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

Graham McDonald

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow


March 2019

# Abstract

More than a hundred countries implement freedom of information laws. In the UK, the Freedom of Information Act 2000 (c. 36) (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, for example the minutes from significant meetings, 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 the international relations of the UK if it was opened to the public. Therefore, all government documents that are to be publicly archived must be *sensitivity reviewed* to identify and *redact* the sensitive information, or *close* the document until the information is no longer sensitive. Historically, government documents have been stored in a structured file-plan that can reliably inform a sensitivity reviewer about the subject-matter and the likely sensitivities in the documents. 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 the automatic classification of documents that contain sensitive information, *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. For example, identifying if information is sensitive or not 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. Furthermore, the vocabulary and entities that are associated to particular types of sensitive information, e.g., confidential information, can vary greatly between different collections.

We propose to address sensitivity classification as a text classification task. Moreover, through a thorough empirical evaluation, we show that text classification 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. 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.

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 review process. Furthermore, our framework can provide the reviewers with useful information to assist them in making their reviewing decisions. Our framework consists of four components, namely the *Document Representation*, *Document Prioritisation*, *Feedback Integration* and *Learned Predictions* components, that can be instantiated to learn from the reviewers' feedback about the sensitivities in a collection or provide assistance to reviewers at different stages of the review. In particular, firstly, the Document Representation component encodes the document features that can be reliable indicators of the sensitivities in a collection. Secondly, the Document Prioritisation component identifies the documents that should be prioritised for review at a particular stage of the reviewing process, for example to provide the sensitivity classifier with information about the sensitivities in the collection or to focus the available reviewing resources on the documents that are the most likely to be released to the public. Thirdly, the Feedback Integration component integrates explicit feedback from a reviewer to construct a representation of the sensitivities in a collection and identify the features of a reviewer's interactions with the framework that indicate the amount of time that is required to sensitivity review a specific document. Finally, the Learned Predictions component combines the information that has been generated by the other three components and, as the final step in each iteration of the sensitivity review process, the Learned Predictions component is responsible for making accurate sensitivity classification and expected reviewing time predictions for the documents that have not yet been sensitivity reviewed.

In this thesis, we identify two realistic digital sensitivity review scenarios as user models and conduct two user studies to evaluate the effectiveness of our proposed framework for assisting digital sensitivity review. 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. 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 in the collection that contain sensitive information can increase the reviewers' accuracy, reviewing speed and agreement.

This is the first thesis to investigate automatically classifying FOIA sensitive information to assist digital sensitivity review. The central contributions of this thesis 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*. The thesis draws insights from a thorough evaluation and analysis of our proposed framework and sensitivity classifier. Our results demonstrate that our proposed framework is a viable technology for assisting digital sensitivity review.

# Acknowledgements

First and foremost I would like to thank my supervisors Iadh Ounis and Craig Macdonald for their endless support and inspiration, for the rigorous and lively debates that kept the work moving forward and for teaching me how to write better. Most importantly, without whom, and their attention to detail, I would not be in the position that I am now submitting this thesis.

A great deal of gratitude is due to my external supervisors from The National Archives. Firstly, to Tim Gollins who's foresight and determination started this ball rolling. I am very happy to be able to submit this work in recognition of your efforts that made the undertaking of this thesis possible. Secondly, to Simon Lovett, who's invaluable experience and insights about sensitive information brought life to the subject. Thirdly, to David Willcox who's boundless enthusiasm for the project meant that no operational difficulty was ever an obstacle. Lastly, to Anthea Seles and Anna Sexton for stepping into the role to make sure that there was always support from The National Archives. It has been a real pleasure working with all of you.

I would like to express my sincere gratitude to all the anonymous reviewers at The National Archives, the Foreign and Commonwealth Office and beyond who played such a vital role in constructing the test collection for this research. This thesis would not have been possible without all of your efforts.

I would also like to thank Douglas W. Oard and Mary Ellen Foster for their insights and thoughtful feedback during my PhD viva. Thanks also to all my friends and colleagues in the TerrierTeam and The University of Glasgow who I have had the pleasure of working with, or just hanging out with, over the years. You have made this a very enjoyable endeavour.

Lastly, I would like to give an extra big thank you to: Mom and John, for always encouraging me to pursue whatever interests me in life; Anita and Martin for all the fun times and fantastic food over the years, you have kept me slightly sane; and Anita for being a true life-long friend.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Introduction

More than a hundred countries around the world implement laws to provide the public a right to access information that has been produced by any public body within the country, e.g., the government (The Centre for Law and Democracy, 2016). Moreover, at least fifty nine of these countries have written their citizen's access to information rights, commonly known as *freedom of information* (FOI), into the county's constitution (Right2INFO.org, 2016). In the United Kingdom (UK), freedom of information laws are enacted through the Freedom of Information Act 2000 (c. 36) (FOIA), with the supplemental Freedom of Information (Scotland) Act 2002 (asp. 13) covering public bodies that are under the jurisdiction of the Scottish Parliament.

FOIA provides members of the public with a general right to access information that is held by public bodies. This right to access is facilitated through two mechanisms. Firstly, FOIA enables anyone to request any specific information that is held by a public sector organisation. The second method of accessing public information in the UK applies to historical public records (e.g., documents, photographs, audio and video etc.). The Public Records Act 1958 (c. 51)[1] legislates that all records that are of historical value, i.e., public records, are to be transferred to a designated public archive. For example, public records from central UK government departments are transferred to The National Archives[2] (TNA) for permanent preservation, within twenty years from the document's creation date (Constitutional Reform and Governance Act 2010, c. 25). Public access to the information within these records is governed by the FOIA.

There is an assumption of *openness* within the FOIA, i.e., it is assumed that all of the information within public records will be made available to the public. However, government documents can contain *sensitive* information, such as personal information or information that would likely damage the national security or international relations of a country if the information was

[1]The UK also has acts for: Scotland, Public Records (Scotland) Act 2011 (asp. 12); Wales, Government of Wales Act 2006 (c. 32); and Northern Ireland, Public Records Act (Northern Ireland) 1923 (c. 20).

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

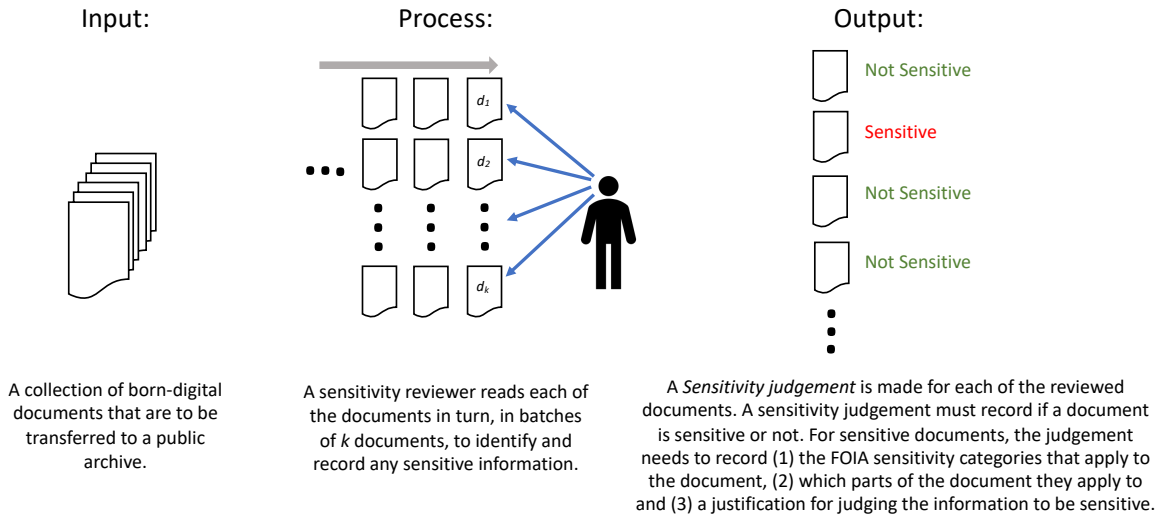



Figure 1.1: The digital sensitivity review input, process and output.

made freely available. Therefore, the FOIA defines twenty five categories of sensitive information that are exempt from the obligation to be released publicly. Moreover, all government documents that are selected to be transferred to a public archive must be *sensitivity reviewed*, to identify any sensitive information and handle it appropriately. For example, the specific information, or the entire document, may be withheld (or *closed*) until it is no longer deemed to be sensitive.

The sensitivity review of paper-based government documents is a manual process. However, it has been widely recognised that there is a need for new technologies to automatically identify sensitive information in *born-digital*[3] government documents to *assist* with the sensitivity review of digital government documents (Allan, 2015; Defense Advanced Research Projects Agency, 2010; Lain, 2013; The National Archives, 2016*a*).

*Digital Sensitivity Review* (DSR) is the process of reviewing a collection of digital government documents to identify any sensitive information in the documents. Figure1.1 illustrates the input, process and the output of the task. The input to the process is a collection of digital documents, $D$, that are to be transferred to a public archive, such as TNA. A sensitivity reviewer reads each document, $d_i \in D$, in turn, in batches of $k$ documents, and records a *sensitivity judgement*, $j_i$, for each of the documents. A sensitivity judgement, $j_i$, needs to record if the document is *sensitive* or *not sensitive*. For a document that is judged as being sensitive, the sensitivity judgement must also include: (1) the FOIA sensitivity categories that apply to the document; (2) which parts (e.g., sentences or paragraphs) of the document they apply to; and (3) a textual description, or justification, about why the information is sensitive and should, therefore, be closed. The output of the DSR process is the collection of reviewed documents and the set of sensitivity judgements, $J$, where for each document, $d_i$, there is a corresponding sensitivity judgement, $j_i$.

[3]Born-digital documents are documents that are originally created in a digital format, such as emails, as opposed to documents that were originally created in a paper format and subsequently digitised, such as scanned or photocopied documents. For brevity, in the remainder of this thesis we will refer to them as digital documents.

Digital sensitivity review is, therefore, an iterative process. A framework to assist digital sensitivity review can, therefore, make decisions, for example, about the order in which (batches of $k$) documents should be reviewed, the information that is provided to a reviewer along with a document (to assist their reviewing task) and the feedback that the reviewer should provide the system about the documents that they review. Moreover, by having the framework make decisions, such as the order that documents are reviewed, digital sensitivity review can be seen as a technology-assisted review (TAR) process. In TAR, a human reviewer and an Information Retrieval (IR) system actively work together to identify and label documents as being either relevant (i.e., sensitive) or not relevant (i.e., not sensitive) to a particular information need (Cormack & Grossman, 2014). In this thesis, we investigate methods for automatically identifying documents that contain sensitive information (we refer to this task as sensitivity classification). Moreover, we show how sensitivity classification can be integrated into a framework to *assist* sensitivity reviewers, and government departments, with the sensitivity review of digital government documents to increase the number of documents that can be reviewed and released to the public, hereby supporting openness. We refer to this as technology-assisted sensitivity review.

## 1.2 Motivation

All documents that are public records of central UK government departments, such as the Foreign and Commonwealth Office[4] (FCO) or the Ministry of Defence[5] (MoD), must be transferred to a designated public archive, such as TNA, for preservation. However, these documents must first be sensitivity reviewed to identify and handle any sensitive information appropriately.

Paper-based government documents are typically stored, and transferred to the archives, within a structured logical file-plan in which each file contains documents that are related to each other. The sensitivity review process for paper documents typically requires a manual review of each file within the file-plan. The contents of the file is checked and if a document contains sensitive information the document is either partially *redacted* or it is removed from the file. If the file is seen to contain a large amount of sensitive information then the whole file is closed (Moss & Gollins, 2017), removing the need to review the documents in the file individually. For many central UK government departments, such as the FCO, this process requires expert reviewers with a good knowledge of the expected sensitivities within the department (The National Archives, 2017).

The sensitivity review process for paper documents is, therefore, a well defined process. However, over the last twenty to thirty years, government departments have been increasingly using digital documents, such as emails, word processing documents, PDFs, web pages and on-line discussions, instead of paper documents. These digital documents are considered to be

[4] https://www.gov.uk/government/organisations/foreign-commonwealth-office
[5] https://www.gov.uk/government/organisations/ministry-of-defence

Table 1.1: The Freedom of Information Act 2000: Categories of exempt information.

| | |
|---|---|
| 21. Information accessible to applicant by other means | 33. Audit functions. |
| 22. Information intended for future publication. | 34. Parliamentary privilege. |
| 22A. Research | 35. Formulation of government policy, etc. |
| 23. Information supplied by, or relating to, bodies dealing with security matters. | 36. Prejudice to effective conduct of public affairs. |
| 24. National security. | 37. Communications with Her Majesty, etc. and honours. |
| 25. Certificates under ss. 23 and 24: supplementary provisions. | 38. Health and safety. |
| 26. Defence. | 39. Environmental information. |
| 27. International relations. | 40. Personal information. |
| 28. Relations within the United Kingdom. | 41. Information provided in confidence. |
| 29. The economy. | 42. Legal professional privilege. |
| 30. Investigations and proceedings conducted by public authorities. | 43. Commercial interests. |
| 31. Law enforcement. | 44. Prohibitions on disclosure. |
| 32. Court records, etc. | |

public records and, therefore, should also be assessed for archiving within twenty years of their creation. In the UK, twenty one government departments are due to transfer digital records to TNA in 2018, this number will rise over the coming years to fifty government departments due to transfer in 2021 (The National Archives, 2016*b*).

However, the process of sensitivity reviewing digital government documents is not yet well defined and, moreover, the paper-based sensitivity review process is not suitable for digital documents, mainly due to two aspects of how digital documents are produced:

1. **Loss of Structure:** Digital documents are not systematically filed in a master file-plan. Therefore, for digital documents, there is no equivalent to the structured logical file-plan of paper documents (Moss & Gollins, 2017). For example, email threads can splinter and create orphan threads (Kerr, 2003). Therefore, sensitivities that may have been contained within a single paper file can be distributed throughout a collection of digital documents.

2. **Volume:** The volume of digital documents that are created means that governments cannot recruit enough resources to sensitivity review all documents that are selected to be archived (The National Archives, 2016*a*). For example, some government departments have reported having up to 190 TB of emails in email servers that will need to be considered for archiving (The National Archives, 2016*b*). A proportion of these emails will be selected to be publicly archived and, therefore, will need to be sensitivity reviewed.

For digital sensitivity review to be effective, sensitivity reviewers need to be assisted by automatic techniques that can address the problems of: 1. the *loss of structure* in how documents are created and stored and 2. the *volume* of documents that are to be reviewed. In this thesis, we propose to address the problem of the loss of structure by automatically *identifying sensitive information* within a document collection to inform the reviewers about which of the documents are likely to contain sensitive information. We propose to address the problem of the volume of digital documents that are to be reviewed by *increasing reviewer productivity*, for example by enabling reviewers to review documents more quickly or prioritising documents for review based on the likelihood that the documents (do not) contain sensitive information, to maximise

the number of documents that can be released with the available reviewing resources.

**Identifying Sensitive Information:** Automatically identifying sensitive information in documents is, therefore, a fundamental task for assisting the digital sensitivity review process. Sensitive information can be defined as information that would likely cause harm to, or prejudice the interests of, a person group or organisation, if the information was made freely available to the public. Table 1.1 presents the twenty five categories of information that are exempt from public release through the Freedom of Information Act 2000 (c. 36). The categories in Table 1.1 are broadly defined. However, each of the categories (except for *21. Information accessible to applicant by other means*) are exempt from public release due to the sensitive nature of the information. To illustrate this point, we shall now consider a scenario relating to exemption 27 (international relations).

Many government operations and decisions are conducted at the international level, for example on matters concerning environmental issues, criminal investigations or counter terrorism activities. However, governments cannot be expected to conduct and maintain these activities without having the privilege of being able to discuss privately, for example, cultural differences between negotiating parties. Moreover, within this scenario it may be necessary for government officials to express opinions that could be offensive to the other parties. There is an expectation within the FOIA that documents relating to the decisions taken within a negotiation will be opened to the public. However, if the *offensive opinions* were released into the public domain it would likely corrode the trust between the parties (i.e., the governments) and damage the relations between the countries, making such negotiations less likely in the future.

This scenario illustrates two important points about sensitive information. Firstly, sensitivity is not *topic-oriented*. It is not the fact that information in the documents is about the negotiations, or the countries involved, that makes the information sensitive. Rather, the sensitivity arises due to the context of the information, e.g., *who said what about whom*. Moreover, a judgement on whether the information is sensitive or not is dependent on the expected *effects* from making the information freely available (Ackerman & Sandoval-Ballesteros, 2006). Secondly, identifying sensitive information (and, moreover, sensitivity review) is a recall-oriented task. There may be many offensive comments, or other sensitive pieces of information, within a collection of documents. However, it only requires one of these pieces of information to be opened to the public for the information to cause harm. Therefore, a reviewer must identify *all* of the sensitive information in the documents that are to be reviewed and released to the public.

**Increasing Reviewer Productivity:** All documents that are opened to the public must first be sensitivity reviewed and the volume of digital documents is expected to be very significantly larger than the volumes that have been seen for paper documents (Allan, 2015). There are many different types and sizes of collections that need to be sensitivity reviewed, from email collec-

tions to public enquiries (Inquiries Act 2005, c. 12). Since governments are not expected to be able to recruit enough reviewing resources to sensitivity review all digital documents (The National Archives, 2016*a*) and, moreover, the volumes of sensitivities vary greatly between different collections and government departments (Allan, 2014), the digital sensitivity review task will likely have different priorities, depending on the available resources and the collections that are being reviewed.

For example, public inquiries (Inquiries Act 2005, c. 12), such as the Iraq Inquiry (Chilcot, J, 2016) or the Al Sweady Inquiry (Forbes, 2014), often deal with a lot of information that is sensitive. Moreover, the information that is created in public inquiries is predominantly in digital documents (The National Archives, n.d.). All of the documents that are created or used in a public inquiry must be sensitivity reviewed before the end of the inquiry, so that they can be transferred to TNA or retained if they are highly sensitive. Therefore, when sensitivity reviewing public inquiries, the main priority is that the review can be carried out quickly so that the public sees that the outcome of the inquiry is being published in a timely manner. However, the order in which documents are reviewed may be less important since the whole collection must be reviewed before the documents can be released.

The order in which documents are reviewed is most likely to be important for reviewing the day-to-day documents of government departments. There is a large backlog of paper documents that are awaiting review in some government departments (Allan, 2014). For example, the Department for Culture, Media and Sport[6] (DCMS) has a backlog of roughly twenty two thousand paper files awaiting review (The National Archives, 2018), although it is hard to make an accurate assessment of the size of the DCMS backlog (Allan, 2014). Other departments that have reported having a reviewing backlog include the Department for Business Innovation and Skills[7] (BIS), the Department for Transport[8] (DfT) and the FCO. Moreover, the FCO, the MoD, the Home Office[9] and the Cabinet Office[10] have a large volume of records containing potentially sensitive material (Allan, 2014). With the increased volumes of digital documents, and shortfall in reviewing resources, this situation is expected to be worse for digital sensitivity review. For departments that have a backlog and are, therefore, not meeting the time-to-transfer obligation of the Public Records Act 1958 (c. 51), ensuring that reviewing can be conducted quickly is clearly a priority. However, to comply with the Public Records Act 1958 (c. 51) a priority is also to release as many of the documents as possible. Therefore, prioritising the documents that are most likely to be the quickest to review and, importantly, that are the most likely to be released to the public will increase the overall productivity of the review process.

The additional pressure to review quickly, that is likely to result from dealing with the back-

[6] https://www.gov.uk/government/organisations/department-for-digital-culture-media-sport
[7] https://www.gov.uk/government/organisations/department-for-business-innovation-skills
[8] https://www.gov.uk/government/organisations/department-for-transport
[9] https://www.gov.uk/government/organisations/home-office
[10] https://www.gov.uk/government/organisations/cabinet-office

log of documents that are to be reviewed, has the potential to negatively impact the quality of the human reviews. Sensitivity reviewing is a labour-intensive process, that requires a judgement about the effect of releasing the information, and one where mistakes are bound to occur (Allan, 2014). Moreover, we have observed that there is only moderate agreement between sensitivity reviewers (McDonald *et al.*, 2014). Indeed, high levels of reviewer disagreement have been observed in a studies about assessing a document's relevance (Scholer *et al.*, 2011; Voorhees, 1998; Webber, 2011). Although assessing if a document is sensitive or not is not exactly the same task as assessing for (topical) relevance, they are both tasks that assess if a document is related to a particular category. Assessing sensitivity is complex and the difficulty of assessing relevance has been shown to have the greatest impact on assessor agreement (Oard *et al.*, 2010). Therefore, a framework that can increase the agreement between reviewers can also increase productivity, since increased agreement will lead to less time and resources being allocated to discussing reviews that are disputed, either by senior sensitivity reviewers in departments such as the FCO or by The Advisory Council[11] when a department applies to have the information closed.

## 1.3 Scope of the Thesis

The transition to digital government documents brings a very wide range of challenges, from human-computer interaction factors to the efficient storage and distribution of documents and reviews. This thesis is a first investigation into how Information Retrieval (IR) and Text Classification (TC) technologies can be deployed to assist with the sensitivity review of digital government documents. The scope of this thesis is bounded by, what we argue to be, the three essential elements of technology-assisted review that need to be addressed when developing a framework to assist with digital sensitivity review. Moreover, we argue that they are essential to be able to ensure that the framework can enable a sensitivity reviewer to review documents more quickly and/or release more documents to the public:

1. **Sensitivity Classification:** A reliable method of modelling sensitive information that is generalisable enough to automatically classify documents by whether they do or do not contain sensitive information that is exempt from being released to the public, due to a Freedom of Information exemption.

2. **Sensitivity Identification:** An effective method of quickly identifying the sensitivities that are in a collection so that an efficient sensitivity classifier can be developed while using minimal reviewing resources. The logical file plan of paper records means that, for paper-based sensitivity review, the topic of a file, i.e., what the documents in the file are

[11]The Advisory Council is the independent body that considers applications for the retention or closure of government records. http://www.nationalarchives.gov.uk/about/our-role/advisory-council

about, can give a reviewer an idea if the documents are likely to contain sensitive information and, if so, what the sensitivities will be (Allan, 2015). However, for digital sensitivity review, the lack of a structured file plan means that the sensitivities are dispersed throughout a collection. Therefore, it is not clear what the sensitivities in a specific collection are, i.e., what they are related to or the vocabulary that is used in the sensitive text. With this in mind, it is important that a framework to assist sensitivity review can quickly learn from the reviewer what the sensitivities in a collection *look like*. The framework should, therefore, deploy an *active learning* strategy to quickly learn to classify the sensitivities in a collection by integrating explicit reviewer feedback about the sensitivities in a collection as the documents are reviewed.

3. **Reviewing Models:** A strategy, or strategies, for using the developed sensitivity classifier's predictions to increase the effectiveness and efficiency of available reviewing resources. Since this is a first investigation into automatically classifying FOI sensitivities to assist with the sensitivity review of digital government documents, we have identified two realistic user models to investigate how our proposed framework can assist sensitivity review in two different scenarios. Firstly, in our *limited review* user model, we investigate how our framework can assist reviewers when there are not enough reviewing resources to review all of the documents in a collection that is due to be publicly archived. Secondly, in our *exhaustive review* user model, we investigate how our framework can assist reviewers when *all* of the documents in a collection are reviewed.

## 1.4 Thesis Statement

The statement of this thesis is that automatic sensitivity classification can be effective for assisting human reviewers with the sensitivity review of digital government documents. Moreover, an effective sensitivity classifier can be learned by identifying the latent vocabulary, syntax and semantic language features of the sensitive information in a corpus. Furthermore, by deploying an active learning strategy to select specific documents to be reviewed and by having a reviewer annotate, or *redact*, any passages of sensitive text in a document as they review, we can identify the most informative annotated terms to construct a representation of the sensitivities in a collection. Assigning the identified informative terms more weight, or importance, in the classifier will result in fewer documents being required to be reviewed to learn an effective sensitivity classifier.

In particular, sensitivity classification can assist with the sensitivity review of digital government documents by predicting which of the documents contain sensitive information in a collection that is to be reviewed. Moreover, automatic sensitivity classification predictions can be used to prioritise specific documents for review to increase the number of non-sensitive documents that can be reviewed and released to the public when the available reviewing time budget

is insufficient to review all of the documents that are due for release. Furthermore, providing the reviewers with sensitivity classification predictions for the documents that are to be reviewed can increase the reviewers' accuracy, speed and agreement.

## 1.5 Contributions

The main contributions of this thesis are the following. Firstly, we introduce the task of automatically classifying documents that contain sensitive information that is exempt from being publicly released through Freedom of Information Act 2000 (c. 36), i.e., international relations and personal information sensitivities. Secondly, we contribute a deployable end-to-end framework for assisting with the sensitivity review of digital government documents. Our framework consists of four components, namely *Document Representation*, *Document Prioritisation*, *Feedback Integration* and *Learned Predictions* that can be instantiated to perform different tasks at different stages of the review process, depending on the current priorities for assisting the review process. For example, developing an effective sensitivity classifier early in the review process when the sensitivities in a collection are not known or, later in the review process, selecting the specific documents that should be reviewed to increase the number of documents that are released to the public within a limited period of time.

Throughout the digital sensitivity review process, the four components of our framework collaborate to address the three essential elements of technology-assisted review that we identified in Section 1.3. In the course of this thesis, we instantiate various functionalities of our framework's components to propose, develop and evaluate novel approaches, to ensure that our framework can assist a sensitivity reviewer to review documents more quickly and release more documents to the public. Our main contributions with respect to the three essential elements of technology-assisted review are as follows:

1. **Sensitivity Classification:** We propose a fully-automatic approach for discovering latent vocabulary, syntactic and semantic features of sensitive information, that are useful for increasing the accuracy of a sensitivity classification. The approach uses natural language processing and analysis of the semantic relations of the terms in a collection of documents, to construct a document representation for classification. The approach can effectively capture the latent features of the context-dependent FOIA sensitivities *international relations* and *personal information*.

2. **Sensitivity Identification:** Early in the digital sensitivity review process, a reviewer will not know the topics that are being discussed in, or the vocabulary that is used in, or even the entities that are related to the sensitive information in the documents that are to be reviewed. Therefore, we investigate methods for constructing a representation of the sensitivities in the collection to improve the effectiveness of sensitivity classification. We

propose to integrate the process of redacting the sensitive text in a document into the digital sensitivity review process. By having reviewers annotate (or redact) the sensitive text in a document, as they review, our approach measures the expected amount of information that terms from the annotations will provide to the classifier and integrates the most informative terms. We present a thorough analysis of the approach and show that it can reduce the amount of reviewing effort (i.e., time) that is required to learn an effective sensitivity classifier.

3. **Reviewing Model:** We present the results of a user study that demonstrates how automatic sensitivity classification can be deployed to assist reviewers when all of the documents in a collection will be reviewed. Our study shows that sensitivity classification can provide reviewers with additional information about the sensitivities in a document at the time of review to reduce the time that is required to review a collection of documents and increase the agreement between reviewers.

   We also propose a novel approach for prioritising documents to be reviewed to increase the total number of documents that can be reviewed and released to the public when there are not enough reviewing time resources to review all of a collection. The approach models a sensitivity reviewer's behaviour, along with a document's complexity, to predict the amount of time that a reviewer will require to review a specific document. Moreover, the approach prioritises documents that are (1) classified as being not sensitive so that reviewing resources are focused on reviewing documents that will be released to the public, *and* (2) that are predicted to take less time to review so that more documents will be reviewed within the available time. We thoroughly evaluate the approach and show that is can increase the number of documents that can be released to the public within a specific period of time.

Additionally, in the course of this thesis, we provide recommendations for government departments that are looking to implement a technology-assisted review process for digital sensitivity review. Moreover, we propose a roadmap for future directions in this emerging and important area of research.

## 1.6 Origins of Material

Some of the work in this thesis is based on a number of conference publications. The following are our publications that form the basis for research detailed in the following chapters:

- Chapter 3: The possibility of applying Information Retrieval (IR) technologies to address the challenges of archival transfer of digital government documents (Selection, Appraisal and Sensitivity Review) was discussed in our paper published at PIR@SIGIR

2014 (Gollins, McDonald, Macdonald & Ounis, 2014). A broader outline of the characteristics of sensitive information that can be identified as features of sensitivity for developing automatic classifiers was published in FDIA 2015 (McDonald, 2015).

- Chapter 5: Our baseline text classification approach for sensitivity classification was published at ECIR 2014 (McDonald *et al.*, 2014). This work also experimented with extending text classification with additional features from external resources, such as knowledge bases, to identify specific sensitivities.

- Chapter 6: Our approach for deploying part-of-speech (POS) sequences as features for identifying passages of sensitive text in government documents was published in ICTIR 2015 (McDonald *et al.*, 2015). Following from this, the work identifying effective kernels for POS sequence classification approaches for sensitivity classification was published in SIGIR 2017 (Mcdonald, García-Pedrajas, Macdonald & Ounis, 2017). Our methodology for extending text classification with semantic features, and extended text sequences, for sensitivity classification was published in ECIR 2017 (McDonald, Macdonald & Ounis, 2017).

- Chapter 7: Our proposed active learning strategies for integrating reviewer feedback into the classification process and generating knowledge of the sensitivities within a collection through annotation features were published in ECIR 2018 (McDonald *et al.*, 2018*a*).

- Chapter 8: Our proposed approach for modelling reviewing times and prioritising documents for review in our *limited review* reviewer model was published in ECIR 2018 (McDonald *et al.*, 2018*b*).

## 1.7 Outline of Thesis

The remainder of this thesis is organised as follows:

- Chapter 2 provides an introduction to the fundamental concepts in document classification, active learning and technology-assisted review that we build on throughout this thesis. In particular, we first introduce the essential components for developing a document classifier: generating a test collection; representing documents as structured data; feature reduction techniques for improving classification effectiveness; supervised machine learning classifiers that are effective for document classification; and methods for evaluating the effectiveness of document classification. We then discuss active learning strategies for selecting informative documents as classification training data before, finally, providing an introduction to technology-assisted review.

- Chapter 3, firstly, reviews the types of sensitive information that most frequently result in document closures within central UK government and identifies the Freedom of Information (FOI) exemptions (sensitivities) that we focus on in this thesis. The chapter then provides a detailed description of the types of information that are likely to be closed due to the identified sensitivities, before discussing existing works on automatically classifying sensitive information and identifying the existing knowledge gaps for classifying FOI sensitivities. Finally, the chapter presents our methodology for creating the test collection that we use throughout this thesis for developing and evaluating our framework for technology-assisted sensitivity review.

- Chapter 4, firstly, provides details of the exhaustive review and limited review user models that we have identified to evaluate the effectiveness of our proposed framework. The chapter also presents our proposed framework for technology-assisted sensitivity review and describes each if its four components, namely Document Representation, Document Prioritisation, Reviewer Feedback and Learned Predictions.

- Chapter 5 evaluates the effectiveness of a document sanitisation approach from the literature for classifying confidential information in documents. We show that document sanitisation is not an appropriate strategy for identifying Freedom of Information Act 2000 (c. 36) sensitivities. Therefore, in this chapter, we propose to address the task of identifying sensitive information (i.e. sensitivity classification) as a document (text) classification task. Moreover, we present our sensitivity classification baseline that we build on in the remainder of the thesis. Furthermore, we evaluate the effectiveness of classifying individual FOI exemptions compared to classifying sensitive information as a single category and an ensemble classification approach for combining sensitivity classifiers.

- Chapter 6, looks at advanced feature engineering techniques for sensitivity classification. In particular, the chapter presents our empirical evaluation of extended text sequences and part-of-speech sequences for sensitivity classification. Moreover, the chapter presents an empirical evaluation of kernel techniques for sequence classification approaches to sensitivity classification. Furthermore, the chapter presents an empirical analysis of our approach for engineering classification features that capture semantic relations within documents. The chapter evaluates the effectiveness of the engineered features for extending text classification for sensitivity identification.

- Chapter 7 presents an empirical analysis of active learning strategies for integrating reviewer feedback into the sensitivity classification process. Moreover, the chapter presents our approach for generating a representation of the sensitivities within a collection. Furthermore, through empirical analysis, the chapter shows how the generated sensitivity representations can be integrated into the classification process to learn an effective sensitivity classifier more quickly by reducing the amount of reviewing effort required for

training a classifier. Lastly, the chapter empirically evaluates methods for predicting when the developed sensitivity classifier has achieved peak performance.

- Chapter 8 investigates how our proposed framework can assist reviewers in the first of the two user models for technology-assisted sensitivity review that we address in this thesis, namely *limited review*. Limited review addresses a scenario when there are not enough reviewing resources available to review a whole collection of documents. In this chapter, we propose a method for prioritising specific documents for review so that we can maximise the number of documents that can be released to the public with the available reviewing resources. We conduct a user study to analyse reviewing behaviour, such as the time taken to review documents. Moreover, we use the log data from the user study to develop and evaluate our proposed approach for prioritising documents for review. Furthermore, we evaluate how the distribution of sensitive information within the collection and how the sensitivity classification effectiveness affects our proposed approach.

- Chapter 9 investigates how our proposed framework can assist reviewers in the second of the two user models for technology-assisted sensitivity review that we address in this thesis, namely *exhaustive review*. This chapter presents the results of a controlled user study that investigates how our framework can assist reviewers when all the documents in a collection will be reviewed. The study evaluates the impact that 1) classification effectiveness has on reviewer accuracy and agreement, and 2) the classifier's confidence levels has on the level of trust that a reviewer has in the classifier's predictions and the impact that this has on reviewing times.

- Chapter 10 closes the thesis by highlighting the contributions drawn from each of the individual chapters and provides our recommendations for deploying technology-assisted sensitivity review within government departments, before discussing possible further research directions and future work.

# Chapter 2

# Background

## 2.1 Introduction

In this chapter, we provide an overview of the classification, active learning and technology-assisted review techniques that this thesis builds on. Indeed, we describe existing works that our framework relies on as a basis for the automatic classification of documents into known categories, selecting documents to have reviewed with the aim of quickly developing an effective sensitivity classifier, and assisting a human reviewer to meet the goals of their review by prioritising documents for review based on the needs of the reviewer or the operational constraints of the government department. The remainder of this chapter is organised as follows:

- Section 2.2 provides the background to automatic document classification, commonly referred to as *text classification*. We provide preliminary details of the components that are required for text classification, i.e., a document test collection, features for document representation, feature reduction techniques, a classification algorithm that is effective for text classification and, finally, the evaluation of text classification.

- Section 2.3 introduces active learning techniques for integrating a reviewer's feedback to the classification process. This section presents the strategies that we evaluate for enabling the classifier to select the most informative documents to have reviewed and, therefore, to develop an effective sensitivity classifier more quickly. Thereby, reducing the number of documents that are required to be reviewed to learn an effective classifier.

- Section 2.4 provides an overview of technology-assisted review, in which automatic document classification and active learning are combined to assist human reviewers by increasing the reviewer's efficiency and effectiveness when performing a reviewing task. Moreover, in this section, we provide a detailed analysis of the similarities and differences between previous technology-assisted review tasks and technology-assisted sensitivity review.

## 2.2 Text Classification

Automatic document classification, also known as *text classification* or *text categorisation*, has been an active area of research since Maron (1961) published his foundational work in which he claimed, and empirically demonstrated, that statistics about the type, order, location and frequency of words in a document can provide reliable enough *clues* to predict the subject category that a document most probably belongs to.

In general, text classification is the process of automatically assigning to a document, $d$, the category, or *class*, labels, $y_1 .. y_n$, $y_i \in Y$, corresponding to a set of pre-defined classes, $C$, $|C| = n$, that correctly identifies which of the classes the document belongs to. Following from Maron (1961), early approaches to text classification relied on rule-based approaches that required manual construction of the rules of classification, for example:

$$if(condition1)\&\&(condition2) \Longrightarrow y_1 : else \Longrightarrow y_0 \tag{2.1}$$

However, these approaches could only make rigid binary decisions about class membership and importantly, as the available text collections grew in size, were typically difficult to modify (Dumais, 1998). Following on from this, from the 1990's onwards, machine learning approaches to text classification started to gain in popularity (Sebastiani, 2002).

A machine learning document classifier for a specific class of interest, $c \in C$, is automatically built, or *trained*, through a process whereby the classifier learns by observing examples of documents that belong to $c$ (i.e., *positive* examples) and examples of documents that do not belong to $c$, denoted as $\bar{c}$, (i.e., *negative* examples). The classifier identifies statistical patterns of document features that are more frequent, and therefore indicative of, the positive or negative class, to predict the most likely class of a new document based on its previous observations.

In this thesis, we investigate the effectiveness of machine learning classifiers for automatically classifying documents by whether they do, or do not, contain sensitive information. Moreover, the classification tasks that we address in this thesis are *binary* classification tasks, i.e., where $n = 2$. Therefore, in the remainder of this chapter we will limit out discussion to the case of binary classification.

### 2.2.1 Test Collection

Machine learning approaches to document classification are *supervised* tasks. Training, and evaluating, a supervised document classifier relies on there being a collection of documents, $D$, that contains both positive and negative examples of the class of documents that we wish to classify. Moreover, each document, $d_i \in D$, must have a corresponding class label, $y_i \in Y$, that states if the document is either a positive or a negative example. The labelled collection, $D$, usually referred to as a *test collection*, is typically separated into three disjoin sets: a set of training documents, $D_{tr}$, that is used for training the classifier; a validation set of documents, $D_{va}$, that is used to

learn parameters of the classifier or other processes; and a set of documents, $D_{te}$, that are used to evaluate, or *test*, the classifier's performance.

A test collection is typically constructed by having humans assessors read each document in the collection and manually assign the appropriate class label to a document. This process can be very labour intensive and, for complex classes such as sensitive information, can require the assessors to have a certain level of expertise in judging whether a document is a positive or negative example (Grossman & Cormack, 2010; The National Archives, 2017).

Moreover, for complex tasks, judging the class or classes that a document belongs to can be subjective and levels of disagreement between assessors can be relatively high (Cleverdon, 1984). Therefore, it is often standard practice to have documents judged by multiple assessors and to assign to a document the label selected by a majority of the assessors. Kappa measures, such as Kohen's $\kappa$ (Cohen, 1960) or Fleiss $\kappa$ (Fleiss & Cohen, 1973), can be used to measure inter-assessor agreement, providing a quantitative measure of how difficult it was for the assessors to manually classify the documents.

### 2.2.2 Document Representation

In Section 2.2.1, we detailed the importance of constructing a representative test collection for developing and evaluating an automatic document classifier. However, documents in a test collection typically contain unstructured, or semi-structured, text. To be able to use these documents as the basis of a classification system the documents need to be transformed to a structured data representation that is suitable and efficient for the classifier (Song *et al.*, 2005).

The most popular representation of a document collection for document classification is known as the *bag of words* (BOW) representation. In the BOW representation, a document is represented by statistics about the words from the training data vocabulary, $V$, that the document contains. In the following, we illustrate the process of document representation by considering an example document collection and vocabulary, presented in Figure 2.1.

The collection, $D$, in Figure 2.1, containing three documents, $d_1 .. d_3$, is split into training, $D_{tr}$, and test, $D_{te}$, sets. The vocabulary, $V$, is the set of unique terms, $t_1 .. t_n$, $t_i \in D_{tr}$. In the example collection, $n = 15$. In the BOW representation, a document is represented as a feature vector, $\mathbf{x}$, where each feature is a statistic of a term in $V$, and $|\mathbf{x}| = |V|$. This corresponds to the Vector Space Model representation of Salton *et al.* (1975).

