test collection of government documents with known sensitivities. The ground truth was generated within the scope of the study, but before the human user study experiment, by having expert sensitivity reviewers from UK government departments review the collection and record the actual sensitivities. The participants review documents from this collection and we store the reviewed documents and associated sensitivity judgements. Moreover, we log the participants' interactions with the reviewing platform by storing a timestamp whenever the reviewer performs an action such as viewing a document or saving a sensitivity judgement. We provide full details of our experimental method in Section 5.

4 HYPOTHESES

We expect that if a sensitivity classifier correctly predicts that a document is either *sensitive* or *not sensitive* (and the reviewer agrees with its prediction) then the classifier's prediction will help the reviewer to make their sensitivity judgement more quickly. As the classifier's accuracy increases and it makes more correct predictions, the reviewers will in-turn be assisted more frequently to make quicker reviewing decisions and the overall (or average) speed of the review will increase. Therefore, we state our first hypothesis as:

H1: As the effectiveness of the classifier increases, the classifier will be of more benefit to reviewers and, therefore, reviewers will:

(a) Make more correct and less incorrect judgements.
(b) Make quicker reviewing decisions (i.e., review documents faster) on average.

Our second hypothesis is concerned with how the level of confidence that a classifier has in its prediction for a particular document affects the reviewers. We postulate that the level of confidence that a classifier has in its prediction will have a direct influence on how much trust the reviewer has in the prediction. The second hypothesis that we investigate in this study is stated as:

H2: Reviewers will rely on the classifier more when the classifier is more confident about a prediction, and therefore when the classifier is confident reviewers will:

(a) Agree with the classifier more as the classifier's confidence increases.
(b) Make quicker reviewing decisions when they agree with the classifier.

In our study, sensitivity classification predictions help to assist human reviewers to sensitivity review documents more quickly while maintaining high levels of reviewing accuracy. Therefore, we argue that this work demonstrates that sensitivity classification is a viable technology to

effectively provide sensitivity reviewers with valuable information about the sensitivities within a collection to assist the sensitivity review process.

5 EXPERIMENTAL METHOD

To test the two hypotheses stated in Section 4, we conducted a controlled user study under laboratory conditions, using the reviewing platform and evaluation framework discussed in Section 3. The study participants, i.e., *reviewers*, were asked to review government documents to identify documents that contained sensitive information relating to either of two FOIA exemptions,[5] namely: Section 27 international relations and Section 40 personal information. In this section, we, first, present details of the test collection and expert ground truth in Section 5.1 before providing details about the reviewing interface and system logging in Section 5.2. We discuss the study participants and instructions in Section 5.3, before detailing our experimental design in Section 5.4. Finally, we present the evaluation and metrics that we use in Section 5.5.

5.1 Test Collection and Expert Judgements

The documents used in this study are sampled from a larger collection of 4,000 government documents. As previously mentioned in Section 3.2, the 4,000 documents in the collection were sensitivity reviewed by expert sensitivity reviewers from UK government departments prior to the start of the user study experiment to generate a set of ground truth judgements. In total, 24 expert reviewers volunteered to sensitivity review the collection of 4,000 documents as a component of the efforts provided by their government departments in support of our project. The experts performed the reviewing task in their own time and at their own pace and were not paid or compensated for providing their reviews.

Each of the documents in this study were sensitivity reviewed by a single expert reviewer. In a previous study, McDonald et al. [2014] reported finding a moderate level of inter-assessor agreement between expert sensitivity reviewers on a small collection of documents. In that work, the authors reported a Cohen's κ of 0.5525 for 150 double-judged documents and a Fleiss' κ of 0.4414 on 50 documents that were assessed by four expert reviewers. We note that, similarly to relevance, sensitivity is to some degree inherently subjective and it would have been desirable to have the documents in this study assessed by multiple expert reviewers; to construct the ground truth from majority vote judgements and reduce the assumption of infallibility that is inherent in judgements from a single-reviewer [Cormack and Grossman 2017]. However, this was not feasible with the expert reviewing resources that were available to us. It is also worth noting that in the current practices of many government departments, documents are routinely sensitivity reviewed by a single expert reviewer and there is an implicit assumption of infallibility. In this work, we use the expert reviewers' ground truth to compare the relative effects of sensitivity classification treatments, e.g., no classifier predictions vs. medium effectiveness classifier predictions vs. perfect classifier predictions. In a study on how variations in relevance judgements affect the robustness of findings with regard to retrieval system effectiveness, Voorhees [2000] showed that the relative performances of different retrieval systems remain stable despite substantial differences in relevance judgements. In other words, if we were to reproduce this study using sensitivity judgements from a different expert sensitivity reviewer, then we would expect to observe a strong correlation in the relative differences between the levels of classification treatments evaluated in this study and in the reproduction study.

The documents are born-digital written internal government communications (as opposed to transcribed verbal communications or digitised paper-based communications). Many of the

[5] https://www.legislation.gov.uk/ukpga/2000/36/part/II.

Fig. 3. Reviewing Interface Information Panel: The panel displays the classification prediction (Sensitive or Not-Sensitive), and the classifier's prediction confidence score. The panel also enables participants to record their sensitivity judgements and provide comments.

documents include discussions of recent (at the time of writing) events or conversations. The use of the collection for the study was facilitated through a non-disclosure agreement. We provide statistics about the documents that are used in the study in Section 5.4. The expert reviewers used the same judging interface to review the documents as the study participates used (we provide more details about the interface in Section 5.2). However, the expert reviewers were not provided with sensitivity classification predictions.

5.2 Reviewing Interface and Logging

Reviewers were provided with a (web-based) interface to navigate the collection and record document sensitivities. The interface had an information panel at the top of the screen, as shown in Figure 3, that displayed the current document's classification prediction (Sensitive or Not-Sensitive) and a prediction confidence score. The document to be reviewed was displayed below the panel in Figure 3. Therefore, the reviewers were presented with the classification prediction and confidence score before they reviewed the document. As we stated in the previous section, the reviewing interface used in the study was identical to the interface used by the expert reviewers to generate the ground truth, except that the experts were not provided with classification predictions.

Figure 3 also shows how reviewers recorded their judgements as to whether a document contained sensitive information. First, participants recorded a sensitivity judgement by selecting one of the four radio button options at the left of the panel. In sensitivity review, any identified sensitivities must be recorded in the sensitivity judgement. To reflect this, the participants were asked to provide a short explanatory comment about their decision in the text box at the centre of the panel. In addition to providing this comment, for documents that were judged to be sensitive, the participants were asked to highlight all of the sensitive text within the document. This is akin to the real sensitivity reviewing task of redacting sensitive content before release. A simple mouse-click and drag functionality facilitated the highlighting of sensitive text. Importantly, having the participants highlight the sensitivities helped to ensure that they did not solely rely on the classification predictions to make their decisions. As part of the training session at the beginning of the experiment, the participants were made aware that they could not expect the classifier to always make a correct prediction and that sensitive information can be a small portion of the text that occurs in any part of a document. Therefore, the participants were advised to read the entire document carefully to ensure that they did not miss any sensitive information. Moreover, if a document was predicted to be sensitive by the classifier, the participants had to *find* the sensitive information before they could agree with the classifier's prediction. We note, however, that

we could not enforce this constraint and it is possible that, having identified only some of the sensitive information, the participants could have chosen to move onto the next document.

In addition to storing the participants' sensitivity judgements, the framework also logged a timestamped record of when a participant loaded a document, saved a judgement, paused or restarted the system. The participants were provided with a button to pause the logging functionality when they wanted to have a comfort break or ask a question to the experimenter. This helped to ensure that we recorded accurate timings of when participants were focused on the reviewing task. The participants could also indicate if a judgement was particularly hard to make.

5.3 Participants, Incentives, and Instructions

We recruited eight participants for the study. To ensure that the participants had the requisite ability to understand the nuances of the FOIA personal information and international relations exemptions, we limited participants to those who had a background in politics or international relations and who were familiar with the FOIA. Additionally, each of the recruited participants had been speaking 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. A detailed sensitivity review training manual was generated with the assistance of the expert sensitivity reviewers to ensure that the expert reviewers and user study participants performed the same reviewing task. The training manual provided definitions of *international relations* and *personal information* sensitivities, along with descriptions of the sensitivities. To provide the participants with a deeper insight into the sensitivities, the sensitivity descriptions were broken down into several categories of information that are likely to be sensitive.[6]

The participants were provided with a copy of the training manual at the beginning of the training session, along with a presentation of the information in the manual. The presentation also contained additional illustrative examples about some of the considerations that need to be taken into account when making a judgement about sensitivity. For example, when considering if it is appropriate for a document to contain the salary details of a named individual, it can be useful to consider the individual's role within the organisation paying the salary. If the named individual is the director of the organisation, then it is more likely that it is appropriate to publish such details than it would be if the individual is a normal employee or a contractor (since, generally, there is an expectation that directors' salaries may be in the public domain, while it is often the case that a general employee's salary is considered to be personal, or private, to the individual). The 1 hour training session also included examples of sensitive and not-sensitive documents and time to answer any questions that participants had. After the training session, participants were given time to review a batch of eight practice documents (the lengths of the practice documents ranged from 104 words to 989 words), and to discuss their reviewing decisions with the study coordinator, before the study began. This practice session also enabled the study coordinator to observe and monitor the participants' performance to make a qualitative judgement evaluating the participants' comprehension of, and ability to do, the task. Through this process, one potential participant was evaluated as not having a sufficiently clear understanding of the sensitivity reviewing task and was not selected to take part in the user study. Throughout the study, the study coordinator was in the same room as the participants but did not interact with a participant unless they paused their system logging to ask a specific question. The study coordinator also

[6]The details of the reviewing instructions and the details of the information that the participants were asked to find are protected under the non-disclosure agreement, which the participants had to agree to before taking part in the study.

Table 1. The Distributions of True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) Sensitivity Classification Predictions for Documents in Batches Representing Different Classification Effectiveness Treatments, *None*, *Medium,* and *Perfect*, along with the Classifier's Resulting Balanced Accuracy (BAC) for the Treatment

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

periodically checked the participants' judgements remotely, using the online interface, to ensure that the participants were engaged in the task and making credible judgements.

Reviewing for sensitivity requires a considerable amount of effort from the study participants. In total, the participants took between 15 and 19 hours to complete the study (including training sessions), split over two or three separate sessions in a 2-week period. There was a 30-minute refresher training session on the task and sensitivities at the beginning of the second session. To control for the possible effects of fatigue and to ensure the well-being of the participants, in line with the findings of McLean et al. [2001], the participants were advised to take regular and frequent short breaks; as previously stated in Section 5.2, the reviewing interface was set up to not include time spent on breaks as part of the reviewing times. The participants were remunerated £7.50 per hour for taking part in the study.

5.4 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 three batches of 20 documents and, for each document, record a sensitivity judgement as to whether the document was "*not-sensitive*" or contained "*Section 27*" (international relations), "*Section 40*" (personal information), or "*Both: Section 27 & Section 40*" sensitive information.

Using the expert sensitivity reviewers ground truth as gold standard judgements, we sampled documents from our collection to fit the distributions of sensitive and not-sensitive documents presented in Table 1. As can be seen from the table, each batch of 20 documents had an associated *treatment level* of sensitivity classification predictions with an overall level of classification accuracy. Similarly to Turpin and Scholer [2006], who simulated IR systems to evaluate the impacts of different Average Precision performances, we *simulate* classifiers to achieve different effectiveness levels, either *None* (no classification predictions were provided), *Medium* (the accuracy of the simulated classifier was 0.7 BAC) or *Perfect* (the simulated classifier's predictions agreed perfectly with the expert gold standard and, therefore, had an accuracy of 1.0 BAC). Table 2 provides an overview of the maximum, minimum and mean lengths of the documents (i.e., the number of words in the documents) for each of the classification effectiveness treatments. The table also presents the standard deviation of the documents' lengths, denoted as Std.

