A Framework for Technology-Assisted Sensitivity Review: Using Sensitivity Classification to Prioritise Documents for Review

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

Abstract

More than a hundred countries implement freedom of information laws. In the UK, the Freedom of Information Act 2000 [1] (FOIA) states that the government's documents must be made freely available, or *opened*, to the public. Moreover, all central UK government departments' documents that have a historic value must be transferred to the The National Archives (TNA) within twenty years of the document's creation. However, government documents can contain *sensitive* information, such as personal information or information that would likely damage international relations if it was opened. Therefore, all government documents that are to be publicly archived must be *sensitivity reviewed* to identify and *redact* the sensitive information. However, the lack of structure in digital document collections and the volume of digital documents that are to be sensitivity reviewed mean that the traditional manual sensitivity review process is not practical for *digital sensitivity review*.

In this thesis, we argue that *sensitivity classification* can be deployed to *assist* government departments and human reviewers to sensitivity review born-digital government documents. However, classifying sensitive information is a complex task, since sensitivity is context-dependent and can require a human to judge on the likely effect of releasing the information into the public domain. Moreover, sensitivity is not necessarily topic-oriented, i.e., it is usually dependent on a combination of what is being said and about whom.

Through a thorough empirical evaluation, we show that a text classification approach is effective for sensitivity classification and can be improved by identifying the vocabulary, syntactic and semantic document features that are reliable indicators of sensitive or non-sensitive text [2]. Furthermore, we propose to reduce the number of documents that have to be reviewed to learn an effective sensitivity classifier through an active learning strategy in which a sensitivity reviewer redacts any sensitive text in a document as they review it, to construct a representation of the sensitivities in a collection [3].

With this in mind, we propose a novel framework for technology-assisted sensitivity review that can prioritise the most appropriate documents to be reviewed at specific stages of the

sensitivity review process. Furthermore, our framework can provide the reviewers with useful information to assist them in making their reviewing decisions. We conduct two user studies to evaluate the effectiveness of our proposed framework for assisting with two distinct digital sensitivity review scenarios, or *user models*. Firstly, in the *limited review* user model, which addresses a scenario in which there are insufficient reviewing resources available to sensitivity review all of the documents in a collection, we show that our proposed framework can increase the number of documents that can be reviewed and released to the public with the available reviewing resources [4]. Secondly, in the *exhaustive review* user model, which addresses a scenario in which all of the documents in a collection will be manually sensitivity reviewed, we show that providing the reviewers with useful information about the documents that contain sensitive information can increase the reviewers' accuracy, reviewing speed and agreement [5].

This is the first thesis to investigate automatically classifying FOIA sensitive information to assist digital sensitivity review. The central contributions are our proposed framework for technology-assisted sensitivity review and our sensitivity classification approaches. Our contributions are validated using a collection of government documents that are sensitivity reviewed by expert sensitivity reviewers to identify two FOIA sensitivities, namely *international relations* and *personal information*. Our results demonstrate that our proposed framework is a viable technology for assisting digital sensitivity review.

Supervisors Prof. Iadh Ounis (University of Glasgow), Dr. Craig Macdonald (University of Glasgow)
Available from: `http://theses.gla.ac.uk/41076`

References

[1] Freedom of Information Act 2000. c. 36. `https://www.legislation.gov.uk/ukpga/2000/36/contents`, 2000. HMSO.

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

[3] Graham McDonald, Craig Macdonald, and Iadh Ounis. Active learning strategies for technology assisted sensitivity review. In *Proceedings of The 40th European Conference on Information Retrieval*, pages 439–453. Springer, 2018.

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

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