There are three methods of generating term statistics that are commonly used for document representations. Firstly, a *binary* representation simply records if a term feature from $V$ is present or absent in the document $d_i$. Secondly, *term frequency* (*tf*) (Salton, 1971) records the frequency of each term feature from $V$ in the document $d_i$. The third statistic, *term frequency inverse document frequency* (TF-IDF) (Salton, 1971), is a *weighted* version of *tf*, where each term in $V$ is weighted by the frequency of the term in the collection. The IDF component, in effect, weights terms by their *discriminative power* / *importance* within the collection. TF-IDF can be

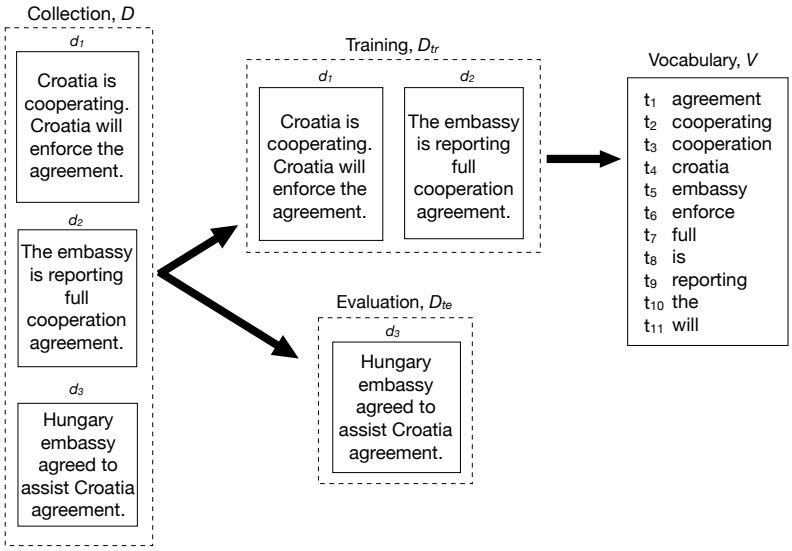



Figure 2.1: An example document collection, $D$, with three documents, $d_1$.. $d_3$. The collection is split into training data, $D_{tr}$, and test data, $D_{te}$. The collection vocabulary, $V$, is all the unique terms in $D_{tr}$.

calculated as:

$$TF\text{-}IDF(t) = tf \cdot log\frac{N}{df_t} \tag{2.2}$$

where $N$ is the number of documents in the training data, $D_{tr}$, and $df_t$ is the number of documents in $D_{tr}$ that contain $t$. Figure 2.2 presents the resulting document representations for document $d_1$ in Figure 2.1, for each of the term statistics: binary, *tf* and TF-IDF.

| | |
|---|---|
| binary | $\mathbf{x}_{d1}$ = [ 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1 ] |
| *tf* | $\mathbf{x}_{d1}$ = [ 1, 1, 0, 2, 0, 1, 0, 1, 0, 1, 1 ] |
| TF-IDF | $\mathbf{x}_{d1}$ = [0.3, 0.3, 0.0, 0.6, 0.0, 0.3, 0.0, 0.0, 0.0, 0.0, 0.3] |

Figure 2.2: The document vector representations, $\mathbf{x}$, for document $d_1$ presented in Figure 2.1, for each of the representation strategies: binary, *tf* and TF-IDF.

The choice of statistical representation can have a notable impact on the classification effectiveness. When comparing text representation schemes for document classification, Song *et al.* (2005) concluded that "*there are strong interactions between text representation factors. And the best text representation schemes are corpus-dependent.*"

The BOW model has many benefits for document classification, such as its simplicity and effectiveness (Aggarwal & Zhai, 2012; Song *et al.*, 2005), however it also has some inherent limitations. Firstly, the number of terms in $V$ can be very large and, hence, can result in document representations with more than tens of thousands of features (Joachims, 1998). This can

result in unmanageable computational complexity (Aphinyanaphongs *et al.*, 2014). Secondly, in the BOW representation, statistics are generated about the occurrence of individual words in a document and, hence, the model results in a loss of information such as the co-occurrence of terms, term proximity or the semantic relationships between terms (Song *et al.*, 2005).

### 2.2.3 Feature Reduction

As previously mentioned in Section 2.2.2, representing documents by the statistics of the distribution of words from the collection vocabulary that the document contains, such as tf or TF-IDF, can result in a document's vector representation being very large, with often hundreds of thousands of features (Joachims, 1998). Very large feature vectors can result in unmanageable computational complexity for the classification algorithms (Aggarwal & Zhai, 2012) and, moreover, can lead to the classifier *over-fitting* to too closely match the characteristics of the training data when the number of features is larger than the number of documents used for training (Joachims, 1998).

Feature reduction reduces the computational complexity of learning over such large feature spaces by identifying and retaining only the most informative term features, while discarding non-informative features, prior to the document representation process presented in Section 2.2.2. However, feature reduction can generally result in a loss of the amount of information that is present in a document representation and, moreover, in text classification tasks there are often very few non-informative terms (Joachims, 1998). Therefore, the choice of what feature reduction technique, if any, to deploy is dependent on the document collection to be classified and the classifier that is deployed.

The simplest feature reduction technique is to remove term features that are *stopwords*. By observing that the frequency of terms in a document collection typically follows a Zipfian distribution, Luhn (1958) noted that high frequency terms that appear very often in many documents, for example prepositions such as "in", "of" or "for", have very low discriminative power and can therefore be discarded without losing significant information.

Another simple method of feature reduction is to reduce each term in a collection to its root form, i.e., *Stemming* (Lovins, 1968; Porter, 1980). For example, plural or past tense forms of a word are formed from the same root and, importantly, are in essence about the same thing (e.g., "stemmer" and "stemmed" are both about the action to "stem"), and therefore these multiple representations can be conflated to a single representation. Stemming can result in a significant reduction in the size of the feature space. However, it can also result in a significant reduction in the amount of information in the document representation (Aphinyanaphongs *et al.*, 2014).

Advanced methods of feature reduction attempt to retain only the term features that are the most important for the classification task by identifying terms that are more correlated to a class distribution (Aggarwal & Zhai, 2012). There are many statistical techniques that can be deployed for advanced feature reduction, for a comprehensive overview see Yang (1995) and Yang &

Pedersen (1997). In the remainder of this section, we present two advanced feature reduction techniques that we deploy in the remainder of this thesis to identify feature distributions that can be indicative of sensitive information, namely *information Gain* (IG) and *Chi-Squared* ($\chi^2$).

In the case of binary classification, the IG or $\chi^2$ for a term, $t$, can be computed from a 2-way contingency table, as illustrated in Table 2.1. The table is constructed from the number of documents in the class $c$ that contain $t$ (A) or that do not contain $t$ (C), and the number of documents in the other class, denoted as $\bar{c}$, that contain $t$ (B) or that do not contain $t$ (D).

Table 2.1: An example 2-way class contingency table.

| | $c$ | $\bar{c}$ | *Total* |
|---|---|---|---|
| Containing $t$ | $A$ | $B$ | $T_t$ |
| Not Containing $t$ | $C$ | $D$ | $T_{\bar{t}}$ |
| *Total* | $T_c$ | $T_{\bar{c}}$ | $N$ |

**Information Gain**: The amount of randomness in the distribution of a term, $t$, within the category, $c$, can be considered as a measure of the amount of information that the term contains and can be calculated as the term's class specific entropy (Shannon, 1948), defined as:

$$H(t,c) = -P(t|c)logP(t|c). \tag{2.3}$$

Information Gain (IG) is a measure of a term's class-specific entropy, with respect to its entropy in the overall collection, and is calculated as follows:

$$\begin{aligned} IG(t) = & -\sum_{i=\{c,\bar{c}\}} P(c_i)logP(c_i) \\ & +P(t)\sum_{i=\{c,\bar{c}\}} P(c_i|t)logP(c_i|t) \\ & +P(\bar{t})\sum_{i=\{c,\bar{c}\}} P(c_i|\bar{t})logP(c_i|\bar{t}) \end{aligned} \tag{2.4}$$

where, referring to the cells in Table 2.1:

$$\begin{aligned} P(c) = \frac{T_c}{N}, P(\bar{c}) = \frac{T_{\bar{c}}}{N}, P(t) = \frac{T_t}{N}, P(\bar{t}) = \frac{T_{\bar{t}}}{N}, \\ P(c|t) = \frac{P(c,t)}{t}, \text{ where } P(c,t) = \frac{A}{N}, \\ P(\bar{c}|t) = \frac{P(\bar{c},t)}{t}, \text{ where } P(\bar{c},t) = \frac{B}{N}, \\ P(c|\bar{t}) = \frac{P(c,\bar{t})}{\bar{t}}, \text{ where } P(c,\bar{t}) = \frac{C}{N}, \\ P(\bar{c}|\bar{t}) = \frac{P(\bar{c},\bar{t})}{\bar{t}}, \text{ where } P(\bar{c},\bar{t}) = \frac{D}{N} \end{aligned} \tag{2.5}$$

**Chi-square Statistic**: Chi-square ($\chi^2$) measures the dependence between a term, $t$, and a class, $c$, by calculating how much the observed frequency of the term diverges from its expected frequency within the collection. Referring to the contingency table cell labels from Table 2.1, the Chi-square score for $t$ is calculated as follows:

$$\chi^2 = \frac{N(AD-BC)^2}{T_t T_{\bar{t}} T_c T_{\bar{c}}} \tag{2.6}$$

An important benefit of $\chi^2$ is that it is a normalised value and, hence, $\chi^2$ scores are comparable for terms within and across classes. However, this advantage does not hold if any cell in the contingency table has a low value (e.g., for low frequency terms) (Yang & Pedersen, 1997).

Having calculated a feature reduction statistic (such as IG or $\chi^2$) for the terms in a corpus, it is necessary to identify an appropriate static value threshold (i.e. the term's score) that separates the terms that should be retained from the terms that should be discarded. The approach that we take in this thesis is to learn the threshold value from classification predictions made on the validation set, $D_{va}$, of the test collection.

### 2.2.4 Effective Classifiers for Text Classification

In this section we briefly present two machine learning classifiers that are widely used in the literature, and have been shown to be effective approaches for text classification (Sebastiani, 2002; Yang & Liu, 1999).

The first classification approach that we present is Support Vector Machines (Vapnik, 1995) (SVM). SVM are a type of supervised learning algorithms that try to learn a linear separating hyperplane that separates documents that belong to one of two classes (i.e., positive or negative) by the widest possible margin within the vector space. SVM achieve this by solving a dual optimisation problem on a set of document vectors, $\mathbf{x}_i$, with corresponding class labels, $y_i$, where $i = 1..m$, $\mathbf{x} \in \mathbb{R}^n$ and $y \in \{\pm 1\}$, that aims to (1) maximise the *distance* between the hyperplane and the *closest* documents in either of the classes, and, (2) minimise the number of documents that are misclassified.

To compute the miss-classification error for a single training instance, i.e, a specific document, SVM typically[1] use the hinge loss function (Vapnik, 1995), defined as:

$$Hinge(y, \rho) = max(0, 1 - y \cdot \rho) \tag{2.7}$$

where $y$ is the document's class label and $\rho$ is the classifier's real value output for the predicted class label. Hinge loss is a one-sided loss function. When the document's class label $y$ and the classifier's prediction $\rho$ have the same sign ($+$ or $-$) and $\rho \geq 1$ the loss assigned to the instance is 0. However, if the document's class label $y$ and the classifier's prediction $\rho$ have opposite signs

[1]Other loss functions can be used for SVM. For example, least squares (Suykens & Vandewalle, 1999) or modified versions of hinge loss, e.g., Wu & Liu (2007).

(i.e., the wrong class is predicted), or, importantly, if $y$ and $\rho$ have the same sign and $|\rho| < 1$, i.e., the correct class is predicted but the margin is not large enough, then hinge loss increases linearly in proportion to the difference between $y$ and $\rho$. This property of hinge loss makes it particularly suitable for maximal-margin classifiers such as SVM. The SVM optimisation problem, defined as:

$$Maximise \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle \tag{2.8}$$

requires learning the optimal weights, $\alpha_i$ for $i = 1..m$, where $\alpha_i \geq 0$. Eq. 2.8 relies only on the inner products $\left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle$, which can be viewed as a distance measure and, moreover, can be *substituted* by a *kernel function*, $K(\mathbf{x}_i, \mathbf{x}_j)$. In theory, this ability to select an appropriate kernel function for a classification task means that, given an appropriate kernel, most text classification problems are linearly separable (Joachims, 1998) and hence solvable by SVM classification.

In practice, for text classification tasks in which a document's feature vector is constructed only from statistics about the words in the document, a linear kernel, defined as:

$$K_{linear}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \tag{2.9}$$

is often the most appropriate kernel choice (Joachims, 1998).

The linear kernel has a couple of advantages for text classification tasks. Firstly, $K_{linear}$ is the simplest kernel to compute and therefore it takes less time to train an SVM model with a linear kernel than if the model implements a more complex kernel. Secondly, as previously mentioned in Section 2.2.3, document representations in text classification problems are often very large, typically with hundreds of thousands of dimensions or features in a document vector. This can be problematic when learning a classifier since, especially if there are more features than example documents, the classifier is likely to *over-fit* the model to the example data. However, it has been shown that a SVM classifier with a linear kernel is robust to over-fitting the learned model to $D_{tr}$ when $|\mathbf{x}|$ is very large (Joachims, 1998).

The second classification approach that we present is Multinomial Naïve Bayes (MNB). *Bayssian* classifiers, originally proposed by Duda & Hart (1973) for their work on pattern recognition, are a set of generative probabilistic classifiers that make strong assumptions about how the objects that are to be classified, in our case documents, were generated. The Naïve Bayes classifier assumes that the features of a document, i.e., its terms, are independent, hence why the classifier is said to be Naïve. This assumption is clearly untrue. However the assumption makes the Naïve Bayes classifier very efficient and, in text classification tasks, it has been shown that the assumption does not generally result in a loss of classification effectiveness (Domingos & Pazzani, 1997; Friedman, 1997).

Naïve Bayes classifiers are commonly based on one of two *event models* (Kibriya *et al.*,

2004; McCallum *et al.*, 1998). In *multi-variate* models, a document is viewed as an event and is represented by a binary vector recording the presence or absence of a term in the document. The multi-variate Naïve Bayes classifier has been a popular approach for text classification, e.g. (Kalt & Croft, 1996; Larkey & Croft, 1996; Lewis, 1992; Robertson & Jones, 1976). However, the event model that we deploy in this thesis is the *multinomial* event model, since it enables our framework to weight individual term features higher if they are more associated to sensitive information. The Multinomial Naïve Bayes classifier is modeled on the frequency of occurrences of terms from a vocabulary, $V$, in a document, $d$, and can be viewed as a unigram language model (McCallum *et al.*, 1998). Multinomial Naïve Bayes has been shown to work well for text classification (Lewis & Gale, 1994; McCallumzy & Nigamy, 1998; Nigam *et al.*, 1998; Rennie *et al.*, 2003).

In Multinomial Naïve Bayes, documents are viewed as a mixture model of classes, with each class having a multinomial distribution of terms. Given a vector $\theta$, where $\theta_j = P(c_j)$ denotes the probability of class $c_j$, and $\theta_{jk} = P(t_k|y_j)$ denotes the probability of generating term $t_k$ given class $c_j$, the likelihood of a document, $d$, being generated by $c_j$ is:

$$P_\theta(d|c_j) = P(|d|)\prod_k (\theta_{jk})^{t_k(d)} \tag{2.10}$$

where $t_k(d)$ is the frequency of the term $t_k$ in $d$. Assuming that the distribution of $P(|d|)$ is independent of $c$ and $|d|$ is fixed, and therefore dropping the first term $P(|d|)$, Bayes' rule can be used to calculate the posterior probability of a class, $c_j$, for a given document, $d$, and, therefore, $d$ can be classified as follows:

$$P_\theta(c_j|d) = \frac{P_\theta(c_j)P_\theta(d|c_j)}{P_\theta(d)} = \frac{\theta_j \prod_k (\theta_{jk})^{t_k(d)}}{Z(d)} \tag{2.11}$$

where $Z(d)$ is a normalisation constant summing over all class labels.

By considering features as events, the Naïve Bayes classifier can learn from the term features independently from the document vectors (Settles, 2011). This property of the Naïve Bayes classifier is particularly suited to the approach that we propose for integrating term-level reviewer feedback about the sensitivities in a collection in Chapter 7.

### 2.2.5 Evaluation and Metrics

When developing a classification system, the ultimate goal is to develop the most effective classifier possible, i.e., we want to maximise the number of correct classifications *and* minimise the number of incorrect classifications. However, within this broad definition of effectiveness, and to establish a true understanding of a classifier's behaviour, a number of classification metrics are typically reported that each focus on a particular aspect of classification effectiveness. In this section, we present the most popular metrics reported for evaluating binary text classification.

When evaluating a classification system, we first need to summarise the agreement, and disagreement, between the classifier's predictions and the correct, *gold standard*, judgements provided by human assessments. The predictions summary is collated in a contingency table that is often referred to as a *confusion matrix* since, particularly in *multi-class* classification, it is easy to identify from the table which classes the classifier is confused about. In the remainder of this thesis, we will refer to this table as a confusion matrix to make the distinction from other contingency tables.

Table 2.2 illustrates the confusion matrix for a binary classification task. The cells in the top row of the confusion matrix contain the total number of *instances*, e.g., documents, that have been assessed by human reviewers as being in the class of interest, $c$, and predicted by the classifier, $\Omega$, as belonging to $c$, i.e., *True Positive* ($TP$) predictions, or not belonging to the class of interest, $\bar{c}$, i.e., *False Negative* ($FN$) predictions. The cells in the bottom row of the matrix contain the total number of instances that have been assessed by human reviewers as not being in the class of interest, $\bar{c}$, and predicted by $\Omega$ as belonging to $c$, i.e., *False Positive* ($FP$) predictions, or not belonging to the class of interest, $\bar{c}$, i.e., *True Negative* ($TN$) predictions.

Table 2.2: A contingency table (matrix) for evaluating a classifier's effectiveness. The matrix cells contain the number of documents that are *true positive* predictions ($TP$), *false negative* predictions ($FN$), *false positive* predictions ($FP$) and *true negative* predictions ($TN$).

| Classified as $\rightarrow$ | $c$ | $\bar{c}$ |
|---|---|---|
| $c$ | $TP$ | $FN$ |
| $\bar{c}$ | $FP$ | $TN$ |

Precision (Kent *et al.*, 1955), defined as $precision = \frac{TP}{TP+FP}$, is the proportion of instances classified as $c$ by the classifier, $\Omega$, that actually belong to $c$. Precision has historically been viewed as the most important classification metric (Sebastiani, 2015). Conversely, recall (Kent *et al.*, 1955), defined as $recall = \frac{TP}{TP+FN}$, provides a measure of the proportion of the class of interest, $c$, that is correctly predicted as $c$.

The $F_1$ (van Rijsbergen, 1979) measure was first proposed as a text classification measure by Lewis & Gale (1994) and over the last two decades, $F_1$ has become the standard evaluation measure for binary classification in Information Retrieval (IR), machine learning (ML) and Natural Language Processing (NLP) (Sebastiani, 2015). $F_1$ is the harmonic mean of precision and recall. However, $F_1$ can be generalised to situations where there is a greater importance on either precision or recall (as is the case with sensitivity classification, where the recall is of greater importance, since there is a far greater penalty from wrongly predicting a document as being not sensitive and, therefore, releasing a sensitive document to the public, than there is from wrongly predicting a document as being sensitive) by parametrising the function with a value, $\beta$:

$$F_\beta = \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} = \frac{(1+\beta^2) \cdot TP}{(1+\beta^2) \cdot TP \cdot FN \cdot FP} \tag{2.12}$$

When $\beta = 0$, $F_0$ is the same metric as precision. $F_{0.5}$ is used for precision-oriented tasks and $F_2$ is usually selected for recall-oriented tasks. In general, the larger the $\beta$ value then the more recall-oriented the $F$ measure is (Sebastiani, 2015). The $F$ measure provides an adaptable metric for evaluating how good the classifier is at predicting the class of interest, $c$. However, it does not provide an overall evaluation of the classifier's performance if the distribution of classes in the collection are not known, since it does not use the number of True Negative predictions in the calculation. Moreover, this can result in a random classifier achieving different scores on collections with different distributions.

Accuracy, defined as $\frac{TP+TN}{TP+FP+FN+TN}$, is a simple measure that can provide an overall summary of the classification effectiveness. However, accuracy is not suitable in the case of binary classification where the proportion of the classes is heavily skewed, e.g., when the negative class is much larger than the positive class. In such a case, a classifier that predicts that every instance is in the negative class would achieve a high accuracy score, when in reality the classifier performs very poorly at identifying the positive class.

Balanced Accuracy (BAC) (Brodersen *et al.*, 2010) addresses the class imbalance problem for binary classification by weighting the true positive and true negative predictions by the total positive and total negative instances respectively. For any distribution of classes, a random classifier will result in 0.5 BAC. BAC is calculated as:

$$BAC = \frac{1}{2} \cdot \left( \frac{TP}{TP+FN} + \frac{TN}{FP+TN} \right) \tag{2.13}$$

We select $F_2$ and BAC as our main metrics when evaluating the effectiveness of sensitivity classification approaches throughout this thesis. We select $F_2$ to account for the fact that, in sensitivity classification, the potential consequences from mis-classifying a sensitive document are greater than the consequences of mis-classifying a non-sensitive document[2]. We select BAC as a general measure of the effectiveness of the classifier for classifying both sensitive and non-sensitive documents.

## 2.3 Active Learning

The approach for developing a classifier that we presented in Section 2.2 relies on there being an available test collection of documents, $d_1..d_n$, $d_i \in D$, with associated class labels, $y_1..y_n$, $y_i \in Y$. Moreover, as stated in Section 2.2.1, generating a labelled test collection by manual assessment

[2]The use of $F_2$ is a conventional standard for evaluating document classification tasks when recall is of greater importance than precision. It may be that in some cases, out-with this thesis, that it is appropriate to assign a larger weight to the importance of recall in evaluation (e.g. $F_4$), for example in risk-averse government departments.

is a resource intensive task that can take a lot of time.

Active Learning is one approach that enables us to reduce the amount of manual labelling that is required to be able to learn an effective classifier. In active learning, the learning algorithm is allowed to be curious and search a set of candidate documents and choose which ones to have labelled first, i.e., to choose the data from which it learns, enabling it to perform better with less training data (Settles, 1995).

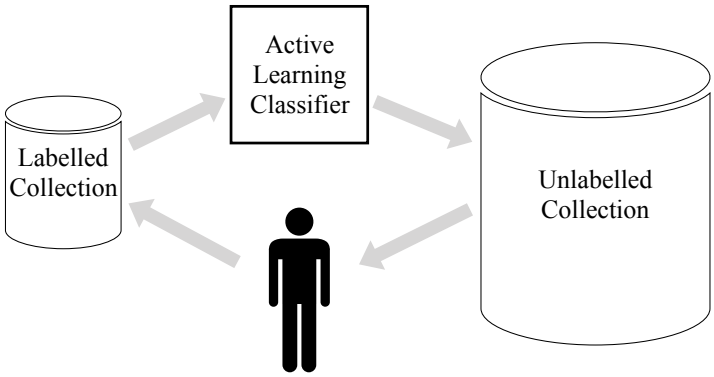



Figure 2.3: Illustration of pool-based active learning cycle.

Pool-based active learning (Lewis & Gale, 1994) is probably the most popular active learning strategy for applied research in text classification (Settles, 2012). As illustrated in Figure 2.3, in pool-based active learning, the classifier, $\Omega$, has access to a large pool of unlabelled documents, $D_{\bar{y}}$, and a small set of labelled documents, $D_y$. The classifier selects the documents from the unlabelled pool that it believes will provide the most useful information about the classification problem, and presents the selected documents to a human reviewer to have them labelled. The aim of the approach is to select the documents from the unlabelled pool that enable the classifier to learn the best possible model while using the least reviewer labelling effort possible. In order to do this, the classifier must deploy a strategy of how to predict a document's informativeness and select which documents to have reviewed by the human assessors.

### 2.3.1 Selecting Informative Documents

As mentioned in Section 2.3, an active learning classifier has to deploy a strategy for predicting the informativeness of documents so that it can choose which documents to have labelled by the human reviewer. Uncertainty sampling (Lewis & Catlett, 1994; Lewis & Gale, 1994) is a well-known set of active learning approaches for evaluating the informativeness of documents in an unlabelled collection. In uncertainty sampling the algorithm tries to identify, and present to a reviewer, the documents in the collection for which the classifier is least certain about their correct class labelling. Intuitively, the documents that the classifier is least certain about should provide the most information about the problem, since the more confident the classifier is about its prediction then the more likely it is that the classifier's prediction is correct.

In general, uncertainty sampling is a popular set of approaches for active learning since they are relatively easy to implement, are not computationally expensive and have been shown to be effective for many classification tasks (Settles, 2012). An additional benefit of uncertainty sampling approaches is that, when they are deployed with a classifier that outputs probabilities or confidence scores, the classifier can be viewed as a *black box*. Therefore, uncertainty sampling approaches can be developed independently from developing the classifier.

In this section, we introduce three uncertainty sampling approaches that have previously been shown to be particularly effective for text classification (Lewis & Gale, 1994; Settles, 2012). We will evaluate the effectiveness of these three approaches for sensitivity classification in Chapter 7. The first uncertainty sampling strategy that we present is *entropy based* uncertainty (Settles, 2012). Entropy uncertainty sampling ranks documents by the sum of their entropy (Shannon, 1948) scores, introduced in Section 2.2.3 (Eq. 2.3), for each of the document's possible labels, $y_i \in Y$.:

$$H(Y) = -\sum_i P(y_i) log P(y_i) \tag{2.14}$$

One way to view the intuition of this approach is that it calculates the number of bits it would take to encode the distribution of possible outcomes for $Y$. Therefore, documents with a high $H(Y)$ score should provide more information about their assigned label.

The second uncertainty sampling strategy that we present is the *margin* uncertainty sampling (Scheffer *et al.*, 2001) approach, defined as:

$$M(d_i, y_1, y_2) = |P(y_1|d_i) - P(y_2|d_i)| \tag{2.15}$$

This approach to uncertainty sampling calculates the margin, or difference, between the classifier's predicted probability scores for a document's first and second most likely classification labels. The intuition of margin sampling is that documents with a small margin between the two most likely class prediction probabilities are more ambiguous and, therefore, knowing the class label of these documents would be most beneficial to the classifier.

The third, and final, uncertainty sampling approach that we evaluate is *relevance* sampling (Lewis & Gale, 1994). Relevance sampling selects the documents that the classifier is *most* confident are examples of the positive class[3], i.e., in our task the positive class is sensitive documents. Therefore, we refer to this approach as sensitivity confidence, denoted as *sConf*, and it is defined as:

$$sConf(d_i, y_i) = P(y_i|d_i) \tag{2.16}$$

[3]Given that the relevance sampling selects documents that the classifier is most certain about, it may be more accurate to say that the approach is based on *certainty* sampling.

## 2.4 Technology-Assisted Review

Technology-assisted review (TAR) is a task in which human reviewers and an Information Retrieval (IR) system actively work together to identify and label documents as being either relevant or not relevant to a particular information need (Cormack & Grossman, 2014). The human reviewers in TAR are often experts in the field of the information that is being searched, and/or searched for, but they are not necessarily experts in search (Cormack & Grossman, 2015). Typically in TAR, the information need is known prior to the start of the review. However, in a TAR task, the process typically begins the system and the reviewers having no knowledge of the dataset that is being searched (Cormack & Grossman, 2014). Moreover, typically in TAR, the information that is being searched for and the collection that is being searched is different, or novel, for each TAR task (Cormack & Grossman, 2015). Cormack & Grossman (2015) recommend that, since the task and collections are unique to each TAR task and the subject matter can vary greatly each time that a TAR system is deployed, then it is probably best to avoid having many system parameters that need to be tuned. TAR is usually a recall-oriented task in that, differently from in ad hoc search, the information need is usually satisfied when *close to* all the relevant documents have been discovered (Cormack & Grossman, 2015).

TAR is a relatively recent name for this type of task. However, the origins of TAR can be traced back to Larkey & Croft (1996) who presented an early example of a semi-automatic classification system, where the system provided a ranking of predicted categories and relied on a human assessor to make the final labelling. Larkey & Croft (1996) ranked ICD9 codes for inpatient discharge summaries, but envisioned that an expert would have to make the final allocation. The authors expected that the benefit of the system would be that it would reduce the number of codes that an expert would have to consider.

More recently, TAR systems have been successful in increasing the productivity of human reviewers in fields such as generating IR test collections without the need for system pooling (Sanderson & Joho, 2004), and conducting systematic reviews in evidence-based medicine (Lefebvre *et al.*, 2008). However, TAR is most notably associated with the process of identifying electronically stored documents that are relevant to a legal litigation case; this field has come to be known as e-discovery (Cormack & Grossman, 2014; Oard *et al.*, 2010).

Research on TAR technologies for e-discovery was pioneered through the Text REtrieval Conference (TREC)[4] Legal Track (Baron *et al.*, 2006). The TREC Legal track's goal was to develop search technologies to help lawyers perform the legal discovery task on collections of digital documents. The Interactive Task (Cormack *et al.*, 2010; Hedin *et al.*, 2009; Oard *et al.*, 2008) of the The TREC Legal Track simulated the process of reviewing a large collection of documents to identify relevant, or *responsive*, documents for a *request for production*, i.e. a textual description of the documents that are to be judged as being relevant, in a class action lawsuit.

[4]https://trec.nist.gov/

Participating teams were provided with a document collection, a complaint and the complaint's associated requests for production. Participants were allowed to use any combination of automatic and manual approaches to identify as many responsive documents as possible (although the size of the document collection and imposed time constraints meant that it was not practical to for participants to perform an exhaustive manual review of the collection). The most effective approaches in the Interactive Task included a combination of *interactive search and judging* and relevance feedback (Cormack & Mojdeh, 2009) and *predictive coding*, in which an SVM classifier is trained using uncertainty sampling.

More recently, research into TAR has been developed through the TREC Total Recall Track (Grossman *et al.*, 2016; Roegiest *et al.*, 2015). The Total Recall Track did not focus on TAR within a legal context. However, it was similar to the Legal Track in that the goal of the Total Recall Track was to find close to all relevant documents for a particular search request (motivational examples given by the track organisers include e-discovery). The track was designed to simulate a human-in-the-loop retrieval task and the participants could submit either automatic or semi-automatic approaches. Participants were given a simple topic description, similar to what is typically used in ad-hoc or Web search, and asked to identify as many relevant documents as possible. Participating systems submitted one document at a time and received feedback on the relevance of the document from a server acting as the human-in-the-loop. Many of the retrieval strategies submitted to the Total Recall Track were enhancements of the most effective approaches in the Legal Track Interactive Task.

Cormack & Grossman (2014) provide an overview of machine learning approaches for TAR in the context of e-discovery. Each of the approaches, firstly, identifies an initial *seed set* of documents that are used to, secondly, train a document classifier to identify the $k$ documents to have reviewed. The reviewed documents are then used to re-train the classifier and the process continues in an iterative cycle until a decision is made that close to all of the relevant documents have been discovered. According to Cormack & Grossman (2014), there are three main TAR approaches. The first approach, simple passive learning (SPL), selects the seed set through random sampling before applying the classifier to the collection. In the second approach, simple active learning (SAL), the classifier is trained through a supervised active learning approach. In the third approach, continuous active learning (CAL) (Cormack & Mojdeh, 2009), keyword search is used to identify the seed set of documents before the classifier is applied to the entire collection and then deploys relevance feedback at each iteration of the review.


Figure 2.4: The typical components of a system for technology-assisted Review (TAR).

Figure 2.4 presents the common components and process of a modern TAR system. As we discussed above, the TAR protocol typically consists of two components. Firstly, a sampling strategy or a keyword search system is deployed to generate a pool of candidate documents to be reviewed and, secondly, a document classifier is trained to predict the relevant, or responsive, documents from the collection. Given a collection of documents, $D$, and a textual description of the documents that are to be identified (e.g. the request for production in e-discovery), the TAR system formulates a query/sampling strategy to retrieve the initial pool of documents to be manually reviewed and labelled, or *coded*. This *seed set* of labelled documents is used to train the learning algorithm and the iterative predict-review-retrain cycle continues until a stoping condition is met.

In the context of e-discovery, Grossman & Cormack (2010) showed that TAR can be more effective and more efficient than an exhaustive manual review to identify the responsive documents. The TAR protocol has the potential to be adapted to assist digital sensitivity review and increase the effectiveness and efficiency of human sensitivity reviewers. However, there are two fundamental differences between the requirements of TAR, for example in e-discovery, and the requirements of assisting sensitivity review.

Firstly, the goal of TAR, in tasks such as for e-discovery, is to identify *close to* all the relevant documents in a collection while minimising the required reviewing effort (Cormack & Grossman, 2014)[5]. Moreover, in most TAR tasks, the reviewer only reviews the documents that the system identifies as being relevant and this is usually a very small portion of the document collection that is being reviewed (Grossman & Cormack, 2010).

However, in sensitivity review, all documents that are released to the public must first be manually reviewed. Moreover, it is generally accepted that this will not change until TAR for sensitivity review has matured enough for governments and reviewers to develop trust in the technology (The National Archives, 2016*a*). Furthermore, even with the adoption of TAR, the volume of documents to be reviewed is expected to be much greater than the reviewing time available (The National Archives, 2016*a*) and documents that cannot be reviewed will be subject to *precautionary closure*, which is difficult for the government to justify (Moss & Gollins, 2017).

Secondly, the sensitive information within a collection is not known at the start of the reviewing process and, therefore, differently from TAR for e-discovery, in sensitivity review there is no equivalent to the request for production (i.e., there is no textual description of the sensitivities in the collection) that can be used as a query to generate an initial pool of documents for training a classifier. In theory, it is possible to manually construct queries with keywords that a reviewer might expect to be related to a specific sensitivity. However, in practice, due to the range of potential sensitivities, manually constructing separate queries for each specific

[5]We note that, in e-discovery, a secondary review is required to identify any responsive documents that are *privileged*, i.e., that should not be returned as relevant due to attorney-client privilege (see Gabriel *et al.* (2013) and Vinjumur *et al.* (2014)). The task of reviewing for privilege shares with sensitivity review the objective of identifying documents that should not be released.

sensitivity would result in an unmanageably large number of results sets. Moreover, identifying sensitivity by manually generating queries is limited to searching for the sensitivities that a reviewer expects to be in a collection. However, since the actual sensitivities are unknown, this approach is likely to result in low recall of sensitive information.

With this in mind, differently from in e-discovery, a framework for technology-assisted sensitivity review must be able to satisfy three basic principles, namely: (1) minimise the number of documents that are required to be reviewed to learn an effective sensitivity classifier; (2) provide reviewers with useful information that can assist them in making reviewing decisions; (3) prioritise specific documents to be reviewed to meet various objectives at different stages of the review, for example increasing the number of documents that can be reviewed and released to the public with the available reviewing resources. In this thesis, we propose a framework that is suitable for technology-assisted sensitivity review since it satisfies the three principles listed above by quickly learning to identify the sensitivities in a collection, providing reviewers with useful information about the sensitivities as they perform the review and prioritising specific documents to be reviewed to maximise openness.

## 2.5 Conclusions

In this chapter we have provided a summary of the key concepts within document classification, active learning and technology-assisted review that we build on in the remainder of this thesis. In particular, we provided an introduction to document classification in Section 2.1 and test collections, document representation and feature reduction techniques in Sections 2.2.1, 2.2.2 and 2.2.3, respectively. We introduced document classification techniques that we deploy in this thesis in Section 2.2.4 and methods for evaluating document classifiers in Section 2.2.5. In Section 2.3, we introduced active learning as a method of integrating the reviewer's feedback to the classification process to learn an effective classifier more efficiently. In particular, in Section 2.3.1 we introduced two methods for selecting informative documents to have reviewed and labelled by reviewers that can enable the classifier to learn a model more quickly. In Section 2.4 we introduced technology-assisted review, whereby a human reviewer and an IR system work together to find documents that are relevant to a particular information need or document finding task. Moreover, we identified why current TAR protocols are insufficient for assisting the sensitivity review of digital government documents. In the next chapter, we discuss the Freedom of Information Act 2000 (c. 36) sensitivities that we focus on in this thesis and how automatic classification techniques have previously been deployed to classify sensitive information. Moreover, we identify the sensitivity classification knowledge gap that we specifically address in this thesis.

# Chapter 3

# Classification of Sensitive Information

## 3.1 Introduction

In Chapter 2, we provided an overview of the fundamental concepts in text classification, active learning and technology-assisted review that we build on in this thesis. In this chapter, firstly, we provide an overview of the types of Freedom of Information Act 2000 (c. 36) sensitive information that account for the most frequent and high volumes of closures within central UK government, before identifying the specific freedom of information (FOI) sensitivities that we address within this thesis. Next, we provide details of the range of information that is likely to be closed from public release due to each of the identified sensitivities. We then discuss how automatic classification techniques have previously been deployed to identify sensitive information in documents, and identify the knowledge gap in sensitivity classification that make classifying FOI sensitivities a challenging task. Next, we present our methodology for constructing the test collection that we use throughout this thesis for developing and evaluating a FOI sensitivity classifier as the basis for our framework for technology-assisted sensitivity review, and provide some statistics about the generated collection. Finally, we conclude the chapter by summarising the sensitivities that we focus on identifying in this thesis and the knowledge gap that we have identified. In particular, the remainder of this chapter is structured as follows:

- Section 3.2 provides an overview of sensitivities within UK central government that have resulted in the most frequent applications for closures in recent years, before identifying the types of sensitive information that we focus on in this thesis, namely *international relations* and *personal information*. We provide a detailed description of international relations sensitivities in Section 3.2.1 and personal information sensitivities in Section 3.2.2.

- Section 3.3 discusses previous approaches from the literature for automatically identifying, and *masking*, sensitive information, before identifying the knowledge gap that make the classification of FOI sensitivities a challenging task.

- Section 3.4 presents our methodology for constructing the test collection that we use for

developing and evaluating our sensitivity classifiers, and provides statistics about the generated collection.

- Section 3.5 concludes this chapter by summarising the identified challenges of classifying sensitive information relating to international relations and personal information FOI exemptions.

## 3.2 Identified Sensitivities

The National Archives (TNA) of the United Kingdom (UK) conducted a business intelligence review (The National Archives, 2016*b*) to assess the digital landscape in UK government departments and their readiness for the imminent transition from transferring paper records to transferring digital records to TNA. The review reported that the challenges faced by government departments and TNA with respect to the imminent transition are mainly due to the volume and limited structure of digital records, a lack of technologies for digital sensitivity review, and the available resources in government departments.

In the UK, public bodies were expected to begin the transfer of digital records to TNA in 2016. Twenty one government departments are expected to transfer digital records to TNA in 2019, rising to fifty departments by 2021 (The National Archives, 2016*b*). Almost all of the government departments that were interviewed for The National Archives (2016*b*) review said that the sensitivity review of digital records is a major challenge for them.

There is an expectation of *openness* within the Public Records Act 1958 (c. 51) and for a document to be closed or redacted, as an outcome of a sensitivity review process, an application must be submitted to The Advisory Council. The expectation of openness is viewed as an essential element of transparent government in the UK and as a means for the public, e.g., through social scientists and historians, to hold the government to account (Moss & Gollins, 2017). Therefore, information will not be restricted or closed unless there is a clear and reasonable argument outlining the expected significant negative effects, or illegality, of releasing the information (Public Records Act 1958, c. 51).

The National Archives (2016*b*) report analysed the applications for closure made to The Advisory Council between 10$^{th}$ February 2005 and 30$^{th}$ April 2014. The analysis found that 75% of closures were related to personal sensitivities (exemptions 38, 40 and 41, cf. Table 1.1), 17% were due to national interest sensitivities (exemptions 24, 27 and 29, cf. Table 1.1), 5% were due to other sensitivities and 2% were due to policy development, Royal Household communications or legal privilege.

In this thesis, we focus on two specific sensitivities that are representative of the two largest groups of closures reported by The National Archives (2016*b*), namely Section 27: international relations and Section 40: personal information of the Freedom of Information Act 2000 (c. 36). The types of information that can be closed under these exemptions are broadly defined.

Moreover, these sensitivities, typically, require a judgement on the potential effects of releasing the information into the public domain. Therefore, assessing if information is sensitive often depends on many factors, for example who provided the information and is there a reasonable expectation that the information would not be released into the public domain, i.e. the sensitivities are context-dependent. Therefore, in Sections 3.2.1 and 3.2.2 we provide details of the range of information that can be closed under Exemptions 27 and 40 respectively.

### 3.2.1 Section 27: International Relations

The Freedom of Information Act 2000 (c. 36) Section 27: *international relations* (Ministry of Justice, 2008*a*) (S27) protects the UK's international relations and its interests abroad. S27 identifies two categories of information that are exempt from public release. S27(1) protects information that would likely prejudice (a) relations between the UK and any other state, international organisation or international court, or (b) the interests of (or protection/promotion of) the UK abroad, if the information was publicly released. S27(2) protects confidential information obtained from another state, an international organisation or an international court.

