Towards Maximising Openness in Digital Sensitivity Review using Reviewing Time Predictions

Graham McDonald[1], Craig Macdonald[2], Iadh Ounis[2]

University of Glasgow, G12 8QQ, Glasgow, UK
[1]g.mcdonald.1@research.gla.ac.uk
[2]firstname.lastname@glasgow.ac.uk

Abstract. The adoption of born-digital documents, such as email, by governments, such as in the UK and USA, has resulted in a large backlog of born-digital documents that must be *sensitivity reviewed* before they can be *opened* to the public, to ensure that no sensitive information is released, e.g. personal or confidential information. However, it is not practical to review all of the backlog with the available reviewing resources and, therefore, there is a need for automatic techniques to increase the number of documents that can be opened within a fixed reviewing time budget. In this paper, we conduct a user study and use the log data to build models to predict reviewing times for an average sensitivity reviewer. Moreover, we show that using our reviewing time predictions to select the order that documents are reviewed can markedly increase the ratio of reviewed documents that are released to the public, e.g. +30% for collections with high levels of sensitivity, compared to reviewing by shortest document first. This, in turn, increases the total number of documents that are opened to the public within a fixed reviewing time budget, e.g. an extra 200 documents in 100 hours reviewing.

1 Introduction

Sensitivity review is the manual process of reviewing government documents that are to be transferred, or *opened*, to the public domain, to ensure that no *sensitive* information is released, e.g. personal or confidential information. However, existing sensitivity review processes are not practical for the review of born-digital documents, such as email, due to the volume of documents that are created. For example, in the UK, some government departments have reported having a backlog of 190 TB of emails [1][1]. A significant portion of this backlog will be selected for transfer to the public archive and, hence, will need to be sensitivity reviewed.

Technology assisted review (TAR), most notably associated with e-discovery [2], has the potential to alleviate some of the barriers to digital sensitivity review [3]. However, it is generally accepted that all government documents that are to be opened will continue to be manually reviewed until reviewers develop trust in TAR technologies [3]. Moreover, even with the adoption of TAR, the volume of documents to be reviewed is expected to be much greater than the available reviewing time [3] and, therefore, there is a need for strategies to prioritise the review of the documents that are most likely to be released, and to increase the overall number of documents that are opened to the public within the available reviewing time budget.

In this work, we conduct a user study and use the log data to study how government archivists sensitivity review born-digital documents. Moreover, we use the reviewers'

[1] In the UK, fifty government departments are expected to transfer born-digital documents to the public archive by 2021 [1].

interactions to predict the time an average reviewer would require to review a specific document. Furthermore, using simulated collections containing varying distributions of sensitive information, we compare the effectiveness of four ranking strategies for maximising openness within an available reviewing time budget. We show that by ranking documents by their predicted reviewing times, we can markedly increase the mean hourly ratio of reviewed documents that are released to the public (+30% for collections with high levels of sensitivity). This, in turn, will enable government departments to release more of the backlog of documents. For example, on a collection in which 70% of documents contain some portion of sensitive information, for 100 hours of reviewing we expect an extra 200 documents to be released. This could substantially increase the total number of documents that can be opened by each government department.

2 Related Work

Assisting the sensitivity review of digital government documents has received some attention in the literature in recent years [4–9]. Most of that work has focused on developing classification algorithms for identifying sensitivity, either at the document level [5, 7] or sensitive text within documents [9]. Berardi *et al.* [8] investigated improving the cost-effectiveness of sensitivity reviewers by deploying a utility-theoretic [10] *semi-automatic* text classification approach to identify a ranking strategy that can maximise the overall classification effectiveness when a reviewer corrects a portion of mis-classified documents, i.e. to minimise the number of mis-classified documents released to the public (when a portion of the released documents are not manually reviewed) by having reviewers review the documents that are most likely to be mis-classified.

Differently from the work of Berardi *et al.*, in this work, we model the time a reviewer is likely to take to review a document, to increase the number of documents that can be released to the public within a fixed reviewing time budget, when all documents that are released must first be manually reviewed.

Predicting reviewing times is a complex task, as there are many variables that can lead to large variations in reviewing times, such as document length, the complexity of documents or a reviewer's reading speed. Jethani and Smucker [11] modeled the average time to review as a function of document length. In that work, the authors learned a linear model to predict reviewing times and found that the model accounted for 26% of the variance in reviewing times, when a reviewer had to review an entire document to make a decision (as is the case for sensitivity review). This is a relatively good result since, in [11], there is a large variance in the times taken to judge documents of similar lengths. In this work, we also use a linear model to predict document reviewing times. However, differently from Jethani and Smucker, we use the reviewing time predictions to select effective ranking strategies for technology assisted digital sensitivity review.