Each batch of documents contained 5 *sensitive* documents and 15 *not-sensitive* documents, resulting in 25% of documents in each batch containing sensitive information. This is slightly higher than the percentage of sensitive documents in the collection that the documents were sampled from, which was 16%. The study participants were informed that we would expect sensitivity to be the minority class, but they were not made aware of the distributions of sensitive documents in the study. We note that it would have been desirable to have had more classification effectiveness treatment levels in the study design, with additional treatment levels between *Medium* and *Perfect* for a finer-grained evaluation of the impact of the classifier's effectiveness on the participants'

Table 2. Maximum, Minimum, Mean and Standard Deviation (Std.) of the Lengths of the Documents (i.e., the Number of Words in the Documents) for Each of the Classification Effectiveness Treatments: *None*, *Medium,* and *Perfect*

	Max	Min	Mean	Std.
None	1136	91	427.35	261.92
Medium	1337	101	517.15	353.78
Perfect	1082	101	514.60	230.84

Table 3. Distributions of *Low*, *Medium,* and *High* Prediction Confidence Scores for the *Medium* and *Perfect* Classification Effectiveness Treatments

Classification	Low	Medium	High
Medium	7	6	7
Perfect	7	6	7

accuracy and agreement. However, we developed the experimental design containing three classification accuracy levels and 25% sensitivity distribution as a reasonable balance between (1) being able to observe levels of classification accuracy that are less than, close to, and better than that of the sensitivity classifiers from the literature (e.g., from McDonald et al. [2017b]) and (2) so that we could reasonably assume that participants would be able to complete the task within 12 hours (including training times).

For batches with *Medium* classification effectiveness, e.g., 0.7 BAC, 3 documents had associated True-Positive (TP) predictions, 2 documents had False-Negative (FN) predictions, 3 documents had False-Positive (FP) predictions, and 12 documents had associated True-Negative (TN) predictions, where *sensitive* is the positive class. To generate the classifier's errors, we randomly added noise to the expert ground truth to identify candidate documents to assign FP and FN predictions to. As a sanity check, we manually checked the candidate FP and FN assignments and selected documents that were credible classification errors based on the documents' contents.

In the treatment batches with either *Medium* or *Perfect* classification effectiveness, each prediction had an associated score in the range $(0, 1)$ that represented the level of confidence the classifier had about the prediction. Each assigned confidence score represented either a *Low*, *Medium*, or *High* confidence, where $Low < 0.35 < Medium < 0.7 < High$. Simulated confidence scores were assigned to the classifier's predictions randomly to fit the distributions presented in Table 3. The simulated confidence scores were manually inspected by the study coordinator prior to the beginning of the user study as a check to ensure that the scores were credible based on the document's content.

Each participant reviewed 1**None*, 1**Medium* and 1**Perfect* batches, i.e., 60 documents each (3*batches of 20 documents). The participants were informed that the classification predictions were intended to provide assistance for identifying sensitive documents but they could not expect the classifier to always be correct. The participants were not made aware of the effectiveness levels of the classification treatments (i.e, the accuracies of the simulated classifiers) before they conducted their review.

The reviewing interface used by the study participants was identical for each of the classification treatments, except that when the participants were reviewing documents with classification effectiveness *None*, the information panel presented in Figure 3 said "Classification Prediction :: Off," and there was no text displayed in the Confidence Score section of the panel. The study design enforced that the participants reviewed batches of documents in a prescribed order. The documents within a batch were presented in random order, consistently between reviewers. The participants were advised to proceed linearly through each batch. However, the participants were able to select documents within a batch in any order to re-visit documents or change previously made judgements. This setup reflects how sensitivity review is conducted within government departments, and the system was set up to record the total time that a participant spent reviewing (viewing) a particular document.

To control for potential effects from the order that the participants interacted with each of the treatment levels, we counterbalanced the allocation of batches to the participants, i.e., we permuted the order in which batches were reviewed by different reviewers. Counterbalancing is standard practice in within-subject studies to minimise the potential for carryover effects, i.e., the potential for an effect of one condition, e.g., a treatment level, having an effect on the participants' behaviour in a later condition. There are three main types of potential carryover effects in a within-subject study such as ours: *learning* effects, *context* effects, and *fatigue*. A learning effect is when a participant learns how to perform the task as they spend more time doing it, and this learning improves the participant's performance over time. A context effect is when the context of one condition, e.g., a treatment level, can influence a participant's behaviour in another condition. For example, in our study, a potential context effect would be when a participant is presented with the *Medium* or *Perfect* classification effectiveness treatment before the *None* treatment level the participant could potentially be *primed* by the classifier's predictions to be more confident that they know what sensitivities *look like* in the *None* treatment. A fatigue effect occurs when a participant becomes tired from (or of) performing the task and their performance deteriorates.

We note that, in our study, there are six possible permutations of the Classification Effectiveness treatment levels (*None*, *Medium* and *Perfect*) and eight participants. Therefore, there is one complete counterbalancing of the treatment levels plus two additional participants. This could be viewed as a potential limitation of the study. However, we have taken steps to mitigate the likelihood of potential carryover effects influencing the findings of our study. First, the additional participants were not presented with the *None* treatment level before the other two levels, *Medium* and *Perfect*. Indeed, the additional participants were presented with the *None* treatment last with the *Medium* and *Perfect* treatments permuted. This design ensures that any potential learning or context effects could not be interpreted as an effect of classification effectiveness. In other words, any learning or context would result in the participants performing better in the *None* treatment (this is akin to having a stronger performing baseline system). In practice, from the observations that we will present in Section 6, this does not appear to have happened. Second, to minimise the potential for fatigue effects, as previously mentioned in Section 5.3, the study participants took regular and frequent breaks throughout the study. Moreover, the participants could take as many breaks as they wished and as frequently as they wanted.

Counterbalancing is typically a best-effort endeavour. If carryover affects, such as priming, do exist, then counterbalancing does not remove them but rather entangles the order effects with the treatment effects [Winer 1962], and a completely successful counterbalancing depends on the ability to rule out the existence of such effects [Campbell and Stanley 2015]. As an alternative, a between-subject experiment design was considered. However, a between-subject design would require many additional participants to achieve a robust study design. As has previously been mentioned, sensitivity review is a complex task, and this study required a substantial commitment

from the study participants (2–3 days over a 2-week period). Moreover, as Birnbaum [1999] argues, the lack of context in between-subject designs is often more of a problem than the potential for context effects in within-subject designs. Therefore, we opted to include the additional participants in a within-subject study, since, we argue, they provide valuable additional observational data. Finally, we note that since the study is a within-subject design, the study is balanced in the sense that each of the treatment levels have the same number of participants and each of the participants were exposed to the same treatment levels.

5.5 Evaluation and Metrics

As previously mentioned in Section 3.2, in this study, we use the expert sensitivity reviewers' judgements as a ground truth when evaluating the performance of the study participants. We evaluate the participants' performance 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 for a reviewer to sensitivity review a document (reviewing speed).

To evaluate the impact that the accuracy of the sensitivity classifier has on reviewer accuracy (compared to the expert ground truth) and reviewing speed, we report the mean reviewer accuracy and reviewing speed (calculated over all reviewers) for each classification treatment, *None*, *Medium*, and *Perfect*. However, when evaluating the effects of the classifier confidence, we report the mean reviewing speed and the reviewer-classifier agreement for the confidence levels, *Low*, *Medium*, and *High*, over the *Medium* and *Perfect* classification batches combined (since the distributions of *Low*, *Medium*, and *High* confidences are the same in the *Medium* and *Perfect* batches, and there are no classification predictions for the *None* batch).

We select BAC and F_2 as our metrics to evaluate reviewer accuracy, since they are particularly suited to evaluating sensitivity classification [McDonald et al. 2017a]. More specifically, we select BAC, since it provides an accuracy score for the performance over both classes when the distributions of classes are heavily skewed. Moreover, 0.5 BAC indicates that a classifier, or a reviewer in our case, is not discriminating between the two classes. For example, a strategy that assigns classification labels based on a fair coin toss would be expected to achieve a BAC score of 0.5. We also report F_2, since it is a recall-oriented metric that accounts for the fact that, in sensitivity review, there are more severe consequences from incorrectly judging/classifying a sensitive document as not-sensitive than there are from incorrectly judging/classifying a not-sensitive document as being sensitive. If a sensitive document is incorrectly classified, and therefore enters into the public domain, then the discovery of the sensitive information could have a detrimental impact for individuals, organisations, or governments that are linked to the information.

When evaluating the participants' reviewing speeds, we use NPS [Damessie et al. 2016] to control for the effects of inter-subject differences in reading speeds and varying document lengths. NPS is calculated as:

$$\frac{|d|}{exp(\log(time) + \mu - \mu_\alpha)}, \tag{1}$$

where $|d|$ is the document length, measured in number of words, 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 6, we plot the participants' performance to show the mean participant score (e.g., in terms of BAC or NPS) and the 95% confidence intervals. We use the Loftus and Masson [1994] method of calculating confidence intervals for within-subject study designs, with the Cousineau [2005] update and the Morey [2008] correction. Using this method (the Cousineau and Morey method), we would expect that in a replication study five of six

participants would be included in this interval [Cumming and Maillardet 2006]. Importantly, this method 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, for the classification effectiveness treatment, we use a one-way repeated measures omnibus ANOVA over the three classification effectiveness treatment levels *None*, *Medium,* and *Perfect*. We use a one-way ANOVA, since there is only one factor to analyse in the *None* treatment level, i.e., there is no classifier confidence factor in the *None* treatment level, since there are no classifier predictions presented to the participants. For the classifier confidence treatment, we perform a two-way repeated measures omnibus ANOVA over the three classification confidence treatment levels (*Low*, *Medium,* and *High*) and the *Medium* and *Perfect* classification treatment levels only (since there are no classification confidence scores for the *None* treatment level). We test that the variances of the differences between all combinations of related groups (treatment levels) are equal using Mauchly's Test of Sphericity and, in tests where sphericity is violated, we report Greenhouse-Geisser [Greenhouse and Geisser 1959] corrected ANOVAs. We report the observed power and Partial Eta Squared (η^2) effect size for our omnibus ANOVAs and follow these up with post hoc tests using paired samples t-tests with the Bonferroni correction for multiple comparisons [Dunn 1961]. We select $p < 0.05$ as our significance threshold.

6 USER STUDY RESULTS

In this section, we present the results of our technology-assisted digital sensitivity review user study. First, in Section 6.1 we investigate hypothesis **H1**, which states that as the effectiveness of the classifier increases, the classifier will be of more benefit to reviewers and, therefore, reviewers will: (a) make more correct and less incorrect judgements and (b) make quicker reviewing decisions (i.e., review documents faster) on average. Second, in Section 6.2, we investigate hypothesis **H2** that reviewers will rely on the classifier more when the classifier is confident about its predictions, and will therefore: (a) agree with the classifier more as the classifier's confidence increases and (b) make quicker reviewing decisions when they agree with the classifier.