The likelihood of information being considered for closure under S27 is dependent on the context of the information, and can be a result of a wide range of communications from formal diplomatic exchanges to informal conversations. S27(1) does not focus on the type of information that the exemption covers but instead focuses on the *effects* of disclosure, i.e., the likely effect of releasing the information into the public domain. The UK's interests abroad cover a wide range of potential subjects relating to, for example, trade, defence, the environment, human rights, international crime or terrorism. Moreover, the UK's interests abroad change over time. Therefore the definition of S27 does not try to define the specific topics of information that can be protected. Instead, the Ministry of Justice (2008*a*) provides examples of situations that are more likely to result in S27 sensitivities. For example:

1. Reports on, or exchanges with, foreign governments or international organisations such as the EU, NATO, or the UN.

2. Information relating to UK citizens or companies' consular or commercial activities abroad.

3. Information about other states' views or intentions provided in the course of diplomatic and political exchanges.

4. Details of state visits and visits by ministers and officials.

5. Information supplied by other states through diplomatic channels.

6. Discussion within the UK government on approaches to particular states or issues.

7. Information relevant to actual or potential cases before an international court.

8. Details of the UK's positions in multilateral or bilateral negotiations.

S27(2) protects confidential information and applies to information that either, (1) the terms on which the information was obtained require it to be held in confidence, or (2) the circumstances in which the information was obtained make it reasonable for one of the parties to expect that it will remain confidential. This expectation of confidentiality could arise from, for example, an explicit agreement or an implicit code of practice. An implied confidentiality requirement is likely to apply to the content of most diplomatic exchanges with other states, or political discussions with Ministers or officials of other governments (Ministry of Justice, 2008*a*).

Information that is already in the public domain is less likely to be sensitive. However, the manner in which the information came into the public domain can impact this. For example, if the information originally came from a non-official source then the subsequent release of the information by the UK government could provide confirmation of the reliability of the source. This is clearly a judgement that requires some inside knowledge of official sources and, therefore, in our work, we assume that any information that is passed or received in private is confidential, unless there is additional contextual information to the contrary. For example, we assume that information that is already in published content, such as in a press report, is not confidential even if it discusses information that was originally passed in private. However, if the document contains the opinion(s) of an official from the government, or other organisation, that provides additional information about the contents of the press release, then the opinion(s) can be considered to be confidential.

The state of general relations between the UK and another country, or the other county's views on freedom of information, can also impact the likelihood of information being sensitive. For example, a state with a more liberal approach to freedom of information may be less likely to take offence at disclosure of some kinds of information and so the risk of prejudice to international relations may be lower (Ministry of Justice, 2008*a*).

Finally, when considering S27, a sensitivity reviewer must also consider if the risk of disclosure is outweighed by the public interest in the information being released. For example, if the disclosure would be unlikely to cause a *significant* negative reaction, then it is likely that the information would be released (Ministry of Justice, 2008*a*). However, a judgement on the intensity of a country's reaction is out-with the scope of our research, and, therefore, we view all information that is likely to cause a negative reaction from another country and prejudice the interests of the UK abroad as being sensitive.

Based on the information provided by Ministry of Justice (2008*a*) and through conversations with the Foreign and Commonwealth Office (FCO), we have identified seven types of sensitive information that are closed from public release through S27. We list the seven subcategories of S27 international relations sensitivities and provide a description of some of the most likely reasons for each of them in Table 3.1. As can be seen from Table 3.1, the range of information and topics that can potentially have international relations sensitivities is very broad. Moreover,

as we have discussed in this section, there are many complexities in making a sensitivity judgement (e.g., is the information public or private, what is the risk of release and how good are the relations between the countries involved). Therefore, automatically identifying, or classifying, international relations sensitivities is clearly a challenging task.

Table 3.1: Exemption S27: International Relations, descriptions of sub-categories.

| | |
|---|---|
| In Confidence | Any indication that the UK was provided information in confidence from an individual or organisation. |
| Sources | Any indication that the UK was in receipt of privileged information from an informer within a government or independent organisation. |
| Damage | Disparaging remarks about a country, e.g., their competence in managing a situation, international role or an important aspect of government.<br>Discussion of a bilateral relationship that it would be inappropriate to reveal e.g., where the existence or nature of the relationship would be unacceptable to another country. |
| Significant Figures | Disparaging remarks or culturally inappropriate references about, for example, a politician, royalty, ambassadors, or a significant historical figure who is held in high regard. |
| Treaty | Indication that the UK, or another nation where relations are sensitive, may be in breach of a treaty or international convention with another nation or nations.<br>Misuse of the diplomatic bag privilege. |
| Corruption | Any reference to bribes paid to heads of state, senior politicians, officials or their relations. |
| Behaviour | Any reference to inappropriate behaviour by a senior official, politician, or royalty e.g., serious sexual misconduct or culturally inappropriate sexual activity. |

To provide a more concrete representation of the S27 sensitivities that our framework aims to automatically classify, Figure 3.1 presents six example excerpts from documents that contain international relations sensitivities. The sensitivities in Figure 3.1 were identified by professional sensitivity reviewers from a central UK government department and are highlighted with a yellow background. However, to protect the sensitive information, the documents in Figure 3.1 have been sanitised by substituting the entities and the subject matter with synonymous text. We will present details of the collection of documents and the reviewing process that was used to identify these sensitivities in Section 3.4.

Figures 3.1(a) and (b) present examples of information that has been supplied in confidence and the name of the source that supplied the information. As can be seen from the examples, there are some similarities between the two excerpts. For example, both of the excerpts recount

In their conversation of 19th June, the special advisor on Liberian military affairs, Ayo Nuru, told our policy advisor, about a mass grave that was uncovered by the army in the Jenne area of Liberia early this year. Nuru said that official documents on the case are secret, but there were reports on the ground that about sixty bodies were found in the area, near the former headquarters of the LURD. It is thought that some of the interred may have passed naturally or possibly in combat. There are, however, thought to be about twenty-five bodies that appeared to have been executed by a single shot to the head.

(a) International Relations: In Confidence / Sources

It was reported in Khmer News last week that reporters had uncovered a possible plot by government employees to assassinate the Japanese Prime Minister during his forthcoming visit to Phnom Penh palace. Ambassador John Franklin was recently asked by the then visiting culture secretary Igu Huii if HMG was in possession of information concerning an expected imminent attempt to overthrow the Japanese government. Ambassador Franklin informed Huii that the government had no knowledge of the potential event or its planning. However, Huii said that they had reason to believe that the information existed.

(b) International Relations: In Confidence / Sources

The parliamentary committee heard a statement from a Saudi minister, Ahmed Ali, during the committee's proceedings. Ali testified that the GOSA had all but eradicated corruption within government affairs. However, the minister later confirmed that the personal information in GOSA passports is often inaccurate. According to Ali, passports are routinely issued to citizens of certain other countries within the continent if they make a sufficient contribution to the government. There are benefits from securing one of these passports due to the its stronger traveling power compared to some other countries in the continent.

(c) International Relations: Damage

There have been reports of casualties from friendly-fire during a recent hostage-taking in Turkey. This incident follows two hijackings that happened after a plane left Turkish airspace. The Turkish airport security did not detect the hijackers and the Turkish ministers and government agencies did not follow their own official procedures in both of the hijacking incidents. There are therefore serious doubts as to the Turk's abilities to detect and respond to future terrorist or hostage incidents. We have no reason to expect them to perform better in the future. We should continue to liaise with the Turkish authorities to moniter their reaction.

(d) International Relations: Damage

Greece has generally provided extensive cooperation whenever HMG has requested assistance and they play a vital role in international counterterrorism activity and intelligence. The GOG recently arrested a lieutenant of Osama Bin Laden who was attempting to pass through the country. The GOG put the lieutenant on trial within a third country at the request of the USG. Around the same time, the GOG also provided extensive coverage of all ports of entry, including sea, land and air, in pursuit of a second member of the OBL organisation who is thought to be an even more important player in the activities of the terrorist outfit.

(e) International Relations: Damage

It is not yet clear what the impact of DRC's multi-billion-dollar arms agreement will be on the country's arms trade. However, the agreement has shown President Kabila will be very difficult to deal with. Kabila and his team of advisors have shown themselves to be aggressive and have acted angrily to defiantly resist calls for an independent investigation into the agreement. We need to maintain relations with this important African leader and communications should be couched within positive and supportive language while building relations so we can lock horns with him at a future date.

(f) International Relations: Significant Figures

Figure 3.1: Examples of international relations sensitivities.

the details of private conversations between two people acting in an official capacity. Moreover, both of the documents discuss a specific event that is linked to a named country. In Figure 3.1(a), the source, Ayo Nuru, is passing-on second-hand information about details of a secret investigation in Liberia and uncorroborated details of how people were killed. The passing of such details could anger the Liberian authorities, or the people involved in the investigation, if it were made public knowledge. Moreover, Ayo Nuru could be put in danger if he was known to be the source of the information. In Figure 3.1(b) it is not quite as obvious that information has been supplied in confidence. In this example, the culture secretary Igu Huii asks if the ambassador has any information that can be supplied about a planned assassination, which the ambassador says he is not aware of. However, the claim that the Japanese government is already *aware* of the existence of the information that is being asked for results in the information being deemed as sensitive. If this information does indeed exist and the ambassador (i.e., the government) lied about its existence then this would likely detriment future relations with Japan. The examples in Figures 3.1(a) and (b) deal with what could be viewed as extreme situations, i.e., a mass killing and a planned assassination. However, S27 information supplied in confidence can be just as likely to arise from more day-to-day operational government discussions.

Figures 3.1(c), (d) and (e) each present examples of the S27 subcategory *Damage*. This subcategory of S27 is concerned with discussions about another country's lack of competence, or details of bilateral relationships that would be inappropriate to release. In Figure 3.1(c), the claim that the Government of Saudi Arabia (GOSA) provides passports with inaccurate, or made-up, information to non-SA citizens and, importantly, that the Saudi government has confirmed this, could be viewed as a statement of the GOSA's incompetence in performing an international role (i.e., issuing passports). This information would likely undermine the trust that the international community has in SA passports, making it more difficult for people to travel on these passports. Relations with the GOSA would likely be damaged if the (UK) government was responsible for making this information public knowledge.

The excerpt in Figure 3.1(d) contains disparaging remarks about the ability of the Turkish authorities to effectively respond to terrorist incidents, such as hijackings. It is likely that the Turkish authorities would refute these claims and this document could be seen as undermining their ability to effectively govern. Figure 3.1(e) discusses specific details of an intelligence co-operation that is due to a bilateral relationship with the government of Greece. Moreover, the document states that, at the request of USG, Greece handed over the suspect to a third country. Intelligence operations, such as these, are often necessarily enabled through a certain level of secrecy and releasing this information into the public domain could impact on the possibility of future such relationships.

Lastly, Figure 3.1(f) presents an example of the S27 subcategory Significant Figures. The excerpt claims that the Congolese president Kabila and his advisors have acted angrily and aggressively and that dealing with Kabila is an issue. The document states that dialogue with

Kabila should be *couched* in positive and supportive language. Relations with Kabila, and The Democratic Republic of Congo (DRC), would likely be damaged if this document were to reveal the true nature of dealings with the president. It is worth noting, however, that the sensitivity of this document would depend on Kabila's reputation in DRC at the time of sensitivity review. If Kabila is seen as a figure of the past and no longer of great significance in DRC, the document would possibly be released. However, if, for example, his son becomes president, then it could increase the sensitivity of this document, as the sensitivity is current, not historic.

### 3.2.2 Section 40: Personal Information

The Freedom of Information Act 2000 (c. 36) Section 40: *personal information* (Ministry of Justice, 2008*b*) (S40) guidance is mainly concerned with releasing information in response to a specific freedom of information request. However, in this thesis, we are interested in what constitutes personal information at the time of transfer to the archive and, therefore, in addition to the Ministry of Justice (2008*b*), the definition of personal information that we use has been drawn from the Information Commissioner's Office (2014) and The National Archive (2007).

S40 exempts the release of personal data, as defined by the Data Protection Act 1998 (c. 29). Personal data is information that is about a living individual from which that individual can be identified. Importantly, this includes any expression of opinion about the individual or an indication of the intentions of any other person with respect to the individual and, therefore, any relevant contextual information is also part of the sensitive personal information. Personal information is exempt from release if its disclosure would be in breach of any of the *data protection principles* (Information Commissioner's Office, 1998) or would be likely to cause substantial unwarranted distress to an individual. The Data Protection act defines sensitive personal data to mean personal data consisting of a person's:

- Racial or ethnic origin.
- Political opinions.
- Religious beliefs or other beliefs of a similar nature.
- (Non-)membership of a trade union.
- Physical or mental health.
- Sexual life.
- Alleged or committed criminal offence or any such related proceedings or sentencing.

In deciding if personal data should be closed under S40, a sensitivity reviewer must decide if it is reasonable to expect that the data would be released. For example, whether the information relates to a person's public or private life has to be taken into consideration. Information about

Table 3.2: Exemption S40: Personal Information, descriptions of sub-categories.

| | |
|---|---|
| Finance | Details of a named individual's claim, e.g., for rent, benefits, bankruptcy, investments, loans, compensation etc. Confidential tax, financial or business-related information. |
| Family Life | References to or claims of inappropriate or personal relationships, medical information, adoption, illegitimacy, maintenance payments, or comments on the morals or behaviour of a named individual. |
| Crime | Discussion of victims of sexual offences, juvenile defendants or defendants with mental illness. Individuals arrested but released without charge. Allegations of criminality, names of a defendant or sentencing. |
| Nationality | Discussions of an application for residency, passport or asylum that include extensive personal details. Discussions of citizenship or placement on the Visa Warning Index. |
| Employment | Comments about the abilities or performance of named officials or staff, including disciplinary action. Employment or biographical details, including pay information or employment rejected on security grounds. |
| Military | Discussions of specific activities of named individuals during wartime, such as involvement special operations or clandestine activities. |
| Health | Discussions about the medical condition, mental health or psychological condition of a named individual, including medical records. |
| Other | Discussions of an individual's defection or intention to defect, religious, terrorist or communist affiliations or sympathies, alleged sexual orientation, or derogatory remarks about the personal qualities of an applicant. |

a person's private life, e.g., their personal finances or medical records, is likely to be protected. However, information relating to a person who is acting in an official or work capacity should normally be released, providing that the release of the information would not be damaging, or distressing, to the individual. However, it is not the case that all information relating to a person's work will be released. For example, information about the names, positions, job functions or decisions that a person has taken would normally be released. However, information about internal disciplinary matters would normally be closed. Moreover, while the bank account details of staff would not be released, it would usually be justified to publish details of claimed expenses, pay grades or, in the case of senior staff, salaries. Although this information relates to staff personally, there is a strong public interest in the transparency of how public authorities spend their money.

We have identified seven types of sensitive information that are closed from public release through S40. We list the seven subcategories of S40 and provide a high-level description of each of them in Table 3.2. As can be seen from Table 3.2, personal information sensitivity is much broader in scope than what is typically considered to be personal data, i.e., attributes of entities, such as names, addresses, telephone or bank account numbers. Moreover, the scope of subject matter that can constitute personal information is vast, from discussions of a person's personal activities, relationships, health and finances to comments about an individual's morality.

Figure 3.2 presents six example excerpts from documents that contain personal information sensitivities that were identified by professional sensitivity reviewers from a central UK government department. The sensitivities are highlighted with a yellow background and the documents have been sanitised in the same way as those in Figure 3.1. The excerpt in Figure 3.2(a) presents an example of personal information due to employment details. The figure illustrates how the seniority of named individual's role can have an impact on whether their details will be released. The mention of the Quabar project director Pat Humphry does not result in S40 exemption since the project director is a senior position and it is reasonable to expect that the name of the person in this senior role would be in the public domain. However, Andy Simmons is a contractor and, as such, there is no clear expectation that details of him and his firm working on this project would be publicly released.

Figure 3.2(b) presents an example of S40: Crime. Although there are no actual crimes, or allegations of crime, reported in this document, the fact that the individual is being *vetted* is enough for the details of the person to be protected. Moreover, it is the case that any listing of a person's personal details, such as is in this document, which includes attributes such as date and place of birth, has the potential to be closed or redacted under S40.

Figure 3.2(c) discusses personal health details of the President Jamal Kanazie and his wife Annette. The fact that President Kanazie recovered quickly from a severe illness may be well-documented in the public domain and, if so, this information on its own would probably not result in an S40 exemption. However, the document speculates that Annette has "serious health

During a visit to the Quabar project last month, the committee was provided with details, from the Project Director Pat Humphry, of what he saw as obstacles to the completion of the project vision. In addition to the theme park, which will feature a multimedia audio and visual instillation, the complex is to include food and retail. During the tour, we met Andy Simmons, a contractor for the project. Simmons said that, as far as he is aware, no tenants had yet been signed up for the complex. Simmons's firm is building the Aquatic Theme Park, part of the Quabar Entertainment Centre Project, which is believed to be on track to be delivered on time.

(a) Personal Information: Employment

The Home Office and Ministry of Defence have reviewed their files. The departments possess no credible information of gross violations of human rights for the individuals identified below.

Anwar Kuratarew, Lieutenant, Patrol Commander, Kijkou provincial reserves, Naga, Philippines National Police; DOB: 14 June 1983; POB: Pili, Naga, Philippines; MALE

To abide by request: HO and MOD verifies that no credible information of gross violations of human rights by the individuals listed above is possessed, as of this date.

(b) Personal Information: Crime

President Jamal Kanazie had no credible alternatives and has had to retain in office many of his former council from before the revolution. This strategy has caused major unrest in large factions of his party and the country. Kanazie speaks English well and is well versed in international affairs, he also speaks three other languages well. His wife, Annette, does not speak English but is said to have appointed a tutor. Kanazie recovered quickly from a severe illness earlier in the year and appears to be fully recovered. However, Annette is thought to have serious health issues and is being treated for a severe illness – it could be cancer.

(c) Personal Information: Health

On May 28th, the ambassador met with Amir Fata Imagari, President of the Union of Retail Workers and Wholesale Industries, at the As-Salt VIZ, located near the city of Amman. (Note: Imagari has been a long-time contact of the embassy since long before he was President of the Union. Imagari will be traveling to Athens on October 14th for consultations with the ILO, labour NGOs and EA officials). Imagari escorted the ambassador through two factories and highlighted conditions that are of concern to the union. Details of the issues raised on the visit are included below.

(d) Personal Information: Other

The embassy responded to the Namibian Foreign Ministry with a diplomatic note of protest with regard to the following incident. On July 28, 1997, a Namibian property belonging to two UK citizens, John and Janet Shank, was served with a Section 8 notice, or Final Notice of Acquisition. The Shanks bought their property in 1981. The Shanks invested close to £50,000 in their property before they formally incorporated the property as a wildlife sanctuary under Namibian law in 1989. The GOM have insisted that land resettlement ceased in May 1996. However, twenty-five properties have reportedly been forcibly acquired since that date. We have not received any response to our protest from the Namibian government.

(e) Personal Information: Finance

Neija Terzic fled her village of Golici early in 1992 with her younger brother after it was attacked by BSA. Neija spent six months in Konjevic Polie before moving on to Srebrenica. Neija is a slight undernourished eighteen-year-old girl who was still receiving schooling before she had to flee. She has only received intermittent schooling since her displacement. The witness agreed to testify on the condition that her and her family's identities were not released publicly. She is clearly still very scared for the safety of other members of her family or relatives that are still being detained in BH.

(f) Personal Information: Family Life

Figure 3.2: Examples of personal information sensitivities.

issues and is being treated for a severe illness - it could be cancer". This information is protected by S40 (health) since there is clearly no reasonable expectation that speculations such as these would be publicly released. As we stated earlier in this chapter, sensitive information is often only a small portion of a document. It is worth noting that this excerpt is from a larger document that is 2945 words in length (17 paragraphs of text). All other mentions of Jamal and Annette Kanazie in the document are not sensitive. Therefore, if the document did not contain these 1 or 2 sentences about the Kanazie's health then the document would not be sensitive.

Figure 3.2(d) is a somewhat subtle example of personal information sensitivity. The Document discusses a meeting between an ambassador and the President of the Union of Retail Workers and Wholesale Industries. Both of these people are acting in an official capacity and there is, in general, no reason to apply S40 to the mention of the president, Amir Fata Imagari, or any other individuals mentioned in the document. However, the reference to Imagari being a long time *contact of the embassy* raises S40 issues. There is no reason to assume that this relationship is in the public domain and it is not clear that Imagari should reasonably expect that the information should be made public. Therefore, the government can not be responsible for putting the information into the public domain in case there are repercussions for Imagari.

Figure 3.2(e) includes specific details about a named individuals being served with an official notice that their property is being forcefully acquired. Moreover, the document discusses when the named individuals purchased the property and specific financial details about the purchase. It is reasonable that the named individuals would not expect that this information would be released to the public by a third party such as the government. Therefore, the information should be protected by S40.

Figure 3.2(f) includes personally identifiable information about a named individual, such as where she is from, a history of where she has lived and her schooling. Details such as these will always have some chance of being protected by S40 but the exemption decision will depend on the context in which the details are included. For example, if the details were part of a previously published document, such as a press report, then they would likely not be protected. However, the excerpt in Figure 3.2(f) is part of a document recounting a witness testimony from the named individual. This context means that the individual's identity should automatically be protected.

Interestingly, this excerpt also provides an example of how information that is not itself sensitive can be a reliable indicator that some other information is sensitive. The document states that "The witness agreed to testify on the condition that her and her family's identities were not released publicly". This sentence shows us that, regardless of the document's context or the topics discussed in the document, it would be a clear breach of confidence if the individuals details were (mistakenly) released to the public.

The examples in Figure 3.2 clearly only cover a small sample of the information that is potentially sensitive as a result of the categories listed in Table 3.2. However, the figure provides some concrete examples of what S40 personal information sensitivities *look like*. In the excerpts

presented in Figure 3.2, most of the named individuals are linked to a passage of sensitive text. However, importantly for the task of sensitivity classification, generally, in a collection of documents that are to be sensitivity reviewed a large majority of the documents that discuss named individuals do not contain S40 personal information sensitivities. Moreover, it is often the case that a document that does contain S40 personal information about a named individual also discusses other named persons that do not have any associated S40 personal information. Furthermore, personal information can be positioned within a document far away from any identifiable named person, i.e., the phrase "he was held on suspicion of indecent assault" may be far away from a definition of who "he" is, and there might not be a good reason to withhold any of the other information that is about that person in the document (as previously mentioned with respect to Figure 3.2(c)). For these reasons, automatically classifying S40 personal information sensitivities is a challenging task.

## 3.3 Previous Approaches for Classifying Sensitive Data

Most of the previous work on automatically detecting sensitive information in documents has focussed on the anonymisation of *personal data*. As previously stated in Section 3.2.2, personal information is broader in scope than what is usually thought of as personal data, which is typically limited to attributes of entities, such as names, addresses, telephone or bank account numbers. Most of the research into automatic anonymisation and redaction of documents has come from within the domain of clinical records (Tveit *et al.*, 2004). Early examples of automatic redaction software used dictionaries (or medical knowledge bases) to term-match known sensitive terms and regular expressions to identify repeated patterns such as postal codes and dates of birth (Gupta *et al.*, 2004; Neamatullah *et al.*, 2008; Sweeney, 1996).

Word lists-based systems such as these are costly, time consuming and *fragile*. Moreover, regular expression based systems are restricted in their application generalisability (Tveit *et al.*, 2004). With this in mind, recent research into detecting sensitive information in documents has tended to focus on more automatic and generalisable approaches.

Named entity recognition (NER) has become a popular approach for detecting personal data in documents. NER is a supervised machine learning process for automatically identifying entities such as *persons*, *organisations* and *locations*. There are many NER implementations available[1][2][3] that have been trained on large collections, for example the CoNLL collections[4]. Although training on such corpora results in domain specific NER algorithms, these *off the shelf* models are generally viewed as being generalisable enough to be applied in other domains. Therefore, specific NER models are not typically learned for identifying sensitive information.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml
[2] http://alias-i.com/lingpipe/index.html
[3] http://opennlp.apache.org
[4] http://ifarm.nl/signll/conll

Abril *et al.* (2011) presented an early example of using NER for identifying personal data. The authors proposed three approaches for disguising an entity's identity, adapted from the fields of Statistical Disclosure Control (Willenborg & De Waal, 2012) and Privacy-Preserving Data Mining (Agrawal & Srikant, 2000). The first approach, called *entity generalisation*, converts entities to less specific forms. For example the entity IBM is replaced by "computer-based company". The authors showed that this approach can retain more of a document's utility than the other two approaches, *entity swapping* and *entity noise reduction*, since they do not necessarily retain a document's original semantics. However, these approaches and other sensitivity identification approaches that are based on NER, e.g., from Cumby & Ghani (2011); Dernoncourt *et al.* (2017); Gardner & Xiong (2008); Guo *et al.* (2006); Uzuner *et al.* (2008); and Wellner *et al.* (2007), assume that *all* named entities are likely to be sensitive. Therefore, NER approaches are prone to large amounts of False Positive predictions. Moreover, these approaches assume that all documents are likely to contain some portion of sensitive information.

Assuming that all documents are likely to contain sensitive information is reasonable for domains in which there is a high chance that a document does contain sensitive information, such as in the medical domain where patients' treatment records all contain sensitive personal data (Tveit *et al.*, 2004). However, we argue that, this is not a reasonable assumption for FOI sensitivities, since in a collection of government documents that are to be sensitivity reviewed, the number of documents that contain sensitive information is often (much) smaller than the number of not sensitive documents (The Advisory Council, 2017). Moreover, we argue that, NER-based approaches are, therefore, not suitable for FOI sensitivities due to three main reasons:

1. NER approaches assume that all named entities likely to be sensitive. This results in a large False Positive rate and a decrease in the document's utility, which does not align well with the expectation of openness for transparent government (Moss & Gollins, 2017).

2. NER approaches only identify attributes of entities, such as names and addresses. However, as noted in Section 3.2.2, the contextual information associated to exempt personal data is part of the sensitive information. This information may not be in close proximity to a named entity within a document, and therefore cannot be identified by the approaches.

3. NER approaches are not capable of identifying features of more complex sensitivities, such as international relations. For example, as listed in Table 3.1, disparaging remarks about a country, e.g., their competence in managing a situation, an international role or an important aspect of government.

More recently, there has been a shift in focus for personal data classification from simple *masking* of named entities to *Document Sanitisation*. Document sanitisation aims to produce a privacy-preserved version of a document that retains the original document's utility. Nettleton & Abril (2012) measured the effect of document sanitisation on retrievability. Using the Wikileaks

cables, the authors derived two sets of queries from the top ten Wikileaks stories on Yahoo! News[5]. The query sets were designed to test (1) information loss (document utility) and (2) risk of disclosure. The authors found that documents sanitised to preserve utility resulted in only a 16% reduction in retrievability. However, Nettleton & Abril (2012) also found that sanitising documents to remove high-risk text resulted in a 47% reduction in retrievability, suggesting that removing high-risk text through sanitisation notably decreased the utility of the documents. This is problematic for FOI sensitivities since, as previously noted in Section 3.2, the expectation that governments can be held to account is an essential element of transparent government, and the large scale loss of specific details in released information would severely negatively affects this (Moss & Gollins, 2017).

Sánchez *et al.* (2012) presented a document sanitisation approach that is more general than the NER approaches. The authors assume that sensitive text is likely to be more specific than non-sensitive text, and use the Information Content (IC) of noun phrases as a measure of how sensitive the phrase is. Sánchez *et al.* found that their approach achieved higher recall of sensitive information than NER approaches. Moreover, although their work focused on identifying *personal information* sensitivities, they also identified *confidential* information. Therefore, their work is more closely aligned to identifying a broader range of sensitivities and has the potential to be useful for identifying the international relations sensitivity *information that has been supplied in confidence*. Therefore, we empirically evaluate the approach of Sánchez *et al.* (2012) for identifying this sensitivity in Chapter 5. We will show, however, that the approach does not perform well for this task.

There has been little previous work that has directly tried to classify sensitive information relating to government sensitivities. 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 meta data, 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 twelve different classification models. The authors selected the best performing seven models to deploy an ensemble classifier to predict security categorisations. Souza *et al.* 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 $F_1$. The work of Souza *et al.* (2016) suggests that deploying a supervised machine learning classifier to identify government sensitivities is a viable approach. However, there are two important points to note when comparing classifying original security categorisations and classifying FOI sensitivities. Firstly, it has been well documented that, as a precaution, government documents are often given a more strict security categorisation than the document requires (Roffman, 1975), so closing documents due to their security categorisations would lead

[5]https://uk.news.yahoo.com

to many documents being closed that should be released. Secondly, sensitivity usually decays over time so, even if the document is sensitive at the time of creation, it is unlikely that it is sensitive at the time of sensitivity review (Moss & Gollins, 2017).

In summary, previous work on classifying sensitive information in documents has focused on masking, or redacting, personal data. The most effective approaches have been the NER-based approaches or document sanitisation. However, these approaches assume that all entities are potentially sensitive and all documents are likely to contain sensitive information, which make them not suitable approaches for identifying FOI sensitivities. Automatically identifying FOI sensitive information has not been well examined in the literature. The most closely related task from the literature is classifying the original security categorisation of government documents. However, this task differs from sensitivity classification, since security classifications are often more strictly applied than is required and sensitivity tends to decay over time. Therefore, the original security categorisations are not a reliable indicator of current sensitivities in historic documents and there is a need for a sensitivity classification approach that can classify a range of current FOI sensitivities in historical government documents.

## 3.4 A Test Collection for Sensitivity Classification

Due to the sensitive nature of the information that we wish to classify in this thesis, there is no publicly available test collection with associated class labels. Therefore, we had to generate a suitable, and representative, test collection for developing and evaluating our sensitivity classifiers. In this section, we provide details of our methodology for generating our test collection.

As our document collection, we use a random selection of documents from a collection of formal government written communications between central government and embassies around the world. The collection contains real sensitivities and ethics approval was obtained to use the collection. However, due to the sensitivities in the collection and to abide by the constraints of the ethics approval, the collection can not be distributed in any form to any parties out-with those who were party to the original non-disclosure agreement[6].

The document collection had not previously been sensitivity reviewed and, therefore, we had to create a ground truth of the sensitive information within the collection. To do this, we enlisted the assistance of personnel from The National Archives, The Foreign and Commonwealth Office, The National Records of Scotland and Northumbria University who had previous experience in sensitivity review, to review the documents and identify any sensitive information. A detailed set of guidelines was provided to the reviewers, to ensure that the reviewing task was consistent between reviewers. The guideline provided reviewers with some background to the project and the aims of generating the test collection, before providing an introduction to the

[6]To avoid disclosing sensitive information, all of the example documents that are presented in this thesis have been sanitised by replacing the entities and subject matter with synonymous text.

document collection, the definitions of the sensitivities that the reviewers were to identify (as we presented in Sections 3.2.1 and 3.2.2) and the process of annotating any identified sensitive information[7].

Reviewers were provided access to the web based reviewing interface presented in Figure 3.3. The interface enables reviewers to navigate the document collection by selecting documents in the left hand panel. Documents are displayed in the right hand panel, where reviewers must record a document level classification judgement by selecting one of four possible options: the document (1) is not sensitive, or (2) contains sensitive information that would be closed due to Section 27: international relations or (3) contains sensitive information that would be closed due to Section 40: personal information or (4) contains both Section 27 and Section 40 sensitive information.

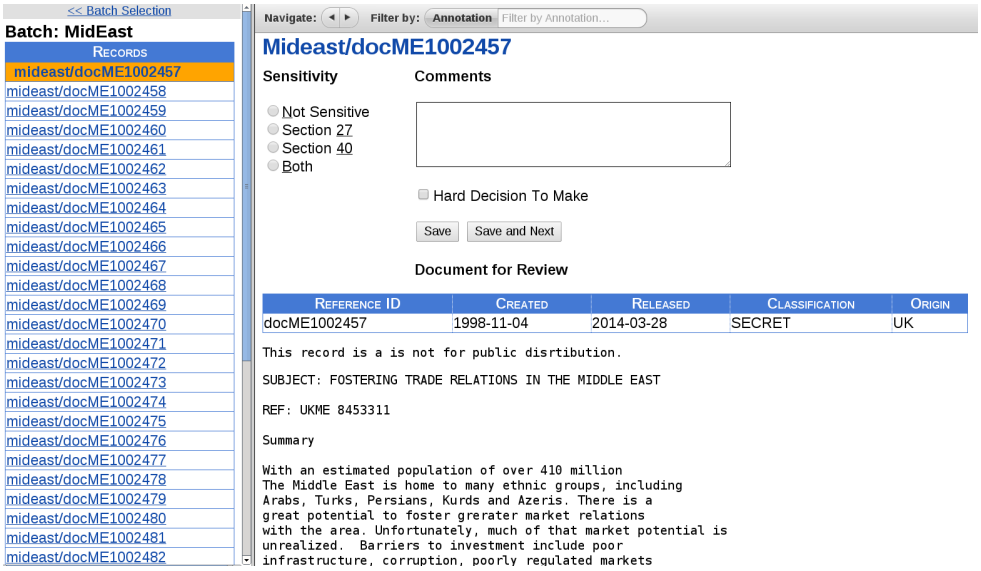



Figure 3.3: The sensitivity reviewing interface used to generate our test collection. The panel on left of the interface enables reviewers to navigate the collection, while the main panel enables reviewers to sensitivity review the documents and to record a document's sensitivity judgement.

In addition to providing document level classifications, the interface enables reviewers to annotate any sensitive text within a document and tag the text with the relevant sensitivity subcategories from Tables 3.1 and 3.2 (the annotation functionality is shown in Figure 3.4). Reviewers were provided with a user manual and were provided with training in how to use the interface.

A total of twenty four reviewers were recruited. The reviewers performed the reviewing task in their own time, and at their own pace. The reviewers were not paid or compensated for providing their reviews, however we have confidence in the quality of the reviews that were provided due to the vested interest of the reviewers in the successful development of assistive technologies for sensitivity review. Reviewers were provided multiple batches of fifty documents and asked to complete as many as was reasonably possible for them to do. 150 documents were

[7]Due to the sensitivities in the guidelines, we can not include the full guidelines in this thesis.

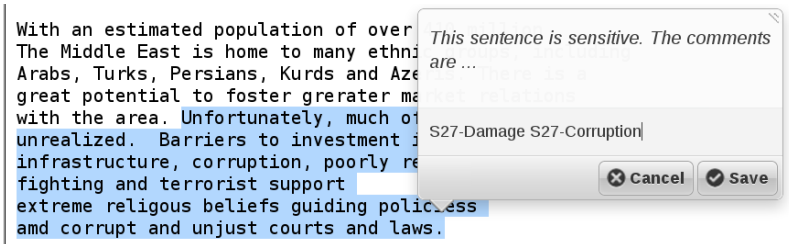



Figure 3.4: The sensitivity reviewing interface annotation functionality for identifying sensitive text within documents and recording the relevant sensitivity sub-categories.

judged by two reviewers and 50 documents were judged by four reviewers. Agreement was found to be 0.5525 measured by Cohen's $\kappa$ (Cohen, 1960) for the double-judged documents and a Fleiss' $\kappa$ (Fleiss & Cohen, 1973) score of 0.4414 for documents which received four judgements each. While these values indicate only moderate agreement (Landis & Koch, 1977), we note that levels of agreement in the paper-based review process is unknown, as only one assessor will routinely judge each document. When generating the final gold standard labels, documents that were double judged were labelled as sensitive if either of the reviewers judged the document to be sensitive. For documents that were judged by four reviewers, labels were assigned using a majority vote where appropriate. If the vote was tied, the document was labelled as sensitive.

Table 3.3 presents the salient statistics of the generated test collection. In total, the reviewed collection contains 3801 documents. 502 documents (13.2% of the collection) contain sensitive information relating to international relations, personal information or both sensitivities. 3299 documents contain no sensitive information. It is difficult to get an accurate figure for the distributions of sensitive information in the government collections that are to be sensitivity reviewed. Moreover, there are large differences in reports of the percentage of information that is transferred to TNA closed. The Advisory Council (2017) stated that 5% of documents that were transferred in 2017 were transferred closed. However, a member of the Advisory Council was recently quoted as saying that 75% of the documents transferred to TNA are closed (Pauli, 2015). We have been assured, however, by sensitivity review professionals that the distribution of sensitive information in our collection is representative. Moreover, the distribution of sensitive information in our collection is close to the average of the distributions of collections that were transferred to TNA in 2015/16 (14.9%). This is based on figures from The Advisory Council (2016) and using the calculation:

$$\frac{TotalReviewed}{Closures} \cdot 100 \tag{3.1}$$

where Total Reviewed = (total transferred + total retention applications - withdrawn retention applications) and Closures = (closure applications + retention applications). We note, however, that not all retention applications are due to sensitive information, with some documents being retained for administrative purposes (The National Archives, 2016*c*).

Table 3.3: The salient statistics of our test collection.

| Total Documents | Not Sensitive | Sensitive | | | | Unique Terms | Avg. Doc Length |
|---|---|---|---|---|---|---|---|
| | | International Relations | Personal Information | Both | Total | | |
| 3801 | 3299 | 231 | 156 | 115 | 502 | 122 348 | 710 terms |

## 3.5 Conclusions

In this Chapter, we introduced the freedom of information (FOI) sensitivities, as defined by the Freedom of Information Act 2000 (c. 36), that we address within this thesis. Moreover, we discussed the properties of context-dependent FOI sensitive information that make a judgement about, or the automatic classification of, the potential sensitivity of information (or a document) a complex task. In particular, firstly in Section 3.2, we provided an overview of the types of sensitive information that account for the largest volume of applications for closure by central UK government departments and identified the sensitivities that we focus on classifying throughout this thesis, namely *international relations* and *personal information*. Moreover, we provided a detailed description of the types of information that are protected through each of these sensitivities and showed that the classification of FOI sensitivities a challenging task. In Section 3.3, we discussed previous approaches from the literature for automatically redacting sensitive information in documents. Most of the previous approaches have been based on NER or have focused on document sanitisation. We argued that these approaches are not suitable for FOI sensitivity classification, since they assume that all documents are likely to contain sensitive information and that all entities are likely to be sensitive. Moreover, we argued that FOI sensitivity classification has not been well examined in the literature and there is a need for automatic classification approaches that can classify a range of current FOI sensitivities in historical government documents. Finally, in Section 3.4, we presented our methodology for constructing the test collection that we use for developing and evaluating our sensitivity classifiers for our framework for technology-assisted sensitivity review. In the following chapter, we present the components of our framework in more detail. Moreover, we describe how each of the components of our framework contributes to assisting human reviewers with the sensitivity review of digital government documents. In particular, by identifying latent vocabulary, syntax and semantic language features sensitive information and incorporating explicit reviewer feedback to develop an effective sensitivity classifier. Moreover, we also show how our framework can prioritise specific documents for review to learn an effective classifier more quickly, reduce the time that is required to sensitivity review a collection of documents and increase the number of non-sensitive documents that can be reviewed and released to the public within the available reviewing time budget.

# Chapter 4

# A Framework for Technology-Assisted Sensitivity Review

## 4.1 Introduction

In the previous chapter, we introduced the two Freedom of Information Act 2000 (c. 36) (FOIA) exemptions, i.e., the sensitivities, that we focus on identifying in this thesis, namely Section 27: International Relations (Section 3.2.1) and Section 40: Personal Information (Section 3.2.2). Moreover, in Section 3.3, we showed that the problem of automatically identifying sensitive information that is exempt from public release through the FOIA has not been thoroughly examined in the literature. Indeed, as we showed in Section 3.3, the previous literature that addresses automatically identifying sensitive information in documents has, almost exclusively, focused on identifying personal data, such as names, addresses and social security numbers. However, information that can be *closed* from public release due to a FOIA exemption is more broadly defined and context-dependent than personal data and, therefore, there is a need for new approaches to automatically identify documents that contain FOIA sensitivities, so that we can reliably assist with the sensitivity review of digital government documents.

In Chapter 1, we argued that, to effectively assist the digital sensitivity review process we must deploy a technology-assisted review (TAR) approach for digital sensitivity review. Moreover, we argued that to be able to effectively assist the sensitivity review process, the TAR approach must be able to adapt the assistance that it provides to sensitivity reviewers to assist them in both of the sensitivity review scenarios that we introduced in Chapter 1, namely: *exhaustive review*, i.e. when all the documents in a collection will be manually sensitivity reviewed; and *limited review*, i.e. when there are insufficient reviewing resources available to review a full collection and, therefore, the objective is to *open* as many documents to the public as possible with the available resources. We argue that technology-assisted sensitivity review will enable human sensitivity reviewers to work in partnership with automatic sensitivity classification and, therefore, to be able to review a collection of digital government documents more quickly and

more consistently.