Damessie *et al.* [12] used a reviewer's dwell time, i.e. the time from a reviewer first viewing a document until the reviewer records a relevance judgment, to study the relationship between the time taken to assess relevance and 1) topic difficulty, 2) the degree of relevance and 3) the presentation order. To normalise for the differences in the reading speeds of reviewers, they proposed *normalised dwell time* (NDT) to measure the reviewing time of an average reviewer. Differently from Damessie *et al.* [12], in this work, we use NDT to predict the number of documents an average reviewer can review within a fixed reviewing time budget and, moreover, to maximise the number of documents that are opened to the public within the available budget.

Table 1. The generated test collection. Document length is measured by number of words. Reviewing time and Normalised Dwell Time (NDT) are measured in seconds.

	docs	%sensitive	Avg. Length	Avg. Review Time	Avg. NDT
Training Data	184	9.63	824.6	321.05	297.88
Test Data	181	17.4	710.3	385.77	333.38

3 Digital Sensitivity Reviewer Study

Study Design and Participants: 16 volunteers from the official UK government archive were asked to sensitivity review a collection of digital government documents. The volunteers were familiar with sensitivity review, however, they were provided detailed guidance regarding 1) the scope of the task that they were being asked to perform and 2) the software deployed in the task to collect sensitivity reviews.

The collection used in the study contains real sensitivities, as defined by the UK Freedom of Information Act. Reviewers were asked to identify any documents containing personal information or international relations sensitivities[2]. In addition to recording judgements at the document level, reviewers were asked to annotate any sensitive text in a document. Non-sensitive documents could simply be identified as such.

Reviewers were provided a web-based interface to navigate the collection and record sensitivities. To ascertain the duration taken to review, we logged the time when a document was loaded to view, t_0, and when a judgement was saved, t_1. The reviewing time, rt, for a document, d, is then calculated as $rt(d) = t_1 - t_0$. Judgements could also be revisited. For revisited documents, we calculate reviewing time as $rt(d) = \sum_{i=1}^{n} t_{1i} - t_{0i}$, where n is the number of times the document was viewed and judged.

461 documents were reviewed in total. 62 documents were judged as being sensitive and 399 as not-sensitive. The mean number of documents reviewed by a reviewer was 28.8, with a range of 5 to 199 and standard deviation of $\sigma = 45.4$.

Generated Test Collection: We use the collected reviews to generate a test collection for developing our models. To ensure that reviewers had committed to the task, we only included reviews from reviewers who 1) made at least 10 judgements, and 2) recorded sensitivity annotations. This resulted in 11 reviewers contributing to the test collection. Additionally, since we could not control for reviewers taking breaks, we removed documents that took longer than 2 hours to review.

Each reviewer's reviews were ordered by the order that they were judged and we then split the reviews so that the first 50% of a reviewer's reviews contribute to the training data and the later 50% contribute to the test data. Table 1 provides an overview of the training and test data for the generated test collection.

4 Predicting Reviewing Time

Developing the reviewing times prediction model: As a measure of the time that an average reviewer would be expected to take to review a particular document, we deploy an approach proposed by Damessie *et al.* [12] that accounts for variations in reviewers reading speeds, namely *normalised dwell time* (NDT). The NDT for a document, d, is defined as $NDT = exp^{(\log(time)+\mu-\mu_\alpha)}$, where $\log(time)$ is the log of the time

[2] Section 40 and Section 27 are representative of the most frequent types of sensitivities in UK government documents. 92% of paper *records* (i.e. documents, photographs, etc.) that were closed between 10/02/05 and 30/04/14 were closed due to Personal or National Interest sensitivities [1].

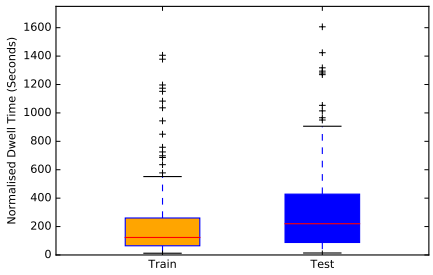

Fig. 1. Normalised Dwell Time (NDT) distributions in seconds for the training and test data.

Table 2. Reviewing Time Predictions. R^2, adjusted R^2 (R^2_{Adj}) and root mean squared error (RMSE) for the test data predictions.