6.1 The Impact of Classification Effectiveness on Reviewer Performance

To evaluate the impact of classification effectiveness on the reviewers' performance, we compare the mean reviewer performance, in terms of reviewer accuracy and reviewing speed, for each of the classification effectiveness levels *None*, *Medium,* and *Perfect*.

First, we evaluate whether the effectiveness of the classifier impacts the correctness of the participants' judgements, when compared to the ground truth of the *expert* sensitivity reviewers' judgements (**H1**(a)). Figure 4 presents the mean participant performance in terms of their BAC for each of the levels of classification effectiveness, while Figure 5 presents the analogous participant performance in terms of F_2.

From Figure 4, 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 for medium classification (+37.9%) effectiveness and 0.8 BAC when the classification predictions agree perfectly with the ground truth (+60.0%).

Importantly, 0.5 BAC indicates that without classification predictions, averaged over sensitive and not-sensitive judgements, the participants' judgements in terms of BAC were effectively random. However, there is a moderate level of inter-assessor agreement between the participants in the *None* treatment, Fleiss' $\kappa = 0.4120$. This is in line with the moderate level of inter-assessor agreement between expert sensitivity reviewers, Fleiss' $\kappa = 0.4414$ (50 documents assessed by four

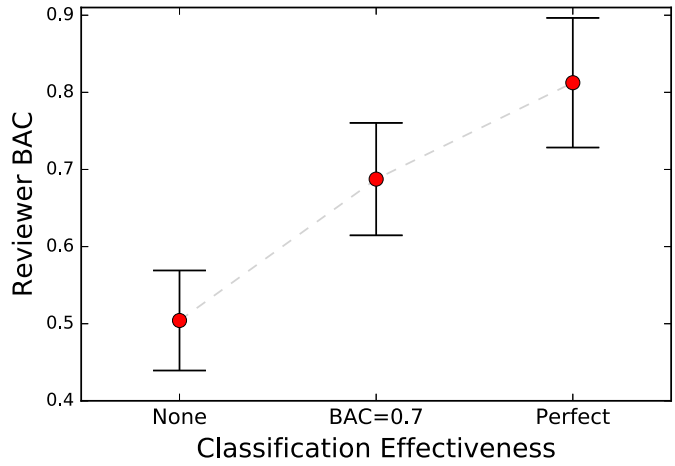

Fig. 4. Mean BAC for the study participants (reviewers) for each of the classification treatments, *None*, *Medium* (BAC=0.7), and *Perfect*, with 95% confidence intervals.

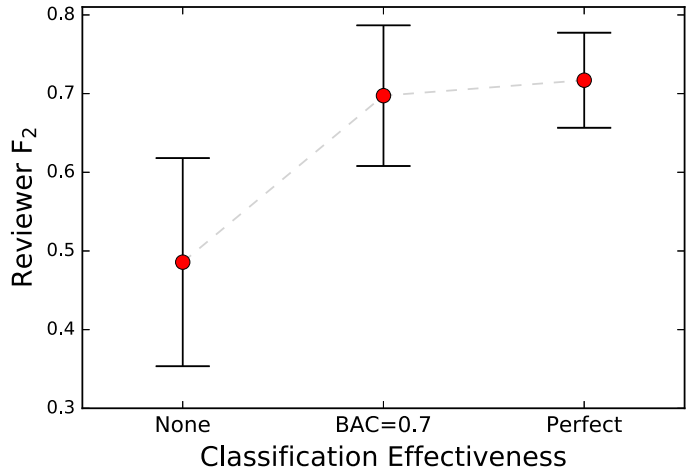

Fig. 5. Mean F_2 for the study participants (reviewers) for each of the classification treatments, *None*, *Medium* (BAC=0.7), and *Perfect*, with 95% confidence intervals.

reviewers) that was previously reported by McDonald et al. [2014].[7] Table 4 presents the True-Positive Rate (TPR) and True-Negative Rate (TNR) of each of the reviewers for each of the classification treatment levels *None*, *Medium*, and *Perfect*. In our study, the participants' mean accuracy scores for sensitive documents, i.e., their mean TPR, in the *None* treatment is 0.3250, while for not-sensitive documents in the *None* treatment level the mean participants' accuracy, i.e., their mean TNR, is 0.6833. In other words, on average, the participants found it easier to judge not-sensitive documents than to judge sensitive documents. This is indicative of the difficulty of the sensitivity reviewing task, which requires a reviewer to make a judgement about the potential consequences of releasing information into the public domain and the level of risk that poses to

[7]McDonald et al. [2014] also reported moderate inter-assessor agreement for expert sensitivity reviewers in terms of Cohen's κ (0.5525) for 150 double-judged documents.

Table 4. The Reviewers' True Positive Rate (TPR) and True Negative Rate (TNR) for Each of the *None*, *Medium,* and *Perfect* Classification Effectiveness Treatments

	None		*Medium*		*Perfect*	
	TPR	TNR	TPR	TNR	TPR	TNR
Reviewer 1	0.4000	0.6666	0.4000	0.8666	0.8000	0.8000
Reviewer 2	0.6000	0.6000	0.6000	0.8000	1.0000	0.7333
Reviewer 3	0.2000	0.6000	0.6000	0.9333	1.0000	0.9333
Reviewer 4	0.4000	0.6666	0.8000	0.7333	0.8000	0.8666
Reviewer 5	0.0000	0.8666	0.4000	0.9333	0.6000	0.7333
Reviewer 6	0.2000	0.8000	0.4000	0.8000	1.0000	0.8000
Reviewer 7	0.4000	0.4666	0.8000	0.6000	0.6000	0.5333
Reviewer 8	0.4000	0.8000	0.6000	0.7333	1.0000	0.8000
Mean	0.3250	0.6833	0.5750	0.7999	0.8500	0.7749

individuals, organisations or governments. Moreover, this underlines why government departments have typically employed expert reviewers for the task, since high domain expertise can often lead to improved efficiencies and task completion rates [Mao et al. 2018]. We note, however, that as the volume of digital documents that need to be sensitivity reviewed increases rapidly over the coming years, it is not expected that government departments will be able to recruit enough (experienced) sensitivity reviewers [The National Archives 2016], and the level of disagreement between the reviewers is likely to increase. Technologies, such as sensitivity classification, are likely to be able to assist in reducing the level of disagreement between reviewers by identifying documents that should be prioritised for review by multiple reviewers to help to come closer to a unified view of what makes information sensitive; indeed, helping the reviewers to make judgements, as we will show.

From Figure 4, we also note that for the *Medium* classification effectiveness treatment, the mean participant performance, in terms of BAC, is almost equivalent to the level of classification effectiveness (participants = 0.69 BAC, classifier = 0.7 BAC). However, although the mean participant performance is the highest when a perfect classifier is deployed (i.e., its predictions are the same as the expert generated ground truth), in this treatment, the reviewers only achieved an accuracy of 0.8 BAC. In fact, none of the participants completely agreed with the classifier when its predictions were the same as the expert reviewers' judgements. This finding is consistent with the work of McDonald et al. [2014], which provides additional evidence that identifying sensitive information is a complex task. A one-way repeated measures omnibus ANOVA with a Greenhouse-Geisser correction shows that there is a statistically significant difference in the mean participant BAC scores between the classification effectiveness treatment levels $F(1.176, 12.33) = 24.892, p < 0.0005$, with effect size $\eta^2 = 0.781$ and observed power of 1.0. A post hoc test shows that there is a statistically significant difference ($p < 0.05$) between the *None* and *Medium* treatment levels (and between *None and Perfect*) in terms of BAC. The difference between the *Medium* and *Perfect* treatment levels is not statistically significant with respect to BAC ($p < 0.05$).

Turning our attention to Figure 5, which presents the participant performance in terms of F_2, we note that the relative mean participant performance increase is much greater between the no classification and *Medium* classification effectiveness than between the *Medium* and *Perfect* classification. A one-way repeated measures omnibus ANOVA with a Greenhouse-Geisser correction shows that there is a statistically significant difference between the mean participant F_2 scores between classification effectiveness treatment levels $F(1.170, 8.192) = 9.46, p = 0.013$, with effect

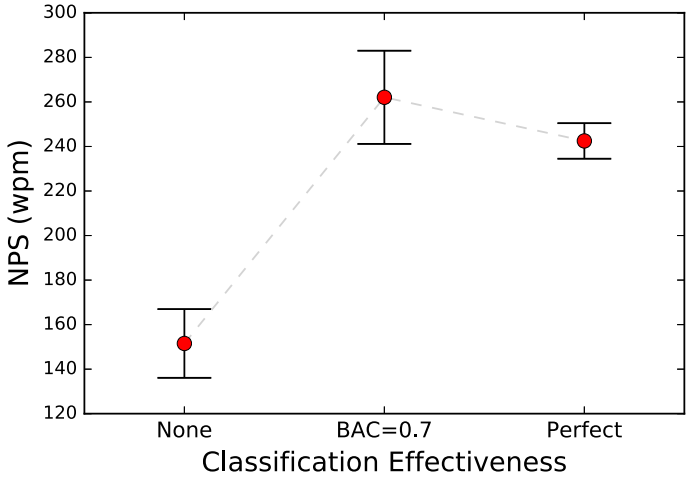

Fig. 6. NPS in wpm for the study participants (reviewers) for each of the classification treatments, *None*, *Medium* (BAC=0.7), and *Perfect*, with 95% confidence intervals.

size $\eta^2 = 0.575$ and observed power of 0.805. A post hoc test shows that there is a statistically significant difference ($p < 0.05$) between the *None* and *Perfect* treatment levels in terms of F_2, but the difference between the *None* and *Medium* treatment levels is not statistically significant with respect to F_2 ($p < 0.05$). The main increase in the participants' performance between the *Medium* and *Perfect* classification effectiveness treatments came as a result of the participants making more True Negative judgements. This is reflected in the observation that the BAC score, which accounts for True Negatives, significantly increased, while for F_2, which does not consider True Negatives, there was no significant increase between the *Medium* and *Perfect* treatment levels in terms of F_2. We postulate that the classifier's True Negative predictions enabled the participants to be more confident about making non-sensitive judgements, when without the classifier's predictions the participants would be more likely to be conservative in their judgements. Indeed, as we will show in Section 7.2 (Table 6), the reviewers agreeing with the classifier's not-sensitive (negative) predictions also leads to the greatest increase in reviewing speeds.

In response to hypothesis **H1**(a), we conclude that improved classification effectiveness does indeed lead to a significantly improved performance of the participant reviewers in terms of BAC and F_2. However, there appears to be diminishing gains in the reviewer performance improvements as the classification effectiveness increases. We leave to future work the identification of a threshold above which the classification effectiveness does not further enhance the reviewers' accuracy.

Turning our attention to hypothesis **H1**(b), which tests if more effective classification predictions will result in the reviewers processing documents faster on average. Figure 6 presents the participants mean NPS [Damessie et al. 2016], in wpm, for each of the levels of classification effectiveness. From Figure 6, we observe that the mean processing speed of reviewers when no classification predictions are provided is 151 wpm. Providing reviewers with classification predictions results in a mean reviewing time increase of 72.2% to 260 wpm, when the classifier predictions have an accuracy of 0.7 BAC. Interestingly, we note from Figure 6 that the mean reviewing speed is slightly less when reviewers are provided with classification predictions that agree perfectly with the ground truth (244 wpm, +61.6% compared to the *None* treatment level) than when the reviewers are assisted by the sensitivity classifier that achieves 0.7 BAC (260 wpm).