In Section 2.4, we reviewed the previous literature on TAR and outlined why the existing TAR approaches are not suitable for assisting with the sensitivity review of digital government documents. In this chapter, we propose a framework for technology-assisted sensitivity review. The framework consists of four components, namely *Document Representation*, *Document Prioritisation*, *Feedback Integration* and *Learned Predictions*. In the remainder of this chapter, we first provide a more detailed discussion of the two user models for technology-assisted sensitivity review that our framework is designed to assist, before providing an overview of our proposed framework and a detailed discussion of each of its four components. The remainder of this chapter is structured as follows:

- Section 4.2 provides a detailed discussion of the two assisted review user models that our proposed framework addresses and, moreover, discusses how the framework can assist sensitivity reviewers in each of the identified models.
- Section 4.3 provides an overview of the framework that we propose for technology-assisted sensitivity review, and the four components of the framework that are required to assist the sensitivity review process. In the following four sections, we then discuss each of the individual components.
- Section 4.4 provides details of the first component of our framework, the document representation component.
- Section 4.5 provides details of the second framework component, which prioritises documents for review depending on the priorities of the review process at a particular point in time.
- Section 4.6 provides details of the framework's third component, which deals with integrating a reviewer's judgements and feedback into the document representation.
- Section 4.7 provides details of the fourth, and final, component of our framework, the learned predictions component.
- Section 4.8 provides a summary of this chapter.

## 4.2 User Models for Technology-Assisted Sensitivity Review

All documents that are opened to the public must first be sensitivity reviewed. However, government departments are not expected to be able to recruit enough resources (The National Archives, 2016*a*) to sensitivity review all of the documents that are expected to be selected for permanent preservation in the public archive. Therefore, any documents that cannot be sensitivity reviewed are likely to be subject to *precautionary closure*, i.e., they will not be released to

the public since opening sensitive information to the public is at least negligent and potentially illegal (Sloyan, 2016). Moreover, the risk of documents that have not been reviewed containing sensitive information, and that information causing damage through public release, is too great.

However, there are many different types and sizes of collections that need to be sensitivity reviewed, from email collections to public enquiries (Inquiries Act 2005, c. 12). Moreover, the amount of documents that need to be reviewed, the distributions and volumes of different types of sensitivities, and the amount of resources that can be deployed for sensitivity review will vary between different government departments. Therefore, assisting the digital sensitivity review task can have different priorities, depending on whether there are enough reviewing resources available to sensitivity review all of the documents in a collection that need to be reviewed.

We have, therefore, identified two user models for assisting digital sensitivity review depending on the available reviewing resources: the *exhaustive review* user model addresses how our proposed framework can assist sensitivity reviewers when all of the documents in a collection are manually sensitivity reviewed; and the *limited review* user model addresses how a technology assisted review approach can best assist the review process when there are not enough reviewing resources to review a full collection. In the remainder of this section, we provide a more detailed discussion of each of these user models:

**Exhaustive Review**: Although it is generally accepted that some form of technology-assisted review is necessary to be able to assist with the sensitivity review of digital documents (Allan, 2015), it is also the case that a fully automated solution to (even parts of) the digital sensitivity review process cannot be adopted until governments and reviewers develop trust in the ability of automatic sensitivity classification techniques to reliably and consistently identify previously unseen examples of sensitivities. Therefore, it is generally accepted that *all* government documents that are released to the public will continue to be *manually* sensitivity reviewed until there is an acceptable level of trust in the classification technologies (The National Archives, 2016*a*).

The exhaustive review user model addresses the question of how a technology-assisted review process can assist sensitivity reviewers if *all* of the documents in the collection that is to be reviewed will be manually reviewed. In this thesis, we argue that by automatically predicting which documents in a collection contain sensitive information and providing the reviewer with this information, our proposed framework will be able to reduce the time that it takes for a reviewer to review a collection of documents, since this will enable a reviewer to identify sensitive documents more quickly and, moreover, to review non-sensitive documents more quickly. Additionally, we argue that providing reviewers with automatic sensitivity predictions will increase the level of judging agreement between sensitivity reviewers. We provide a thorough analysis of the effectiveness of automatic sensitivity classification predictions for increasing the speed of, and the agreement between, sensitivity reviewers in Chapter 9.

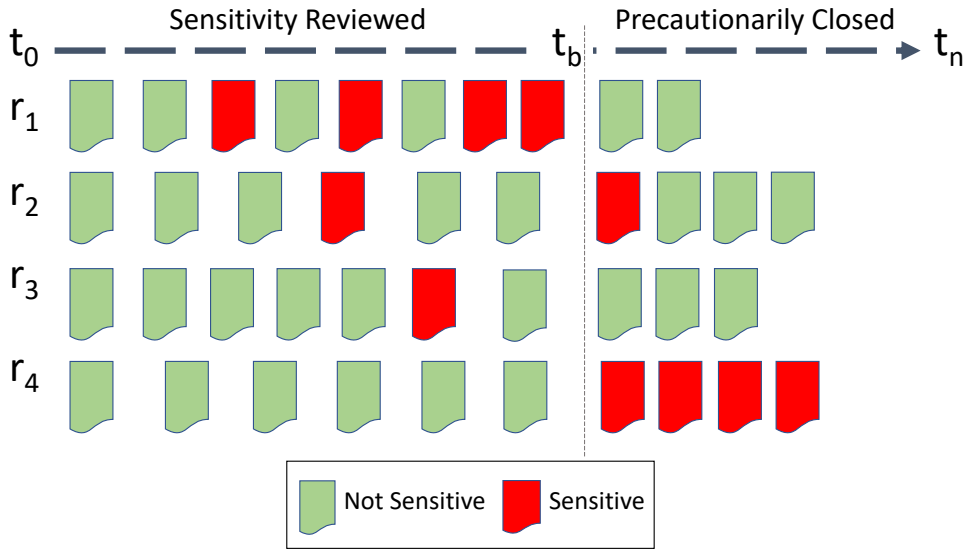



Figure 4.1: Limited Review user model. Examples of the possible ordering of documents to be reviewed and their resulting *openness*.

**Limited Review**: The *limited review* user model addresses a scenario in which there are insufficient reviewing resources available to review all of the documents that are being considered for public release. In this model, in addition to the benefits that sensitivity classification predictions can provide reviewers (as addressed in the exhaustive review model), there is an opportunity for technology-assisted sensitivity review to play a role in ensuring that the available reviewing resources are used effectively to meet the objectives of the review. For example, it may be the case that, when there are insufficient resources to review a whole collection of documents, the main priority is to maximise the number of documents that are opened to the public within the available reviewing time budget, i.e., to maximise *openness*. Therefore, in the limited review user model, the objective of the proposed framework is to: firstly, identify sensitive documents that are not to be released; and secondly, prioritise for review the documents that should be released.

While achieving these objectives, the proposed framework must aim to minimise the amount of reviewing resources that review documents that contain sensitive information and are, therefore, closed from public release, since this will reduce the resources available for opening documents to the public. With this in mind, in the limited review user model for assisting digital sensitivity review, the framework must be able to satisfy two basic principles, namely: (1) Maximise Openness; and (2) Minimise Precautionary Closure.

Figure 4.1 illustrates how the order that documents are sensitivity reviewed can have an impact on the number of documents that are opened to the public and the number of documents that are precautionarily closed. The figure shows four different orderings, or rankings $r_1$ to $r_4$, of documents in a collection consisting of ten documents. Six of the documents in the collection are not sensitive (the green documents) and four of the documents are sensitive (the red documents). Along the top edge of the figure, the dashed arrow shows the time taken to review the documents.

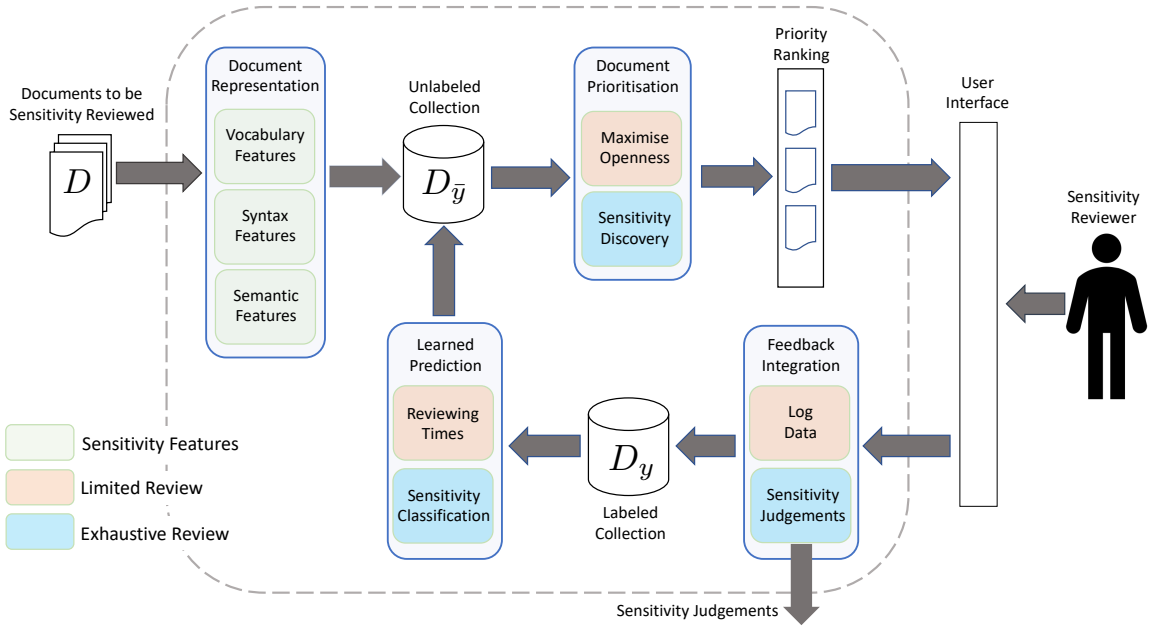



Figure 4.2: An overview of our proposed technology-assisted sensitivity review framework.

The reviewing session starts at time $t_0$ and the length of time that the reviewer would require to review the whole collection is $t_n - t_0$. However, the available reviewing time budget only allows the reviewer to review from $t_0$ until $t_b$ and, therefore, not all of the documents in the collection can be reviewed.

As we can see from Figure 4.1, in the first ranking, $r_1$, the reviewer manages to review the greatest number of documents, i.e., eight of the ten documents are reviewed. However, although this ranking has the fewest number of documents precautionarily closed, the ranking actually results in the fewest number of documents being released to the public, since four of the reviewed documents are sensitive. Each document in the collection takes a different amount of time to review and, therefore, only six documents are reviewed in $r_2$ and seven are reviewed in $r_3$. The openness of $r_2$ and $r_3$ is greater than that of $r_1$. However, in $r_2$ and $r_3$, three not sensitive documents are precautionarily closed since there is not enough time to review them and, therefore, they are not an optimal rankings. For the collection in Figure 4.1, $r_4$ is the best of the presented rankings, since the reviewer only spends time reviewing not sensitive documents and the only documents that are precautionarily closed are sensitive.

The challenge for a technology-assisted sensitivity review framework when there are insufficient reviewing resources available, is the generation of a ranking of documents that closely satisfies the objectives discussed above. Our limited review user model addresses this scenario and we evaluate our proposed approach in Chapter 8.

## 4.3 Framework Overview

Figure 4.2 presents an overview of the framework that we propose for the sensitivity review of digital government documents. In Figure 4.2, the framework is enclosed within the dashed grey line and each of the framework's four components, *Document Representation*, *Document Prioritisation*, *Feedback Integration* and *Learned Predictions*, are represented by rectangles with rounded corners and a solid blue outline. The rounded green, red and blue rectangles inside the components, in Figure 4.2, represent how each component can be instantiated to meet the component's objectives at a particular point in the review process. The green rectangles show the types of sensitivity features that are generated by the Document Representation component. The Document Prioritisation, Feedback Integration and Learned Predictions components can be instantiated differently to meet the needs of each of the exhaustive review (blue rounded rectangles) or limited review (red rounded rectangles) user models. Figure 4.2 also presents the inputs and output of the framework and the direction of the flow of information is represented by grey arrows. In the remainder of this section, we provide a high-level overview of our proposed framework, before in Sections 4.4, 4.5, 4.6 and 4.7 discussing each of the components individually and, moreover, how the component can be instantiated.

Our proposed framework consumes two types of inputs. Firstly, a collection of documents, $D$, that are to be sensitivity reviewed and, secondly, the sensitivity judgements with the log data from the user interface that a human reviewer uses to sensitivity review the documents in $D$. From Figure 4.2, we observe that, starting from the left-hand side of the figure, the initial input to the framework is the collection of documents that are to be sensitivity reviewed, $D$. The collection is passed to the Document Representation component to be transformed into a format that is efficient for a classifier to process and, moreover, that encodes the documents' features that will enable a classifier to learn an effective model. The output from the Document Representation component is the transformed unlabelled collection, $D_{\bar{y}}$, i.e., the collection is said to be *unlabelled* since the documents in the collection do not have any associated *labels* that identify an appropriate classification of the document (e.g., *sensitive* or *not-sensitive*).

The remainder of the framework forms an iterative process that we shall refer to as the *review cycle*. In the review cycle, the Document Prioritisation component takes as input the unlabelled collection, $D_{\bar{y}}$, and generates a ranking of documents in which the documents that should be prioritised for review in the current iteration are ranked closer to the top of the ranking. The output of the Document Prioritisation component is the top $k$ documents from the generated ranking, which are provided to a human reviewer, via a user interface, to be sensitivity reviewed.

As previously discussed in Section 1.2, when a sensitivity reviewer reviews a document, the reviewer reads the document and records a judgement as to whether the document contains any sensitive information that should not be publicly released. When recording a sensitivity judgement, the reviewer assigns an appropriate *class label*, $y_i \in \{sensitive, nonSensitive\}$, to a reviewed document. As discussed in Section 1.2, documents that are judged to contain sensitive

information are either closed from the public for a specified period of time or, alternatively, the sensitive information in a document is redacted so that the document can be released without disclosing the sensitive information. For paper-based sensitivity review, the redaction of sensitive information in a document is a separate process that is not done at the time of review (Allan, 2014). However, in this thesis, we propose to have reviewers perform the redaction as part of the sensitivity review, to provide additional evidence to the classifier about the sensitivities in the collection. Therefore, in addition to an associated class label, a sensitivity judgement also contains a record of any text that a reviewer judges to be sensitive within a document, and additional feedback from the reviewer about their decision.

After the reviewer has sensitivity reviewed the $k$ documents presented to them during the current iteration of the review cycle, the log data from the user interface and the reviewer's sensitivity judgements, are integrated into the document representations of the $k$ reviewed documents by the Feedback Integration component. The updated document representations are output from the Feedback Integration component and added to the collection of documents that have associated class labels, $D_y$. The size of the labelled collection, $D_y$, increases in each iteration of the review cycle and, therefore, provides the framework with more information about the sensitivities in the collection.

The labelled collection, $D_y$, is used as input to train a sensitivity classifier in the Learned Predictions component and this component is deployed to make predictions about the documents in $D_{\bar{y}}$, for example to predict if a document contains sensitive information or to predict the length of time that a reviewer would require to review a specific document. The output predictions from the Learned Predictions component provide additional information that can be used by the Document Prioritisation component in the following iteration.

Our proposed framework, presented in Figure 4.2, is designed to be able to learn quickly and adapt to the needs of the reviewer, by making use of the information that is provided by the sensitivity reviewer and by intelligently prioritising the documents that should be reviewed at each iteration of the review cycle.

## 4.4 Document Representation

Figure 4.3 presents the functionality of the Document Representation framework component, and the component's input and output. As previously mentioned in Section 4.3, the component takes as input a collection of documents, $D$, that are to be sensitivity reviewed. As we previously discussed in Section 2.2.2, to be able to use documents, such as those in $D$, as the basis of a classification system, the documents need to be transformed into a structured data representation that is suitable and efficient for a classifier to learn from (Song *et al.*, 2005). The output from the Document Representation component is the documents from $D$, where each document is transformed into a feature vector representation, $\mathbf{x}$. Moreover, the vector representation, $\mathbf{x}$, con-

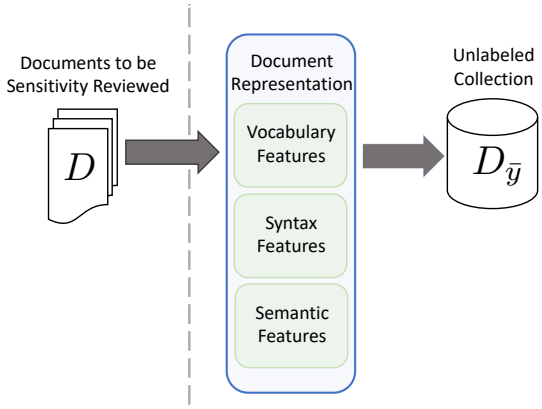



Figure 4.3: The Document Representation component of our technology-assisted sensitivity review framework.

tains additional *engineered* document features that are effective for enabling a learned classifier to identify documents that do, or do not, contain sensitive information. The resulting document collection, $D_{\bar{y}}$, is *unlabelled*, i.e., at this stage in the process, the documents do not have any associated class labels (e.g., *sensitive* or *not-sensitive*) or any additional information that has been supplied by a human reviewer.

The Document Representation component is decoupled from the individual user models. Its main purpose, in addition to the basic transformation of documents, is to engineer useful document features for sensitivity classification. The expected volumes of individual types of sensitivity vary between specific government departments (The National Archives, 2016*b*). For example, in the UK, the Foreign and Commonwealth Office (FCO) encounters many more international relations sensitivities than, for example, the Department of Health. As previously stated in Section 3.4, the collection of documents that we use in this thesis is a set of formal communications between central government and embassies around the world. The text of these documents is prose and, therefore, has no inherent structure beyond typical structures such as sentences or paragraphs. The feature engineering approaches that the Document Representation component deploys are designed to identify latent structures and relationships within the documents in $D$ that can be used to help to identify the sensitivities. Moreover, the feature engineering approaches depend only on the terms in a collection and, therefore, they can be deployed as a *first line of defence* across government departments.

The component focuses on identifying vocabulary, syntactic and semantic features[1]:

**Vocabulary Features**: This set of features are based on term $n$-grams (Sebastiani, 2002). Term $n$-grams are derived from the distributions of terms in the vocabulary, $V$, of the collection $D$. They are designed to capture the context in which a term appears by treating multiple terms that appear in close proximity to each other in a document as a single token, before calculating a

[1]We note that, other features could also be useful for sensitivity classification, for example document metadata such as a document's author or creation time. However, we leave this to future work.

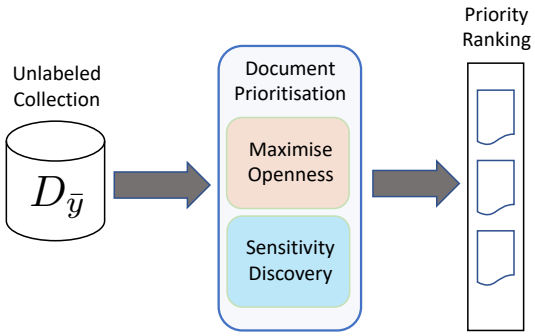



Figure 4.4: The Document Prioritisation component of our technology-assisted sensitivity review framework.

distribution statistic such as TF-IDF.

**Syntax Features**: Natural language processing (NLP) techniques have been a very popular, and effective, way to have a computer try to understand the underlying structures and meaning of written text for the last thirty years (Johnson, 2009). One of the main techniques of NLP is Parts-of-speech (POS) tagging. POS, e.g., nouns, pronouns, verbs, adverbs, adjectives, conjunctions, prepositions, and interjections, are categories of words that have similar grammatical properties and, typically, follow similar syntactic roles within the structure of sentences. This set of features are derived from the *sequences* of parts of speech tags that are within a document and, moreover, the document collection.

**Semantic Features**: As we previously discussed in Section 1.2, sensitive information is often a product of a combination of factors, such as *who said what about whom*. Sentences that share this type of structure can be said to be *semantically similar*. This set of features identifies latent semantic relations that appear frequently in sensitive text and that can, moreover, be effective for predicting if a document contains sensitive information.

## 4.5 Document Prioritisation

The Document Prioritisation component identifies the documents that should be prioritised for review, at any particular stage of the reviewing process. Throughout the sensitivity review process, our proposed framework aims to (1) learn from the sensitivity judgements that the reviewer makes and (2) increase the productivity of a reviewer by, for example, increasing the number of documents that can be released to the public within a specified time period. However, this can result in competing objectives that require the framework to prioritise different documents for review at any particular stage of the review process. For example, the objective may be to learn an effective sensitivity classifier quickly so that it can inform the reviewer which documents

contain sensitive information. In this case, the documents that are the most informative for the classifier should be prioritised for review. However, if the objective is to maximise the number of documents that are reviewed and released to the public, then the documents that are not sensitive should be prioritised for review.

The input to the Document Prioritisation component is the unlabelled collection $D_{\bar{y}}$ that is output by the Document Representation component, previously discussed in Section 4.4. The Document Prioritisation component is responsible for ranking the documents in $D_{\bar{y}}$ and selecting $k$ high priority documents to present to the reviewer. The component selects the most appropriate documents to have reviewed by deploying a document selection strategy based on the current knowledge of the sensitivity classifier and the appropriate reviewing user model. For the exhaustive review user model, the component can be instantiated to prioritise discovering the sensitivities in the collection (to improve the sensitivity classifier). Alternatively, for the limited review user model, the component can be instantiated to maximise openness.

**Sensitivity Discovery**: Sensitive information is broadly defined and context dependent. Moreover, much of the information that is likely to be sensitive in each of the collections that have to be sensitivity reviewed is likely to be very different, and possibly unrelated. Therefore, when a collection is to be sensitivity reviewed, the initial role of our proposed framework is to ensure that the framework classifier's can quickly learn to accurately predict the sensitivities that are in the specific collection that is being reviewed. Our proposed framework deploys an active learning approach to accomplish this. As previously discussed in Section 2.3, in active learning, the learning algorithm is allowed to select which documents to have sensitivity reviewed such that it can learn an effective classification model more quickly. Moreover, the goal is to do this using the least reviewer labelling effort possible. An active learning classifier has to deploy a strategy for predicting the informativeness of documents so that it can choose which documents to have labelled by the human reviewer.

The responsibility of the sensitivity discovery instantiation of the Document Prioritisation component is, therefore, to select and deploy an effective active learning strategy to enable the framework to quickly learn to identify the particular sensitivities within the collection that is to be sensitivity reviewed. Moreover, to do this using the least reviewing effort possible.

**Maximise Openness**: When the Document Prioritisation component is instantiated to maximise openness, the objective is to prioritise for review documents that are not sensitive. The first step of this process is to know when an effective (*enough*) sensitivity classifier has been learned, so that the component can switch the ranking (document selection) strategy from sensitivity discovery to maximising openness. Freund *et al.* (1992) showed that, for a collection of $m$ documents, the number of labelled example documents that are needed to be able to train an effective classifier is $O(\frac{1}{m})$, assuming that there are (1) no noisy examples in the training data,

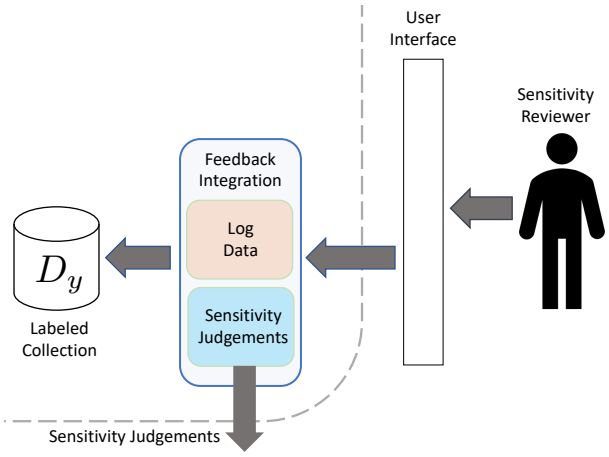



Figure 4.5: The Feedback Integration component of our technology-assisted sensitivity review framework.

(2) a perfect deterministic classifier, and (3) the possibility to select classifiers randomly from the *version space* of a *query by committee* active learning strategy. In practice, the three listed assumptions are not possible for real world classification tasks. It is, however, possible to learn a classifier that is optimal, given the available data, without having to have associated class labels for *all* of the examples in the collection. This component deploys a *stopping criteria* (Lewis & Gale, 1994; Schohn & Cohn, 2000) when it detects that an effective classifier has been learned and uses the predictions from the classifier, along with additional information obtained from the reviewers' log data, to prioritise the documents that are not sensitive and that are expected to take less time to review.

## 4.6 Feedback Integration

Figure 4.5 illustrates the third component of our proposed technology-assisted sensitivity review framework, the Feedback Integration component. One of the challenges of developing a technology-assisted review framework for sensitivity review is how to learn from the reviewers to (1) improve the framework's effectiveness and (2) assist the reviewer as best as possible. The Feedback Integration component addresses this challenge. The component takes as input the sensitivity reviewer's feedback in the form of sensitivity judgements and the log data from the user interface that is used to review the documents. The role of the component is to integrate the reviewer's feedback into the document vector representations, in a way that provides additional useful information to the Learned Predictions component, which we will discuss in the following section.

The feedback from a reviewer's sensitivity judgements are primarily used as additional evidence in the exhaustive review user model to improve active learning strategies. The reviewers log data is used for the limited review user model to predict how long a reviewer will take to review a document so that the Document Prioritisation component can prioritise the documents

that are quicker to review to increase the number of documents that can be reviewed and released to the public.

**Sensitivity Judgements**: As we have previously discussed, when a human sensitivity reviewer reviews a document, the reviewer provides a sensitivity judgement as to whether the document contains any sensitive information. A sensitivity judgement can contain multiple pieces of information, or *attributes*. Table 4.1 provides a description of the attributes of a sensitivity judgement, that we propose. For each document that is reviewed, the reviewer provides a class label, $l_i, l \in \{sensitive, nonSensitive\}$. Moreover, for sensitive documents, i.e., for documents that the reviewer labels as $l_{sensitive}$, the sensitivity judgement also contains a set of *annotations* that identify the passages of text that the reviewer believes to be sensitive. Therefore, a sensitivity judgement for a sensitive document can contain 0, 1 or many annotations (since there may be many sensitive passages in a document) and an annotation can vary in size between a single term and all of the terms in the document.

Table 4.1: The attributes of a sensitivity judgement.

| Attribute | Description |
|---|---|
| DocID | A unique identifier for a document |
| Class Label | A class label, $l_i, l \in \{sensitive, nonSensitive\}$, is supplied for each document that is reviewed. |
| Annotations | For a document, $d_i$, that the reviewer has labelled $l_{sensitive}$, a sensitivity judgement also contains a set of text-level annotations, $a_{di}, \lvert a_d \rvert \in \{0..\ \lvert d_i \rvert\}$, that indicate which text within the document led to the reviewer's decision that the document is sensitive and which of the sensitivity subcategories (from Tables 3.1 and 3.2) the sensitivities relate to. |

The Feedback Integration component is responsible for updating the document vectors to incorporate the reviewer feedback in a sensitivity judgement, in a way that provides the most useful information to a sensitivity classifier. For example, the component can re-weight the importance of specific terms in a document's vector representation, based on a feedback integration strategy.

**Log Data** The Feedback Integration component is also responsible for integrating the log data from the user interface (i.e., the reviewer's interactions with the reviewing interface) to model the reviewers behaviour. Figure 4.6 shows an example of the log data that the component takes as input. As can be seen from Figure 4.6, a time-stamp is logged each time that a document is loaded (DOCUMENT_LOADED), a reviewing judgement is made (0 = Not Sensitive, 1 = Section 27, 2 = Section 40 or 3 = Both 27 and 40) and if the reviewer pauses (PAUSED) or restarts (RESTARTED) the reviewing task.

Table 4.2 presents the attributes of a reviewer's interactions that the Feedback Integration component generates from the log data. As can be seen from Table 4.2, the component generates two reviewer interaction attributes, the reviewer's dwell times (i.e., the length of time the reviewer took to review a document) and a record of documents that the reviewer judged prior

```
Sun Mar 18 12:58:02 2018 Review_0_8 review8/doc622 DOCUMENT_LOADED
Sun Mar 18 13:10:06 2018 Review_0_8 review8/doc622 1
Sun Mar 18 13:10:06 2018 Review_0_8 review8/doc1281 DOCUMENT_LOADED
Sun Mar 18 13:13:51 2018 Review_0_8 review8/doc1281 PAUSED
Sun Mar 18 13:14:38 2018 Review_0_8 review8/doc1281 RESTARTED
Sun Mar 18 13:18:55 2018 Review_0_8 review8/doc1281 1
Sun Mar 18 13:18:55 2018 Review_0_8 review8/doc204472 DOCUMENT_LOADED
Sun Mar 18 13:21:44 2018 Review_0_8 review8/doc204472 0
Sun Mar 18 13:21:44 2018 Review_0_8 review8/doc215 DOCUMENT_LOADED
Sun Mar 18 13:21:54 2018 Review_0_8 review8/doc204472 DOCUMENT_LOADED
Sun Mar 18 13:22:04 2018 Review_0_8 review8/doc204472 0
Sun Mar 18 13:22:04 2018 Review_0_8 review8/doc215 DOCUMENT_LOADED
Sun Mar 18 13:22:09 2018 Review_0_8 review8/doc215 DOCUMENT_LOADED
Sun Mar 18 13:27:06 2018 Review_0_8 review8/doc215 PAUSED
Sun Mar 18 13:29:03 2018 Review_0_8 review8/doc215 RESTARTED
Sun Mar 18 13:30:16 2018 Review_0_8 review8/doc215 3
Sun Mar 18 13:30:16 2018 Review_0_8 review8/doc3384 DOCUMENT_LOADED
```

Figure 4.6: Example of the Feedback Integration component's log data input.

Table 4.2: The attributes of a reviewer's interactions from the reviewing interface log data.

| Attribute | Description |
|---|---|
| DocID | A unique identifier for a document |
| Dwell times | A list of the length of time that the reviewer spent reviewing the document each time that the reviewer viewed the document. |
| Previously Judged | A record of documents that the reviewer judged before the current document. |

to judging the current document.

The Feedback Integration component integrates the log data information from multiple reviewers, along with additional information about the documents in the collection and the predicted sensitivity of documents, to model reviewer behaviour. Modelling reviewer behaviour enables the Learned Predictions component, discussed in the following section, to predict how long a reviewer will require to review a specific document from the unlabelled collection.

## 4.7 Learned Predictions

The fourth, and final, component of our proposed framework is the Learned Predictions component. This component is a collection of supervised machine learning algorithms that are responsible for making predictions about the documents that have not yet been reviewed, i.e., the unlabelled collection $D_{\bar{y}}$. The final step of each iteration of the process presented in Figure 4.2 is to extend the document representations in $D_{\bar{y}}$ to incorporate the classifier's predictions (in the current iteration of the review cycle) about each of the documents. The component is responsible for two types of predictions: For the exhaustive review user model, the component is responsible

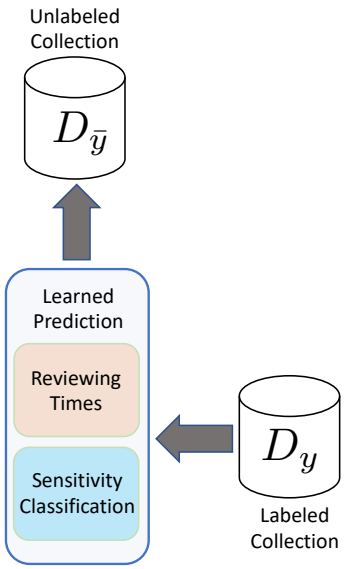



Figure 4.7: The Learned Predictions component of our technology-assisted sensitivity review framework.

for predicting the sensitivity classification of documents to provide additional information that can be (1) used by the Document Prioritisation component, and (2) provided to a reviewer to assist them to make quicker decisions; for the limited review user model, the component uses the classification predictions and information about reviewers' interactions to predict the length of time that will be required by a reviewer to review a document.

**Sensitivity Classification**: A sensitivity classifier that can accurately and reliably predict if a document contains any sensitive information, that should not be released to the public, is a fundamental component of our proposed framework. Moreover, developing a sensitivity classifier is a central component of this thesis. In our proposed framework, sensitivity classification predictions provide additional information that the Document Prioritisation component can use, in the exhaustive review user model, to make more intelligent decisions when selecting documents to have reviewed to identify the sensitivities in a collection (and in turn improve the performance of the sensitivity classifier). Moreover, the sensitivity predictions can inform a reviewer if a document is likely to be sensitive or not, to help to reduce the time that it takes to review a collection of documents and increase the agreement between reviewers. Furthermore, these predictions are also used in the limited review user model to prioritise documents that are not sensitive.

**Reviewing Times**: Predicting the length of time that a reviewer is likely to take to review a document enables our proposed framework to prioritise specific documents to be reviewed to increase the number of documents that can be reviewed *and* released to the public with the available reviewing resources in the limited review user model.

## 4.8 Conclusions

In this chapter, we proposed a framework for technology-assisted sensitivity review. Our proposed framework consists of four components, namely *Document Representation*, *Document Prioritisation*, *Feedback Integration* and *Learned Predictions* and is designed to adapt the assistance that it provides to sensitivity reviewers for two realistic digital sensitivity review scenarios that we have identified as user models for technology-assisted sensitivity review: *exhaustive review* and *limited review*.

In particular, in Section 4.2, we discussed the identified user models that our proposed framework addresses, before providing an overview of the framework and its components in Section 4.3. We then discussed the role of each of the components individually in Sections 4.4 to Section 4.7 and how each component can be instantiated to address the identified user models. Moreover, we discussed how each component contributes to validating the objectives of our thesis statement, namely: to incorporate explicit reviewer feedback to learn an effective sensitivity classifier more quickly; to provide reviewers with classification predictions to reduce the time that is required to sensitivity review a collection of documents and increase the agreement between reviewers; and to prioritise non-sensitive documents for review to increase the number of documents that can be reviewed and released to the public with the available reviewing resources.

Automatically identifying documents that contain sensitive information is fundamental to our proposed framework. In the following chapter, we, firstly, evaluate the effectiveness of a document sanitisation approach, that we discussed in Section 3.3, for identifying the S27 international relations sensitivity *information supplied in confidence*. We will show that document sanitisation is not an appropriate approach for identifying the sensitivities that we address in this thesis. Therefore, we propose to address this problem as a document (text) classification task. Moreover, we present and empirically evaluate the baseline sensitivity classification approach that we build on in the remaining chapters of this thesis.

# Chapter 5

# Sensitivity Classification Baseline

## 5.1 Introduction

In the previous section, we introduced our framework that we propose for technology-assisted sensitivity review. Moreover, we discussed the two realistic digital sensitivity review scenarios that we have identified as user models to evaluate the effectiveness of our proposed framework for assisting digital sensitivity review, namely the limited review user model and the exhaustive review user model (see Section 4.2). We introduced each of the four components of our framework, namely the Document Representation component (see Section 4.4), the Document Prioritisation component (see Section 4.5), the Feedback Integration component (see Section 4.6) and the Learned Predictions component (see Section 4.7). Moreover, we described how each of the four components can be instantiated to assist sensitivity reviewers, and government departments, in the limited review and exhaustive review user models.

Our proposed framework is built upon our argument that sensitivity reviewers can be *assisted* by automatically identifying which of the documents in a collection contain sensitive information that is exempt from public release through the Freedom of Information Act 2000 (c. 36) (FOIA), we refer to this as *sensitivity classification*. However, classifying sensitive information is a complex task. As we previously discussed in Chapter 3 FOIA sensitive information is context-dependent. For example, identifying if information is exempt from release through the FOIA can require a human to make a judgement on the possible effect of releasing the information to the public. Moreover, sensitivity is not necessarily topic-oriented, it is more often due to a combination of what is being said and about whom.

As we previously discussed in Chapter 3, most of the previous literature on identifying, or classifying, sensitive information, such as *document sanitisation*, has focused on identifying personal data. In this chapter we, evaluate the effectiveness of a document sanitisation approach from the literature for classifying the international relations sensitivity *information that has been supplied in confidence*. Through our evaluation, we will demonstrate that document sanitisation is not an effective approach for identifying FOIA sensitivities. Therefore, we propose to address

the problem of sensitivity classification as a text classification task. In this chapter, we empirically evaluate the baseline sensitivity classification approach that we build on in the remainder of this thesis. Furthermore, we compare the effectiveness of learning to classify sensitive information at different levels of granularity (i.e., learning to classify sensitive information as a single category vs. learning to classify the individual FOI exemptions *international relations* and *personal information*) and evaluate combining sensitivity classifiers in an ensemble classification approach for sensitivity classification. The remainder of this chapter is structured as follows:

- Section 5.2 evaluates the effectiveness of a document sanitisation approach from Sánchez *et al.* (2012) for classifying information that has been supplied in confidence. Document sanitisation tries to *mask*, i.e., redact or hide, personal data in documents while retaining the document's utility. The approach from Sánchez *et al.* (2012) has previously been shown to be effective for masking confidential information in on-line personal profiles and has the potential to be effective for identifying *in-confidence* sensitivities. We will empirically show, however, that the approach is not an effective approach for identifying this FOI sensitivity.

- In Section 5.3, we propose to address FOI sensitivity identification as a text classification task. Moreover, we evaluate the effectiveness of text classification for classifying documents by whether they do or do not contain sensitive information (we refer to this task as sensitivity classification). In particular, in this section, we evaluate appropriate document representation and feature reduction strategies for sensitivity classification.

- In Section 5.4, we evaluate the effectiveness of learning to classify the individual sensitivities *international relations* and *personal information*. Moreover, we evaluate the effectiveness of extending the classifiers for individual sensitivities with additional *hand-crafted* features of sensitivity.

- In Section 5.5, we evaluate the impact of classifying sensitive information at different levels of granularity on the overall classification effectiveness. In particular, we compare the effectiveness of classifying sensitive information as a single category of information against classifying individual FOI exemptions, i.e. *international relations* and *personal information*). Moreover, we evaluate the effectiveness of combining multiple sensitivity classifiers in an ensemble classification approach for sensitivity classification.

- In Section 5.6, we summarise the conclusions of this chapter.

## 5.2 Masking Information that is Supplied in Confidence

Document sanitisation tries to *mask* personal data in documents, while retaining the document's utility. The document sanitisation task is, therefore, a classification task at the *term-level*, i.e. the

(a) The text that a human reviewer judged as being sensitive.

Records indicate that many ethnic minority passengers have been turned around since the flight commenced. The resulting negative perception will make the goal of attracting a carrier a difficult task. **Michael agreed to assist with efforts to establish a magistrate's court within the airport to target mala fide passengers**.

(b) The text that a document sanitisation classifier predicted as being sensitive.

Records indicate that many **ethnic minority passengers** have been turned around since the flight commenced. The resulting negative perception will make the goal of attracting a carrier a difficult task. Michael agreed to assist with efforts to establish a **magistrate's court within the airport** to target mala fide passengers.

(c) Confusion

| | |
|---|---|
| TP | magistrate's court within the airport |
| FP | ethnic minority passengers |
| FN | Michael agreed to assist with efforts to establish a to target mala fide passengers |
| TN | Records indicate that many have been turned around since the flight commenced. The resulting negative perception will make the goal of attracting a carrier a difficult task |

Figure 5.1: Document sanitisation analysis. The figure shows a document with (a) the text that a human sensitivity reviewer judged as being sensitive shown in red, and (b) text that the document sanitisation classifier predicted as being sensitive. The figure also show the resulting confusion matrix (c) for the individual terms in the document.

task is to classify individual terms or sequences of terms (e.g.,, sentences) so that specific text in a document can be masked. Therefore, document sanitisation has the potential to effectively identify specific passages of text that should be redacted in government documents. Moreover, the approach could potentially be used for engineering features, or weighting term features, in a classifier that classifies sensitive documents. Figure 5.1 illustrates the document sanitisation classification task and how it is evaluated. The figure presents a document with (a) the text that a human reviewer judged as being sensitive shown in red and (b) the terms that are predicted as being sensitive by a document sanitisation classifier shown in green. The figure also shows the resulting confusion matrix (c) with True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) term predictions. The effectiveness of the classifier can then be measured using standard classification evaluation metrics.