Feature Set	R^2	R^2_{Adj}	RMSE
Decision	0.0537	0.0483	297.72
Surface	0.1095	0.0942	288.81
Complexity	-0.0639	-0.0822	315.68
Decision+Surface	0.2599	**0.2385**	263.29
Decision+Complexity	0.0898	0.0635	291.97
Surface+Complexity	0.1087	0.0722	288.94
All Features	**0.2714**	0.2326	**261.23**

taken to review d, μ is the global mean reviewing time calculated over all documents for all reviewers, and μ_α is the mean reviewing time for the reviewer who reviewed d. However, since calculating NDT relies on the means μ and μ_α, we learn a linear regression model to predict a document's NDT using three sets of features, as follows:

The first set of features represent aspects of a reviewer's *decision* process when making a sensitivity judgement: 1) the number of documents that a reviewer has reviewed prior to the current document; and 2) whether the document is sensitive or not[3]. The second set of features are document *surface* features: 1) the number of sentences in a document; 2) total prepositions, such as *at*, *with* or *from*; 3) total number of syllables; and 4) the ratio of unique words / total words.

The last set of features that we test are standard readability metrics that represent the *complexity*, or reading difficulty, of a document: 1) Simple Measure of Gobbledygook (SMOG) [13] is a simple readability metric based on the number of polysyllabic words per sentence within a 30-sentence sample from a document; 2) the Automated Readability Index (ARI) [14] is a weighted sum of the mean words per sentence and the mean number of characters per word; 3) the Coleman-Liau Index [15] is a weighted sum of the avg. number of characters per 100 words and the average number of sentences per 100 words; 4) the Gunning Fog Index [16] is a weighted sum of the avg. sentence length and the percentage of *complex* words. In total, ten features were used to build our reviewing time prediction model.

Model Effectiveness: Table 2 presents the results of our reviewing time predictions. We select root mean squared error (RMSE) as our main metric as it provides an absolute measure of variance, in seconds, for our predictions. Additionally, we report R^2, defined as $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, where y is a document's NDT, $\bar{y}$ is the mean NDT of all documents and $\hat{y}$ is a document's predicted NDT. R^2 measures the amount of variation in the data that is explained by the learned model. It has an upper bound of 1, obtained by a perfect model, and can be negative since the model can be arbitrarily worse. We also report adjusted R^2, $R^2_{Adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$, where n is the number of documents and k is the number of features. R^2_{Adj} enables a fair comparison between models with different numbers of features, i.e. when a new feature is added to a model R^2_{Adj} increases only if the model improves more than would be expected by chance.

[3] In a production environment, when predicting a document's reviewing time, this feature must be supplied by a sensitivity classifier, e.g. [7].

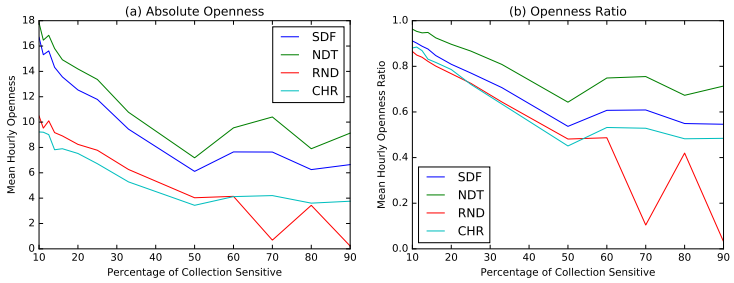

Fig. 2. (a) Number of documents opened per hour. (b) Ratio of reviewed documents opened.

As can be seen from Table 2, deploying all three feature sets results in a RMSE of 261.23 ($\sim$4 mins). 261.23 RMSE provides relatively good predictions since, as can be seen from Figure 1 which presents the distribution of NDT in the training and test data, although the median NDT in the test data is $\sim$200 seconds, there are many outlier documents with NDT in the range of 600 to 1600 seconds and, therefore, the model performs well at predicting the reviewing time for documents that take longer to review. Table 2 also shows that R^2_{Adj} for our model deploying all three feature sets is 0.23, i.e. 23% of the variance in NDT in the test data is explained by the model. This is in line with the 0.26 R^2_{Adj} observed by Jethani and Smucker [11] when reviewers were required to read an entire document to make a relevance judgment. This gives us additional confidence that our model provides relatively good predictions and we, therefore, select this configuration for evaluating ranking strategies in the remainder of this paper.