A one-way repeated measures omnibus ANOVA with a Greenhouse-Geisser correction shows that there is a statistically significant difference in the mean participant NPS scores between the classification effectiveness treatment levels, $F(1.131, 7.915) = 78.89, p < 0.0005$, with effect size $\eta^2 = 0.919$ and observed power of 1.0. Post hoc tests show that there is a statistically significant difference ($p < 0.0005$) between the *None* and *Medium* and the *None* and *Perfect* treatment levels in terms of NPS. However, the difference (decrease) in mean NPS between the *Medium* and *Perfect* treatment levels is not statistically significant with respect to NPS ($p = 0.238$). Therefore, the significant gains in reviewing speeds from providing the reviewers with classification predictions are sustained over both levels of classification predictions accuracy. In response to **H1**(b), we conclude that providing reviewers with classification predictions leads to significant increases in reviewing speeds. However, the observed increased reviewing speeds do not continue to increase when the classifier predictions agree perfectly with the ground truth, since reviewers must make their own reviewing decisions and, therefore, sometimes disagree with the classifier.

In this section, we have shown that providing the reviewers with sensitivity classification predictions can lead to increased reviewer accuracy (**H1**(a)) and increased reviewing speeds (**H1**(b)). As with any technology-assisted decision support system, there is a potential for the classifier's predictions to influence the reviewers' decisions about what is or is not sensitive. In practice, as the field of technology-assisted sensitivity review develops, classifiers will be able to learn from multiple reviewers and perform additional checks on documents that are difficult to classify, or that a reviewer disagrees with the classifier about their sensitivities. This, in-turn, has the potential to develop a more common or shared view of sensitivity than is the case in the current practice of sensitivity judgements often being made by a single sensitivity reviewer. As we previously discussed in Section 5.1, sensitivity is to some degree inherently subjective and reviewers can disagree on sensitivity judgements. However, as is the case with differing judgements having little effect on the relative effectiveness of systems [Voorhees 2000] we would not expect differing sensitivity judgement to impact our findings on the relative effects of the classification treatments that we evaluate in this work. As future work, we will investigate if the accuracy of the classifier has an impact on the amount of, or choice of, text within a document that the reviewers annotate, i.e., label as being sensitive.

6.2 The Impact of Classification Confidence on Reviewer Performance

We now evaluate the impact that the confidence level, *Low*, *Medium* or *High*, of a classification prediction has on the reviewers' performance (**H2**). When evaluating the effects of classifier confidence, we analyse the mean participant performance for the relative classifier confidence levels over the *Medium* and *Perfect* classification batches.

Addressing **H2**(a), Figure 7 presents the mean Cohen's κ scores for the agreement between participants and the classification predictions for each of the classifier confidence levels *Low*, *Medium*, and *High*. From the figure, we note that there is a clear and steady trend showing increased mean participant-classifier agreement as the classifier's confidence level increases. A two-way repeated measures omnibus ANOVA, calculated over the classification confidence (*Low, Medium*, *High*) and classifier effectiveness (*Medium*, *Perfect*) treatment levels (sphericity = $\chi^2(2) = 1.1, p = 0.577$), shows that, in our study, there is no statistically significant two-way interaction between classification confidence and classification effectiveness in terms of mean participant-classifier Cohen's κ agreement, $F(2, 14) = 0.61, p = 0.557$[8] with effect size $\eta^2 = 0.08$ and observed power of 0.132. We therefore move to evaluating the main effects of the classification confidence and classification effectiveness factors individually.

[8]This finding also holds when the Greenhouse-Geisser correction is applied, $F(1.713, 11.992) = 0.61$, $p = 0.535$.

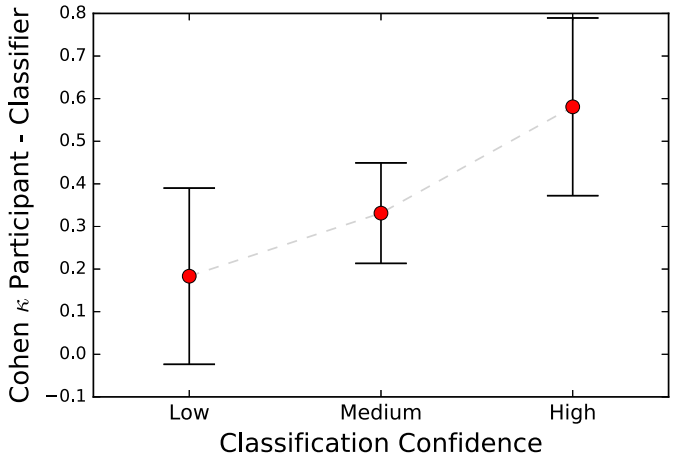

Fig. 7. Cohen's κ for participant and classifier agreement for each of the classifier confidence levels, *Low*, *Medium*, and *High*, and 95% confidence intervals.

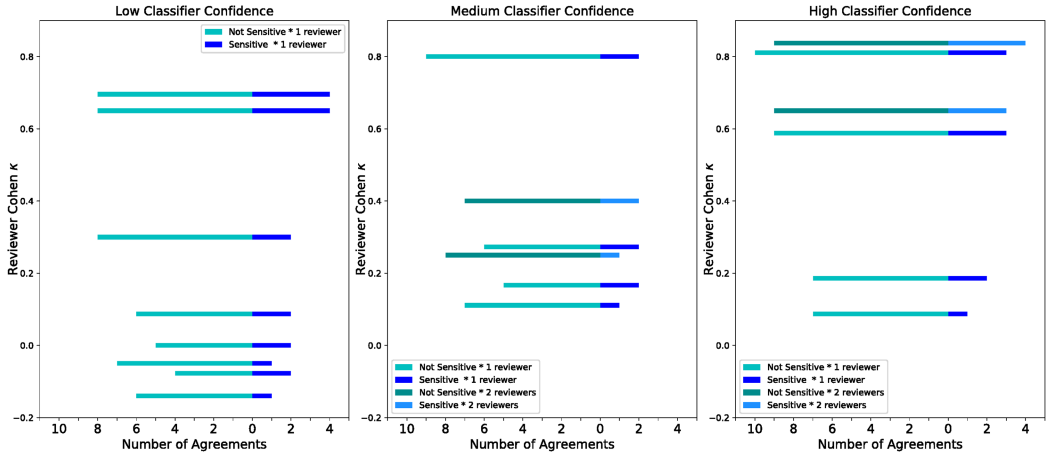

Fig. 8. Number of participant and classifier agreements (x-axis) for sensitive (blue) and not-sensitive (green) documents and the participants' Cohen's κ scores (y-axis) for each of the classifier confidence levels, *Low*, *Medium*, and *High*. Bars represent reviewers. Where two reviewers obtained the same κ score due to an identical agreement, agreements for sensitive documents are light blue and agreements for not-sensitive documents are dark green.

The classification confidence factor meets the test of sphericity ($\chi^2(2) = 2.885, p = 0.236$), and the main effect of this treatment shows that there is a statistically significant difference in the participant-classifier agreement (Cohen's κ) between the classification confidence treatment levels, $F(2, 14) = 5.793, p = 0.015$. However, in post hoc tests adjusted for multiple comparisons, we find that the differences between individual treatment levels are not significant, $p = 0.218$ *Low* vs. *Medium*, $p = 0.095$ *Low* vs. *High*, and $p = 0.264$ *Medium* vs. *High*. As an additional analysis, it is interesting to investigate whether the increase in reviewer-classifier agreement that we observe is due to the reviewers' agreeing with the classifier more for sensitive or not-sensitive documents. Figure 8 presents bar charts for each of the classifier confidence levels *Low*, *Medium*, and *High* showing the number of agreements with the classifier on the x-axis for either sensitive documents

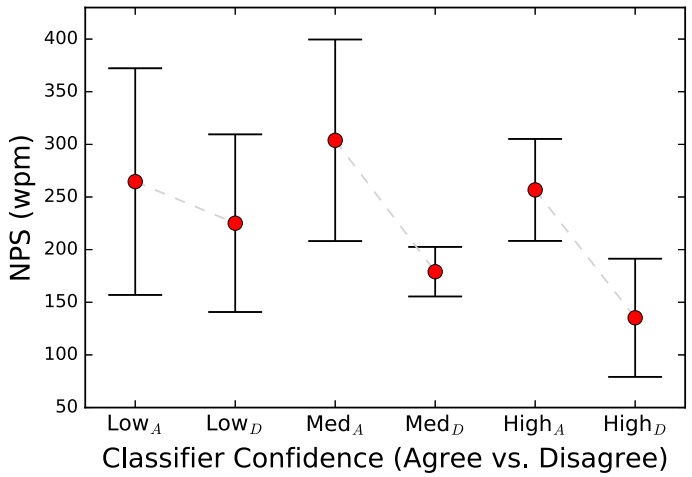

Fig. 9. NPS in wpm for the study participants (reviewers), when participants either agree (subscript *A*) or disagree (subscript *D*) with the classifier for each of the classifier confidence levels, *Low*, *Medium*, and *High*, and 95% confidence intervals.

(blue) or not-sensitive documents (green), with the resulting Cohen's κ score on the y-axis. From Figure 8, we can see that six of the eight reviewers achieved $\kappa > 5$ when the classifier's confidence was high, compared to two reviewers when the classifier had low confidence and only one reviewer when the classifier confidence was medium. However, we also note from Figure 8 that this increased agreement when the classifier's confidence was high was due to reviewers agreeing with the classifier more for both sensitive and not-sensitive documents.

The classification effectiveness factor has only two treatment levels for the two-way ANOVA (since there are no classifier confidence levels associated to the *None* classification effectiveness treatment level), and, therefore, the test of sphericity is not required. The post hoc test shows that the difference in participant-classifier agreement in the *Medium* and *Perfect* classification effectiveness treatment levels is statistically significant ($p = 0.038$). Therefore, for hypothesis **H2**(a), we conclude that classifier confidence can have a significant effect on reviewer agreement (according to the main effect of classification effectiveness in our study), although this significant difference is not reflected in our post hoc tests.

Turning our attention to hypothesis **H2**(b), Figure 9 presents the mean participant NPS for each of the classifier confidence levels *Low*, *Medium,* and *High*, when participants either agree (subscript *A*) or disagree (subscript *D*) with the classifier's predictions. First, we investigate whether there is a two-way interaction effect between the classifier's confidence and effectiveness in terms of NPS. The two-way omnibus ANOVA shows that, in our study, there is no statistically significant interaction effect in terms of NPS, $F(2, 14) = 0.992, p = 0.396$ with effect size $\eta^2 = 0.124$ and observed power of 0.188 (sphericity = $\chi^2(2) = 4.523, p = 0.104$). The two-way ANOVA also shows that the main effects of classification confidence and classification effectiveness (over the two effectiveness treatment levels *Medium* and *Perfect*) in terms of NPS are not statistically significant, ($\chi^2(2) = 1.142, p = 0.565$) $F(2, 14) = 0.205, p = 0.817$ with effect size $\eta^2 = 0.028$ and an observed power of 0.076 and $F(1, 7) = 2.212, p = 0.181$ with effect size $\eta^2 = 0.240$ and observed power of 0.252, respectively. Therefore, the classifier's confidence in its predictions does not have a statistically significant impact on reviewing speeds in terms of NPS.

From Figure 9, we note, however, that the participants review documents faster when they agree with the classifier's predictions. This trend is consistent for each of the classifier confidence levels.