The document sanitisation approach of Sánchez *et al.* (2012) measures the specificity, or Information Content (IC), of a noun phrase (NP) as an indication of potential sensitivity. A NP is a word or string of contiguous words within a sentence that has a noun as the subject, object or preposition, for example *a magistrates court* and *a magistrates court within the airport* are two NPs that could be embedded within a larger sentence. The intuition of the IC approach (Sánchez *et al.*, 2012) is that sensitive information is more specific than information that is not sensitive and, therefore, more specific NPs are more likely to be sensitive.

Sánchez *et al.* (2012) showed that their IC approach can be effective for masking personal information in on-line personal profiles. Moreover, the authors showed that their approach can also be effective for identifying *confidential* information in the profiles. This shows that the approach of Sánchez *et al.* (2012) has the potential to be suitable for identifying the international relations sensitivity *information that has been supplied in confidence*. Therefore, in this section, we evaluate the effectiveness of the approach for classifying this sensitivity. We will show that

the IC approach from Sánchez *et al.* (2012) is not effective for classifying the international relations sensitivity, information that has been supplied in confidence.

### 5.2.1 Experimental Methodology

Following the methodology of Sánchez *et al.* (2012), we evaluate the effectiveness of the IC of NPs for classifying information that has been *supplied in confidence*. To evaluate the approach, we use 143 documents that contain *supplied in confidence* sensitivities, from our test collection that we previously presented in Chapter 3 (Table 3.3). For our ground truth, any terms that were annotated by a sensitivity reviewer, and that the reviewer tagged as being supplied in confidence, was labelled as sensitive. All other terms in the 143 documents were labelled as not sensitive. This results in 10838 terms labelled as *sensitive* and 221055 terms labelled as not sensitive. To calculate the IC of NPs in a document, the document is first parsed to extract its syntactic structure. NPs are then extracted from the resulting syntax tree and submitted to a Web search engine[1] as a query. The IC of the noun phrase is calculated using the number of returned results as an indication of the phrase's specificity. The IC of a NP is computed as:

$$IC_{(NP)} = -\log_2 p_{(NP)} = -\log_2 \frac{res(NP)}{totalpages} \tag{5.1}$$

where $res(NP)$ is the number of returned search results and $totalpages$ is the number of sites indexed by the search engine. Following Sánchez *et al.* (2012), we set the number of total pages to 3.5 Billion. We note that, the IC scores produced by this approach reflect a snapshot in time and could be affected by temporal variations in the state of the search engine (Fetterly *et al.*, 2003; Ntoulas *et al.*, 2004). To mitigate the effects of temporal variations, an alternative approach would be to use a static corpus to calculate the IC scores, for example the Google N-Grams corpus[2]. In our experiments, each term within a noun phrase with an IC score greater than an threshold, $\beta$, is classified as being sensitive, while all other terms are classified as non-sensitive. We empirically evaluate threshold values $\beta \geq \{1,2,3,4,5,6,7,8,9,10,20,30,40,50,75,100\}$. Statistical significance is measured by McNemar's non-parametric test (McNemar, 1947) ($p < 0.05$), to evaluate if the approach is statistically significantly better than random.

### 5.2.2 Results and Discussion

Table 5.1 presents the performance of the IC approach for classifying information that has been supplied in confidence. Firstly, we note from Table 5.1, that when the threshold, $\beta$, is set at $\beta \geq 40$ the approach classifies all of the terms in the collection as being not sensitive and the classifier is effectively random (0.5 BAC). (We note that exactly the same result is obtained if $\beta$ is set at any value higher than $\beta \geq 40$. For brevity, we omit these results from Table 5.1). As the

[1] https://azure.microsoft.com/en-gb/services/cognitive-services/bing-web-search-api
[2] https://catalog.ldc.upenn.edu/LDC2006T13

Table 5.1: Information Content (IC) document sanitisation results. The table shows the resulting precision, recall, $F_1$, $F_2$ and Balanced Accuracy (BAC) scores. Threshold values, $\beta$, that are statistically significantly better than random (McNemar's test, $p < 0.05$) are denoted by †.

| | | Precision | Recall | $F_1$ | $F_2$ | BAC |
|---|---|---|---|---|---|---|
| $IC_{\beta \geq 7}$ | † | **0.2564** | **0.0116** | **0.0222** | **0.0143** | **0.5048** |
| $IC_{\beta \geq 8}$ | † | 0.2388 | 0.0103 | 0.0197 | 0.0127 | 0.5042 |
| $IC_{\beta \geq 9}$ | † | 0.2281 | 0.0094 | 0.0180 | 0.0116 | 0.5038 |
| $IC_{\beta \geq 10}$ | † | 0.2466 | 0.0093 | 0.0179 | 0.0115 | 0.5038 |
| $IC_{\beta \geq 20}$ | | 0.1911 | 0.0055 | 0.0108 | 0.0069 | 0.5021 |
| $IC_{\beta \geq 30}$ | | 0.0429 | 0.0004 | 0.0008 | 0.0005 | 0.5000 |
| $IC_{\beta \geq 40}$ | | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.5000 |

threshold value is reduced, we start to see a statistically significant improvident in performance when $\beta \geq 10$. However, at this $\beta$ value, the approach only achieves 0.5038 BAC. The approach performs best when $\beta \geq 7$ (the approach performs identically if the threshold is set lower than $\beta \geq 7$. Again, for brevity, we omit these results from Table 5.1.). However, even for this best performing $\beta$ value, the approach only identifies a very small portion of the sensitive text (0.0116 recall). Moreover, setting $\beta \geq 7$ results in a precision score of 0.2564, so only a small portion of the terms that are predicted as being sensitive were actually identified as being sensitive by the human assessor. This shows that the approach does not provide an effective method of identifying information that has been supplied in confidence.

## 5.3 Sensitivity Classification Baseline

We propose to address the task of identifying documents that contain FOI sensitivities as a text classification task. We argue that deploying a text classification strategy is a reasonable starting point for developing sensitivity classification, since it has been shown to be effective for many other document classification tasks (Sebastiani, 2002). Moreover, text classification is often an effective approach for discovering latent patterns in distributions of text (Allahyari *et al.*, 2017). Therefore, we argue that it has the potential to be effective at discovering, and classifying, the more challenging FOI sensitivities, such as international relations. In this section, we evaluate the effectiveness of text classification as a baseline sensitivity classification approach that we build on in the remainder of this thesis. In particular, we evaluate appropriate document representation and feature reduction strategies for sensitivity classification. We present our experimental methodology in Section 5.3.1 and discuss the results in Section 5.3.2.

### 5.3.1 Experimental Methodology

In our experiments, we deploy a Support Vector Machine (Vapnik, 1995) (SVM) with a linear kernel as our classifier. We select SVM as our classifier due to three observations that make it particularly suitable for our task. Firstly, SVM has been shown to be the most effective classifier for many other classification tasks (Sebastiani, 2002). Secondly, SVM's default parameter settings are theoretically motivated and have previously been shown to be the most effective for text classification (Joachims, 1998)[3]. Thirdly, any algorithm that is deployed by a government department to make, or assist in making, decisions is likely to become part of the public record and, hence, it needs to have an acceptable level of transparency so that the department can adequately explain the decisions that are made. By deploying a SVM classifier with a linear kernel, we can use the perpendicular distance of a term feature from the separating hyperplane as a heuristic to measure the importance of the term for the classifier's prediction decision (Guyon *et al.*, 2002). This, in-turn, provides a reasonable level of explainability since we can: (1) rank the terms in the document collection that the classifier was trained on by how important they are, or the amount of influence that they have on, a deployed classification model; and (2) rank the terms in a document that is predicted as being (not) sensitive by how much they contribute to the prediction. Additionally, it is worth noting that SVM can be a robust approach when documents are represented by vectors with a very high number of dimensions (Joachims, 1998), which also makes it particularly suited to text classification tasks. Moreover, this means that, with a SVM classifier, there is often no need to apply very aggressive feature reduction techniques that can result in a loss of information.

In this section, we wish to answer three research questions:

- RQ5.1: What is that most effective term distribution statistic for representing documents for sensitivity classification?
- RQ5.2: What is the most appropriate feature reduction approach, if any, for sensitivity classification?
- RQ5.3: What is the most effective document representation and feature reduction combination?

As term statistics for document representations, we evaluate Binary (BIN), tf (TF) and TF-IDF statistics, where TF-IDF is calculated using the formula that we presented in Eq. 2.2. For feature reduction techniques we evaluate stopword removal and stemming as *basic* feature reduction approaches. Moreover, we evaluate combinations of the basic approaches with two *advanced* feature reduction techniques. Firstly, the *information Gain* (IG) feature reduction approach that we previously presented in Chapter 2 (Eq. 2.4). Secondly, the *Chi-Squared* ($\chi^2$)

[3]We note, however, that this is dependent on the chosen evaluation metric and may not necessarily consistently be the case for sensitivity classification. We will investigate this as future work.

Table 5.2: The feature reduction combinations that we evaluate for each of the document representation approaches binary (BIN), tf (TF) and TF-IDF. Each of the document representations is evaluated with combinations of *Basic* feature reduction (stopword removal and stemming), Basic plus Information Gain (IG) and Basic plus Chi-Squared ($\chi^2$) feature reduction.

| | Stopwords Removed, No Stemming | Stopwords Retained, No Stemming | Stopwords Removed, Stemming Applied | Stopwords Retained, Stemming Applied |
|---|---|---|---|---|
| Basic | noSpNoSm | stopNoSm | noSpStem | stopStem |
| Basic + IG | noSpNoSm_IG | stopNoSm_IG | noSpStem_IG | stopStem_IG |
| Basic + $\chi^2$ | noSpNoSm_$\chi^2$ | stopNoSm_$\chi^2$ | noSpStem_$\chi^2$ | stopStem_$\chi^2$ |

feature reduction approach that we also presented in Chapter 2 (Eq. 2.6). Table 5.2 presents the abbreviations that we use in Section 5.3.2 to represent the combinations of feature reduction techniques that we evaluate for each of the document representation strategies.

As our test collection, we use the collection of 3801 documents that we presented in Chapter 3 (Table 3.3). We label all of the documents in the collection that were judged as containing international relations or personal information as *sensitive*, all other documents are labelled *not sensitive*. We, therefore, perform a binary classification *sensitive* vs *not sensitive*. We split the test collection into training, validation and test sets and perform 5-fold Cross Validation. For the IG, $\chi^2$ and Freq feature reduction approaches, we learn a reduction threshold on the validation set of each of the 5-fold Cross Validation folds. We test for threshold values that retain 1%, 5%, 10%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 75%, 80%, 90% and 95% of the terms in the collection and optimise for the $F_2$ metric, since it is a recall oriented weighted mean of precision and recall. We use the Scikit-learn[4] machine learning library for all of our experiments in this chapter. When removing stopwords, we remove all terms that are in Scikit-learn's standard English stopword list. To apply stemming, we use the Natural Language Toolkit[5] implementation of Porter stemmer. For classification, we set the SVM C parameter, which provides a trade-off between the importance of minimising the number of mis-classified documents in the training data and maximising the margin between the support vectors that lie on the margin boundaries and the separating hyperplane of the learned model, to its default 1.0.

In Table 5.3, we report the precision, recall, $F_1$, $F_2$, balanced accuracy (BAC) and the area under the receiver-operator characteristic curve (auROC) metrics. The receiver-operator characteristic (ROC) (Fawcett, 2006) is an technique for visualising and analysing the performance of a classifier. Classifiers, such as SVM, that output a numeric score to indicate the classifier's confidence about (or the probability of) the instance's class membership can be used with a *threshold* to create a binary classifier. Instances with a classification score that is above the threshold are classified as *positive* and instances with a score below the threshold are classified as *negative*. Conceptually, the classification threshold can be varied to produce a *conservative* classifier that

[4]http://scikit-learn.org/stable/index.html
[5]https://www.nltk.org/

emphasises precision (i.e., only high scoring instances are classified as positive) or a *liberal* classifier that emphasises recall (i.e., the threshold is low and all documents with scores above the threshold are classified as positive). Moreover, the True Positive Rate (TPR), defined as $\frac{PositivesCorrectlyClassified}{TotalPositives}$, and the False Positive Rate (TPR), defined as $\frac{NegativesIncorrectlyClassified}{TotalNegatives}$, can be calculated for any threshold value. In a ROC analysis, the TPR is plotted on the *y* axis and the FPR is plotted on the *x* axis to produce a step-function (or curve, if the number of instances is large enough) that plots the performance of the classifier as its threshold is varied. When comparing the performance of multiple classifiers, it is often appropriate to compare how the classifiers' relative performance is affected by the precision recall trade-off as the classification threshold is varied. However, in our experiments to identify a baseline sensitivity classification approach, varying the classification threshold introduces uncertainty to our evaluation process, since we would need to make assumptions about how the precision recall trade-off affects the human sensitivity reviewers as they perform their task. Therefore, in our experiments, we use the default SVM threshold as our fixed operating point and report the area under the ROC curve (auROC). The auROC denotes the probability that a randomly selected positive instance is ranked above a randomly selected negative instance, when the classified documents are ranked by the output of the classifier's decision function.

Statistical significance is measured by McNemar's non-parametric test (McNemar, 1947) ($p < 0.05$). When testing for significance, we perform two sets of significance tests. Firstly, we evaluate the document representation approaches for a single feature reduction combination, e.g.,, $\text{BIN}_{\text{noSpStem\_IG}}$ vs. $\text{TF}_{\text{noSpStem\_IG}}$ vs. $\text{TF-IDF}_{\text{noSpStem\_IG}}$, we refer to these tests as *between-doc* tests. Secondly, we test for significant differences between basic and advanced feature reduction techniques for each of the document representation and simple feature reduction pair, e.g.,, $\text{BIN}_{\text{stopStem}}$ vs. $\text{BIN}_{\text{stopStem\_}\chi^2}$ vs. $\text{BIN}_{\text{stopStem\_IG}}$, we refer to these tests as *within-doc* tests. In Table 5.3, any approach that performs statistically significantly better than the next best performing approach, according to BAC, is denoted as † in between-doc tests, and as ‡ in within-doc tests.

### 5.3.2 Results and Discussion

Table 5.3 (on the following page) presents the results of our experiments combining document representations and feature reduction approaches for sensitivity classification. Firstly, addressing RQ5.1, we wish to know what is the best performing document representation strategy. We note, from Table 5.3, that in the *between-doc* statistical significance tests, denoted as † in Table 5.3, the TF-IDF document representation strategy results in the greatest number of statistically significant improvements when compared with the BIN or TF representations using the same feature reduction approach. Eight combinations of TF-IDF and a feature reduction approach ($\text{TF-IDF}_{\text{noSpNoSm}}$, $\text{TF-IDF}_{\text{noSpStem}}$, $\text{TF-IDF}_{\text{noSpStem\_}\chi^2}$, $\text{TF-IDF}_{\text{noSpStem\_IG}}$, $\text{TF-IDF}_{\text{stopNoSm}}$, $\text{TF-IDF}_{\text{stopStem}}$, $\text{TF-IDF}_{\text{stopStem\_}\chi^2}$ and $\text{TF-IDF}_{\text{stopStem\_IG}}$) statistically significantly

Table 5.3: Stopword removal, stemming, Information Gain (IG) and Chi-Squared ($\chi^2$) feature reduction combinations for binary (BIN), *tf* (TF) and TF-IDF document representations. The table presents precision, recall, $F_1$, $F_2$ and Balanced Accuracy (BAC) scores. Statistical significance is denoted by † for *between-doc* tests and ‡ for *within-doc* tests (McNemar's test $p < 0.05$).

| | | | Precision | Recall | $F_1$ | $F_2$ | BAC | auROC |
|---|---|---|---|---|---|---|---|---|
| $BIN_{noSpNoSm}$ | | ‡ | 0.2462 | 0.6613 | 0.3583 | 0.4937 | 0.6763 | 0.7455 |
| $BIN_{noSpNoSm\_\chi^2}$ | † | | 0.2323 | 0.6414 | 0.3408 | 0.4739 | 0.6593 | 0.7285 |
| $BIN_{noSpNoSm\_IG}$ | † | | 0.2381 | 0.6454 | 0.3473 | 0.4800 | 0.6652 | 0.7404 |
| $BIN_{noSpStem}$ | | ‡ | 0.2335 | 0.6433 | 0.3424 | 0.4758 | 0.6607 | 0.7329 |
| $BIN_{noSpStem\_\chi^2}$ | | | 0.2276 | 0.6553 | 0.3378 | 0.4761 | 0.6581 | 0.7173 |
| $BIN_{noSpStem\_IG}$ | | | 0.2304 | 0.6614 | 0.3417 | 0.4812 | 0.6624 | 0.7143 |
| $BIN_{stopNoSm}$ | | | 0.2445 | 0.6553 | 0.3556 | 0.4896 | 0.6734 | 0.7414 |
| $BIN_{stopNoSm\_\chi^2}$ | | ‡ | 0.2401 | 0.6692 | 0.3528 | 0.4920 | 0.6729 | 0.7284 |
| $BIN_{stopNoSm\_IG}$ | † | | 0.2411 | 0.6493 | 0.3511 | 0.4842 | 0.6688 | 0.7384 |
| $BIN_{stopStem}$ | | | 0.2349 | 0.6274 | 0.3415 | 0.4698 | 0.6576 | 0.7248 |
| $BIN_{stopStem\_\chi^2}$ | | ‡ | 0.2299 | 0.6613 | 0.3410 | 0.4805 | 0.6617 | 0.7161 |
| $BIN_{stopStem\_IG}$ | | | 0.2329 | 0.6633 | 0.3445 | 0.4839 | 0.6649 | 0.7152 |
| $TF_{noSpNoSm}$ | | ‡ | 0.2243 | 0.6452 | 0.3323 | 0.4681 | 0.6526 | 0.7040 |
| $TF_{noSpNoSm\_\chi^2}$ | | | 0.2238 | 0.6572 | 0.3333 | 0.4725 | 0.6549 | 0.6970 |
| $TF_{noSpNoSm\_IG}$ | | | 0.2200 | 0.6353 | 0.3262 | 0.4601 | 0.6459 | 0.7018 |
| $TF_{noSpStem}$ | | | 0.2253 | 0.6553 | 0.3347 | 0.4731 | 0.6555 | 0.7041 |
| $TF_{noSpStem\_\chi^2}$ | | | 0.2186 | 0.6434 | 0.3256 | 0.4620 | 0.6456 | 0.6907 |
| $TF_{noSpStem\_IG}$ | | | 0.2246 | 0.6553 | 0.3339 | 0.4724 | 0.6547 | 0.7035 |
| $TF_{stopNoSm}$ | † | | 0.2293 | 0.6532 | 0.3391 | 0.4763 | 0.6588 | 0.7069 |
| $TF_{stopNoSm\_\chi^2}$ | | | 0.2252 | 0.6453 | 0.3333 | 0.4689 | 0.6524 | 0.7097 |
| $TF_{stopNoSm\_IG}$ | | | 0.2271 | 0.6513 | 0.3365 | 0.4736 | 0.6565 | 0.7036 |
| $TF_{stopStem}$ | † | | 0.2272 | 0.6573 | 0.3374 | 0.4762 | 0.6584 | 0.7084 |
| $TF_{stopStem\_\chi^2}$ | | | 0.2248 | 0.6434 | 0.3329 | 0.4682 | 0.6522 | 0.7011 |
| $TF_{stopStem\_IG}$ | | | 0.2247 | 0.6533 | 0.3341 | 0.4723 | 0.6551 | 0.7063 |
| $TF\text{-}IDF_{noSpNoSm}$ | † | | 0.2376 | 0.6612 | 0.3489 | 0.4863 | 0.6678 | 0.7394 |
| $TF\text{-}IDF_{noSpNoSm\_\chi^2}$ | | | 0.2359 | 0.6612 | 0.3470 | 0.4847 | 0.6659 | 0.7375 |
| $TF\text{-}IDF_{noSpNoSm\_IG}$ | | | 0.2386 | 0.6612 | 0.3499 | 0.4870 | 0.6689 | 0.7386 |
| $TF\text{-}IDF_{noSpStem}$ | † | | 0.2399 | 0.6312 | 0.3469 | 0.4747 | 0.6630 | 0.7316 |
| $TF\text{-}IDF_{noSpStem\_\chi^2}$ | † | | 0.2371 | 0.6291 | 0.3437 | 0.4717 | 0.6603 | 0.7294 |
| $TF\text{-}IDF_{noSpStem\_IG}$ | † | | 0.2405 | 0.6332 | 0.3477 | 0.4759 | 0.6637 | 0.7314 |
| $TF\text{-}IDF_{stopNoSm}$ | † | ‡ | **0.2546** | **0.6831** | **0.3701** | **0.5098** | **0.6882** | **0.7518** |
| $TF\text{-}IDF_{stopNoSm\_\chi^2}$ | | ‡ | 0.2499 | 0.6811 | 0.3648 | 0.5050 | 0.6835 | 0.7503 |
| $TF\text{-}IDF_{stopNoSm\_IG}$ | | | 0.2519 | 0.6731 | 0.3657 | 0.5029 | 0.6833 | 0.7513 |
| $TF\text{-}IDF_{stopStem}$ | † | | 0.2491 | 0.6312 | 0.3562 | 0.4814 | 0.6699 | 0.7390 |
| $TF\text{-}IDF_{stopStem\_\chi^2}$ | † | | 0.2452 | 0.6353 | 0.3534 | 0.4812 | 0.6680 | 0.7340 |
| $TF\text{-}IDF_{stopStem\_IG}$ | † | | 0.2444 | 0.6312 | 0.3513 | 0.4778 | 0.6659 | 0.7342 |

improve upon the BAC score of the next best performing BIN or TF representation with the same feature reduction technique. Moreover, the TF-IDF representation with stopwords retained and no stemming applied ($\text{TF-IDF}_{\text{stopNoSm}}$) achieves the best overall scores for all of the reported metrics. Therefore, in response to RQ5.1, we conclude that TF-IDF is the most effective document representation strategy for sensitivity classification, from the strategies that we evaluate on our corpus. Turning our attention to RQ5.2, we wish to know what is the most effective feature reduction technique. Firstly, focusing on the advanced feature reduction techniques IG and $\chi^2$, we note from Table 5.3 that not applying these techniques results in the highest $F_2$ and BAC scores for seven combinations ($\text{BIN}_{\text{noSpNoSm}}$, $\text{BIN}_{\text{stopNoSm}}$, $\text{TF}_{\text{noSpStem}}$, $\text{TF}_{\text{stopNoSm}}$, $\text{TF}_{\text{stopStem}}$, $\text{TF-IDF}_{\text{stopNoSm}}$ and $\text{TF-IDF}_{\text{stopStem}}$) out of the twelve document representation and *basic* feature reduction combinations. However, from the results of the within-doc statistical significance tests, denoted as ‡, we note that only two of these approaches are statistically significant improvements ($\text{BIN}_{\text{noSpNoSm}}$ and $\text{TF-IDF}_{\text{stopNoSm}}$) compared to when IG or $\chi^2$ feature reduction is added to the approach. Deploying IG or $\chi^2$ feature reduction does not result in a statistically significant, best performing feature reduction combination (according to any metric in Table 5.3) for any of the document representation strategies. Therefore, the advanced feature reduction techniques IG and $\chi^2$ do not improve sensitivity classification on our corpus. Turning our attention to the basic feature reduction techniques, stopword removal and stemming, (without IG or $\chi^2$) we note that retaining stopwords and not applying stemming (stopNoSm) is the most effective combination, according to $F_2$ and BAC, for the TF and TF-IDF document representations (and the second best combination for BIN). Therefore, in response to RQ5.2, we conclude that retaining stopwords, not applying stemming and not applying IG or $\chi^2$ is the best feature reduction combination for sensitivity classification on our corpus. Finally, turning our attention to RQ5.3, as we previously stated, the TF-IDF document representation with stopwords retained and no stemming applied ($\text{TF-IDF}_{\text{stopNoSm}}$) achieves the best scores for all of the reported metrics. Moreover, we have already shown (in response to RQ5.1) that TF-IDF is, overall, the most effective document representation strategy and (in response to RQ5.2) that stopNoSm is the most effective feature reduction strategy. Therefore, in response to RQ5.3 we conclude that there is clear evidence that selecting TF-IDF as our document representation and not applying any feature reduction (i.e. retaining stopwords) is the most effective combination for sensitivity classification on our corpus. It is worth noting that removing stopwords is often assumed to be an important step in text classification (Sebastiani, 2002; Silva & Ribeiro, 2003). However, there have been a number of notable studies in which retaining stopwords increases the effectiveness of text classification algorithms, e.g., (Nigam *et al.*, 2000; Riloff, 1995; Song *et al.*, 2005).

In conclusion, as our baseline text classification approach for sensitivity classification, we select the combination of TF-IDF for term features, we retain the stopwords in our collection and we do not apply stemming or any other feature reduction techniques. In this section, we

have evaluated classifying *sensitive* information as a single category of information. However, another viable approach for sensitivity classification is to classify each FOI exemption, i.e. international relations and personal information, individually. We evaluate classifying individual FOI sensitivities in the following section.

## 5.4 Classifying Individual Sensitivities

In the previous section, we investigated classifying sensitive information as a single category of information. It is reasonable to assume that considering sensitive information as a single *complex* category of information might be a more challenging classification task than classifying a more specific category of sensitive information, such as personal information. Therefore, if this assumption held, it would be reasonable to assume that classifying individual FOI sensitivities would be a more effective approach for assisting digital sensitivity review. In this section, we evaluate the effectiveness of classifying the individual FOI exemptions *international relations* and *personal information*. Moreover, we postulate that, when learning to classify specific sensitivities, it may be beneficial to engineer *hand-crafted* features that are tailored to the sensitivities. For example, personal information and international relations sensitivities are often related to topical entities, such as people or countries. Therefore, we evaluate the effectiveness of extending text classification with additional hand-crafted features for classifying individual FOI sensitivities. We present the features that we extend our sensitivity classifiers with in Section 5.4.1 and our experimental methodology in Section 5.4.2, before discussing the results of our experiments in Section 5.4.3.

### 5.4.1 Hand-Crafted Sensitivity Features

A document's sensitivities are likely to be anchored by topical entities, such as people or countries. For example, personal information is intrinsically linked to a person and international relations sensitivities are linked to one or more countries. However, some entities are more likely to be an indicator that information is not sensitive, for example mentions of a media organisation can be an indicator of a press release. These links can be implicit within a document, which makes the task of identifying sensitivity-entity links very challenging. For these reasons, for our features, we chose to focus on identifying specific types of person, country and organisation entities within a document, and language features (i.e. terms) that are likely to be related to specific entities, e.g., verbs.

We define four groups of feature types that we evaluate in our experiments. We evaluate the effectiveness of each feature group, each of the features individually (nineteen in total) and all of the feature groups combined. Table 5.4 lists the features in each of the feature groups. We now discuss each of the feature groups and their features to provide an intuition of why they may be beneficial for sensitivity classification:

Table 5.4: The Feature groups that we evaluate for classifying individual FOI exemptions.

| Feature Group | Features |
|---|---|
| Entity | PersonName, Country, NamedEntity |
| Entity Role | Ambassador, Dictator, Diplomat, Media, Military, Monarch, Politician, PrimeMinister, Royals |
| Risk | countryRisk, FrmToRisk |
| Language | AllVerbs, SupplyVerbs, Negation, UNAcronyms |

- **Entity** features are simply a count of the number of occurrences of a type of named entity in a document. We define three entity features:
  - *PersonName*: The number of named persons in a document. If a document is about a single person it may be more likely to contain personal or confidential information about the person.
  - *Country*: The total number of countries mentioned in a document. Documents that mention many countries may be more likely to cover general topics relating to the countries and not contain sensitive information.
  - *NamedEntity*. NamedEntity includes occurrences of person, country and organisation entities. The number of named entities in a document could be a good indicator of whether the topic(s) of the document is/are wide-ranging or not. For example, a press release summary that covers a lot of topics, or a general policy related document that covers many aspects of a policy, are less likely to be sensitive.
- **Entity Role** features are counts of the occurrences of names of people or organisations that fulfil a specific role of employment. Sensitive information is more likely to be associated to some employment roles than others. For example, the mention of military personnel is more likely to be associated to sensitivity than a media organisation is. We define nine entity role features;
  - *Ambassador*: The names of ambassadors in a document may be an indication of the reporting of specific conversations or operations abroad that could contain sensitive details.
  - *Dictator*: Countries that are governed by a dictator are less likely to have good international relations with the UK. Therefore, information that discusses them could be more likely to be sensitive.
  - *Diplomat*: The names of diplomats in a document may be an indication of the reporting of specific conversations or operations abroad that could contain sensitive details.
  - *Media*: Mentioning media organisations, such as the BBC, in a document can be an indication that the information in the document is already in the public domain and, therefore, not sensitive.

  - *Military*: Discussions of military operations are likely to contain sensitive information, unless they are discussed within a press release.
  - *Monarch*: Any details of discussions about specific monarchs, that are not already in the public domain, are likely to be sensitive.
  - *Politician*: Documents that mention many politicians are likely to be about general political topics and not sensitive.
  - *PrimeMinister*: Documents that discuss a prime minister are likely to be official documents and not sensitive.
  - *Royals*: Any details of discussions about specific royals, that are not already in the public domain, are likely to be sensitive.

- **Risk** features model the risk associated to specific countries. Relations between countries are not all on par and, therefore, the accidental release of documents has varying potential for damaging the international relations between a country that produced the document and a referenced country or a third-party. For example, a country that a government has a strong and long lasting relationship with would be less likely to have an extreme reaction to a small indiscretion by the government than a country that the government has a fragile relationship with. Therefore, different countries have different levels of risk associated to them. To model this notion of country risk, we assign a risk score to each of the identified countries. The real nature of these relations is privileged information. Therefore, we model this fragility using *our* perception of current international relations. We were assisted in doing this by the guidance of a sensitivity review professional. We define two risk features:

  - *countryRisk* is defined as $countryRisk(r) = \sum_{c \in r} risk(c)$, where $c$ is a country occurring in document $d$ and *risk* is the risk score from the set {1 (None), 2 (Moderate), 3 (High)} associated with country $c$.
  - *FrmToRisk* is the sum of the countryRisk scores for the country that produced a document, $d$, and all of the countries that $d$ was sent to.

- **Language** features are a simple count of the occurrences of specific language features. We define four language features:

  - *AllVerbs* is the total number of verbs in a document. International relations sensitivities often arise from reports of particular actions of individuals.
  - *SupplyVerbs* is the total number verbs in document $d$ from a manually curated subset of verbs, presented in Figure 5.2, that indicate a verbal action of giving something. Some sensitivities, such as information supplied in confidence, are often related to someone saying or giving something to another individual.

acted advise advised agree agreed announce announced answer answered apologize apologized apologise apologised arrange arranged ask asked ascertain ascertained assist assisted assure assured clarify clarified communicate communicated complain complained confess confessed confront confronted consider considered consult consulted critique critiqued deliver delivered demonstrate demonstrated describe described detail detailed detect detected determine determined develop developed devise devised diagnose diagnosed evaluate evaluated explain explained give gave given identify identified illustrate illustrated influence influenced inform informed interview interviewed offer offered persuade persuaded plead pleaded question questioned receive received say said speak spoke summarize summarized summarise summarised telephone telephoned tell told verbalize verbalised

Figure 5.2: Supply Verbs. Verbs that are associated with an action of giving something.

- *Negation* is a count of the number of occurrences of negated words in a document. Recognising that something did not happen or was not said can be a good indication that the information is not sensitive.
- *UNAcronyms* is a count of acronyms that are used by the United Nations[6]. Documents that contain many instances of UN acronyms, such as NATO (North Atlantic Treaty Organisation), OPEC (Organisation of the Petroleum Exporting Countries) or PAHO (Pan American Health Organization), can be an indication of an official document that is less likely to be sensitive.

### 5.4.2 Experimental Methodology

In this section, we wish to answer the following research questions:

- RQ5.4: How effective is text classification for classifying the individual FOI sensitivities personal information and international relations?
- RQ5.5: Does extending text classification with additional *hand-crafted* features improve the effectiveness of sensitivity classification for specific sensitivities?

When evaluating the effectiveness of classifiers for the individual sensitivities, personal information and international relations, we deploy the document representation and feature reduction strategy that we identified as being most effective in Section 5.3[7], namely TF-IDF with stopwords retained and no stemming applied (denoted as TF-IDF$_{\text{stopNoSm}}$).

For entity identification, firstly, we use a dictionary of 43,286 named entities of interest (Politicians, Prime Ministers, Presidents, Royals, Monarchs and Dictators), constructed from the DBpedia[8] knowledge base. We also use a dictionary of 131,232 person names, constructed from the Drupal Name Database[9] and from the lists of unambiguous names supplied with *deid* (Nea-

[6] http://www.un.org/en/index.html

[7] We performed a separate evaluation to ensure that this is an appropriate choice for classifying the individual FOI exemptions. The results are in line with those reported in Section 5.3. To save space, we do not report these results.

[8] http://dbpedia.org

[9] https://drupal.org/project/namedb

matullah *et al.*, 2008), removing duplicates and non-Latin names (because they do not appear in the corpus), to extract generic instances of person entities from the records. We use LingPipe to match dictionary entries with mentions of the entities in a document $d$. When extending the document representations with additional features, to generate a document representation, we append the vector of additional features values to the text classification term features vector. The additional features are scaled in the range [0, 1].

We evaluate the approaches on the test collection that we previously presented in Chapter 3 (Table 3.3). To generate the ground truth for the experiments, when evaluating variants of the international relations classifier, any document that was judged to contain international relations sensitivities are labelled as positive examples and all other documents are labelled as being negative examples, this results in 346 positive and 3455 negative examples. Correspondingly, when evaluating variants of the personal information classifier, a document that was judged as containing personal information sensitivities is labelled as being positive and all other documents are labelled negative, resulting in 271 positive and 3530 negative examples. We evaluate the classifiers in this section using the same 5-fold Cross Validation folds that are used in Section 5.3. In Tables 5.5 and 5.6, in Section 5.4.3, classifiers that are statistically significantly different than, and achieve a higher BAC score than, their respective baseline approaches are denoted as †. We test for statistical significance using McNemar's non-parametric test (McNemar, 1947) with $p < 0.05$.

### 5.4.3 Results and Discussion

Table 5.5 presents the results of the international relations classifier using only term features, denoted as TF-IDF$_{\text{stopNoSm}}$, and the classifier extended with each of the feature groups, denoted +Entity, +Entity Role, +Risk and +Language respectively. Table 5.5 also presents the results for the classifier when it is extended with all nineteen features, denoted +All Features, and each of the features individually. Table 5.6 follows the same structure to present the results of the personal information classifier. We will, firstly, discuss the results of the international relations classifier, before, secondly, moving on to discuss the results of the international relations classifier.

From Table 5.5, firstly, we note that the baseline international relations classifier achieves 0.6837 BAC when it is not extended with additional hand-crafted features (TF-IDF$_{\text{stopNoSm}}$). This suggests that deploying classifiers for individual FOI exemptions may be an appropriate strategy to deploy in certain circumstances. Moving on to the results of extending the approach with hand-crafted features, we note from Table 5.5 that extending the classifier with the +Language feature group results in an increase in $F_2$ score from 0.4389 to 0.4407. Moreover, the increase in performance is statistically significant (denoted as †). This provides good evidence that specific language features can be useful for classifying international relations sensitivities. The addition of other feature groups performs less well. Adding the features groups +Entity Role or +Risk, results in a drop in the classifier's precision and recall scores (and, therefore,

$F_1$ and $F_2$ scores). This is somewhat surprising, since these feature groups were designed with international relations sensitivity in mind, and illustrates that engineering features to identify sensitivity is not an easy task. Moreover, it suggests that a classification approach that does not rely on hand-crafted feature engineering may be a more desirable solution. Extending the approach with all nineteen features results in a similar decrease in precision but a slight increase in recall. Overall, extending the approach with groups of features did not result in any notable increase in classification effectiveness (+Language is the only feature group that results in statistically significant improvements).

Surprisingly, we see the biggest increases in effectiveness when the approach is extended with individual features, with the features +Verbs, +Country, +DiplomaticRole, +NegationCount, +PersonName, +President and +Royals each resulting in significant increases in precision and recall scores. Most notably, the number of negated words in a document (+NegationCount) appears to be a good feature for classifying international relations sensitivities. The addition of this feature results in increased recall, $F_2$, BAC and auROC scores. This feature appears to have helped the classifier to identify sensitivities relating to claims of other countries' incompetence in which the conversation focuses on what they have *not* done.

Moving on to personal information classification, Table 5.6 presents the results of the personal information classifiers. From Table 5.6, firstly, we note that the personal information baseline classifier achieves 0.6480 BAC when it is not extended with additional hand-crafted features (TF-$IDF_{stopNoSm}$). It is, therefore, reasonable to expect that extending the approach could produce good results.

Firstly, reviewing the performance of the feature groups, we note that all feature groups and all of the features combined results in a notable drop in precision, from 0.25 for the baseline (TF-$IDF_{stopNoSm}$) to 0.15 for +Entity Role and 0.14 for all other groups. However, recall is notably increased for all feature groups and the BAC score is statistically significantly increased for each of the feature groups. In general, the feature groups result in the classifier being a lot more aggressive in predicting sensitivity and, therefore, although it makes more correct predictions it also makes more incorrect predictions.

Moving to when the classifier is extended with the features individually, for all of the features we observe the same trend as for the feature groups. Each of the features resulted in a decrease in precision score and an increase in recall, i.e., a more aggressive classifier that over-predicts sensitivity. These results lead to a statistically significant increase in BAC scores for all of the individual features, except for +MilitaryPerson.

In response to RQ5.4, we conclude that classifying individual FOI exemptions appears to be a reasonable approach to deploy for sensitivity classification. Therefore, we will evaluate the effectiveness of this approach, compared to and combined with, classifying sensitivity as a single category of information in the next section. In response to RQ5.5, we conclude that engineering hand-crafted features for sensitivity has the potential to improve the performance

Table 5.5: Hand crafted features s27. The table presents the precision, recall, $F_1$, $F_2$ and Balanced Accuracy (BAC) scores of the TF-IDF$_{\text{stopNoSm}}$ baseline and the baseline extended with hand crafted features. Statistical significance is denoted as † (McNemar's test, $p < 0.05$).

| | | Precision | Recall | $F_1$ | $F_2$ | BAC | auROC |
|---|---|---|---|---|---|---|---|
| TF-IDF$_{\text{stopNoSm}}$ | | 0.1700 | 0.7364 | 0.2751 | 0.4389 | 0.6837 | 0.7606 |
| +Entity | | 0.1706 | 0.7342 | 0.2758 | 0.4392 | 0.6846 | 0.7609 |
| +Entity Role | | 0.1698 | 0.7337 | 0.2746 | 0.4377 | 0.6837 | 0.7504 |
| +Risk | | 0.1695 | 0.7286 | 0.2739 | 0.4359 | 0.6818 | 0.7609 |
| +Language | † | 0.1729 | 0.7307 | 0.2783 | 0.4407 | 0.6856 | 0.7593 |
| +All Features | | 0.1669 | 0.7052 | 0.2689 | 0.4259 | 0.6733 | 0.7530 |
| +AllVerbs | † | 0.1731 | 0.7369 | **0.2791** | 0.4429 | 0.6880 | 0.7600 |
| +Ambassador | | 0.1697 | 0.7338 | 0.2745 | 0.4376 | 0.6822 | 0.7583 |
| +CountryRisk | | 0.1714 | 0.7480 | 0.2778 | 0.4442 | 0.6893 | 0.7593 |
| +Country | † | 0.1710 | 0.7402 | 0.2768 | 0.4415 | 0.6866 | 0.7604 |
| +Dictator | † | 0.1724 | 0.7272 | 0.2775 | 0.4392 | 0.6844 | 0.7609 |
| +DiplomaticRole | † | 0.1710 | 0.7396 | 0.2769 | 0.4416 | 0.6864 | 0.7598 |
| +FrmToRisk | † | 0.1727 | 0.7333 | 0.2785 | 0.4418 | 0.6870 | 0.7622 |
| +Media | † | 0.1708 | 0.7360 | 0.2762 | 0.4400 | 0.6847 | 0.7592 |
| +MilitaryPerson | | 0.1710 | 0.7419 | 0.2768 | 0.4419 | 0.6868 | 0.7585 |
| +Monarch | | 0.1699 | 0.7377 | 0.2750 | 0.4391 | 0.6843 | 0.7587 |
| +NamedEntity | † | 0.1718 | 0.7364 | 0.2774 | 0.4412 | 0.6857 | 0.7611 |
| +NegationCount | † | 0.1721 | **0.7484** | 0.2787 | **0.4452** | **0.6895** | **0.7643** |
| +PersonName | † | 0.1718 | 0.7428 | 0.2780 | 0.4432 | 0.6882 | 0.7606 |
| +Politician | † | 0.1725 | 0.7335 | 0.278 | 0.4409 | 0.6854 | 0.7587 |
| +President | † | **0.1732** | 0.7428 | 0.2797 | 0.4449 | **0.6895** | 0.7609 |
| +PrimeMinister | | 0.1698 | 0.7252 | 0.2740 | 0.4352 | 0.6801 | 0.7574 |
| +Royals | † | 0.1714 | 0.7392 | 0.2771 | 0.4416 | 0.6861 | 0.7608 |
| +SupplyVerbs | † | 0.1716 | 0.7307 | 0.2766 | 0.4391 | 0.6841 | 0.7613 |
| +UNAcronyms | | 0.1710 | 0.7446 | 0.2770 | 0.4427 | 0.6870 | 0.7603 |

of classifiers for specific sensitivities. However, feature design for sensitive information is not an easy task and it is not obvious which features are likely to benefit the classifier. In general, in our experiments, extending a classifier with multiple features, or groups of features, did not result in a more effective classifier. Therefore, we argue that engineering hand-crafted features for sensitivity classification is not the best approach for developing sensitivity classification and, in the remainder or this thesis, we propose to focus on classifying sensitivity by pure automatic approaches.