5 Strategies for Maximising Openness

In this section, we present how our reviewing time prediction model can be used to increase the number of documents that a reviewer can review and release in a given time period, we denote this as the achieved *openness*. Moreover, we evaluate how effective our model is depending on the amount of sensitivity that is in a collection. To do this, we simulate collections with varying distributions of sensitivity by sampling with replacement from the test data to fit the desired sensitivity distribution. We generate nine separate collections, ranging from 10% - 90% sensitive data, where for each collection, C, $\sum_{i=0} NDT(d_i) = 1hour, d_i \in C$. We select one hour as our reviewing time budget, since it is straightforward to reason about larger time periods from this basis. Moreover, to ensure the generalisability of our findings, we generate 100 example collections for each distribution. Therefore, in this section we report mean values over 100 * 1 hour samples of the test data presented in Section 3 and Figure 1.

We evaluate our shortest predicted reviewing time approach (NDT) against three baseline approaches, namely: random (RND); shortest document first (SDF), this strategy naively assumes that shorter documents take less time to review; and chronological (CHR), a strategy currently deployed by sensitivity reviewers.

Figure 2 presents the effectiveness of each of the four ranking strategies on collections of varying sensitivity distributions. Firstly, from Figure 2(a), we note that ordering documents by their expected time to review (NDT), consistently results in more documents being released to the public than the next best approach, i.e. shortest document first (SDF). This shows that the complexities of reviewing a document for sensitivity are not strongly correlated with document length. Secondly, we note that the improvements

in openness are fairly consistent when $< 50\%$ of the collection is sensitive. However, when the collection has high levels of sensitivity, NDT can result in higher relative gains in openness. Figure 2(b) presents the ratio of reviewed documents that were released. As can be seen from Figure 2(b), for a collection that is 60%-70% sensitive, NDT results in a 30% increase in the ratio of reviewed documents that are actually opened, e.g. for a collection in which 70% of documents contain some portion of sensitive information our NDT ranking strategy would result in an extra 200 documents being released for 100 hours of reviewing time. This, in turn, will enable government departments to substantially reduce the backlog awaiting review by increasing the total number of documents that can be opened to the public within the available reviewing time budget.

6 Conclusions

In this work, we presented an approach for predicting the time taken to sensitivity review digital government documents. Moreover, we showed that by using these reviewing time predictions to select the order that documents are presented to reviewers, we can notably increase the rate at which documents are released to the public. Presenting documents based on their predicted reviewing times resulted in a 30% increase in the proportion of reviewed documents that were released when the collection contained 60%-70% sensitive documents. As future work, we will expand this approach to meet other reviewing objectives, such as quickly identifying specific types of sensitivity.

References

1. TNA: The digital landscape in government 2014-2015: business intelligence review (2016)
2. Oard, D.W., Baron, J.R., Hedin, B., Lewis, D.D., Tomlinson, S.: Evaluation of information retrieval for e-discovery. Artificial Intelligence and Law **18**(4) (2010) 347–386
3. TNA: The application of technology-assisted review to born-digital records transfer (2016)
4. Gollins, T., McDonald, G., Macdonald, C., Ounis, I.: On using information retrieval for the selection and sensitivity review of digital public records. In: Proc. PIR@SIGIR. (2014)
5. McDonald, G., Macdonald, C., Ounis, I., Gollins, T.: Towards a classifier for digital sensitivity review. In: Proc. ECIR. (2014)
6. Elragal, A., Päivärinta, T.: Opening digital archives and collections with emerging data analytics technology: A research agenda. Tidsskriftet Arkiv **8**(1) (2017)
7. McDonald, G., Macdonald, C., Ounis, I.: Enhancing sensitivity classification with semantic features using word embeddings. In: Proc. ECIR. (2017)
8. Berardi, G., Esuli, A., Macdonald, C., Ounis, I., Sebastiani, F.: Semi-automated text classification for sensitivity identification. In: Proc. CIKM. (2015)
9. McDonald, G., Macdonald, C., Ounis, I.: Using part-of-speech n-grams for sensitive-text classification. In: Proc. ICTIR. (2015)
10. Berardi, G., Esuli, A., Sebastiani, F.: A utility-theoretic ranking method for semi-automated text classification. In: Proc. SIGIR. (2012)
11. Jethani, C.P., Smucker, M.D.: Modeling the time to judge document relevance. In: Proc. SIGIR. (2010)
12. Damessie, T.T., Scholer, F., Culpepper, J.S.: The influence of topic difficulty, relevance level, and document ordering on relevance judging. In: Proc. ADCS. (2016)
13. Mc Laughlin, G.H.: SMOG grading - a new readability formula. Journal of reading **12**(8) (1969) 639–646
14. Senter, R., Smith, E.A.: Automated readability index. Technical report, DTIC (1967)
15. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology **60**(2) (1975) 283
16. Gunning, R.: The technique of clear writing. McGraw-Hill, New York (1952)