We perform a two-way repeated measures omnibus ANOVA calculated over the three classification confidence treatment levels (*Low*, *Medium* or *High*) and whether the reviewer either *agrees* or *disagrees* with the classification prediction. The ANOVA shows that, in our study, there is no *overall* statistically significant two-way interaction effect between the level of confidence that the classifier has in its prediction and whether the reviewer agrees or disagrees with the classifier's prediction, $F(2, 14) = 1.782, p = 0.204$ (sphericity = $\chi^2(2) = 0.147, p = 0.929$). In other words, the reviewers' processing speed does not significantly change as the classifier's confidence increases dependent on whether the reviewer agrees or disagrees with the classifier.

Figure 9 shows that the reviewers' agreement with the classifier appear to impact their processing speed at each individual level of classifier confidence. Therefore, we now evaluate the main effects of each of the classification confidence levels and conduct post hoc tests to investigate if there is a statistically significant difference in NPS at each of the individual confidence levels dependent on if the reviewer agrees or disagrees with the classifier's prediction.

The two-way ANOVA shows that there is no statistically significant difference in reviewer NPS over the different levels of classifier confidence, $F(2, 14) = 1.219, p = 0.325$ with effect size $\eta^2 = 0.148$ and observed power of 0.223. However, when evaluating the main effect of reviewer agreement in terms of NPS, the two-way ANOVA shows that there is indeed a statistically significant difference in NPS between agree and disagree over the classification confidence levels, $F(1, 7) = 10.44, p = 0.014$ with effect size $\eta^2 = 0.599$ and observed power of 0.791. The post hoc test shows that the mean difference between agree and disagree over the different classifier confidence levels is 95.26 wpm (95%CI, 25.55 to 164.9, $p = 0.014$).

The observed trend in increased reviewing speed when participants agree with the classifier's predictions at each of the classifier confidence levels is in line with **H2**(b) and is observed over all levels of classifier confidence. Therefore, for hypothesis **H2**(b), we conclude that reviewing speeds are indeed significantly increased when the reviewer agrees with the sensitivity classification predictions. Moreover, this significant difference is observed over all of the levels of classifier confidence. However, the classifier's confidence does not have a significant impact on reviewing times. We note that, although we observe a clear trend that reviewers review faster when they agree with the classifier for each level of classifier confidence, this does not lead to an *overall* increase in reviewing speed as the classifier gets more confident. From Figure 9, we observe that there is a slight decrease in reviewing speeds when the classifier's confidence level is *High*. This observation suggests that, for a reviewer, disagreeing with a classification prediction when the classifier has a high level of confidence in its prediction has a greater negative impact on reviewing speed than when a reviewer disagrees with a classification prediction that the classifier is less confident about. When the classifier's confidence is high, we postulate that reviewers have taken more time to ensure that they fully reviewed the documents, so as not to rely solely on the classifier's prediction, thereby reducing their reviewing speed.

To complete the analysis in this section, it is interesting to investigate if there is a main effect of classifier confidence on the accuracy of the reviewers in terms of BAC and F_2. Therefore, we conduct two two-way repeated measures omnibus ANOVAs between classification confidence and classifier effectiveness. In the ANOVA analyses presented here, the classification effectiveness treatment has only two levels, *Medium* and *Perfect*, since there is no classifier confidence factor in the *None* classification effectiveness treatment level. First, we find that there is no two-way interaction effect between classifier confidence and classification effectiveness in terms of BAC $F(2, 14) = 0.328, p = 0.726$, with effect size $\eta^2 = 0.045$ and observed power of 0.092 (sphericity = $\chi^2(2) = 1.27, p = 0.529$). Moreover, there is no statistically significant main effect of classification confidence on reviewer performance in terms of BAC, $F(2, 14) = 1.954, p = 0.179$, with effect size $\eta^2 = 0.218$ and observed power of 0.336 (sphericity = $\chi^2(2) = 4.873, p = 0.87$). However, we find

Table 5. Summary Table of Our Hypotheses Conclusions and the Sources of Supporting Evidence

	Treatment	Metric	Validated?	Source
H1(a)	Effectiveness	Increased judgement accuracy	✓	Figures 4 and 5
H1(b)	Effectiveness	Increased reviewing speed	✓	Figure 6
H2(a)	Confidence	Increased reviewer-classifier agreement	–	Figure 7
H2(b)	Confidence	Increased reviewing speed	✓	Figures 9

that there is a statistically significant main effect of classification effectiveness on reviewer performance in terms of BAC, $F(1,7) = 9.051, p = 0.020$, with effect size $\eta^2 = 0.564$ and observed power of 0.733 (the test of sphericity is not required). This observation is also reflected in the post hoc test ($p = 0.020$). In terms of F_2, the two-way ANOVA, with the Greenhouse-Geisser correction (sphericity = $\chi^2(2) = 6.075, p = 0.048$), shows that there is no two-way interaction effect between classifier confidence and classification effectiveness in terms of F_2, $F(1.222, 8.554) = 1.349, p = 0.291$, with effect size $\eta^2 = 0.162$ and observed power of 0.234. With regard to a main effect, the ANOVA shows that there is a statistically significant main effect of classification confidence on reviewer performance in terms of F_2, $F(2,14) = 4.196, p = 0.037$, with effect size $\eta^2 = 0.375$ and observed power of 0.638 (sphericity = $\chi^2(2) = 0.088, p = 0.957$). However, this significant effect is not reflected in a post hoc test adjusted for multiple comparisons ($p = 0.231$). Finally, the ANOVA shows that there is no statistically significant main effect of classification effectiveness on reviewer performance in terms of F_2, $F(1,7) = 1.723, p = 0.231$, with effect size $\eta^2 = 0.198$ and observed power of 0.207 (the test of sphericity is not required). Overall, this analysis suggests that there is a potential for classifier confidence to have an effect on the reviewers' accuracy but we do not see clear evidence of this in our study.

6.3 User Study Conclusions

As shown by the results reported in Section 6, our study provides evidence that supports both of our stated hypotheses. Table 5 provides a summary of our hypotheses conclusions and the sources of supporting evidence from Section 6. In short, providing classification predictions to the sensitivity reviewers increases the accuracy (Figures 4 and 5) and speed (Figure 6) of the reviewers (**H1**). For **H2**, we found that the level of confidence that the classifier has in its predictions can result in a statistically significant difference in reviewer agreement. However, we found that the reviewers did not statistically significantly agree with the classifier more as the classifier's confidence in its predictions increased. We also found that reviewing speeds are indeed increased when the reviewer agrees with the classifier (Figure 9).

We argue that our findings from this study demonstrate that sensitivity classification predictions are a viable technology to effectively provide sensitivity reviewers with valuable information about the sensitivities within a collection of documents, which can increase the speed and accuracy of conducting the sensitivity review task. We note that our study participants are not expert sensitivity reviewers. This could be viewed as a limitation of the study, and more research is needed to evaluate the potential benefits and effects of sensitivity classification predictions for expert sensitivity reviewers. However, we argue that our findings are important, since they 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.

Our analysis in this section has focused on the overall benefits of providing sensitivity reviewers with sensitivity classification predictions when they are reviewing a collection of documents (i.e.,

a batch). Moreover, we have evaluated the impact that the effectiveness of the classifier and the classifier's confidence can have on such benefits. In addition to this, it is important to be able to reason about how different aspects of a reviewer's decision making, or judgement, process impacts the amount of time it takes to sensitivity review a document (with and without classification predictions). In the following section, we provide an analysis of the impact that judging a document to contain sensitive information has on the time taken to review. We investigate whether there is an additional reviewing time overhead from judging documents that contain sensitive information. Moreover, we evaluate whether providing reviewers with sensitivity classification predictions can reduce any such overhead. Furthermore, we analyse the impact that the *correctness* of the classifier's predictions has on reviewing times when a reviewer judges a document to contain, or not contain, sensitive information.

7 ANALYSIS OF THE REVIEWERS' SENSITIVITY JUDGEMENTS

In Section 6, we showed that providing human reviewers with automatic sensitivity classification predictions can lead to an overall increase in reviewing speed, while maintaining (or improving upon) the accuracy of the reviewer's sensitivity judgements. However, it is also clear from Section 6 that not all sensitivity classification predictions are of equal benefit to the reviewers. For example, we can see from Figure 6 that different levels of classification accuracy can result in a variation in the amount of time that a reviewer requires to review documents. Indeed, it may be the case that some types of sensitivity classification (e.g., correct *non-sensitive* predictions) have a negligible impact on reviewing times, while others (e.g., incorrect *sensitive* predictions) may in fact result in reviewers taking more time to review a document. In practice, a sensitivity classification prediction can have a *positive*, *negligible*, or *negative* impact on the required reviewing time for a specific document.

In this section, we analyse the log data from our user study to provide additional insights into the differences in reviewing times between documents that are judged to be sensitive or not-sensitive. Moreover, we evaluate how the reviewing times of such judgements are impacted by the correctness of the classifier's predictions. In line with Section 6, when analysing reviewing times, we use NPS [Damessie et al. 2016]. However, differently from Section 6, in this section we do not use the user study's expert reviewers' ground truth *gold standard* judgements. Instead, we focus on whether a (participant) reviewer judged a document to be sensitive or not and whether the classifier is correct, or not, with respect to their judgement (i.e., if the reviewer agrees or disagrees with the classifier).

As we presented in Section 5, the study participants reviewed three batches of documents, each with a classification effectiveness treatment: *None*, i.e., no classification predictions; *Medium*, i.e., a classifier that achieved 0.7 BAC; or *Perfect*, i.e., a perfect classifier. When evaluating the differences between sensitive and non-sensitive judgements without classification assistance, we use the *None* batch only. When evaluating the impact of classification predictions, we use the *Medium* and *Perfect* batches combined. Table 1 provided an overview of the distributions of classification predictions. We first analyse the additional reviewing time overhead required when making *sensitive* judgements and whether sensitivity predictions can reduce this overhead before, second, evaluating the impact that the correctness of *sensitive* or *not-sensitive* predictions have on reviewing times.

7.1 Sensitivity Judgement Reviewing Times

In this section, we provide an analysis of the difference in the amount of time that is required to review *sensitive* and *not-sensitive* documents. As outlined by Baly et al. [2016], formulating a deep understanding of textual information (as is required for sensitivity review) requires the reader to integrate background information with the text that is being read, through a combination of

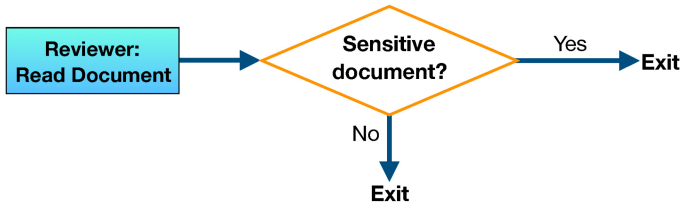

Fig. 10. A sensitivity review user model. The time to sensitivity review a document is dependent on the reviewer reading the document and evaluating if the document is sensitive or not.

low-level processing (such as lexical parsing) and high-level cognitive processes (to develop inferences about the context of the text). To assist with our analysis of the reviewer's sensitivity judgements, in this section and the following section, we formulate simple user models of the inputs to a reviewer's decision making process and the possible decision outcomes. Similarly to other work, for example Azzopardi et al. [2018], the conceptual models that we present are intended to provide a starting point for the development of a complete interaction model as the technology-assisted sensitivity review literature develops.

A simple user model of a reviewer's decision making process, when they are not assisted by classification predictions, can be separated into two distinct activities that reflect this separation between high- and low-level processing. First, the reviewer has to read the document, and, second, the reviewer has to evaluate any potential sensitivities in the document (for example, this may include checking whether a named individual is deceased or checking whether specific information is already in the public domain).