In response to these findings, in the following section, we use the specific sensitivity classifiers without additional hand-crafted features to evaluate the effectiveness of combining the outputs from multiple classifiers to provide a single set of sensitivity predictions.

Table 5.6: Hand crafted features s40. The table presents the precision, recall, $F_1$, $F_2$ and Balanced Accuracy (BAC) scores of the TF-IDF$_{\text{stopNoSm}}$ baseline and the baseline extended with hand crafted features. Statistical significance is denoted as † (McNemar's test, $p < 0.05$).

| | | Precision | Recall | $F_1$ | $F_2$ | BAC | auROC |
|---|---|---|---|---|---|---|---|
| TF-IDF$_{\text{stopNoSm}}$ | | **0.2505** | 0.3874 | **0.3027** | 0.3477 | 0.6480 | **0.7358** |
| +Entity Role | † | 0.1500 | 0.5607 | 0.2363 | 0.3613 | 0.6570 | 0.7166 |
| +Risk | † | 0.1448 | 0.5536 | 0.2291 | 0.3527 | 0.6503 | 0.7065 |
| +Entity | † | 0.1491 | 0.5496 | 0.2341 | 0.3564 | 0.6531 | 0.7134 |
| +Language | † | 0.1472 | 0.5567 | 0.2323 | 0.3565 | 0.6532 | 0.7103 |
| +All Features | | 0.1421 | 0.5358 | 0.2241 | 0.3436 | 0.6433 | 0.7043 |
| +AllVerbs | † | 0.1497 | 0.5529 | 0.2350 | 0.3579 | 0.6538 | 0.7153 |
| +Ambassador | † | 0.1486 | 0.5637 | 0.2347 | 0.3604 | 0.6564 | 0.7122 |
| +CountryRisk | † | 0.1476 | 0.5676 | 0.2338 | 0.3606 | 0.6566 | 0.7104 |
| +Country | † | 0.1451 | 0.5637 | 0.2303 | 0.3563 | 0.6530 | 0.7091 |
| +Dictator | † | 0.1470 | 0.5674 | 0.2331 | 0.3601 | 0.6562 | 0.7139 |
| +DiplomaticRole | † | 0.1513 | 0.5724 | 0.2387 | **0.3662** | 0.6612 | 0.7144 |
| +FrmToRisk | † | 0.1437 | 0.5498 | 0.2274 | 0.3503 | 0.6483 | 0.7086 |
| +Media | † | 0.1510 | 0.5719 | 0.2384 | 0.3660 | **0.6615** | 0.7249 |
| +MilitaryPerson | | 0.1430 | 0.5487 | 0.2264 | 0.3489 | 0.6464 | 0.7105 |
| +Monarch | † | 0.1505 | 0.5571 | 0.2364 | 0.3603 | 0.6558 | 0.7132 |
| +NamedEntity | † | 0.1505 | 0.5461 | 0.2355 | 0.3569 | 0.6535 | 0.7167 |
| +NegationCount | † | 0.1464 | 0.5602 | 0.2317 | 0.3568 | 0.6532 | 0.7107 |
| +PersonName | † | 0.1478 | **0.5681** | 0.2340 | 0.3606 | 0.6564 | 0.7144 |
| +Politician | † | 0.1478 | 0.5576 | 0.2333 | 0.3578 | 0.6547 | 0.7131 |
| +President | † | 0.1486 | 0.5673 | 0.2351 | 0.3618 | 0.6578 | 0.7127 |
| +PrimeMinister | † | 0.1457 | 0.5633 | 0.2311 | 0.3570 | 0.6533 | 0.7121 |
| +Royals | † | 0.1488 | 0.5645 | 0.2350 | 0.3609 | 0.6566 | 0.7142 |
| +SupplyVerbs | † | 0.1465 | 0.5570 | 0.2314 | 0.3555 | 0.6528 | 0.7158 |
| +UNAcronyms | † | 0.1450 | 0.5533 | 0.2293 | 0.3528 | 0.6502 | 0.7118 |

## 5.5 Ensemble Sensitivity Classification

In this section, we evaluate the effectiveness of classifying individual FOI exemptions, compared with classifying sensitivity as a single category of information. Providing a reviewer with classification predictions for multiple sensitivities in a single document increases the probability of making incorrect predictions. For example, predicting that a document contains personal information sensitivities, when the document actually contains international relations sensitivities only, would be an incorrect prediction. This would unnecessarily negatively affect the accuracy of our framework, since the document is in-fact sensitive. However, as we have previously seen in Section 5.4, our individual sensitivity classifiers can achieve comparable balanced accuracy (BAC) scores to our sensitivity classifier presented in Section 5.3. Therefore, it may be the case that learning separate classifiers for individual sensitivities actually results in more accurate predictions overall if we present all of the predictions, or combine the predictions from the individual classifiers to form a single *sensitive* vs. *not sensitive* prediction. In this section, we

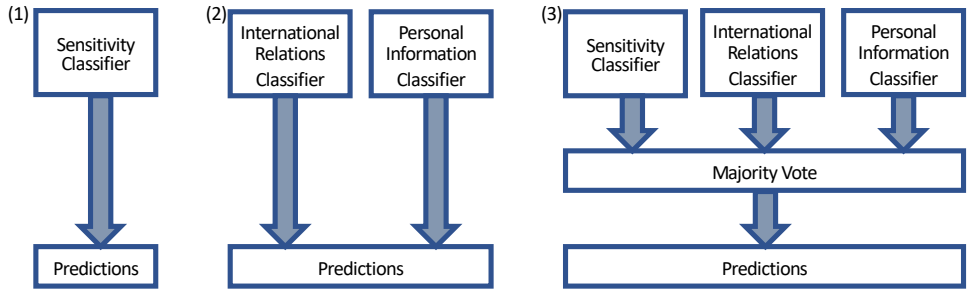



Figure 5.3: An illustration of the methods that we evaluate for combining sensitivity classifiers.

evaluate how predicting individual sensitivities impacts the overall effectiveness of sensitivity classification. Moreover, we evaluate a method of combining multiple sensitivity classifiers in an ensemble approach for sensitivity classification.

### 5.5.1 Combining Sensitivity Classifiers

We evaluate two strategies for presenting the results of the individual sensitivity classifiers and compare the results against learning to predict sensitivity as a single category. Figure 5.3 illustrates the baseline sensitivity classifier (1) and the two approaches that we evaluate for combining sensitivity classifiers, (2) and (3). To make a fair comparison, when comparing approaches, we treat all classification predictions as *sensitive or not sensitive* predictions, i.e. in approaches 2 and 3 any predictions from the international relations and personal information classifiers are treated as sensitive or not sensitive. In Figure 5.3, approach (1) is the sensitivity classifier that simply predicts each document as being either sensitive or not sensitive. We refer to this approach as Sensitive in Table 5.7. Approach (2) is the predictions from the individual sensitivity classifiers. As previously noted, when evaluating this approach to make a fair comparison with the baseline (i.e., approach (1)) the predictions from the classifiers are considered to be *sensitive or not sensitive*. We refer to this approach as $\text{Individual}_{27/40}$ in Table 5.7. Finally, approach (3) is an ensemble classification (Ditterrich, 1997) approach. Ensemble classification methods combine the decisions from a *committee* of individual classifiers with a view to improving the overall classification performance. The simplest, but often most effective, of these approaches combines the predictions from the committee classifiers by viewing each classifier's prediction as a vote for the class of a document (Kuncheva & Rodríguez, 2014). We deploy a majority vote strategy to integrate the sensitivity classifier with the individual international relations and personal information classifiers. In this approach, any document that receives a positive prediction from two of the classifiers is predicted as being *sensitive*, equivalently if a document receives a negative prediction from two of the classifiers the document is predicted as being *not sensitive*. Our hypothesis is that this approach will result in a higher precision classifier, since the rule for classifying a document is more strict, due to the majority vote approach. We refer to this approach as $\text{Majority Vote}_{\text{sense}/27/40}$ in Table 5.7.

### 5.5.2 Experimental Methodology

The research question that we wish to answer is as follows:

- RQ5.6: Which sensitivity classification approach results in the most effective sensitivity classifier, combining the predictions from multiple sensitivity classifiers or predicting sensitivity as a single category?

As our sensitivity classifier, we deploy the document representation and feature reduction strategy that we identified as being most effective in Section 5, namely TF-IDF with stopwords retained and no stemming applied (denoted as TF-IDF$_{\text{stopNoSm}}$ in Section 5). We denote this approach as Sensitive in Table 5.7. For the individual sensitivity classifiers, we use the classifiers presented in Section 5.4 without additional hand-crafted features, since the additional hand-crafted features did not result in consistent increases in precision and recall. As our test collection, we use the collection of 3801 documents that we presented in Chapter 3 (Table 3.3). We label all of the documents in the collection that were judged as containing international relations or personal information as *sensitive*, all other documents are labelled *not sensitive*. We, therefore, perform a binary classification *sensitive* vs *not sensitive*. Moreover, we retain the same 5-fold Cross Validation folds that are used in Sections 5.3 and 5.4.

### 5.5.3 Results and Discussion

Table 5.7 presents the results of the combined classifiers compared with the classifier that predicts sensitivity as a single classification category. Firstly, we note that each approach performs best for at least one of the presented metrics. Therefore, an argument could possibly be made for deploying each of the approaches in very specific situations. However, we are interested in which approach should be selected as a general strategy for assisting with the sensitivity review of government documents. Therefore, we select $F_2$ as our primary measure as it is a recall oriented weighted average of precision and recall.

Combining all of the classifiers by a majority vote strategy (Majority Vote$_{\text{sense/27/40}}$) results in 0.2705 precision (+ 6.8% compared to the Sensitive classifier). This is a statistically significant increase that is well aligned with our expectation that the strict majority vote would result in a more precise classifier. However, the approach only manages to correctly identify 59% of sensitive documents (0.5916 Recall). This result shows that this strategy is not the most appropriate to deploy for the (recall oriented) sensitivity classification task.

Presenting all of the sensitive predictions from the classifiers that identify individual sensitivities (Individual$_{27/40}$) correctly identifies the most sensitive documents, i.e., 74.9% of the sensitive documents in our collection (0.7490 Recall). This is a statistically significant result and this approach may be the most appropriate to deploy in more *risk averse* government departments. However, the approach's precision score is notably less than the other two approaches.

Table 5.7: Ensemble classification results. The table shows the precision, recall, $F_1$, $F_2$, Balanced Accuracy (BAC) and auROC scores. Statistical significance is denoted as † (McNemar's test, $p < 0.05$)

| | | Precision | Recall | $F_1$ | $F_2$ | BAC | auROC |
|---|---|---|---|---|---|---|---|
| Sensitive | | 0.2546 | 0.6831 | 0.3701 | **0.5098** | **0.6882** | **0.7518** |
| $\text{Individual}_{27/40}$ | † | 0.2118 | **0.7490** | 0.3303 | 0.4970 | 0.6625 | 0.7340 |
| $\text{Majority Vote}_{\text{sense/27/40}}$ | † | **0.2705** | 0.5916 | **0.3712** | 0.4781 | 0.6744 | 0.7377 |

This could have a negative affect on a reviewer's perception of how good the system is. Overall, we conclude that the best performing approach, in terms of a general strategy for different government departments, is to learn to identify sensitivity as a single category of information (denoted as Sensitive). Importantly, this approach achieved the highest BAC score and the highest $F_2$ score. This demonstrates that the approach provides an effective balance between (1) correctly identifying sensitive *and* not-sensitive documents, which is important for government departments' and reviewers' perception of how effective the classifier is, and (2) the fact that sensitivity classification is a recall-oriented task, i.e., there is a greater penalty from not identifying a sensitive document than there is from falsely identifying a document as being sensitive (This is intrinsically measured by the $F_2$ metric.)

In response to RQ5.6, we conclude that, on our collection, combining the predictions from multiple sensitivity classifiers is a viable approach that may be more appropriate in certain scenarios, for example in risk-averse government departments. However, learning to identify sensitivity as a single category of information results in the most effective classifier overall and, therefore, we propose to use this strategy as the basis for developing our framework. We will use this strategy in the remainder of this thesis.

## 5.6 Conclusions

In this chapter, we proposed to address the problem of sensitivity classification as a text classification task. Moreover, we presented our baseline sensitivity classification approach that we build on in the remainder of this thesis a basis for developing our proposed framework. In particular, firstly in Section 5.2, we evaluated the effectiveness of a document sanitisation approach (Sánchez *et al.*, 2012) for identifying the international relations sensitivity *information that has been supplied in confidence*. We empirically showed that document sanitisation, which has previously been shown to be effective at identifying sensitive information in other domains, is not a suitable approach for identifying FOI sensitivities. Secondly, in Section 5.3 we proposed to address sensitivity classification as a text classification task. Moreover, we empirically evaluated document representation and feature reduction strategies for sensitivity classification. We showed that, on our collection, a TF-IDF document representation with retained stopwords and

no stemming applied (denoted as TF-IDF$_{\text{stopNoSm}}$ in Table 5.3) resulted in the most effective sensitivity classifier. Next, in Section 5.4, we evaluated the effectiveness of classifying individual FOIA sensitivities and, moreover, the effectiveness of engineering additional hand-crafted features for specific sensitivities. We showed that, in our experiments, extending a classifier with additional hand-crafted did not, in general, result in a more effective classifier (see Tables 5.5 and 5.6). Therefore, we argue that engineering hand-crafted features of sensitivity is not the best approach for sensitivity classification. Finally, in Section 5.5, we evaluated the effectiveness of predicting sensitive information at different levels of granularity, and combining sensitivity predictions from multiple classifiers. We showed that learning to identify sensitivity as a single category of information results in the most effective overall classifier (see Table 5.7). Moreover, we propose to use this strategy as the basis for developing the sensitivity classification component of our framework. In the following chapter, we build on the findings from this chapter to propose more advanced methods of automatically engineering classification features that are specifically designed to improve the effectiveness of sensitivity classification.

# Chapter 6

# Enhanced Sensitivity Classification

## 6.1 Introduction

In the previous chapter, we proposed to identify documents that contain FOIA sensitive information as a text classification task (see Section 5.3). Moreover, we empirically evaluated document representation and feature reduction techniques for sensitivity classification (see Table 5.3). Furthermore, we evaluated the effectiveness of learning to identify sensitive information as a single category of information, compared with separately learning to classify specific FOI exemptions, i.e., international relations and personal information, and combining the predictions from multiple sensitivity classifiers in an ensemble approach to sensitivity classification (see Section 5.5). We showed that learning to identify sensitive information as a single category of information can be more effective for classifying sensitive information than combining multiple classifiers that each identify a single FOIA exemption (see Table 5.7). Moreover, we proposed to deploy this sensitivity classification approach as the basis for our framework for technology-assisted sensitivity review.

As we have previously discussed, in Chapters 3 and 5, sensitivity is not necessarily topic oriented and is often context-dependent. For example, in our test collection of Chapter 3, seventeen written communications to central government from the Nigerian embassy prior to 2010 are about the then Nigerian president Umaru Musa Yar'Adua. The documents cover the president's official duties and announcements and eleven of the documents are not sensitive. Six of the seventeen documents about President Umaru Musa Yar'Adua are sensitive. However, the sensitivity is related to President Umaru Musa Yar'Adua in only one of these sensitive documents. The sensitivities in the other five documents about President Umaru Musa Yar'Adua are a result of reporting information that has been supplied from another individuals that include disparaging remarks about the activities of the government. These sensitivities are not topic-oriented, since the main topics of these documents are the duties and announcements of the president.

We argue that for sensitivity classification to be able to more effectively identify these non-topic oriented, and subtle, sensitivities we need more advanced feature generation techniques

that can identify latent features of sensitive information. In this chapter, we propose to extend sensitivity classification with *automatically* generated document features that can identify latent structures or patterns, in the vocabulary, syntax (grammar) and the semantics of documents that can be reliable indicators of sensitive and non-sensitive text.

As we have previously discussed, in Chapter3, the expectation that governments can be held to account is an essential element of transparent government (Moss & Gollins, 2017). In our discussions with sensitivity review experts, there has been a consensus in the opinion that this accountability and transparency must also apply to any automatic approaches for (assisting with) decision making. Moreover, there has been a consensus in the opinions expressed to us that for governments to develop trust in automatic sensitivity classification, it is important that any method for automatically classifying sensitive information has a reasonable level of transparency, since governments will ultimately be held accountable for any decisions that are made. With this in mind, our proposed extensions to sensitivity classification rely solely on the distributional statistics of words within a collection. Therefore, it is possible to trace back from any automatic classification decision and identify the terms within a document, or the collection of documents, that led to the classifier's prediction of (non-)sensitivity. The remainder of this chapter is structured as follows:

- Section 6.2 presents an introduction to the vocabulary features that we evaluate in this chapter. As vocabulary features, we evaluate the effectiveness of *large* term $n$-gram features, where $n \leq 10$, for reliably identifying sensitive or non-sensitive text.

- Section 6.3 presents the syntactic features that we evaluate. As syntactic features, we evaluate the effectiveness of sequences of Part-of-Speech (POS) tags for capturing latent syntactic, i.e. grammatical, patterns that can be a reliable indicator of sensitive, or not sensitive, information. Representing documents as sequences of POS tags has the potential to capture grammatical patterns that are associated with particular sensitivities, such as information supplied in confidence. Moreover, representing documents as POS sequences results in the possibility of developing sensitivity classification approaches that are based on sequence classification (Xing *et al.*, 2010) techniques, as opposed to the text classification techniques that we have discussed thus far. Therefore, in this section, we evaluate the effectiveness of sequence classification techniques for classifying sensitivity using POS sequences. In particular, we evaluate kernel functions for POS sequence classification in Section6.3.1. Moreover, we evaluate ensemble approaches for combining POS sequence classification with text classification in Section 6.3.2.

- In Section 6.4, we present the approach that we deploy to identify the latent semantic relations in documents that can be a reliable indicator of sensitive or non-sensitive text. As semantic features, we generate a document representation from word embeddings (Balikas

& Amini, 2016; Mikolov, Chen, Corrado & Dean, 2013), to capture semantic relations between the documents in a collection.

- In Section 6.5, we empirically evaluate the effectiveness of language, syntactic and semantic features for extending our baseline text classification approach for sensitivity classification from Chapter 5.
- Section 6.6 presents an analysis of how the feature for extending our sensitivity classification baseline improve the effectiveness of sensitivity classification. In particular we discuss the vocabulary features in Section 6.6.1, the semantic features in Section 6.6.2 and the impact of these additional features for sensitivity review in Section 6.6.3.
- In Section 6.7, we summarize our conclusions from this chapter.

## 6.2 Vocabulary Features

The first automatically generated features that we evaluate for sensitivity classification are vocabulary features. The sensitivity classification approaches that we presented in Chapter 5 were based on the bag of words (BOW) model, where the classification features are the individual terms in the documents and the terms are represented by a distribution statistic such as TF-IDF. One possible drawback of this model is that it does not capture the proximity of terms in a document. For example, it may be the case that a document in which the terms *please* and *protect* appear adjacent to each other, and in that order, is more likely to be sensitive than another document in which the terms are in separate paragraphs, and not necessarily with the same ordering.

One method for capturing the proximity and ordering of terms within the classification model is to include term $n$-gram features (Sebastiani, 2002) in the document representation[1]. A term $n$-gram is a totally ordered set of $n$ contiguous terms in a document. To integrate $n$-gram features to the classification model, each set of $n$ terms in a document are extracted as a single token before calculating a distribution statistic, such as TF-IDF.

In practice, term $n$-grams are not very commonly used in text classification tasks (Wang & Manning, 2012). There are two main reasons for this. Firstly, for topic-based classification tasks, $n$-gram features have been shown to have limited utility. This is likely to be due to the fact that most topics have certain topic keywords that are indicative by themselves (Wang & Manning, 2012). Secondly, extending text classification with additional $n$-gram features can result in very large document representations, often hundreds of thousands of features, that result in unmanageable computational complexity (Joachims, 1998).

Term $n$-gram features have, however, been shown to be beneficial for certain document classification tasks where the context the terms appear in is important, such as sentiment analy-

[1] In text classification, character $n$-grams can be used as features (Cavnar & Trenkle, 1994). However, as textual features for sensitivity classification, we focus only on term $n$-grams.

sis (Wang & Manning, 2012). A term's context, or the proximity and order that terms appear in, is an important factor in how likely the term is to be part of a passage of sensitive information. For example, the terms *do not distribute* are more likely to indicate sensitivity if they appear next to each other and in the order shown. Therefore, we hypothesis that term n-grams will be a strong feature for sensitivity classification. When extending text classification with term $n$-grams in some other tasks, setting $n \leq 4$ has been shown to result in better effectiveness, for example when classifying newswire articles (Fürnkranz, 1998). However, for sensitivity classification, we expect larger values of $n$ to be more effective, since they have the potential to capture document structures that, in turn, can be an indicator of potential sensitivity. For example, table headings, such as *Name, Date of Birth, Residence*, can be a reliable indicator of personal information sensitivities. Therefore, we propose to evaluate the effectiveness of larger term $n$-gram sequences (we refer to these as extended term $n$-grams), along with additional combinations of smaller values of $n$ for identifying sequences of text that are indicative of sensitive information.

## 6.3 Syntactic Features

The second type of features that we propose to enhance sensitivity classification with identify latent grammatical, or syntactic, patterns that can be a reliable indication of sensitive information. Natural language processing (NLP) techniques are a widely used, and effective, way to have a computer try to understand the underlying structures and meaning of written text (Johnson, 2009). One of the main techniques of NLP is Parts-of-speech (POS) tagging. POS, e.g., nouns, pronouns, verbs, adverbs, adjectives, conjunctions, prepositions, and interjections, are categories of words that have similar grammatical properties and, typically, follow similar syntactic roles within the structure of sentences. POS tagging is the process of parsing a document to identify the correct POS for each word in the document.

Similarly to the terms that POS tags are derived from, the distributions of POS tags in a collection can potentially be used to identify latent grammatical or syntactic patterns within a collection of documents. Lioma & Ounis (2006) showed that the distributions of POS $n$-grams in a corpus can indicate the amount of information that they contain. More specifically, Lioma and Ounis showed that high frequency POS $n$-grams are typically *content rich* and removing *content poor* POS $n$-grams from search engine queries can improve the overall retrieval performance.

In this thesis, we postulate that documents that contain the same types of sensitive information will have a similar distribution of POS $n$-grams. For example, the text sequences "an informer gave him", "the ambassador said she" and "a detainee showed us" are good indicators of the international relations sensitivity *information that has been supplied in confidence*. The example text sequences discuss different entities and actions. However, each of the sequences results in the POS tags "DT NN VB PR" and representing these sequences as POS 2-grams re-

Table 6.1: Overview of the kernel functions that we evaluate for classifying sensitive information using POS sequences. The table shows the *type* of kernel, i.e. either *Vector Space* or *String*, and the definition of the kernel function.

| | Type | Definition |
|---|---|---|
| Linear | Vector Space | $K_{linear}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$ |
| Gaussian | Vector Space | $K_{gaussian}(\mathbf{x}_i, \mathbf{x}_j) = exp\left(\frac{-\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2}{2\sigma^2}\right)$ |
| Spectrum | String | $K_{spectrum}(\mathbf{x}, y) = \langle \Phi_k(\mathbf{x}), \Phi_k(y) \rangle$ |
| Mismatch | String | $K_{(k,m)}(x, y) = \left\langle \Phi_{(k,m)}(\mathbf{x}), \Phi_{(k,m)}(y) \right\rangle$ |

sults in the POS sequence "DTNN NNVB VBPR". Therefore, the POS sequence could capture a grammatical structure that might be indicative of sensitivity. We postulate that the distributions of such POS sequences can be a reliable feature of sensitivity.

Representing documents by an abstraction, such as the POS tags they contain, has an additional attractive by-product. In effect, a document's tokens (POS n-grams) can be viewed as a sequence of symbols from an alphabet, rather than terms from a vocabulary and, hence, gives rise to the possibility of developing techniques based on sequence classification (Xing *et al.*, 2010). Sequence classification has been shown to be effective in fields such as Bioinformatics (e.g., classifying protein sequences (Deshpande & Karypis, 2002; Leslie *et al.*, 2002)) and Cyber-Security (e.g., intrusion detection (Lane & Brodley, 1999)), in addition to Information Retrieval (IR) tasks (e.g., generating query suggestions from concept sequences in query logs (Cao *et al.*, 2008)). The SVM classification approaches that we have deployed thus far have used a linear kernel, since this combination is particularly suited to text classification tasks (Joachims, 1998). However, an intrinsic component of sequence classification is selecting a classification *kernel function* that is suitable for the classification task being attempted, for example, sequence-similarity kernels such as the Spectrum kernel (Leslie *et al.*, 2002).

In the remainder of this section, we evaluate SVM kernel functions for POS sequence sensitivity classification, to identify an appropriate strategy for combining POS sequence classification with text classification. Moreover, this section provides insights that we use for comparing the effectiveness of syntactic features with textual and semantic features in Section 6.5. In particular, in Section 6.3.1, we evaluate the effectiveness of four SVM kernel functions for *stand-alone* sensitivity classifiers using POS sequences, i.e., the stand-alone sequence classifiers do not use any term features. Having evaluated POS sequence classification as a stand-alone technique, in Section 6.3.2, we select the SVM kernel functions that perform well for stand-alone POS sequence classification to evaluate ensemble classification approaches that combine POS sequence classification with text classification for classifying sensitive information.

### 6.3.1 Kernel Functions for Sensitivity Classification with POS Sequences

The first kernel function that we evaluate for POS sequence classification is the linear kernel that we have deployed in our SVM classifiers thus far. As previously stated, the linear kernel, defined as $K_{linear}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$, is the simplest kernel function. However, $K_{linear}$ has desirable properties in that it is very fast to train and does not tend to *over-fit* the learned model to the set of training instance vectors, $D_{tr}$, when $|\mathbf{x}|$ is very large (Joachims, 1998). The second kernel that we evaluate is more suitable for non-linearly separable data. The Gaussian kernel is defined as $K_{gaussian}(\mathbf{x}_i, \mathbf{x}_j) = exp\left(\frac{-||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}\right)$, where $\sigma$ is a parameter that determines the *width* of the Gaussian function, i.e., the region of influence for an instance in vector space. A properly tuned Gaussian kernel will always be able to learn the optimal decision of a linear kernel (Keerthi & Lin, 2003), yet tuning $\sigma$ can be expensive and does not guarantee obtaining a better model.

As previously stated in Chapter 2, the SVM classifier solves the following optimisation problem:

$$Maximise \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle \tag{6.1}$$

By substituting $\left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle$ with a kernel function, we effectively create a *feature* map, $\phi$, which maps an instance, $\mathbf{x}$, to a new (possibly higher dimensional) space. For the linear and Gaussian kernels, $\phi$ is implicit within the dot products defined in the functions. Often, however, kernels explicitly define this mapping as the input to the kernel function. String kernels operate on finite sub-sequences of strings and the third kernel function that we evaluate, the Spectrum kernel (Leslie *et al.*, 2002), is a simple string kernel defined by its map $\phi$ over all sub-sequences in an alphabet $A$. For a given alphabet, $A$, a document's feature map, $\Phi_k(\mathbf{x}) = (\phi_a(\mathbf{x}))_{a \in A^k}$, is the frequency weighted set of all contiguous subsequences of length $k \geq 1$, that the document contains, i.e., its $k$-spectrum, and where $\phi_a(\mathbf{x})$ is the frequency of $a$ in $\mathbf{x}$. The Spectrum kernel is then defined as $K_{spectrum}(\mathbf{x}, y) = \langle \Phi_k(\mathbf{x}), \Phi_k(y) \rangle$.

One limitation of the Spectrum kernel is that it is constrained to exact matches when calculating the similarity of instances. The fourth, and final, kernel that we evaluate, the Mismatch kernel (Eskin *et al.*, 2002), addresses this by allowing for a pre-defined number of mismatched symbols within sequences. For a given sequence $\alpha = a_1 .. \ a_k, a \in A$, $N(k,m)(\alpha)$ is the set of all $k$-length sequences, $\beta = b_1 .. \ b_k, b \in A$ that differ from $\alpha$ by $\leq m$ mismatches. The Mismatch kernel's feature map is then defined as $\Phi_{(k,m)}(\alpha) = (\phi_\beta(\alpha))_{\beta \in A^k}$, where $\phi_\beta(\alpha) = 1$ if $\beta \in N(k,m)(\alpha)$, else $\phi_\beta(\alpha) = 0$. From this feature map, the (k,m)-mismatch kernel is defined as $K_{(k,m)}(x,y) = \left\langle \Phi_{(k,m)}(\mathbf{x}), \Phi_{(k,m)}(y) \right\rangle$.

For complex sequence classification tasks, a single SVM kernel may not provide an optimal solution. One method of addressing this is to combine multiple simpler kernels as a hybrid kernel, with the aim of considering multiple aspects of an instance vector. We would expect that sequence-based kernels, such as string kernels, will identify different features of sensitivity than

Table 6.2: The total unique POS *n*-gram tokens in each collection representation.

| | 1-gram | 2-gram | 3-gram | 4-gram | 5-gram | 6-gram | 7-gram | 8-gram | 9-gram | 10-gram |
|---|---|---|---|---|---|---|---|---|---|---|
| Unique Tokens | 15 | 209 | 1877 | 11408 | 51238 | 172109 | 441251 | 888837 | 1465215 | 2052063 |

vector space kernels, i.e., the linear or Gaussian kernels. Therefore, we also evaluate two hybrid kernels that are a linear combination of the scores from the best performing string kernel and each of the linear and Gaussian kernels.

#### 6.3.1.1 Experimental Methodology

In this section, the research question that we wish to answer is:

- RQ6.1: Which SVM kernel functions are effective for sensitivity classification using POS Sequences?

To answer this question, we use the collection of 3801 documents that we presented in Chapter 3 (Table 3.3) and retain the same 5-fold Cross Validation folds from Chapter 5. We perform a binary classification, *sensitive* vs. *not sensitive*.

To generate POS sequence representations of the documents in our collection, following Lioma & Ounis (2006), we use the TreeTagger[2] part-of-speech tagger to POS tag the documents using a reduced set of 15 POS tags. We create separate *n*-gram sequence representations of the collection, resulting in individual *n*-gram sequence collections for $n = \{1...10\}$. Table 6.2 presents the number of observed unique tokens in the alphabet, $A$, for each size of $n$.

We use scikit-learn and extend LibSVM[3] with the Spectrum and Mismatch kernels. For the linear and Gaussian kernels, we represent documents as token frequency vectors, where a token is a POS *n*-gram. For the Spectrum and Mismatch kernels, we count the frequency of $k$ length sub-sequence matches in a pair of documents. Parameter values are selected using a 10-fold Cross Validation on the training data, for each of the 5-fold Cross Validation folds. We vary SVM's $C$ parameter exponentially in the range [0.001,10000], and similarly for the $\gamma$ parameter in the range [0.0001,10]. Sub-sequences are varied for $k = \{3,6,9,12\}$. We optimise parameters to maximise the area under the Receiver Operating Characteristic curve (auROC) since, when documents are ranked by the classifier's decision function, it maximises the probability that a randomly selected positive instance is ranked above a randomly selected negative instance.

#### 6.3.1.2 Results and Discussion

Table 6.3 presents the results for each of the SVM kernels for the POS sequence classifiers. The table shows the best performing size of *n*-gram for each of the individual kernels, according to auROC, and for two hybrid kernels, namely *Spectrum+Linear* and *Spectrum+Gaussian*,

[2]http://www.cis.uni-muenchen.de/∼schmid/tools/TreeTagger/
[3]https://csie.ntu.edu.tw/∼cjlin/libsvm/

Table 6.3: SVM Kernels for POS sequence classification. The table shows the best performing size of *n*-gram. The highest values for each metric are in bold. We denote kernels that perform statistically significantly better than random by † and statistically significantly better than the next best performing kernel (according to BAC) by Δ. We test statistical significance using McNemar's non-parametric test, $p < 0.01$.

| | | | *n* | Precision | Recall | $F_1$ | $F_2$ | BAC | auROC |
|---|---|---|---|---|---|---|---|---|---|
| *Individual* | | | | | | | | | |
| Linear | † | Δ | 5 | 0.2185 | 0.6155 | 0.3225 | 0.4514 | 0.6403 | **0.6897** |
| Gaussian | † | Δ | 4 | 0.2070 | 0.6494 | 0.3139 | **0.4550** | 0.6354 | 0.6820 |
| Spectrum | † | Δ | 1 | 0.1868 | **0.6574** | 0.2909 | 0.4370 | 0.6109 | 0.6636 |
| Mismatch | † | | 1 | 0.1847 | 0.4833 | 0.2673 | 0.3387 | 0.5420 | 0.5415 |
| *Hybrid* | | | | | | | | | |
| Spectrum+Linear | † | Δ | 4 | **0.2266** | 0.6178 | **0.3278** | 0.4384 | **0.6417** | 0.6779 |
| Spectrum+Gaussian | † | Δ | 2 | 0.2245 | 0.5995 | 0.3251 | 0.4361 | 0.6388 | 0.6764 |

that are a linear combination of the Spectrum string kernel and the linear and Gaussian kernels respectively.

Firstly, we note from Table 6.3, that all of the kernels perform significantly better than random, denoted as †. Moreover, each of the kernel functions performs significantly better than the next best performing approach in terms of BAC (denoted as Δ). The linear kernel achieves the best auROC score (0.6897). However, the Gaussian and Spectrum kernels perform competitively with the linear kernel, achieving 0.6820 and 0.6636 auROC respectively. Moreover, the Gaussian and Spectrum kernels perform well for recall-oriented metrics, the highest $F_2$ (0.4550) and TPR (0.6574) scores are achieved by the Gaussian and Spectrum kernels respectively. This is important for sensitivity classification, since the cost of mis-classifying a sensitive document is greater than that of mis-classifying a not-sensitive document, i.e mis-classifying a sensitive document can lead to the accidental release of sensitive information. The Mismatch kernel performs less well, achieving the lowest scores for each of the reported metrics. Therefore, in the remainder of this section, we focus on the Spectrum, Gaussian and linear kernels.

Turning our attention to the hybrid kernels, when evaluating the effectiveness of kernels, we are interested in notable differences in the correctness of predictions for sensitive documents. As shown in Table 6.4, there is substantial Fleiss' $\kappa$ agreement between the linear and Gaussian kernels (0.7301), but only moderate agreement between the Spectrum and linear or Spectrum and Gaussian kernels (0.4502 and 0.4122 respectively). This is in line with our expectation that sequence-based kernels, such as String kernels, can identify different features of sensitivity than

Table 6.4: Fleiss' $\kappa$ agreement between the linear, Gaussian and Spectrum kernels for predictions on sensitive documents, i.e., True Positive or False Negative predictions.

| | Lin-Gau-Spec | Lin-Gau | Lin-Spec | Gau-Spec |
|---|---|---|---|---|
| Fleiss' $\kappa$ | 0.5312 | 0.7301 | 0.4502 | 0.4122 |

vector space kernels, such as linear or Gaussian. Therefore, we select the Spectrum kernel as our base kernel for the hybrid kernels. As can be seen from Table 6.3, the hybrid kernels achieve 0.67 auROC. This is slightly less than the 0.68 auROC achieved by the linear and Gaussian kernels individually. However, in terms of balanced accuracy, the hybrid kernels improve overall performance (0.6417 Spectrum+Linear vs. 0.6109 Spectrum and 0.6403 Linear, 0.6388 Spectrum+Gaussian vs. 0.6109 Spectrum and 0.6354 Gaussian).

In response to RQ 6.1, we conclude that, for individual kernels, the Linear, Gaussian and Spectrum each perform best in terms of auROC, $F_2$ and Recall respectively. However, the Spectrum+Linear hybrid kernel performs best in terms of BAC and $F_2$. Therefore, in the following section, we select the Linear, Gaussian, Spectrum and Spectrum+Linear kernels for evaluating ensemble approaches for combining POS sequence classification with our sensitivity classification baseline from Chapter 5.

### 6.3.2 Combining Text Classification and POS Sequence Classification

In the previous section, we identified that the Linear, Gaussian, Spectrum and Spectrum+Linear kernels can be effective for sensitivity classification using POS sequences. In this section, we evaluate combining each of the kernels with the baseline sensitivity classification approach that we presented in Chapter 5. The baseline approach deploys a linear kernel for text classification, since the linear kernel is the most appropriate solution for text classification. Therefore, to combine the baseline approach with either of the Gaussian, Spectrum and Spectrum+Linear kernels requires that we evaluate ensemble classification approaches that can combine two separate classifiers. We, firstly, introduce the ensemble approaches that we evaluate, before presenting our experimental methodology in Section 6.3.2.1, and discussing our findings in Section 6.3.2.2.

Ensemble classification (Ditterrich, 1997) methods combine the decisions from a *committee* of individual classifiers with a view to improving the overall classification performance. In Chapter 5, we deployed a simple majority vote ensemble when combining individual sensitivity classifiers, where each classifier's prediction is a vote for the predicted class of a document. A simple extension to this approach is weighted majority vote (Kuncheva & Rodríguez, 2014) (WMV), where one or more of the committee classifiers is believed to be a more authoritative source and is, therefore, assigned more importance, or weight, in the model. Majority vote ensembles are the simplest ensemble approach. However, WMV has been shown to perform well for problems with a small number of unbalanced classes (Kuncheva & Rodríguez, 2014), as is the case for sensitivity classification with sensitive being the minority class. Another ensemble classification approach that is widely used in the literature is the stacking (Wolpert, 1992) approach. In a stacking ensemble, a separate combiner function, or *meta-learner*, is trained from the predictions of the committee classifiers. The final classification predictions are the predictions from the meta-learner.

To evaluate ensemble classification approaches, we combine the predictions of the text clas-

sification $p_t$ with the predictions of each of $n$ sequence classifiers $p_{si}$. We evaluate four ensemble approaches. Firstly, in Weighted Majority Vote (WMV), to predict a document's class, the prediction from the text classification model, $p_t$ is assigned a weight $w$ as the authoritative classifier, and the document's overall prediction score is calculated as:

$$\frac{(p_t \cdot w) + \sum_{i=1}^{n} p_{si}}{n+1} \tag{6.2}$$

resulting in $n+1$ votes for each document's class prediction.

The remaining three combination methods are *stacking* approaches. Stacking requires an intermediate step where the text classification predictions, $p_t$, and the predictions from each of the sequence classifiers, $p_{si}$, are used to train a separate meta-learner classifier. To train the meta-learner, the predictions for each single document, $p_{ti}$ and $p_{si}$ are concatenated to form a single document feature vector, resulting in $n+1$ document features, $f, f \in \{p_t, p_s\}$. The resulting document vectors are used to train the combiner classifier. We test three classifiers as combiners, namely Logistic Regression (LR), SVM and Random Forests (RF).

#### 6.3.2.1 Experimental Methodology

The research question that we wish to answer is:

- RQ6.2: What is the most effective SVM kernel function and ensemble classifier combination for combining text classification and POS sequence classification when classifying sensitive documents?