The first activity, reading the document, must be done for all of the documents that are to be reviewed. However, the second activity, evaluating potential sensitivities, only has to be done for documents that have passages of text that are potentially sensitive. Logging accurate values for the time taken to review a *potential* sensitivity is not feasible without an experimental design in which the reviewers explicitly declare when they are either reading the document or considering a potential sensitivity. With this in mind, we assume that a reviewer does not have to evaluate any potential sensitivities in non-sensitive documents.[9] The resulting user model is illustrated in Figure 10 and, from this user model, we form our third hypothesis:

H3: *In the absence of sensitivity predictions, sensitive judgements will require more time to make than not-sensitive judgements.*

H3 expects that there will be a notable additional reviewing time overhead for sensitive judgements due to the reviewer having to evaluate the sensitive information in the document. To investigate this, we aim to determine whether the average time for reviewers to reach the two Exit decisions in Figure 10 is different.

Next, as we previously showed in Section 6 (Figure 6), assisting reviewers with sensitivity classification predictions can lead to an overall increase in reviewing speeds. In its simplest form, the user model for assisted sensitivity review is presented in Figure 11. In this user model, the reviewer is assisted in making their judgement by the classifier's sensitivity prediction. This assistance has the potential to reduce any additional reviewing time overheads that arise from judging sensitivities (either by alerting the reviewer to the sensitivity or by increasing *the reviewer's* confidence that the document is indeed sensitive). Alternatively, the overall increase in reviewing speed (NPS)

[9]We note that, in practice, this will not be the case for all non-sensitive documents.

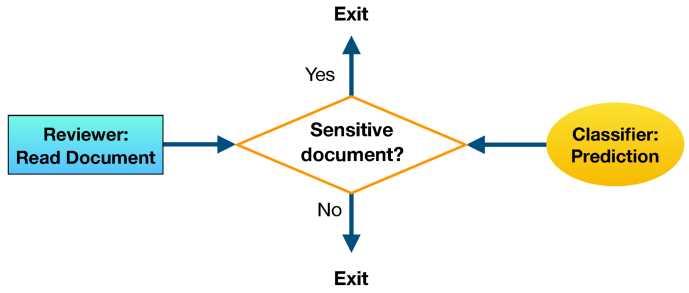

Fig. 11. A simplified *assisted* sensitivity review user model. The time to sensitivity review a document is dependent on the reviewer reading the document assisted by the classifier's prediction.

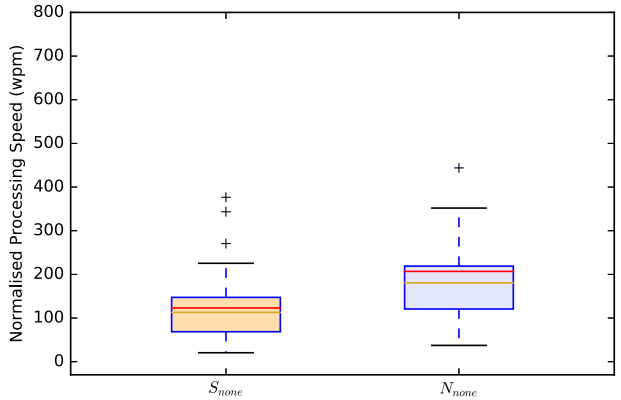

Fig. 12. NPS in wpm for sensitive (S_{none}) and not-sensitive (N_{none}) judgements without sensitivity predictions. Mean reviewing times are shown by a red horizontal line, while median reviewing times are shown by a gold line.

in Figure 6 may result only from the reviewers making *non-sensitive* judgements more quickly. Our fourth hypothesis is derived from this simple assisted review user model:

H4: *In the presence of sensitivity predictions, the additional time that is required to make sensitive judgements will be reduced compared to when there are no sensitivity predictions provided for the reviewers.*

H4 states that sensitivity classification predictions (yellow input in Figure 11) will reduce the difference between the average times for reviewers to reach the two Exit decisions. We now report our analysis to investigate **H3** and **H4**.

Figure 12 presents the distributions of NPS, in wpm, for sensitive and non-sensitive judgements when no sensitivity predictions are provided to the reviewer, denoted as S_{none} and N_{none} respectively (**H3**), while Figure 13 presents the NPS distributions for sensitive (S_{pred}) and non-sensitive judgements (N_{pred}) when the reviewer is assisted by sensitivity predictions (**H4**). In this section, we are not evaluating the effects of the different treatments in our study (i.e., classification effectiveness or classification confidence). Rather, we are analysing the actual judgements that are made by the participants. Therefore, differently from Section 6, in this section we present the distributions of reviewing times as box plots, where a box shows the range of observed NPS values that are within the lower and upper quartile of the observed values, and the mean and median values are denoted by a red and gold horizontal line, respectively.

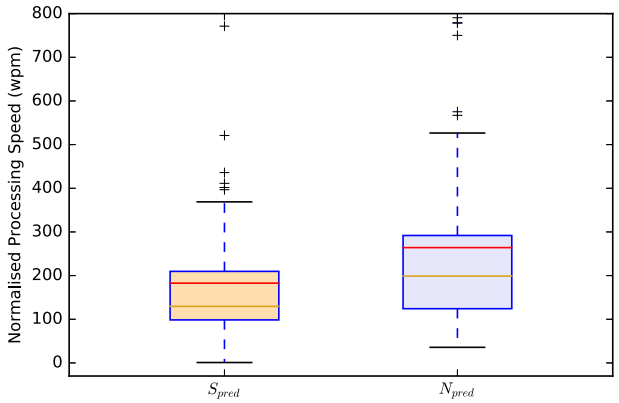

Fig. 13. NPS in wpm for sensitive (S_{pred}) and not-sensitive (N_{pred}) judgements assisted by sensitivity predictions. Mean reviewing times are shown by a red horizontal line, while median reviewing times are shown by a gold line.

First, to investigate **H3**, we begin our analysis by comparing the NPS values for sensitive (S_{none}) and non-sensitive (N_{none}) judgements when the reviewers are not assisted by sensitivity predictions. From Figure 12, we observe that the NPS for *not-sensitive* judgements (N_{none}) is notably higher than for *sensitive* judgements (S_{none}), when the reviewers are not assisted by sensitivity predictions. From Figure 12, we see that the mean normalised processing speed for sensitive judgements is 123.05 wpm compared to a mean NPS of 206.81 wpm for not-sensitive judgements. Therefore, in our study, when averaged over the entire reviewing time, sensitive judgements take 40.5% longer to make than not-sensitive judgements.

Mean processing speed is an important metric for managing the sensitivity review process, e.g., for resource allocation. However, as can also be seen from Figure 12, the processing speeds range from 40 to 395 wpm for sensitive judgements and from 45 to 425 wpm for not-sensitive judgements. Therefore, it is also important to know whether there is a notable difference in processing speeds between sensitive and not-sensitive judgements for *most* of the judgements that are made. From Figure 12, we observe that the median NPS for sensitive judgements, S_{none}, is 112.98 wpm, while the median NPS for not-sensitive judgements, N_{none}, is 180.80 wpm. Therefore, for most of the judgements that are made, when normalising for variations in reading speeds and document lengths, the reviewers are 37.5% slower in making sensitive judgements than not-sensitive judgements.

In this section, to test for statistical significance we conduct paired samples t-tests where each document is paired by the mean time that reviewers required to judge the document either as being sensitive or as being not-sensitive. In other words, each of the documents has been *judged* as being both *sensitive* and *not-sensitive* by different reviewers. We are interested in whether sensitive judgements require more time to make than not-sensitive judgements. Therefore, for each document, we evaluate whether it took longer to judge the document if a reviewer thought that it was sensitive compared to when a reviewer thought that it was not-sensitive. We select $p < 0.05$ as our significance threshold and report Cohen's d [Cohen 1977] as our effect size. Comparing the reviewing times for sensitive (S_{none}) and not-sensitive (N_{none}) judgements when the reviewers are not assisted by sensitivity classification predictions, the paired samples t-test shows that the difference in reviewing times for sensitive and not-sensitive judgements is not statistically significant, $t(12) = 2.131, p = 0.057, d = 0.62$, with observed power of 0.538. With respect to **H3**, in our study, sensitive judgements do indeed require more time to make than not-sensitive judgements. The results from our study showed a difference that was very close

to statistical significance ($p < 0.05$), suggesting a possible effect, i.e., that there is indeed an additional reviewing time overhead from judging sensitivity.

Moving onto **H4**, Figure 13 presents the NPS distributions for sensitive (S_{pred}) and non-sensitive judgements (N_{pred}) when the reviewer is assisted by sensitivity predictions. Figure 13 is plotted on the same scale as Figure 12 and, therefore, we can easily see that assisting the reviewers with sensitivity predictions leads to increased processing speeds for both sensitive and not-sensitive predictions. From Figure 13, we can see that the mean NPS for sensitive judgements is 182.67 wpm, while for not-sensitive judgements the mean NPS is 264.10 wpm. When the reviewers are assisted by sensitivity predictions, the mean time that they take to make a sensitive judgement is 30.8% longer than that of a not-sensitive judgement. However, this 30.8% difference in mean processing speeds is markedly (~10%) less than the 40.5% difference observed when no sensitivity predictions are provided (Figure 12). Therefore, in our study, the mean difference in time required to make sensitive judgements compared to not-sensitive judgements was reduced when the reviewers are assisted by sensitivity predictions.

Figure 13 also shows the median processing speed for sensitive judgements when assisted by sensitivity predictions (129.56 wpm). This is 34.8% slower than when not-sensitive judgements are made with assistance. In line with the findings for mean processing speeds, this indicates that, in our study, assisting the reviewers with sensitivity classification predictions reduces the additional overhead that arises from making sensitive judgements from 37.5% (Figure 12) to 34.8%. Comparing the reviewing times for sensitive (S_{pred}) and not-sensitive (N_{pred}) judgements when the reviewers are assisted by sensitivity classification predictions, the paired samples t-test shows that there is a statistically significant difference in terms of NPS, $t(179) = 2.556, p = 0.011, d = 0.2$, with observed power of 0.761. In response to **H4**, we conclude that assisting reviewers with sensitivity classification predictions can indeed reduce the additional reviewing overhead that arises from judging sensitive information.

In this section, we have provided an overall analysis of the additional reviewing time overhead that is required to make sensitive judgements and how classification predictions can reduce this overhead. However, we expect that the benefit of classification predictions will be dependent on the sensitivity of the document *and* the correctness of the prediction. In the following section, we evaluate the impact on reviewing times from such combinations.

7.2 How Different Classifier Predictions affect Reviewing Times

We expect that not all sensitivity predictions will be of equal benefit to reviewers. More specifically, following from Figure 9, we would expect a greater reduction in reviewing times when the reviewer agrees with the prediction. Moreover, we would expect that the benefits from sensitivity predictions will vary depending on if a document is judged to be sensitive or not.