We retain the same test collection, *sensitive* vs. *not sensitive* classification and 5-fold Cross Validation set-up as is used in Section 6.3.1. We retain the kernel parameters from Section 6.3.1. For WMV, the predictions on the test set of the test collection, from each of the $n$ POS sequence classifiers ($n = 10$) and the text classification classifier, are considered as votes for the document's class. As weights for the text classification vote, we test for $w = \{1..100\}$.

For the stacking approaches, we use predictions on the validation set of the test collection ($n+1$ per document, where $n = 10$) to construct document representations to train the meta-learner classifiers. The reported results are the predictions of the meta-learners on the test set of the test collection. We vary SVM's $C$ parameter exponentially in the range [0.001,10000], and similarly for the $\gamma$ parameter in the range [0.0001,10]. For LR, we select L1 as our loss function and vary $C$ in the same range as for SVM. For RF, we test with number of trees $t = \{100, 250, 500, 750, 1000\}$. We optimise for area under the Receiver Operating Characteristic curve (auROC).

We report Precision, Recall, $F_1$, $F_2$, Balanced Accuracy (BAC) and auROC metrics. We report statistical significance using McNemar's non-parametric test, with $p < 0.05$. Significant improvements compared to the text classification baseline are denoted by † in Table 6.5.

Table 6.5: Results for POS sequence and Text Classification ensembles. The table shows the precision, recall, $F_1$, $F_2$, Balanced Accuracy (BAC) and auROC scores. Statistical significance compared to the Text Classification (TC) baseline is denoted as † (McNemar's test, $p < 0.05$).

| | # Votes | | Precision | Recall | $F_1$ | $F_2$ | BAC | auROC |
|---|---|---|---|---|---|---|---|---|
| Text Classification (TC) | | | 0.2546 | 0.6831 | 0.3701 | 0.5098 | 0.6882 | 0.7518 |
| Weighted Majority Vote (WMV) | | | | | | | | |
| $TC+POS_{Linear}$ | 11 | † | 0.2610 | 0.6853 | 0.3780 | **0.5171** | 0.6950 | **0.7659** |
| $TC+POS_{Gaussian}$ | 11 | † | 0.2631 | 0.6813 | **0.3796** | 0.5169 | **0.6954** | 0.7633 |
| $TC+POS_{Spectrum}$ | 11 | | 0.2412 | **0.6932** | 0.3578 | 0.5042 | 0.6807 | 0.7588 |
| $TC+POS_{LinGausSpec}$ | 31 | | 0.2578 | 0.6554 | 0.3701 | 0.5009 | 0.6842 | 0.7616 |
| $TC+POS_{Spectrum+Linear}$ | 11 | | 0.2211 | 0.6295 | 0.3273 | 0.4597 | 0.6461 | 0.7033 |
| Logistic Regression (LR) | | | | | | | | |
| $TC+POS_{Linear}$ | 11 | | 0.2505 | 0.6752 | 0.3646 | 0.5028 | 0.6837 | 0.7584 |
| $TC+POS_{Gaussian}$ | 11 | | 0.2437 | 0.6513 | 0.3537 | 0.4865 | 0.6718 | 0.7492 |
| $TC+POS_{Spectrum}$ | 11 | | 0.2364 | 0.6495 | 0.3462 | 0.4805 | 0.6650 | 0.7502 |
| $TC+POS_{LinGausSpec}$ | 31 | | 0.2447 | 0.6594 | 0.3559 | 0.4908 | 0.6749 | 0.7531 |
| $TC+POS_{Spectrum+Linear}$ | 11 | | 0.2451 | 0.6733 | 0.3587 | 0.4978 | 0.6789 | 0.7502 |
| Support Vector Machine (SVM) | | | | | | | | |
| $TC+POS_{Linear}$ | 11 | | 0.2461 | 0.6695 | 0.3589 | 0.4964 | 0.6785 | 0.7506 |
| $TC+POS_{Gaussian}$ | 11 | | 0.2385 | 0.6235 | 0.3436 | 0.4691 | 0.6599 | 0.7398 |
| $TC+POS_{Spectrum}$ | 11 | | 0.2410 | 0.6256 | 0.3463 | 0.4717 | 0.6623 | 0.7385 |
| $TC+POS_{LinGausSpec}$ | 31 | | 0.2435 | 0.6236 | 0.3488 | 0.4730 | 0.6631 | 0.7307 |
| $TC+POS_{Spectrum+Linear}$ | 11 | | 0.2455 | 0.6335 | 0.3519 | 0.4782 | 0.6673 | 0.7452 |
| Random Forest (RF) | | | | | | | | |
| $TC+POS_{Linear}$ | 11 | | **0.3858** | 0.2629 | 0.3091 | 0.2791 | 0.5993 | 0.7124 |
| $TC+POS_{Gaussian}$ | 11 | | 0.3531 | 0.2250 | 0.2715 | 0.2412 | 0.5807 | 0.6975 |
| $TC+POS_{Spectrum}$ | 11 | | 0.3190 | 0.2349 | 0.2672 | 0.2463 | 0.5780 | 0.6697 |
| $TC+POS_{LinGausSpec}$ | 31 | | 0.3557 | 0.2230 | 0.2718 | 0.2400 | 0.5800 | 0.6888 |
| $TC+POS_{Spectrum+Linear}$ | 11 | | 0.3522 | 0.2429 | 0.2860 | 0.2583 | 0.5875 | 0.6974 |

#### 6.3.2.2 Results and Discussion

Table 6.5 presents the results for the four ensemble combination approaches, i.e., weighted majority vote (WMV) and the three stacking approaches: Logistic Regression (LR), SVM and Random Forests (RF). For each approach, the table presents the results for text classification combined with the sequence classification predictions from each of the kernels individually, Linear, Gaussian and Spectrum, and from all of the sequence classification kernels (LinGausSpec). The table also shows the results for the hybrid kernel (Spectrum+Linear), along with the text classification baseline (TC).

Firstly, reviewing the effectiveness of the kernel methods in the ensemble approaches, we note from Table 6.5 that the linear kernel performs best for ensemble approaches, since it (1) achieves significant improvements (denoted as †), and performs better for all measures, compared to the text classification baseline for WMV, and (2) achieves a higher $F_1$, $F_2$, BAC and auROC scores than the other kernel methods for the stacking approaches LR, SVM, and RF. This is surprising, since the Gaussian and Spectrum+Linear kernels perform well for $F_1$, $F_2$ and BAC for stand-alone classifiers. This appears be due to the liner kernel model being more similar to the (better) text classification model than the other kernel models are, while having enough uncorrelated variations to enhance the TC predictions. Turning our attention to the combinator

methods, we note that WMV performs better than the stacked approaches since it achieves the highest Recall, $F_1$, $F_2$, BAC and auROC scores.

Overall, in response to RQ6.2, we conclude that combining text classification with linear kernel POS sequence classification (TC+$POS_{Linear}$) and WMV performs best for sensitivity classification, from the combinations we tested. This approach achieves significant improvements, according to McNemar's test with $p < 0.05$, compared to the text classification baseline (TC). Moreover, this combination achieves the highest $F_2$ and auROC scores from all the combinations that we tested.

These results show that POS sequence classification can be an effective approach for improving the effectiveness of a text classification approach for sensitivity classification. The simple weighted majority vote combination strategy and a linear SVM kernel for POS sequence classification has the advantage of requiring much shorter training times compared to the other approaches. This, potentially, is beneficial for deploying sensitivity classification with POS sequence classification within our framework, since it means that the approach is more suited to being deployed early in the sensitivity review process and being re-trained as more data (i.e., examples of the sensitivities in the collection) become available, compared with the stacking approaches which require additional development time to train the combinator.

Moreover, the fact that the linear SVM kernel is the most effective for combining POS sequence classification with text classification suggests that information from the POS sequences can possibly be directly integrated as features to extend the text classification approach. We will evaluate this strategy and compare its effectiveness with vocabulary and semantic features in Section 6.5.

## 6.4 Semantic Features

As we previously discussed in Section 1.2, sensitive information is often a product of a combination of factors, such as *who said what about whom*. Sentences that share this type of structure can be said to be *semantically similar* and, therefore, we propose that identifying latent semantic relations that appear frequently in sensitive text can be an effective approach for identifying documents that contain sensitivity information.

The common factors of semantically similar sensitivities are two-fold: Firstly, relations between terms are often preserved over multiple sensitivities. For example, in the sentences "the source denied offering the plans for the attack" and "The informant provided us the names of the suspect" the relation of Entity A giving something to Entity B is common to both sentences; The second common factor in semantically similar sensitivities is that the entities or actions in the sensitivities often have similar meaning. For example, in the previous example the entities *informant* and *source* are both people who gave information, while the terms *provided* and *offering* tell the reader that something has been given.

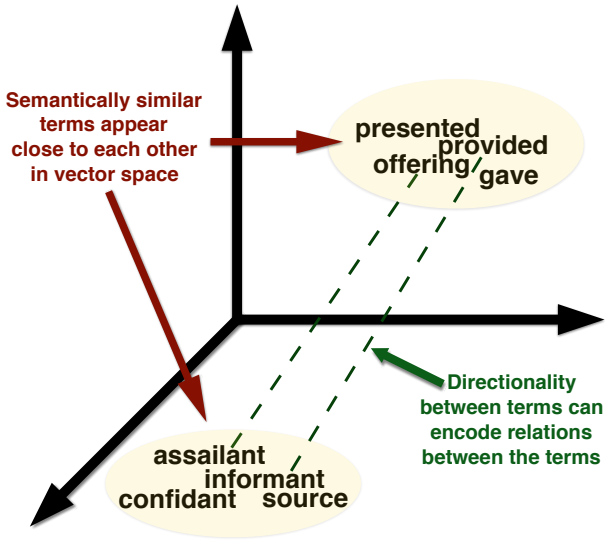



Figure 6.1: An illustration of a word embedding vector space.

One approach that has been shown to be effective at capturing the semantic relations between terms is word embeddings (Harris, 1954; Mikolov, Chen, Corrado & Dean, 2013). A word embedding model is trained by observing the contexts in which terms usually appear within a large collection of documents, with the assumption that words occurring within similar contexts are semantically similar. The resulting word embedding model is a vector space representation of the terms, in which the embedding vectors have low dimensionality, compared to the sparse vector representations more traditionally used in text classification. Each dimension of a word embedding vector maps to a latent feature of the word. The dense vector formation of word embedding models allow them to capture semantic qualities of, and relations between, terms in a collection. This, in turn, means that semantically similar relations tend to have similar values in specific dimensions of their embedding representations (Mikolov, Sutskever, Chen, Corrado & Dean, 2013).

Figure 6.1 illustrates two fundamental properties of word embeddings that can help us to identify semantically related sensitivities. Firstly, semantically *similar* terms are positioned close to each other within the vector space (e.g., informant/assailant) and, secondly, the directionality between multiple terms in the vector space can encode relations between the terms (e.g., the direction of *assailant* to *offering* is close to parallel with *informant* to *provided*). Therefore, relations such as the previous example, *who said what about whom*, can have their relations preserved in specific dimensions of a vector representations.

The properties of word embeddings illustrated in Figure 6.1 have resulted in word embeddings becoming very popular in natural language processing tasks, e.g., (Ghosh *et al.*, 2015; Pavlick *et al.*, 2015). Moreover, word embeddings have been shown to be effective in Information Retrieval and classification tasks, e.g. (Yang *et al.*, 2018; Zheng & Callan, 2015; Zuccon *et al.*, 2015). However, for classification, they have mostly been used for classifying short spans of text, such as tweets or sentences (Joulin *et al.*, 2016; Yang *et al.*, 2018). Typically, word embed-

dings have been used as an initialisation step for neural networks. However, Balikas & Amini (2016) proposed an approach for generating document features for text classification from word embedding models. The authors showed that a vector representation of a document could be derived from a word embedding model by deploying simple composition functions (e.g. min, average or max) (Collobert *et al.*, 2011; Socher *et al.*, 2011) to construct vector representations of combinations of words, such as phrases or sentences, from term vector models (Mitchell & Lapata, 2010). The authors showed that these compositional document vectors could be effectively used as features to extend text classification and improve classification performance.

To identify the latent semantic features of sensitive information, we follow the approach of Balikas & Amini (2016) to construct a document representation from word embeddings using composition functions. For a given word embedding model, $W$, of term vectors, $V^{\text{term}} \in W$ and a document collection, $C$, a vector representation, $V^{\text{doc}}, |v^{\text{doc}}| = |v^{\text{term}}|$, for a document, $d \in C$, is composed by applying a composition function, $f \in \{min, mean, max\}$ to the term vectors of each of the terms, $t_i$, in $d$. For example, using the composition function $f_{max}$, the value of the $n$th dimension of a document representation, denoted as $V^{doc}_{d,n}$, is:

$$V^{doc}_{d,n} = max(V^{term}_{i,n}) \forall t_i \in d \tag{6.3}$$

Each dimension of $V^{\text{doc}}$ can then be used as a single feature for the purposes of classification. Moreover, in addition to the composition functions $f_{min}$, $f_{mean}$ and $f_{max}$, following Balikas & Amini (2016), we also deploy the compound function *concat*, where the resulting document representation for $d$ is the concatenation of the document representations for $d$ for each of the composition functions $f_{min}$, $f_{mean}$ and $f_{max}$, as follows:

$$Concat(d) = [f_{min}(d), f_{mean}(d), f_{max}(d)] \tag{6.4}$$

Word embedding models capture the semantic relations of terms *within a collection*. Therefore, it is possible that semantic relations which are important for identifying sensitivities within our test collection may not be present in our chosen model. To address this, we construct document representations using two separate word embedding models that have been trained on different domains, namely Google News[4] and Wikipedia[5]. To do this, for a document $d$, we apply the selected function, $F \in \{min, mean, max, concat\}$, to $k$ word embedding models, $w_i$, to obtain $k$ separate document representations, where $1 \leq k \leq 2$. The document representations produced by the selected function $F$ for each of the $k$ word embedding models are concatenated together and each of the vector's dimensions are used as a separate classification feature. The

[4]https://code.google.com/archive/p/word2vec/

[5]http://nlp.stanford.edu/projects/glove/

final semantic document representation is therefore:

$$semantic\_representation(d) = [F_j(w_1,d), ... F_j(w_k,d)] \tag{6.5}$$

In the following section, we provide a thorough analysis of the effectiveness of extending sensitivity classification with semantic features derived from word embeddings models. Moreover, we evaluate the effectiveness of this approach compared with the vocabulary and syntactic features that we introduced in Sections 6.2 and 6.3 respectively.

## 6.5 Extending Sensitivity Classification with Vocabulary, Syntactic and Semantic Features

In the previous sections of this chapter we have introduced the methods that we deploy to automatically identify latent features of sensitive information and extend our baseline sensitivity classification approach. In particular: firstly, in Section 6.2 we presented extended term $n$-gram features for identifying language features of sensitivity; Secondly, in Section 6.3 we presented POS sequences for identifying syntactic features of sensitivity. Moreover, we empirically showed that a linear kernel is the most effective choice of SVM kernel for classifying sensitive documents using POS sequences; and, lastly, in Section 6.4 we presented the approach that we deploy for identifying latent semantic features of sensitive information.

In this section, we compare the effectiveness of each of the automatic feature generation approaches, from Sections 6.2, 6.3 and 6.4, for extending our baseline sensitivity classification approach that we presented in Chapter 5 (i.e., text classification). To extend text classification, for each document, $d_i$, we generate a document representation from a feature generation approach, $\mathbf{x}_{fi}$, and concatenate $\mathbf{x}_{fi}$ with the document representation that is used for text classification, $\mathbf{x}_{ti}$. Resulting in the document representation, $\mathbf{x}_i = \mathbf{x}_{ti} + \mathbf{x}_{fi}$. We evaluate each of the feature generation approaches individually, in pairs and using all of the three approaches together.

In the following section, we present our experimental methodology in more detail before discussing our results in Section 6.5.2. We provide an analysis of our findings in Section 6.6.

### 6.5.1 Experimental Method

In this section we present our experimental methodology for evaluating the effectiveness of vocabulary, syntactic and semantic features for sensitivity classification. Table 6.6 presents the combinations of feature sets that we evaluate, and the abbreviations that we use to denote each combination in the remainder of this chapter. The research questions that we address are two-fold:

- RQ6.3: Which automatically generated features, vocabulary, syntax or semantic, are most

Table 6.6: Extending sensitivity classification with language, syntax and semantic features: The feature set combinations that we evaluate and the abbreviations that we use to denote them.

| **Feature Set** | **Stand Alone** | **Extending Baseline** |
|---|---|---|
| Text Classification (baseline) | Text | - |
| Term *n*-grams | TN | Text+TN |
| Grammatical | POS | Text+POS |
| Semantic | WE | Text+WE |
| Term & Grammatical | TN+POS | Text+TN+POS |
| Term & Semantic | TN+WE | Text+TN+WE |
| Grammatical & Semantic | POS+WE | Text+POS+WE |
| Term & Grammatical & Semantic | TN+POS+WE | Text+TN+POS+WE |

effective for extending text classification for sensitivity classification?

- RQ6.4: Do multiple word embedding models trained on different domains further improve the effectiveness of semantic features for sensitivity classification?

To answer these questions, we use the same test collection and 5-fold Cross Validation folds that we use in the previous sections of this chapter and perform a similar binary classification *sensitive* vs. *not sensitive* experiment. For our baseline sensitivity classifier, we use the text classification configuration that we identified as the most effective in Chapter 5, i.e we use TF-IDF as our document representation, we retain stopwords and we do not perform any feature reduction. We denote this approach as Text Classification (Text) in Tables 6.8 and 6.9. For vocabulary features, presented in Section 6.2, we test for term *n*-grams where $n = \{2..10\}$. When testing for values of $n$, we include *n*-grams for all values $< n$, i.e., when $n = 3$ feature vectors are constructed from all bi-grams and tri-grams. In the remainder of this paper, we denote term features as $TN_n$ (i.e., for the previous example, $TN_3$). For syntactic features, presented in Section 6.3, we use the TreeTagger[6] part-of-speech tagger to POS tag documents and use a reduced set of 15 POS tags (Lioma & Ounis, 2006). We test for POS *n*-grams where $n = \{1..10\}$. Following the experimental setup for vocabulary features, when testing for values of $n$, we include *n*-grams for all values $< n$. Syntactic features are denoted as $POS_n$. For generating the semantic features that we presented in Section 6.4, we use *pre-trained* word embedding models and test if using two word embeddings models trained on different domains improves the effectiveness of semantic features for sensitivity classification. Table 6.7 presents the word embedding models that we test. For each of the models, we evaluate each of the composition functions *min*, *mean*, *max* and *concat*. As can be seen from Table 6.7, the models have 300 dimensional vectors and, hence, the functions *min*, *mean* and *max* result in 300 document features (900 for *concat*).

We use scikit-learn[7] for pre-processing and classification. As our classifier, we use SVM with a linear kernel and $C = 1.0$, since this theoretically motivated, default, parameter setting has been

[6]http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
[7]http://scikit-learn.org/

shown to provide the best effectiveness for text classification (Joachims, 1998; Sebastiani, 2002). We report precision, Recall, $F_1$, $F_2$, BAC and auROC scores. We test statistical significance, $p < 0.05$, using McNemar's non-parametric test (McNemar, 1947). Significant improvements compared to the text classification baseline (Text) are denoted with †. Additionally, in Table 6.9, significant improvements compared to the text classification with additional term features (Text+TN) are denoted with ‡.

### 6.5.2 Results

In this section, to answer research questions RQ6.3 and RQ6.4, we present the results of our classification experiments over two tables: Firstly, Table 6.8 presents the classification performance for each combination of vocabulary, syntax and semantic feature sets as *stand-alone* features; Table 6.9 presents the performance of each combination of feature sets extending the text classification baseline.

The baseline text classification approach (Text) is shown at the top of Tables 6.8 and 6.9, followed by sections for single, paired and triple feature sets respectively. We present results for vocabulary (i.e., term *n*-gram) features (TN), syntactic features (POS) and semantic features (WE). For WE, we present the results of the single word embedding models, Wikipedia ($WE_{wp}$) and Google News ($WE_{gn}$), and when used together ($WE_{wp}$+$WE_{gn}$). In Tables 6.8 & 6.9, we use $F_2$ as our preferred metric and present the best performing size of *n*-grams for TN and POS. For semantic features, we present the best performing composition function (*min*, *max*, *mean* or *concat*).

Firstly, from Table 6.8, we observe that when the feature sets are deployed individually the language features are the only feature set that perform better than the text classification baseline. Moreover, the language features ($TN_6$) perform better for all of the reported measures and the improvements are statistically significant (denoted as †). This result is not surprising, since the feature set is actually term n-grams without unigram features. The result does, however, provide good evidence that larger term *n*-grams are useful for identifying sensitive information. Moreover, we would expect this result to improve with the addition of unigram features. Secondly, we note that the syntactic and semantic features perform less well than the text classification baseline. Moreover, we can see that the *concat* composition function consistently performs best for semantic features. These findings are in line with the findings of Balikas & Amini (2016) who also observed that word embedding features did not perform as well as text classification on

Table 6.7: The pre-trained word embedding models that we use for deriving semantic features.

| Model | Architecture | Vocabulary Size | #Dimensions | Training | Context Window | Ref |
|---|---|---|---|---|---|---|
| Google News | word2vec | 3M | 300 | Negative Sampling | BoW5̃ | $WE_{gn}$ |
| Wikipedia+Gigaword5 | Glove | 400,000 | 300 | AdaGrad | 10+10 | $WE_{wp}$ |

Table 6.8: Results for combinations of *vocabulary*, *syntax* and *semantic* feature sets, compared against the text classification (Text) baseline. The table shows the precision, recall, $F_1$, $F_2$, Balanced Accuracy (BAC) and auROC scores. Statistical significance compared to the baseline is denoted as † (McNemar's test, $p < 0.05$).

| Configuration | | Precision | Recall | $F_1$ | $F_2$ | BAC | auROC |
|---|---|---|---|---|---|---|---|
| Text | | 0.2546 | 0.6831 | 0.3701 | 0.5098 | 0.6882 | 0.7518 |
| $TN_6$ | † | 0.2607 | 0.6970 | 0.3786 | 0.5207 | 0.6972 | 0.7626 |
| $POS_{10}$ | | 0.2149 | 0.6095 | 0.3177 | 0.4456 | 0.6353 | 0.6861 |
| $WE_{wp}$(concat) | | 0.2019 | 0.6055 | 0.3025 | 0.4321 | 0.6203 | 0.6801 |
| $WE_{gn}$(concat) | | 0.1959 | 0.6034 | 0.2956 | 0.4258 | 0.6130 | 0.6434 |
| $WE_{wp}$+$WE_{gn}$(concat) | | 0.2106 | 0.6235 | 0.3146 | 0.4474 | 0.6334 | 0.6962 |
| $TN_{10}$+$POS_{10}$ | | 0.2647 | 0.5974 | 0.3632 | 0.4724 | 0.6706 | 0.7407 |
| $TN_{10}$+$WE_{wp}$(concat) | † | 0.2634 | 0.7130 | 0.3839 | 0.5302 | 0.7039 | **0.7797** |
| $TN_9$+$WE_{gn}$(concat) | † | 0.2552 | 0.7208 | 0.3761 | 0.5267 | 0.6993 | 0.7638 |
| $TN_8$+$WE_{wp}$+$WE_{gn}$(concat) | † | **0.2657** | **0.7309** | **0.3890** | **0.5401** | **0.7110** | 0.7772 |
| $POS_{10}$+$WE_{wp}$(concat) | | 0.2174 | 0.6512 | 0.3241 | 0.4619 | 0.6458 | 0.7120 |
| $POS_{10}$+$WE_{gn}$(concat) | | 0.2081 | 0.6275 | 0.3117 | 0.4455 | 0.6315 | 0.6956 |
| $POS_{10}$+$WE_{wp}$+$WE_{gn}$(concat) | | 0.2199 | 0.6552 | 0.3280 | 0.4670 | 0.6507 | 0.7202 |
| $TN_{10}$+$POS_{10}$+$WE_{wp}$(concat) | † | 0.2592 | 0.6931 | 0.3760 | 0.5171 | 0.6942 | 0.7585 |
| $TN_{10}$+$POS_{10}$+$WE_{gn}$(concat) | † | 0.2474 | 0.6651 | 0.3584 | 0.4937 | 0.6757 | 0.7472 |
| $TN_9$+$POS_{10}$+$WE_{wp}$+$WE_{gn}$(concat) | † | 0.2531 | 0.6850 | 0.3679 | 0.5078 | 0.6868 | 0.7599 |

their own and that the concat function was consistently the best composition function. Adding semantic features to the text *n*-grams results in additional improvements, compared to the baseline, and $TN_8$+$WE_{wp}$+$WE_{gn}$(concat) achieves the best overall performance in Table 6.8.

From Table 6.9, we can see that extending text classification with semantic features significantly improves classification performance (Text+$WE_{wp}$+$WE_{gn}$(concat)). Moreover, this configuration performs better for all of the reported metrics compared with the baseline. However, extending text classification with term *n*-grams (Text+$TN_9$) achieves the best classification performance for single feature sets.

Overall, the best performance is achieved when text classification is extended with additional vocabulary and semantic features combined, Text+$TN_7$+$WE_{wp}$+$WE_{gn}$(concat). This combination achieves 0.5425 $F_2$ and 0.7229 TPR, correctly classifying ~6% more sensitive documents than the text classification baseline. Notably, this combination also results in significant improvements compared to extending text classification with only term *n*-gram features (Text+$TN_9$), denoted as ‡ in Table 6.9.

In response to RQ6.3, firstly, we find that semantic word embedding features are, indeed, useful features for sensitivity classification. This is shown by the observation of significant improvements to classification effectiveness when they are added to the next best performing feature set, denoted by ‡ in Table 6.9. However, we conclude that the best overall classification performance is achieved when text classification is extended with additional language and semantic features. Moving to RQ6.4, Tables 6.8 & 6.9 show that using multiple embedding mod-

Table 6.9: Results for extending the text classification (Text) baseline with combinations of *language*, *syntax* and *semantic* feature sets. The table shows the precision, recall, $F_1$, $F_2$, Balanced Accuracy (BAC) and auROC scores. Statistical significance compared to the baseline is denoted as †, and compared to the text classification with additional term features (Text+TN) are denoted with ‡ (McNemar's test, $p < 0.05$).

| Configuration | | | Precision | Recall | $F_1$ | $F_2$ | BAC | auROC |
|---|---|---|---|---|---|---|---|---|
| Text | | | 0.2546 | 0.6831 | 0.3701 | 0.5098 | 0.6882 | 0.7518 |
| Text+$TN_9$ | † | | 0.2667 | 0.7010 | 0.3858 | 0.5279 | 0.7035 | 0.7782 |
| Text+$POS_{10}$ | | | 0.2596 | 0.6532 | 0.3707 | 0.4999 | 0.6846 | 0.7498 |
| Text+$WE_{wp}$(concat) | | | 0.2474 | 0.6692 | 0.3609 | 0.4984 | 0.6799 | 0.7584 |
| Text+$WE_{gn}$(concat) | | | 0.2435 | 0.6653 | 0.3560 | 0.4933 | 0.6752 | 0.7459 |
| Text+$WE_{wp}$+$WE_{gn}$(concat) | † | | 0.2557 | 0.6891 | 0.3725 | 0.5138 | 0.6919 | 0.7594 |
| Text+$TN_6$+$POS_{10}$ | † | | **0.2780** | 0.6751 | 0.3920 | 0.5224 | 0.7029 | 0.7725 |
| Text+$TN_9$+$WE_{wp}$(concat) | † | | 0.2678 | 0.7090 | 0.3881 | 0.5322 | 0.7070 | **0.7874** |
| Text+$TN_6$+$WE_{gn}$(concat) | † | | 0.2699 | 0.7169 | 0.3913 | 0.5371 | 0.7107 | 0.7784 |
| Text+$TN_7$+$WE_{wp}$+$WE_{gn}$(concat) | † | ‡ | 0.2730 | **0.7229** | **0.3956** | **0.5425** | **0.7149** | 0.7859 |
| Text+$POS_{10}$+$WE_{wp}$(concat) | | | 0.2507 | 0.6493 | 0.3609 | 0.4913 | 0.6767 | 0.7620 |
| Text+$POS_{10}$+$WE_{gn}$(concat) | | | 0.2515 | 0.6571 | 0.3626 | 0.4950 | 0.6796 | 0.7546 |
| Text+$POS_{10}$+$WE_{wp}$+$WE_{gn}$(concat) | | | 0.2504 | 0.6532 | 0.3612 | 0.4930 | 0.6779 | 0.7634 |
| Text+$TN_4$+$POS_{10}$+$WE_{wp}$(concat) | † | | 0.2674 | 0.6811 | 0.3827 | 0.5181 | 0.6979 | 0.7789 |
| Text+$TN_9$+$POS_{10}$+$WE_{gn}$(concat) | † | | 0.2634 | 0.6830 | 0.3786 | 0.5154 | 0.6955 | 0.7747 |
| Text+$TN_6$+$POS_{10}$+$WE_{wp}$+$WE_{gn}$(concat) | † | | 0.2657 | 0.6910 | 0.3825 | 0.5214 | 0.6995 | 0.7798 |

els, $WE_{wp}$+$WE_{gn}$, consistently out performs either of the single models, $WE_{wp}$ or $WE_{gn}$, when they are used individually. Therefore, we conclude that using multiple word embedding models trained on different domains does, indeed, improve the effectiveness of semantic features for sensitivity classification.

## 6.6 Analysis

In this section, we provide analysis of the findings from our classification experiments. Firstly, in Section 6.6.1 we review the term features that our classification models believe to provide the most evidence of potential sensitivity. Next, in Section 6.6.2, we discuss the classification predictions that are correct solely due to the word embedding features. Lastly, in Section 6.6.3, we discuss the benefits for the sensitivity review process from extending text classification with semantic and term *n*-gram features.

### 6.6.1 Important Vocabulary Features

In this section, we present some of the important vocabulary (term) features that provide the most evidence of potential sensitivity to our learned models. As outlined in Section 6.5, for our experiments we use a SVM classifier with a linear kernel. An attractive property of a linear model such as this, is that we can use the model's feature coefficients as an estimate of how

123 12356 129 1355 13th abiola abidjan accelerate accurate activities convert criminals criticizes cull customers dimension dob domain dpa drago dublin douala email embassy empower encounter ends enforcing engage england exme february fifth financial finished forefront given goa gonzalez gos humanr information infusions injuries insider khmer leap ljubljana locations logistical looming marc marr massive masters migrant miliband ministry netherlands newspaper ni nicon nigeria night nil ninety nodding nominees non norwegians outdoor place placements pledge poe policy preferred prop protesting psc pter ranging react rebuilding received reopen season seasons semitic sensitive sovereign taped targeted tenant terrorist tribal vung wagp wearing wha

Figure 6.2: The 100 single term (uni-gram) features with the largest coefficient scores, i.e., "most important", for the text classification (Text) model presented in Tables 6.8 and 6.9.

nato permitted participate food distribution programs united
security advise visitors avoid making pejorative comments members
completely destroyed sophisticated firebombing late june 2002 short
100 percent foreign ownership companies outside petroleum sector
facilitated using double cycle gas turbines fueled
motivated violence with incidents attributed armed local muslim
male workers requires factory division ministry labor
cover charges incidents crime patrons establishments drug
ministry foreign affairs socialist republic vietnam wins
verifies department country possesses credible information
classified sensitive materials chancery accommodation residences visitors
nuclear stone bed cluster reactor units 320 coal
groups including bombings public places indicates potential threat british
blow international legitimacy disrespect muslim arab world paper
period see information completeness obviously technologies time schedules change

Figure 6.3: The 15 highest ranked n-gram features, where $6 \leq n \leq 9$, for the text classification + term $n$-gram model (Text+$TN_9$) presented in Table 6.9. Each row presents 1 $n$-gram feature. Stopwords are removed from the collection prior to feature generation.

*important* a classification feature is. It is important to note, however, that the important features presented in this section are a snapshot of a particular classifier's interpretation. Changes to classification parameters, e.g., the SVM $C$ parameter, or data preprocessing, e.g., binary vs. term frequency vectors, can result in some variation in the importance estimation of features. This is especially true of text classification, where features are individual terms. However, over time, through many different classifications, we can build up an understanding of which term features a particular classifier, e.g., SVM, believes are salient to sensitivity classification.

Figure 6.2 presents the 100 most important uni-gram term features for the text classification model (Text) presented in Tables 6.8 and 6.9[8]. From reviewing the terms in Figure 6.2, we see some terms that regularly appear in our analysis of important classification terms, such as "dob"

[8]To protect any sensitive information, Figure 6.2 has been sanitised by replacing some of the terms with synonymous terms.

(date of birth) and "poe" (place of education) which tend to be associated with personal information sensitivities and "information", "react" and "received", which are often associated to international relations sensitivities within our collection. Also, we note that in our experiments, numbers also frequently appear to be important classification features. This is mainly due to an artefact of our test collection, in that the numbers often reference other documents. For example, if many non-sensitive documents reference the non-sensitive document ref: 1355, then the token 1355 can be a strong classification feature. Interestingly, on our collection, a linear SVM model often tends to identify terms that would instinctively appear to be controversial as being useful features of sensitivity. For example, in Figure 6.2 we see terms such as "semitic" and "terrorist". Analysing the use of these terms in documents that have been judged to contain sensitivity, we find that the terms are more related to the context that the sensitive information appears in, rather than the information that was actually judged as being sensitive by the sensitivity reviewers. Useful contextual term features, such as these, are likely to vary substantially between different collections that contain different subject matter. Therefore, it is important that a framework for assisting digital sensitivity review can quickly learn to identify the useful term features for sensitivity classification early in the review process. We will investigate this further in the following chapter.

Moving to $n$-gram term features, Figure 6.3 presents 15 n-gram term features (1 $n$-gram feature per row, where $6 \leq n \leq 9$) that the best performing text classification + term $n$-gram model, Text+$TN_9$ presented in Table 6.9, believes to be most important for sensitivity classification[9]. The features in Figure 6.3 are all within the top 50 "most important" classification features within their 5-fold Cross Validation fold for Text+$TN_9$.

From reviewing the term $n$-gram features in Figure 6.3, we can see that larger values of $n$ can, in effect, provide a short summary of passages that the classifier believes to be most associated to sensitivity, e.g., "verifies department country possesses credible information". As $n$ increases, however, the likelihood of obtaining an exact match of an $n$-gram feature when the model is applied to new previously unseen documents decreases. As previously outlined in Section 6.5.1, when testing for values of $n$, we include $n$-grams for all values $< n$, i.e., when $n = 3$ feature vectors include all bi-grams and tri-grams. Therefore, in practice, larger values of $n$ can provide a good approximation of the "importance" of their contained $n$-gram features for smaller values of $n$.

The term and $n$-gram features presented in this section, thus far, have been directly identified as being important for sensitivity classification by the learned models. Additionally, we can potentially identify important terms for sensitivity classification by combining evidence of feature importance with the word embedding document representations presented in Section 6.4.

As presented Section 6.4, when deriving semantic features we construct a document repre-

[9]To protect any sensitive information, Figure 6.3 has been sanitised by replacing some of the terms with synonymous terms.

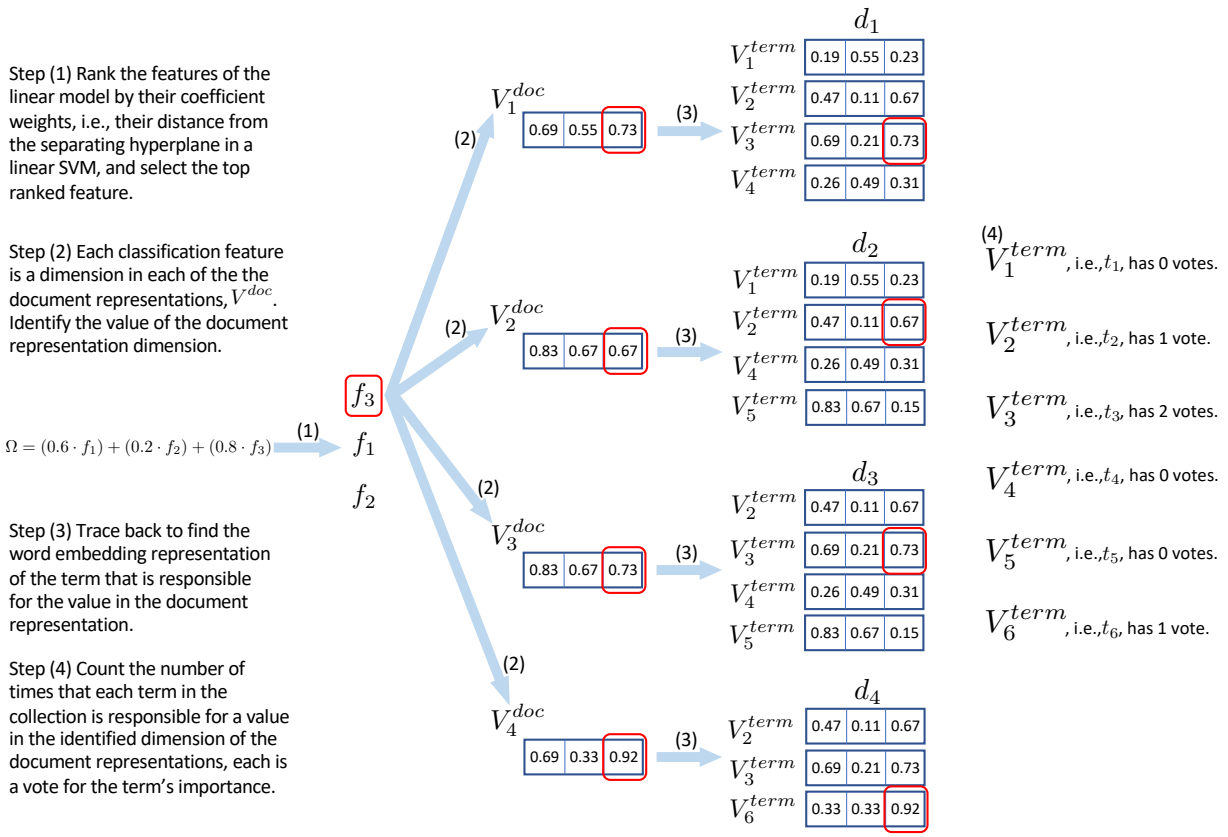



Figure 6.4: An illustration of how we identify the terms that are most associated with important semantic word embedding features.

sentation from word embeddings using composition functions (Collobert *et al.*, 2011; Socher *et al.*, 2011). For example, using the composition function $F_{\max}$, the value of the $n$th dimension of the document representation, $V^{doc}_{d,n}$, is: $V^{doc}_{d,n} = max(V^{term}_{i,n}) \forall i \in C_d$. In other words, the value of the $n$th dimension of the document representation $V^{\text{doc}}$ is the value of the $n$th dimension of the word embedding, $V^{term}_{i,n}$, for the term in $d$ with the largest value in dimension $n$. Therefore, we can trace back to discover the term in $d$ that was responsible for the value of the $n$th dimension of $V^{\text{doc}}$. Moreover, each dimension of $V^{\text{doc}}$ is used as a single feature in classification and, following the same feature importance procedure outlined above, we can therefore identify the word embedding dimensions that are most important to sensitivity classification and the terms associated with these dimensions.

Figure 6.4 illustrates our approach for identifying the importance, or frequency of association, of terms to the semantic word embedding feature that has the largest weight, or importance, in a linear classification model, $\Omega$. In Figure 6.4, the model, $\Omega$, has been trained on semantic word embedding document representations, $V^{doc}$, that are constructed using the composition function $F_{\max}$ on a collection of four documents, $d_1, d_2, d_3$ and $d_4$. Each of the documents $d_1..d_4$ are represented by the word embedding representations, $V^{term}$, of the terms in the document. In total there are six terms, $t_1..t_6$, in the collection and their word embedding representations, $V^{term}$, have three dimensions, $|V^{term}| = 3$. Therefore, the semantic document representations also have

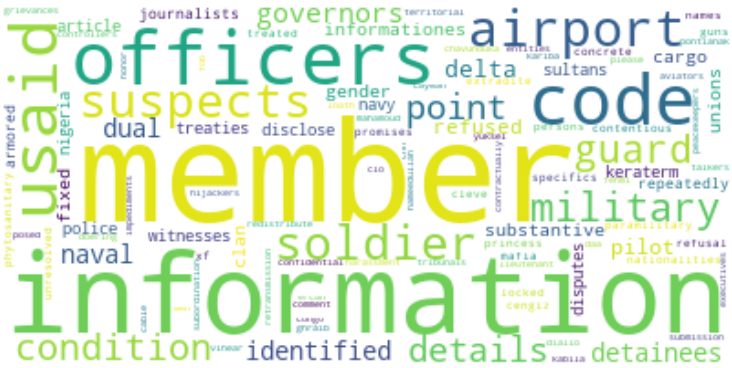



Figure 6.5: Word cloud representation of the terms associated with the most important word embedding dimension classification feature for the *max* function of the best performing model from Table 6.9 (Text+TN$_7$+WE$_{wp}$+WE$_{gn}$(concat)).

three dimensions, $|V^{doc}| = 3$ and the linear model, $\Omega$, has three weighted features, $f1..f3$.