Figure 14 presents a user model for assisted sensitivity review that reflects the possible reviewer-classifier (dis)agreement outcomes that can impact the utility of sensitivity predictions. In this user model, the reviewer *judges* if the document is sensitive or not and the classifier *predicts* if the document is sensitive or not. Therefore, there are four possible (dis)agreement outcomes that can impact the reviewing times, labelled in Figure 14 as follows:

- *SA*: The reviewer judges the document to be sensitive and the classifier predicts that the document is sensitive (agree)
- *SD*: The reviewer judges the document to be sensitive and the classifier predicts that the document is not-sensitive (disagree)
- *NA*: The reviewer judges the document to be not-sensitive and the classifier predicts that the document is not-sensitive (agree)

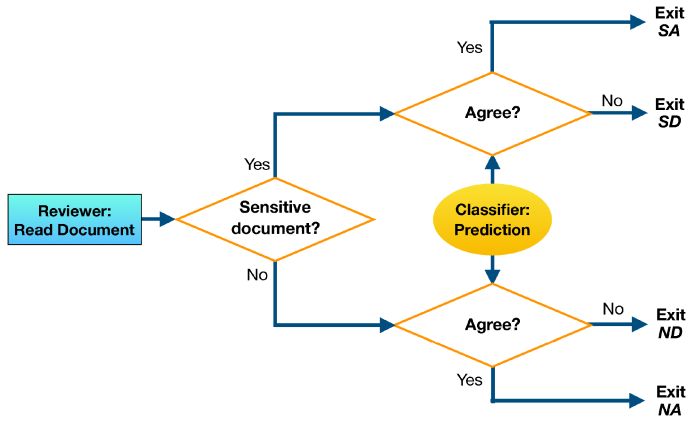

Fig. 14. An assisted sensitivity review user model. The time to sensitivity review a document is dependent on the reviewer reading the document and judging if the document is sensitive or not. Moreover, the amount of increase/decrease in reviewing speed that is a result of the sensitivity classifier is conditioned on if the reviewer judges the document to be sensitive or not and if the reviewer agrees with the classifier's prediction or not.

Table 6. The Reviewers' Mean NPS in Words Per Minute for *Sensitive* and *Not-sensitive* Judgements with No Classification Predictions (NPS_S, NPS_N), or When the Reviewer Agrees or Disagrees with the Classifier (NPS_{SA}, NPS_{SD}, NPS_{NA}, NPS_{ND}), and the Corresponding Δ Values (Δ_{SA}, Δ_{SD}, Δ_{NA}, Δ_{ND})

	Sensitive					Not-sensitive				
	None	Classifier-Agree		Classifier-Disagree		None	Classifier-Agree		Classifier-Disagree	
	NPS_S	NPS_{SA}	Δ_{SA}	NPS_{SD}	Δ_{SD}	NPS_N	NPS_{NA}	Δ_{NA}	NPS_{ND}	Δ_{ND}
Reviewer 1	134.22	155.70	+21.48	109.13	−25.09	301.14	334.09	+32.96	157.64	−143.50
Reviewer 2	103.22	182.61	+79.39	141.18	+37.96	172.23	301.45	+129.23	118.28	−53.95
Reviewer 3	126.54	105.91	−20.63	91.76	−34.78	367.21	247.32	−119.89	71.62	−295.59
Reviewer 4	99.03	183.82	+84.79	151.15	+52.11	216.28	292.83	+76.56	339.35	+123.07
Reviewer 5	52.70	125.35	+72.66	100.81	+48.11	158.69	322.46	+163.77	137.54	−21.15
Reviewer 6	142.21	257.29	+115.08	95.78	−46.43	202.59	317.19	+114.61	342.39	+139.80
Reviewer 7	129.83	114.73	−15.10	94.01	−35.82	178.08	262.83	+84.75	196.18	+18.10
Reviewer 8	171.13	194.75	+23.62	142.18	−28.94	198.93	226.85	+27.91	144.03	−54.90
Grand Mean	119.86	165.02†	+45.16	115.75	−4.11	224.39	288.13	+63.74	188.38	−36.01

Statistically significant differences compared to NPS_S or NPS_N, are denoted as † (t-test, $p < 0.05$).

- *ND*: The reviewer judges the document to be not-sensitive and the classifier predicts that the document is sensitive (disagree)

To investigate the impact on reviewing times from each of the possible (dis)agreement outcomes, we evaluate the difference (Δ) in a reviewer's processing speed when they either agree or disagree with the classifier for sensitive or not-sensitive documents compared to when no classification predictions are provided.

Table 6 presents the reviewer's mean NPS in wpm for: documents that they judged to be sensitive when no classification predictions are provided (NPS_S) and when they either agree (NPS_{SA}) or disagree (NPS_{SD}) with the classifier's prediction; and documents that they judged to be not-sensitive when no classification predictions are provided (NPS_N), and when the reviewer either agrees (NPS_{NA}) or disagrees (NPS_{ND}) with the classifier's prediction.

Table 6 also presents the difference in a reviewer's NPS when they agree or disagree with the classifier for sensitive (Δ_{SA}, Δ_{SD}) or not-sensitive (Δ_{NA}, Δ_{ND}) documents compared to when no classification predictions are provided (NPS_S, NPS_N), calculated as follows:

$$\Delta_{SA} = NPS_{SA} - NPS_S$$
$$\Delta_{SD} = NPS_{SD} - NPS_S$$
$$\Delta_{NA} = NPS_{NA} - NPS_N$$
$$\Delta_{ND} = NPS_{ND} - NPS_N$$

Differently from Section 6, where we measured and evaluated the effects and potential interactions of two independent variables, i.e., classification effectiveness and classifier confidence, each with multiple treatment levels, *None*, *Medium*, *Perfect* and *Low*, *Medium*, *High*, respectively, the variables that we are evaluating in this section, i.e., if a reviewer agrees or disagrees with the classifier's prediction compared to when the reviewer is not provided with a prediction, have a conditional existence. It is not possible for a reviewer to agree or disagree with the classifier if the classifier does not make a prediction. Moreover, the within-subject design of our user study ensured that each participant was exposed to each of the treatment levels through the same data instances. Whereas, in this section, we are evaluating the effect of a conditional variable (agree or disagree) through data as selected by the study participants, i.e., by whether the participant judges a document to be sensitive or not-sensitive. For these reasons, a two-way ANOVA is not an appropriate statistical test for the analysis in this section. Therefore, to evaluate whether the observed differences in reviewing times (NPS) are statistically significant, we perform four paired samples t-tests: NPS_S vs. NPS_{SA}, NPS_S vs. NPS_{SD}, NPS_N vs. NPS_{NA}, and NPS_N vs. NPS_{ND}, where the data is paired for each reviewer's sensitive and not-sensitive judgements. We set our significance threshold as $p < 0.05$. We report Cohen's d as our effect size and denote significant differences (compared to NPS_S or NPS_N) as † in the Grand Mean row of Table 6.

On analysing Table 6, we first note that all of the reviewers in our study are faster at reviewing non-sensitive documents than sensitive documents when no classification predictions are provided (i.e., $NPS_N > NPS_S$) and when the reviewer agrees with the classifier's prediction (i.e., $NPS_{NA} > NPS_{SA}$). This is in line with our findings from **H3** and provides additional quantitative evidence that there is a reviewing time overhead from making *sensitive* judgements—this is expected, as reviewers have to highlight the sensitive information that they identify.

Turning our attention to the impact on reviewing times from the reviewers' (dis)agreement with the classifier's predictions, we will discuss the impact on reviewing times from each of the (dis)agreement outcomes *SA*, *SD*, *NA*, and *ND* in turn. When the reviewers and classifier agree that the document is sensitive (*SA*), we can see from Table 6 that sensitivity predictions led to an increase in NPS compared to when no predictions were provided (Δ_{SA}), for six of the eight reviewers. We note, however, that reviewers 3 and 7 were actually slower at making sensitive judgements when they agreed with the classifier's predictions compared to when no predictions were provided.

The grand mean Δ values in Table 6 provide us with a measure of the overall increase/decrease in reviewing times that result from each of the (dis)agreement outcomes. We can see from the grand mean value for Δ_{SA} (+45.16) that *on average* the reviewing speed (NPS) of our study participants increased when they agreed with the classifier's *sensitive* prediction, compared to when no predictions were provided (119.86 NPS_S vs. 165.02 NPS_{SA}). Indeed, this is a 37.7% increase in reviewing speed (NPS). Therefore, we conclude that in our study *sensitive* classification predictions led to an increase in reviewing speeds when the reviewers agreed with the classifier's prediction. The paired samples t-test shows that the difference in reviewer NPS between NPS_S and NPS_{SA} is statistically significant, $t(7) = 2.564, p = 0.037, d = 0.91$, with observed power of 0.601.

Moving on to *SD*, when the reviewer judges the document to be sensitive and disagrees with the classifier's prediction. For five of the eight reviewers, this outcome led to a reduction in NPS compared to when no predictions were provided (Δ_{SD}). This result is somewhat intuitive, since disagreeing with the classifier can result in the reviewer taking some time to question their own judgement, in addition to the time required to make the judgement. We note that three reviewers increased their reviewing speeds in this outcome. Those reviewers were the slowest of all the reviewers when making sensitive judgements without sensitivity predictions (NPS_S). This potentially suggests that the reviewing times of the reviewers that make sensitive judgements quickly when not assisted by classification predictions are more likely to be negatively affected when the reviewers disagree with the classifier's prediction.

The grand mean value for Δ_{SD} is −4.11 wpm. This is a 3.4% decrease in NPS (119.86 NPS_S vs. 115.75 NPS_{SD}). We conclude that a reviewer's processing speed is likely to decrease if the reviewer judges a document to be sensitive and disagrees with the classifiers prediction. However, the paired samples t-test shows that the difference in NPS that we observe between NPS_S and NPS_{SD} is not statistically significant, $t(7) = 0.276$, $p = 0.791$, $d = 0.1$, with observed power of 0.057.

We now move on to the outcomes where the reviewer judges the document to be not-sensitive. When the reviewer agrees with the classifier (*NA*), we can see from the Δ_{NA} values in Table 6 that for seven of the eight reviewers sensitivity predictions led to an increase in NPS. The reviewer that did not benefit from sensitivity predictions in this outcome also did not benefit when they agreed with the classifier for sensitive predictions, i.e., reviewer 3.

The grand mean for Δ_{NA} is +63.74 wpm. This is a 28.1% increase in NPS (224.39 NPS_N vs. 288.13 NPS_{NA}). We conclude that, in our study, not-sensitive predictions led to an overall increase in mean reviewer NPS when the reviewer agreed with the classifier's prediction. However, the paired samples t-test shows that the observed difference between NPS_N and NPS_{NA} is not statistically significant, $t(7) = 2.063, p = 0.078, d = 0.73$, with observed power of 0.430. It is worth noting that the 28.1% increase in this outcome is less than the 37.7% for outcome *SA*. This suggests that, when a reviewer agrees with the classifier, sensitivity predictions are more beneficial for sensitive judgements than for not-sensitive judgements.

For the final (dis)agreement outcome, *ND*, we can see, from the Δ_{ND}, that five of the eight reviewers were slower when they were provided with sensitivity predictions than when no predictions were provided (i.e., $NPS_{ND} < NPS_N$). Moreover, the grand mean Δ_{ND} in this outcome is −36.01 wpm. This is a 16.0% decrease in NPS (224.39 NPS_N vs. 188.38 NPS_{ND}) and notably more than the 3.4% decrease for *SD*. However, the observed difference in NPS between NPS_N and NPS_{ND} is not statistically significant, $t(7) = 0.723, p = 0.493, d = 2.83$, with observed power of 0.681. If the classifier says a document is sensitive but the reviewer cannot see any sensitivity (ND), then the reviewer is likely to spend more time trying to ensure that they have not missed a sensitivity, than the time that they would likely spend re-reviewing a document that they have judged to be sensitive when the classifier says that the document is not sensitive (SD). Moreover, in the ND outcome the reviewer must *find* the predicted sensitivity in the document for them to judge the document as being sensitive. Therefore, we conclude that when a reviewer judges a document to be not-sensitive but the classifier predicts the document to be sensitive, the reviewer is likely to spend additional time checking their judgement to make sure that they do not accidentally release sensitive information.

8 CONCLUSIONS

We conducted a within-subject digital sensitivity review user study to evaluate the benefits of automatic sensitivity classification predictions for sensitivity reviewers. We investigated how the accuracy of sensitivity classification predictions and the confidence that the classifier has in its

individual predictions affects two key aspects of sensitivity review, namely 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).

Our findings showed that automatic sensitivity classification, with an effectiveness in line with sensitivity classifiers from the literature (e.g., McDonald et al. [2017b]), led to a significant (+37.9%) improvements in reviewer accuracy compared to when no predictions were provided (repeated measures ANOVA, $p < 0.05$) (**H1**(a)). Moreover, we found that assisting reviewers with sensitivity classification predictions resulted in a 72.2% increase in the mean reviewing speed of the reviewers (**H1**(b)). We also found that the level of confidence that the classifier has about its predictions can result in a significant difference in reviewer-classifier agreement (**H2**(a)). Moreover, we found that reviewing speeds were significantly increased when the reviewers agreed with the sensitivity classification predictions (**H2**(b)).

We also performed an in-depth analysis of the log-data from our user study. Our analysis investigated whether there is an additional reviewing time overhead from judging documents that contain sensitive information (compared to documents that are not-sensitive) and whether providing reviewers with sensitivity classification predictions can reduce any such overhead. We showed that, in our study, there was indeed an additional overhead when reviewing sensitive documents, with our study participants taking 40.5% longer to review sensitive documents compared to non-sensitive documents (**H3**). We also showed that sensitivity classification predictions can reduce this additional reviewing time overhead by ~10% (**H4**). However, this value is conditioned on whether a document is sensitive or not *and* if the reviewer agrees with the classifier or not. In particular, we found that for documents that are judged to be sensitive, sensitivity classification predictions increased mean reviewing speeds by 37.7% when the reviewers agreed with the classifier's predictions. This increase in reviewing speed was statistically significant according to a paired samples t-test ($p < 0.05$). However, for sensitive judgements, mean reviewing speeds actually decreased by 3.4% when reviewers disagreed with the classifier. This decrease in reviewing speed was not statistically significant according to a paired samples t-test ($p < 0.05$).

This work is the first user study to evaluate the benefits of automatic sensitivity classification for assisting human reviewers to find sensitive information in digital government documents, so that the Government can comply with freedom of information laws. Overall, our findings provide strong evidence that sensitivity classification is a viable and valuable technology for assisting digital sensitivity review. Moreover, our study demonstrates that assisting human reviewers with sensitivity classification predictions could enable governments to increase the number of digital documents that can be sensitivity reviewed, while maintaining high levels of reviewing accuracy. We suggest, however, that more work needs to be done in evaluating different methods for portraying the classifier's decisions (e.g., highlighting sensitivities [McDonald et al. 2015]) and explaining classification decisions. In essence, this work demonstrates to governments and other stakeholders the importance of conducting further sensitivity classification research both to continue to improve the effectiveness of sensitivity classifiers and to identify additional ways in which they can assist sensitivity review.

ACKNOWLEDGMENTS

We acknowledge the efforts of the study participants and the expert sensitivity reviewers, and their UK Government departments, for providing the ground truth sensitivity judgements. We thank Jonathan Gudgeon, FCDO Services (UK Government), for his support of this work. We acknowledge the financial support of SVGC Ltd. through Project Cicero, EPSRC IAA and The National

Archives. We thank Tim Gollins and Douglas W. Oard for comments on the paper. We also thank the Associate Editor and the four peer reviewers for their comments and suggestions.

REFERENCES

Sir Alex Allan. 2015. Government Digital Records and Archives Review. Cabinet Office. Retrieved from https://www.gov.uk/government/publications.

Sir Alex Allan. 2014. Records Review. Cabinet Office. Retrieved from https://www.gov.uk/government/publications.

Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (CAIR'18)*.

Ramy Baly, Roula Hobeica, Hazem M. Hajj, Wassim El-Hajj, Khaled Bashir Shaban, and Ahmad Al Sallab. 2016. A meta-framework for modeling the human reading process in sentiment analysis. *ACM Trans. Inf. Syst.* 35, 1 (2016), 7:1–7:21.

Giacomo Berardi, Andrea Esuli, Craig Macdonald, Iadh Ounis, and Fabrizio Sebastiani. 2015. Semi-automated text classification for sensitivity identification. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM'15)*. 1711–1714.

Giacomo Berardi, Andrea Esuli, and Fabrizio Sebastiani. 2012. A utility-theoretic ranking method for semi-automated text classification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 961–970.

Michael H. Birnbaum. 1999. How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychol. Methods* 4, 3 (1999), 243.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*. 3121–3124.

Donald T. Campbell and Julian C. Stanley. 2015. *Experimental and Quasi-Experimental Designs for Research*. Ravenio Books.

Jacob Cohen. 1977. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Gordon V. Cormack and Maura R. Grossman. 2017. Navigating imprecision in relevance assessments on the road to total recall: Roger and me. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 5–14.

Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. 2010. Overview of the TREC 2010 Legal Track. In *Proceedings of the 19th Text REtrieval Conference*.

Denis Cousineau. 2005. Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutor. Quant. Methods Psychol.* 1, 1 (2005), 42–45.

Geoff Cumming and Robert Maillardet. 2006. Confidence intervals and replication: Where will the next mean fall?*Psychol. Methods* 11, 3 (2006), 217.

Tadele T. Damessie, Falk Scholer, and J. Shane Culpepper. 2016. The influence of topic difficulty, relevance level, and document ordering on relevance judging. In *Proceedings of the 21st Australasian Document Computing Symposium*. 41–48.

DARPA. 2010. DARPA, New Technologies to Support Declassification. (2010). Retrieved from http://fas.org/sgp/news/2010/09/darpa-declass.pdf.

Olive Jean Dunn. 1961. Multiple comparisons among means. *J. Am. Stat. Assoc.* 56, 293 (1961), 52–64.

Freedom of Information Act 2000. c. 36. Retrieved from https://www.legislation.gov.uk/ukpga/2000/36/contents.

Benjamin Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *Comput. Surv.* 42, 4 (2010), 1–53.

Samuel W. Greenhouse and Seymour Geisser. 1959. On methods in the analysis of profile data. *Psychometrika* 24, 2 (1959), 95–112.

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. J. Law Technol.* 17, 3 (2010), 1.

Victoria Lain. 2013. Digital Records Sensitivity Review. Retrieved from https://blog.nationalarchives.gov.uk/blog/digital-records-sensitivity-review.

Geoffrey R. Loftus and Michael E. J. Masson. 1994. Using confidence intervals in within-subject designs. *Psychon. Bull. Rev.* 1, 4 (1994), 476–490.

Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How does domain expertise affect users' search interaction and outcome in exploratory search? *ACM Trans. Inf. Syst.* 36, 4 (2018), 1–30.

Graham McDonald, Nicolás García-Pedrajas, Craig Macdonald, and Iadh Ounis. 2017a. A study of SVM kernel functions for sensitivity classification ensembles with POS sequences. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1097–1100.

Graham McDonald, Craig Macdonald, and Iadh Ounis. 2015. Using part-of-speech n-grams for sensitive-text classification. In *Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval*. 381–384.

Graham McDonald, Craig Macdonald, and Iadh Ounis. 2017b. Enhancing sensitivity classification with semantic features using word embeddings. In *Proceedings of the 39th European Conference on Information Retrieval*. 450–463.

Graham McDonald, Craig Macdonald, and Iadh Ounis. 2018. Towards maximising openness in digital sensitivity review using reviewing time predictions. In *Proceedings of the 40th European Conference on Information Retrieval*. 699–706.

Graham McDonald, Craig Macdonald, and Iadh Ounis. 2019. How sensitivity classification effectiveness impacts reviewers in technology-assisted sensitivity review. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 337–341.

Graham McDonald, Craig Macdonald, Iadh Ounis, and Timothy Gollins. 2014. Towards a classifier for digital sensitivity review. In *Proceedings of the 36th European Conference on Information Retrieval*. 500–506.

Linda McLean, Maureen Tingley, Robert N. Scott, and Jeremy Rickards. 2001. Computer terminal work and the benefit of microbreaks. *Appl. Ergon.* 32, 3 (2001), 225–237.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*. 3111–3119.

Richard D. Morey. 2008. Confidence intervals from normalized data: A correction to cousineau (2005). *Reason* 4, 2 (2008), 61–64.

Douglas W. Oard, Jason R. Baron, Bruce Hedin, David D. Lewis, and Stephen Tomlinson. 2010. Evaluation of information retrieval for e-discovery. *Artificial Intelligence and Law* 18, 4 (2010), 347–386.

Douglas W. Oard and William Webber. 2013. Information retrieval for e-discovery. *Found. Trends Inf. Retriev.* 7, 2–3 (2013), 99–237.

Public Records Act 1958. c. 51. Retrieved from http://www.legislation.gov.uk/ukpga/Eliz2/6-7/51.

Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, Anubrata Das, Asia Biega, Bart Voorn, Claudia Hauff, Damiano Spina, David Lewis, Douglas W. Oard, Emine Yilmaz, Faegheh Hasibi, Gabriella Kazai, Graham McDonald, Hinda Haned, Iadh Ounis, Ilse van der Linden, Jean Garcia-Gathright, Joris Baan, Kamuela N. Lau, Krisztian Balog, Maarten de Rijke, Mahmoud Sayed, Maria Panteli, Mark Sanderson, Matthew Lease, Michael D. Ekstrand, Preethi Lahoti, and Toshihiro Kamishima. 2019. FACTS-IR: Fairness, accountability, confidentiality, transparency, and safety in information retrieval. *SIGIR For.* 53, 2 (December 2019).

David Sánchez and Montserrat Batet. 2016. C-Sanitized: A Privacy model for document redaction and sanitization. *J. Assoc. Inf. Sci. Technol.* 67, 1 (2016), 148–163.

Mahmoud F. Sayed and Douglas W. Oard. 2019. Jointly modeling relevance and sensitivity for search among sensitive content. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 615–624.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.* 34, 1 (2002), 1–47.

Renato Rocha Souza, Flavio Codeco Coelho, Rohan Shah, and Matthew Connelly. 2016. Using artificial intelligence to identify state secrets. *Arxiv:1611.00356* (2016) Retrieved from https://arxiv.org/abs/1611.00356.

Latanya Sweeney. 2002. K-anonymity: A model for protecting privacy. *Int. J. Uncert. Fuzz. Knowl.-Based Syst.* 10, 5 (2002), 557–570.

The National Archives. 2016. The Application of Technology-assisted Review to Born-digital Records Transfer, Inquiries and Beyond. The National Archives. Retrieved from http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf.

Dale Thompson and Michelle Kaarst-Brown. 2005. Sensitive information: A review and research agenda. *J. Assoc. Inf. Sci. Technol.* 56, 3 (2005), 245–257.

Alistair G. Tough. 2018. The scope and appetite for technology-assisted sensitivity reviewing of born-digital records in a resource poor environment: A case study from malawi. In *Handbook of Research on Heritage Management and Preservation*. IGI Global, 175–182.

Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–18.

Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manage.* 36, 5 (2000), 697–716.

Ben James Winer. 1962. *Statistical Principles in Experimental Design*. McGraw-Hill, New York, NY.

Received November 2019; revised July 2020; accepted August 2020