There are four steps to our proposed process, as illustrated in Figure 6.4. In step 1, the semantic features from $\Omega$ are ranked by their coefficient weights and the top ranked feature is selected as the most important classification feature. The identified feature corresponds to a specific dimension in the semantic document representations, $V^{doc}$, and, in Step 2 of the process, we identify the feature value for the corresponding dimension in each of the document representations, $V^{doc}$. Once we have identified the dimension of interest and the value of the dimension in a document, in Step 3, we can trace back to the word embedding representations of the terms in the document. We then identify which term is responsible for, or associated with, the value in the document representation, $V^{doc}$. In Step 4 of the process, we count the number of times that a specific term, $t_i$, is responsible for a value in the identified dimension of the document representations. Each time that a term is responsible for a value in the important dimension of a document representation counts as a vote for the term's importance. Identifying salient term features from word embedding dimensions is a novel approach to identifying important classification term features and, to the best of our knowledge, there is no existing literature that deploys this approach.

Figure 6.5[10] presents the terms associated to the most important word embedding dimension feature for the *max* composition function in the model Text+TN$_7$+WE$_{wp}$+WE$_{gn}$(concat) presented in Table 6.9. The size of the text in Figure 6.5 is proportional to the number of documents represented by the term. It is the combined distribution of all of these terms within the collection that results in the importance of the semantic feature. As can be seen from Figure 6.5, many of the terms that are strongly associated with this word embedding classification feature are very interesting for potentially predicting sensitivity, for example "information", "suspects" or "officers". Moreover, many of the associated terms that appear in relatively fewer documents also appear to be of interest, for example "identified", "details" or "detainees".

[10]This figure has been sanitised by replacing some of the terms with synonymous terms.

### 6.6.2 Semantic Features

We now provide a short analysis of the documents we can correctly classify due to the inclusion of the semantic features. We compare the best performing system, Text+TN+$WE_{wp}$+$WE_{gn}$, against the next best performing system, i.e. text classification extended with term *n*-gram features, Text+TN.

Additional semantic features (from multiple domains) enable the classifier to convert 23 False Negative predictions to True Positive predictions, and 144 False Positive predictions to True Negative predictions. 13.77% of these converted predictions were sensitive documents. From the 23 converted sensitive documents, 15 are sensitive with respect to *International Relations*, 4 are sensitive with respect to *Personal Information* and 4 are sensitive with respect to both sensitivities.

Each of the documents with International Relations sensitivity contain multiple paragraphs that recount interactions and conversations between people and, moreover, the document's sensitivity is directly linked to these. Figure 6.6 presents example excerpts from two of these documents. The document shown in Figure 6.6(a) reports an informant's recount of inappropriate interrogations and harassment of activists in Cambodia by the police, while the document presented in Figure 6.6(b) recounts disparaging remarks regarding the levels of corruption and efficiency throughout the Cameroon political establishment. Identifying these conversational sensitivities is in line with how we expect semantic features to enhance sensitivity classification, since these relations can be preserved within multiple dimensions of the vector representations.

Interestingly, the personal information sensitivities that we are able to classify correctly due to the semantic features also relate to actions, such as booking hotels, forced resignations and visa bans. This shows that the semantic features enable the classifier to identify personal information sensitivities that arise within a specific context that the classifier performs less well on without the semantic features.

### 6.6.3 Sensitivity Review

It is useful to provide sensitivity reviewers with a reliable way to predict how many sensitive documents remain in a partially reviewed collection. One way to approach this is to rank documents by a classifier's decision function output and review the ranking sequentially. We can then ask "how conservative does a classifier have to be, to correctly predict a certain percentage of sensitive documents?" In line with this, Figure 6.7 presents the Receiver Operating Characteristic curve, and True Positive Rate vs. classification threshold for our classifier with additional term and semantic features, compared against the baseline text classification.

As can be seen from Figure 6.7(a), the additional features increase the True Positive Rate throughout the ranking. Therefore, a reviewer can have increased confidence in the system. Additionally, Figure 6.7(b), shows that semantic and term features enable the classifier to be less

(a)

Embassy was informed on October 19th, from the reform activist Sin Boran, that fellow activists Duong Davuth, Tok Makara and Vang Jorani were visited by police on the evening of October 3rd. The men were taken to Sangkum police department where they were questioned for several hours. The following day, eight other activists were were also detained and interrogated before being released later in the day.

Sin Boran informed the embassy that Duong Davuth and Tok Makara had been coerced to sign a statement but had refused to eat or drink while they were detained. The pair were instructed to return to the police station on October 10th. We do not know details of the events of that day yet.

Sin said that it is his belief that the men were interrogated due to the specific request of the Cambodian General Secretary, Mau Meaker. The request is believed to be prompted by a desire to establish an association to “generate party support and eradicate corruption in the system”. The translated transcript of the testimony is included in Section 4 below.

(b)

Golavech said the GOC stated that good governance would be a primary issue for many African countries, adding that "nations cannot be strong in today's world without good governance." When asked whether Cameroon would put forward a plan or proposal, Golavech stated that the nations of Africa had to reach a common understanding or approach and would be discussing with other African nations in the coming months.

Cameroon faces many challenges in achieving good governance as the country has had decades of dictatorship and disputed elections resulting in corrupt institutions and patchy infrastructure. The judiciary is corrupt and incapable of efficient processing. Also, there is insufficient capacity to enforce national laws and regulations and the physical infrastructure of the country has deteriorated. Public bodies appear to be transparent but this is a sham. Reform will not be speedy and it is not clear that there is the will to do it.

Figure 6.6: Excerpts from two documents containing sensitivities linked to conversations that the classifier could only identify with the addition of semantic features. Document (a) reports an informant’s recount of inappropriate interrogations and harassment of activists in Cambodia by the police. Document (b) recounts disparaging remarks regarding the levels of corruption and efficiency throughout the Cameroon political establishment.

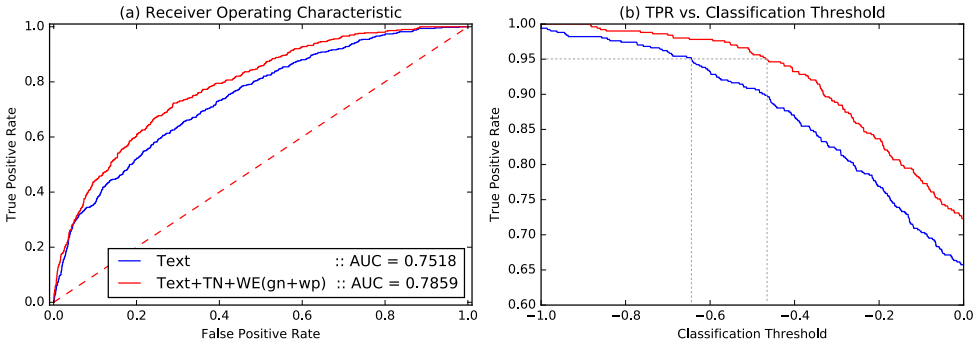



Figure 6.7: (a) Receiver Operating Characteristic Curve. (b) True Positive Rate vs. Classification Threshold. The blue line shows the baseline text classification (Text) and the red line shows Text+$TN_7$+$WE_{wp}$+$WE_{gn}$(concat). The dashed line in (a) shows a random classifier. The dashed lines in (b) show the classification threshold required to achieve 0.95 TPR.

conservative. For example, the gray dashed lines in Figure 6.7(b) show that, with the additional features, we can correctly classify 95% of all sensitive documents by lowering the classification threshold to -0.46, whereas, the baseline would need to be set at -0.645. By using our approach, on this test collection, a reviewer would need to review 262 fewer documents to identify 95% of all sensitive documents. In practice, this could benefit the review process in allocating reviewing resources. For example, as we previously discussed in Chapter 1, government departments are not expected to be able to recruit enough reviewing resources to review all of the digital documents that are to be archived (The National Archives, 2016*a*). One approach for selecting documents to be reviewed is to set the classifier's threshold more conservatively to identify a subset of the document collection that is expected to contain only a small percentage of sensitive documents and review these documents first. We will investigate prioritising documents for review to make the best use of reviewing resources in Chapter 8.

## 6.7 Conclusions

In this chapter, we presented the three approaches for automatically identifying latent language, syntactic and semantic features of sensitive information. Moreover, we empirically evaluated the effectiveness of each of the feature sets for enhancing sensitivity classification. In particular, in Section 6.3 we evaluated ensemble approaches for combining POS sequence classification with text classification for classifying sensitive documents. We showed that by deploying POS sequence classification with a linear SVM kernel and combining the approach with text classification by a weighted majority vote ensemble led to significant improvements in sensitivity classification effectiveness. In Section 6.5 we evaluated extending the sensitivity classification baseline with each of the automatically generated feature sets. We showed that extended term *n*-grams can be effective features for extending sensitivity classification to significantly improve classification effectiveness. Moreover, we showed that semantic features derived from word

embedding models that were trained on different domains led to additional significant improvements for sensitivity classification. These findings provide good evidence that an effective sensitivity classifier can be learned by a purely automatic approach to feature engineering. Moreover, we showed that the approach for deriving semantic features, or document representations, from word embedding models retains the ability to identify the terms in a document that are important for the classifier's prediction. The approaches that we discussed in this chapter rely on there being a collection of documents that contain sensitive information that is representative of the sensitivities that are to be classified. Moreover, that collection needs to have been sensitivity reviewed, so that there is a set of judgements that can be used to train the classifiers. However, as we have previously discussed in Chapters 1 and 4, the sensitivities in a collection are not known *a-priori* and, therefore, we need a method of learning an effective sensitivity classifier while using the least reviewer judging effort possible. In the following chapter, we will investigate methods to reduce the amount of reviewing effort that is required to be able to develop an effective sensitivity classifier.

# Chapter 7

# Active Learning for Sensitivity Classification

## 7.1 Introduction

In the previous Chapter, we introduced three approaches for automatically identifying latent features of sensitive information that we propose to extend our sensitivity classification baseline with, namely vocabulary features (extended term *n*-grams) (see Section 6.2), syntactic features (POS sequences) (see Section 6.3) and semantic features (word embeddings) (see Section 6.4). Moreover, in the previous chapter, we empirically demonstrated how language, syntactic and semantic features can be used to improve the effectiveness of sensitivity classification (see Sections 6.3.2 and 6.5). The sensitivity classification approaches that we have presented in Chapters 5 and 6 rely on there being an available test collection of documents that have previously been sensitivity reviewed. Moreover, the available test collection needs to contain sensitive information that is *representative* of the sensitivities in the collection of documents that is to be sensitivity reviewed. However, sensitivity is context-dependent and, therefore, the sensitive information that is closed due to a specific FOIA exemption, e.g., international relations, can be very different in separate document collections. For example, the international relations sensitivities in a collection of documents that discuss political dealings in a Middle East war zone could be mostly due to specific pieces of information supplied by individual civilians, while the international relations sensitivities in a collection from a political ally, such as the USA, could be mostly due to details of bilateral relationships. These sensitivities are likely to *look* very different from each other. For example, the vocabulary that is used and the entities that are mentioned in the sensitive text will likely vary between different collections, i.e., some of the terms that are used often in sensitive information in one collection might be used often but not associated with sensitive information in another collection.

We argue that, to assist digital sensitivity review, early in the review process (i.e., when none, or only a small portion, of the collection has been reviewed), a framework for technology-

assisted sensitivity review should deploy an active learning strategy to prioritise for review the documents in the collection that will provide the framework's sensitivity classifier with the most information about the sensitivities in the collection. Moreover, the framework should integrate explicit feedback about the vocabulary that is used in the sensitive information in the collection to construct a representation of what the sensitivities look like. We argue that this will reduce the number of documents that have to be reviewed to learn an effective sensitivity classifier. Thereby, enabling the framework and sensitivity classifier to assist the reviewers earlier in the review process. For example, by providing the reviewers with useful information about which documents in the collection contain sensitive information, to assist the reviewers in making reviewing decisions.

In this chapter, we examine how to incorporate explicit feedback from a sensitivity reviewer about the sensitive information in a collection that is to be reviewed so that we can quickly learn to classify the sensitivities in the collection. In particular, we evaluate active learning as a strategy for prioritising specific documents to have reviewed so that the documents that are the most *informative* for developing a sensitivity classifier are reviewed before less informative documents. Moreover, we propose to further reduce the number of documents that are required to be reviewed to develop an effective sensitivity classifier by having a reviewer *annotate* any sensitive text in a document, to generate a representation of the sensitivities within the collection, as they perform the review. Our proposed approach is analogous to integrating the redaction process into the digital sensitivity review process by having a reviewer perform both tasks simultaneously. The remainder of this chapter is structured as follows:

- In Section 7.2, we provide an overview of how each of the components of our framework are instantiated, and the roles that they perform, to deploy our active learning strategies.
- In Section 7.3 we introduce the active learning strategies that we evaluate for selecting informative documents to have sensitivity reviewed. In particular, in Section 7.3.1 we present the three uncertainty sampling approaches that we evaluate. In Section 7.3.2, we present a Semi-Automated Text Classification (SATC) approach (Berardi *et al.*, 2012) from the literature that Berardi *et al.* (2015) have previously shown to be effective for increasing the cost-effectiveness of reviewers. We evaluate the approach from Berardi *et al.* (2012) as an active learning strategy for selecting informative documents to be reviewed.
- In Section 7.4, we propose to integrate reviewer feedback about the sensitive information in a collection by having the reviewer annotate, or redact, any sensitive information as they review. We present three approaches that we evaluate for integrating our proposed sensitivity annotation features into the active learning process for sensitivity classification.
- In Section 7.5, we discuss our choice of classifier for our active learning experiments.
- Section 7.6 presents our experimental methodology.

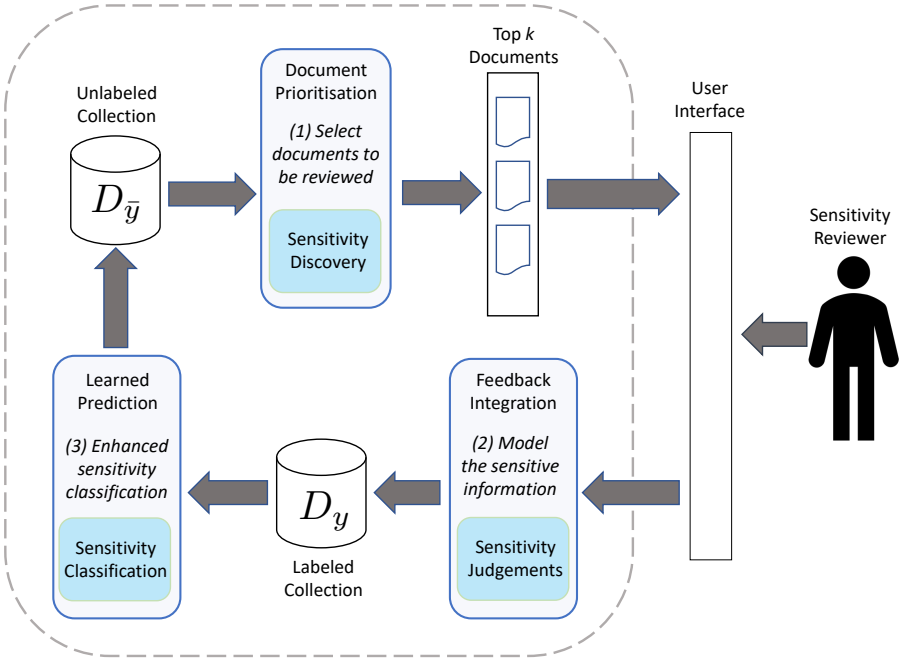



Figure 7.1: The roles of our framework's components for selecting informative documents to be reviewed and constructing a representation of the sensitivities from the reviewer's feedback.

- Section 7.7 presents the results of our experiments. In particular, we present the results of prioritising informative documents for review in Section 7.7.1 and for extending active learning to incorporate a reviewer's sensitivity annotation features in Section 7.7.2. Lastly, we discuss the impact that the *batch size*, i.e., the number of documents that are reviewed at each iteration of the active learning cycle, has on our proposed approach in Section 7.7.3.

- Finally, in Section 7.8, we summarize our conclusions from this chapter.

## 7.2 Active Learning for Sensitivity Classification

We propose to use an active learning approach to quickly learn to classify the sensitivities in a collection of documents, $D$, that is to be sensitivity reviewed. At the beginning of the review process, we do not have any *a-priori* knowledge of the sensitivities in $D$, i.e., the documents are *unlabelled*, $D_{\bar{y}}, |D_{\bar{y}}| = |D|$. Learning an effective sensitivity classifier is an iterative process in which at each iteration a sensitivity reviewer reviews $k$ documents and assigns each of the documents a class label, $y_i \in \{sensitive, nonSensitive\}$. The labelled documents, $D_y$, can then be used to train a sensitivity classifier, $\Omega$. We refer to this iterative process as the review cycle. At the beginning of the review cycle $|D_y| = 0$. The size of $D_y$ increases by $k$ at each iteration and at all times $|D_{\bar{y}}| + |D_y| = |D|$. Figure 7.1 shows how the components of our framework are instantiated to implement the active learning strategy. As can be seen from Figure 7.1, the

Document Prioritisation, Feedback Integration and Learned Predictions framework components work together to learn an effective sensitivity classifier and each component performs a single specific role, as follows:

1. The Document Prioritisation component is responsible for selecting $k$ unlabelled documents from $D_{\bar{y}}$ to have sensitivity reviewed at each iteration of the review cycle. The Document Prioritisation component deploys a document selection strategy to generate a ranking, $r_i$, of document, $d_1 .. \, d_{|D_{\bar{y}}|}, d_i \in D_{\bar{y}}$, to prioritise for review the documents that would likely provide the most useful information for training the sensitivity classifier, $\Omega$, if the document's associated class label, $y_i$, was known. These top $k$ documents in $r_i$ are then presented to the reviewer in rank order, $d_1 .. \, d_k$, to be sensitivity reviewed.

2. The Feedback Integration component is responsible for integrating the sensitivity judgements, $j_1 .. j_k$, that are provided by the sensitivity reviewer into the document representations, $\mathbf{x}_1 .. \mathbf{x}_k$. A sensitivity judgement for a document, $d_i$, contains the document's manually assigned class label, $y_i$ and, for documents that are labelled $y_{sensitive}$, a set of sensitivity annotations, $a_{di}, |a_{di}| \in \{0 .. \, |d_i|\}$, for any passages of sensitive text in the document.

3. The Learned Predictions component uses the documents with associated class labels and annotation information, $D_y$, to train the sensitivity classifier, $\Omega$, to predict the class label of each of the documents in $D_{\bar{y}}$. The predicted class labels, $\hat{y} \in \{sensitive, nonSensitive\}$, with a corresponding prediction confidence score, $\rho_i$, for the documents in $D_{\bar{y}}$ are used by the Document Prioritisation component to prioritise documents for review, based on the classifier's current knowledge about the sensitivities in $D_{\bar{y}}$.

## 7.3 Selecting Documents to be Reviewed

The first task that our active learning strategy has to perform is to prioritise for review the documents that will provide the most useful information for training a sensitivity classifier. This task is performed by the Document Prioritisation component of our framework (labelled as (1) Select select documents to be reviewed in Figure 7.1). Most of the active learning literature for document classification has focused on selecting informative documents for the classifier, e.g. (Lewis & Gale, 1994; McCallumzy & Nigamy, 1998; Yang *et al.*, 2009). As we previously discussed in Chapter 2 (Section 2.3), uncertainty sampling has been shown to be an effective document selection strategy for many active learning tasks (Settles, 2012). In this chapter, we evaluate the effectiveness of four document ranking strategies for selecting informative documents to train a sensitivity classifier. Three of the approaches that we evaluate are uncertainty sampling approaches and we provide the details of these techniques in Section 7.3.1. The fourth document selection strategy that we evaluate, *Utility-Theoretic* (Berardi *et al.*, 2012), is a Semi-Automated Text Classification (SATC) document ranking technique that has been shown to be effective for

increasing the cost-effectiveness of sensitivity reviewers (Berardi *et al.*, 2015). We provide the details of the Utility-Theoretic strategy in Section 7.3.2.

### 7.3.1 Uncertainty Sampling

As we previously discussed in Chapter 2, we evaluate three uncertainty sampling (Lewis & Catlett, 1994; Lewis & Gale, 1994) approaches for selecting informative documents to have reviewed. In this section, we provide a brief reminder of the three approaches that we evaluate.

The first uncertainty sampling strategy that we evaluate is *entropy based* uncertainty (Settles, 2012). Entropy uncertainty sampling ranks documents by the sum of their entropy (Shannon, 1948) scores for each of the document's possible labels, $y_i \in Y$. Entropy uncertainty sampling can be viewed as a measure of the amount information, i.e., entropy, that the document contains. Therefore, documents with a high $H(Y)$ score should provide more information to the classifier, $\Omega$, about their assigned label. Entropy uncertainty is defined as:

$$H(Y) = -\sum_i P(y_i) log P(y_i) \tag{7.1}$$

The second uncertainty sampling strategy that we present is the *margin* uncertainty sampling approach (Scheffer *et al.*, 2001). Margin uncertainty sampling calculates the difference in a classifier's predicted probability scores for a document's first and second most likely classification labels. The intuition of margin sampling is that the documents that the classifier is the most unsure about their correct class label will provide the most useful information to the classifier. Margin uncertainty sampling is defined as:

$$M(d_i, y_1, y_2) = |P(y_1|d_i) - P(y_2|d_i)| \tag{7.2}$$

The third, and final, uncertainty sampling approach that we evaluate is *relevance* sampling (Lewis & Gale, 1994). Relevance sampling selects the documents that the classifier is *most* confident are examples of the positive class. For this reason, this approach is usually referred to as a *certainty sampling* approach. In sensitivity classification, the positive class is sensitive documents and, therefore, we refer to this approach as sensitivity confidence, denoted as *sConf*, defined as:

$$sConf(d_i, y_i) = P(y_i|d_i) \tag{7.3}$$

### 7.3.2 Utility-Theoretic

*Semi-Automated Text Classification* (Berardi *et al.*, 2012) (SATC) addresses a scenario in which the state-of-the-art classifier for a particular classification task is not effective enough to meet an organisation's strict accuracy constraints. Moreover, in the SATC scenario, obtaining more training data for the classifier (if that is possible) is not expected to increase the classifier's

accuracy sufficiently. The aim of SATC is to produce an optimal ranking of documents, based on the predictions of the classifier, $\Omega$, such that if a reviewer was to start from the top of the ranking and proceed down the list correcting any mis-classifications until an available reviewing budget had expired, i.e., reviewing the top $k$ documents, the overall accuracy of the classifier's predictions would be maximised.

We evaluate the *Utility-Theoretic* (Berardi *et al.*, 2012) SATC approach as a document selection active learning strategy for developing a sensitivity classifier. SATC differs from active learning since, in SATC, the classifier is not re-training after the reviewer has reviewed the documents and corrected the classifier's mis-classifications. However, we argue that the Utility-Theoretic approach should perform well as an active leaning strategy for sensitivity classification since it has previously been shown to improve the cost-effectiveness of sensitivity reviewers, when applied to our classifiers from Chapter 5 for classifying individual FOIA exemptions (see Section 5.4) (Berardi *et al.*, 2015; McDonald *et al.*, 2014). Moreover, by feeding the corrected classifications back into the learning process to re-train the sensitivity classifier we are, in effect, just closing the loop in the active learning cycle.

The intuition of the approach from Berardi *et al.* (2012) is that in text classification problems where there is an imbalance in the distributions of classification categories (as is the case with sensitivity classification), and a metric is chosen to account for this imbalance (e.g., $F_2$), the improvements in effectiveness, or *gain*, that are derived from correcting a false positive prediction is not the same as that for correcting a false negative prediction. This is important for sensitivity, since the consequences of mis-classifying a sensitive document are much greater than that of mis-classifying a non-sensitive document.

Berardi *et al.* (2012) provided a thorough examination of the approach for a *multi-class multi-label* text classification scenario. However, in the case of binary classification, as is the case in our experiments, the utility-theoretic measure is defined as:

$$U(d_i) = \sum_{e} P(e)G(e) \tag{7.4}$$

where $P(e)$ is the probability of an event, $e$, occurring, i.e., a false negative (FN) or a false positive (FP) prediction, and $G(e)$ is the gain that can be obtained if that event does occur.

To calculate the probability of an event occurring, the approach relies on the underlying classifier's label predictions, $\hat{y}$, on documents in $D_{\bar{y}}$ to be reliable. The probability of a false negative prediction, given that the classifier has made a negative prediction, is then calculated as:

$$P(FN(d_i)|\hat{y}_i = neg) = 1 - \frac{e^{\sigma\rho_i}}{e^{\sigma\rho_i}+1} \tag{7.5}$$

where $\frac{e^{\sigma\rho_i}}{e^{\sigma\rho_i}+1}$ is a generalised logistic function that monotonically converts a classifier's prediction confidence score, $\rho$, in the range $(-\infty, +\infty)$ to real values in the range $[0.0, 1.0]$. The probability of a false positive occurring is computed analogously.

$G(e)$ is calculated on $D_{\bar{y}}$ and $G(FN) \neq G(FP)$. This inequality is reflected in the definitions of the gain functions:

$$G(FN) = \frac{1}{FN}\left(\frac{2(TP+FN)}{2(TP+FN)+FP} - \frac{2TP}{2TP+FP+FN}\right) \tag{7.6}$$

and

$$G(FP) = \frac{1}{FP}\left(\frac{2TP}{2TP+FN} - \frac{2TP}{2TP+FP+FN}\right). \tag{7.7}$$

To compute $G(FN)$ and $G(FP)$ the $TP, FP$ and $FN$ frequency counts are derived by performing a $k$-fold cross validation on $D_y$. The corresponding frequencies are then obtained by the maximum-likelihood estimation $\hat{\alpha}^{ML} = \alpha^{D_y} \cdot |D_y|/|D_{\bar{y}}|$, $\alpha \in \{TP, FP, FN\}$. To avoid zero counts when calculating the $\hat{\alpha}^{ML}$ values, Laplace smoothing is applied to each $\hat{\alpha}^{ML}$ in an *on-demand* fashion if any $\hat{\alpha}^{ML} < 1$, resulting in $\hat{\alpha}^{ML} + 1$. In the remainder of this thesis, we refer to this approach as *Utility*.

## 7.4 Incorporating Reviewer Feedback

The second task that our active learning strategy has to perform is to integrate feedback from the reviewer about the sensitivities in the documents that are reviewed. This task is performed by the Feedback Integration component of our framework (labelled as (2) "Model the sensitive information" in Figure 7.1). The active learning strategies presented in Section 7.3 use the predictions from the classifier, $\Omega$, as evidence of the classifier's confidence in correctly classifying the unlabelled documents, $D_{\bar{y}}$. The predictions from $\Omega$ are used to prioritise for review the documents that are expected to provide the most useful information for re-training $\Omega$. However, the only feedback from the reviewer that the approaches from Section 7.3 make use of is the documents' manually assigned class label, $y$.

We argue that using only the document-level reviewer feedback (i.e., class labels) is a sub-optimal approach for quickly learning to classify the sensitivities in a collection. Sensitive information is often only a small passage of text within a document. Moreover, sensitive information is often sensitive due to the context that the information appears in, e.g., it is *what* the information says (and often who provided the information) that makes the information sensitive, not just what that information is about. However, the sensitivities within a collection are often related or similar in some respect, e.g., discussions about related events in a geographical location at a particular point in time. It is likely that most discussions about the event or location etc. within the collection will not be sensitive. Therefore, we argue, that only using document-level reviewer feedback is not likely to capture the context-specific details of the sensitivities. Moreover, to quickly learn to classify the sensitivities within a collection, we need to make use of feedback from the sensitivity reviewer about the specific vocabulary that is used in the sensitive information within documents. With this in mind, we propose to have sensitivity reviewers *an-*

Amir Shekah (STRICTLY PROTECT), hereafter referred to as *the witness*, is an Afghan Muslim from the Kandahar province. He was born on the 14th June 1976 and claimed to be an unemployed civilian at the time of his arrest on 20th May 2000.

———————————-

The Witness's Account

———————————-

The witness was in his house in Kabul on 20th May 2000 when an armed man with his face covered entered and ordered everyone in to the street and to line up against the wall. They were all taken to house on the outskirts of the city.

Figure 7.2: An example of a reviewer's sensitivity annotations. The document contains three annotated sensitive passages, shown with a yellow background. Sensitivity annotations are analogous to redacting the sensitive text.

Table 7.1: Summary of the active learning strategies that we evaluate and how we denote them. We evaluate the four *document prioritisation* strategies from Section 7.3 as *Raw* active learning strategies. Moreover, we evaluate each of the document prioritisation strategies *Extended* with one of the three *sensitivity annotations* strategies from Section 7.4 (sixteen strategies in total).

| *Raw* | *Extended* with Sensitivity Annotations | | |
|---|---|---|---|
| | Simple | Information Gain | Annotation Pool |
| Entropy | Entropy+*Anno* | Entropy+*InfAnno* | Entropy+*AnnoPool* |
| Margin | Margin+*Anno* | Margin+*InfAnno* | Margin+*AnnoPool* |
| sConf | sConf+*Anno* | sConf+*InfAnno* | sConf+*AnnoPool* |
| Utility | Utility+*Anno* | Utility+*InfAnno* | Utility+*AnnoPool* |

*notate* the sensitive text within a document as they perform the review. Our proposed approach is analogous to having a sensitivity reviewer redact the sensitive text in a document as they review it. We argue that integrating term-level features of sensitivity into an active learning approach to sensitivity classification will result in the classifier, $\Omega$, making better sensitivity predictions. Moreover, this, in turn, should enable the active learning strategy to select more informative documents to be reviewed.

In this section, we present three strategies, inspired by Settles (2011), that we evaluate for integrating term-level sensitivity features into the active learning process. As shown in Figure 7.2, when a document, $d_i$, is judged to be sensitive, the reviewer annotates the sensitive text within the document, $a_{di}, |a_{di}| \in \{0..\ |d_i|\}$. The *sensitivity annotations* strategies presented in this section utilise these document annotations to *extend* the four document prioritisation strategies presented in Section 7.3 with informative term-level sensitivity features. Table 7.1 provides a summary of the combinations of document prioritisation and sensitivity annotations strategies that we evaluate for active learning sensitivity classification (in Section 7.7). In the remainder of this section, we provide details of the annotations features strategies that we evaluate.

Our first sensitivity annotations strategy assumes that all the terms that a reviewer annotates are equally useful for identifying sensitive information. To incorporate the additional information provided by the reviewer's annotations, we simply increase the importance, or weight, of each of the annotated terms by a constant value, $\alpha$, in the classifier, $\Omega$ (we provide specific details of this term weighting in the following section). We refer to this as *simple* sensitivity annotations, denoted as *+Anno* in the remainder of this thesis.

The remaining two sensitivity annotations strategies make use of the labelled collection of documents, $D_y$, and the classifier's predictions, $\hat{y}$, on the unlabelled documents in $D_{\bar{y}}$ to calculate the expected information gain:

$$IG(t_k) = \sum_{F_k} \sum_{i} P(F_k, c_i) log \frac{P(F_k, c_i)}{P(F_k)P(c_i)} \tag{7.8}$$

for each of the term features in the unlabelled collection $D_{\bar{y}}$, where $F_k \in \{0,1\}$ indicates the presence or absence of a term feature $t_k$ in the class $c_i, c_i = y_i \cup \hat{y}_i$.

The first information gain sensitivity annotations strategy that we present considers all the term features that are in the intersection of the terms identified by $IG(t_k)$ and the terms annotated by a reviewer, in the current batch of documents being reviewed, as good sensitivity features and increases the weight of the feature in $\Omega$, by $\alpha$. We refer to this strategy as *information gain* sensitivity annotations, denoted as *+InfAnno* in the remainder of this thesis.

The final sensitivity annotations strategy that we evaluate, *annotation pool*, identifies useful sensitivity features through the same process as the previous information gain strategy, except that instead of only considering annotation terms from the current batch of documents being reviewed, a pool of potential sensitivity features is built from all previous annotations and any terms that are in the intersection of the terms identified by $IG(t_k)$ and terms in the annotation pool are considered as being good sensitivity features. We denote the annotation pool strategy as *+AnnoPool* in the remainder of this thesis.

## 7.5 Selecting an Appropriate Classifier

For the experiments that we present in this chapter, as our classifier, $\Omega$, we deploy the *Multinomial* Naïve Bayes classifier (Duda & Hart, 1973) (MNB) that we previously presented in Chapter 2. The MNB classifier is modelled on the frequency of occurrences of terms from a vocabulary, $V$, in a document, $d$, and can be viewed as a unigram language model (McCallum *et al.*, 1998). We choose to deploy MNB due to three properties of the algorithm. Firstly, MNB has been shown to work well for text classification tasks (Lewis & Gale, 1994; McCallumzy & Nigamy, 1998; Nigam *et al.*, 1998; Rennie *et al.*, 2003). Secondly, MNB is very quick to train, compared to other classifiers such as SVM. This is a very important property for learning a sensitivity classifier to assist sensitivity review. When developing a sensitivity classifier we

want to be able to re-train $\Omega$ at each iteration of the review cycle, so that we can quickly learn to classify newly discovered sensitivities and provide updated sensitivity predictions to reviewers. Thirdly, MNB can be easily adapted to integrate different sources of feature evidence by simply weighting the underlying feature's multinomial (McCallumzy & Nigamy, 1998; Settles, 2011). For annotation term features that are identified as being important for classifying sensitivity, following Settles (2011), we simply increase the probability, $P(f_k|c_i)$, of the term appearing in the sensitive class by increasing the prior for the corresponding multinomial in $\Omega$, by a constant value $\alpha = 50$.

We note that the choice of MNB as our classifier in this chapter differs to the SVM that we selected to deploy in Chapters 5 and 6. In addition to the reasons for selecting MNB that we listed above, we make this choice since the experiments in this chapter are to evaluate the active learning strategies and validate whether the annotations features strategy, i.e., having a reviewer annotate or redact the sensitive text in a sensitive document, enable us to reduce the number of documents that are required to be reviewed to learn an effective sensitivity classifier. Moreover, if so, which of the strategies for identifying the informative annotated terms is the most effective strategy. Therefore, identifying the *best* underlying classifier is not the main objective, or research question, of this chapter. This research question could be a valuable study as future work.

## 7.6 Experimental Methodology

The three research questions that we wish to answer in this section are as follows:

- RQ7.1 Which document prioritisation active learning strategy is most effective for selecting documents to be reviewed to quickly learn an effective sensitivity classifier?
- RQ7.2 Is extending the document prioritisation strategies with our proposed sensitivity annotations strategies effective for reducing the number of documents that have to be reviewed to learn an effective sensitivity classifier?
- RQ7.3 Which combination of document prioritisation and sensitivity annotations strategies is the most effective approach for reducing the number of documents that have to be reviewed to learn an effective sensitivity classifier?

To answer research questions RQ7.1 and RQ7.2, we use the test collection of 3801 government documents that we presented in Chapter 3 (Table 3.3) to *simulate* iterations of the review cycle. The collection was assessed for two UK FOI exemptions, namely international relations and personal information. All documents that the sensitivity reviewers judged as containing any exempt information are labelled *sensitive*. The remaining documents are labelled *non-sensitive*, resulting in 502 sensitive documents (~13%) and 3299 non-sensitive (~87%).

To ensure the generalisability of our findings, we run our experiments 25 times over different samples of the collection $D$. For each run, we sample 2500 documents from $D$ as a training set

$D_{tr}$, which we use for the active learning simulation i.e., $|D_{\bar{y}}| + |D_y| = D_{tr} = 2500$. Additionally, for each run we sample a separate 500 documents from $D$ as a held out test set, $D_{te}$, for evaluating the performance of the classifier, $\Omega$. We retain the distributions of sensitive and non-sensitive documents from $D$ when generating $D_{tr}$ and $D_{te}$, resulting in $D_{tr} = \{2150 \text{ non-sensitive}, 325 \text{ sensitive}\}$ and $D_{te} = \{435 \text{ non-sensitive}, 65 \text{ sensitive}\}$. We perform a binary classification, *sensitive* vs. *non-sensitive* and report mean scores over the 25 runs. To test for statistical significance when evaluating reviewer effort, following Cormack & Grossman (2014), we use a sign test with $p < 0.01$.

For each iteration of the active learning simulation, we present the reviewer a new batch of $k$ documents. For our experiments, we set $k = 20$. We evaluate the impact of varying $k$ in Section 7.7.3. To counteract the potential learning effect due to the class-imbalance in our collection (i.e., the classifier over-predicting the majority class), when integrating newly labelled documents to $D_y$, we introduce the following constraint: $|\text{non-sensitive}| \in D_y \leq (k/2) + |\text{sensitive}| \in D_y$. We discard newly reviewed non-sensitive documents that violate this constraint. In practice, this means that we randomly down-sample the classifier's training data to loosely match the class frequencies. In preliminary experiments this led to uniform improvements across all tested approaches of ~+0.4 Balanced Accuracy, after all documents had been reviewed.

For the Utility-Theoretic approach, presented in Section 7.3.2, we use the JaTeCS implementation (Esuli *et al.*, 2017). When estimating $G(FN)$ and $G(FP)$, following Berardi *et al.* (2012), we select $F_2$ as our metric and perform a $k$-fold cross validation, setting $k = 10$. As previously mentioned in Section 7.5, for the sensitivity annotations approaches, presented in Section 7.4, when integrating feature importance to the classifier we set $\alpha = 50$, following Settles (2011).

## 7.7 Results

To answer RQ7.1 and RQ7.2, firstly, in Section 7.7.1, we evaluate the performance of the four document prioritisation strategies that we presented in Section 7.3, namely *Entropy*, *Margin*, *sConf* and *Utility* (we refer to these strategies collectively as the *raw* document prioritisation strategies). Next, in Section 7.7.2, we evaluate the performance the raw document prioritisation strategies *extended* with each of the three annotations features strategies that we presented in Section 7.4, namely *+Anno*, *+InfAnno* and *+AnnoPool*. We compare the performance of the best performing extended strategy with the raw document prioritisation strategies. In practice, we view the raw document prioritisation strategies as baseline approaches and evaluate if extending these strategies with the sensitivity annotations strategies results in a significant reduction in the number of documents that have to be reviewed to learn an effective sensitivity classifier (according to a sign test with $p < 0.01$).

We present the results of our experiments over three figures. Firstly, Figure 7.3 presents the raw document prioritisation strategies. Secondly, Figure 7.4 presents the document prioritisa-

tion strategies extended with each of the sensitivity annotations strategies. Thirdly, we make the comparison between the raw document prioritisation strategies and extended with the best performing sensitivity annotations strategy in Figure 7.5. The plots in Figures 7.3, 7.4 and 7.5 present the Precision, Recall, $F_1$, $F_2$ and BAC scores achieved on the held out set $D_{te}$ (on the $y$ axis). The $x$ axis shows the required reviewer effort, in terms of the number of documents reviewed. We select $F_2$ and BAC as our main metrics when evaluating the effectiveness of learned sensitivity classifier. When viewing the plots in Figures 7.3, 7.4 and 7.5, the best performing approaches are those that result in data points that are closest to the upper left hand corner of the plot, i.e., we aim to learn an effective classifier using the least amount of reviewing effort possible.

### 7.7.1 Selecting Informative Documents

Figure 7.3 presents the performances of the raw document prioritisation strategies Entropy, Margin, Utility and sConf. As can be seen from Figure 7.3(a), there is a large variance in the precision scores achieved by each of the approaches when $\leq 250$ documents have been reviewed. However, when $\geq 800$ documents have been reviewed each of the approaches shows a more consistent performance, achieving precision scores between ~0.28 and ~0.38. There are not any notable statistically significant improvements in precision scores between the approaches when reviewing effort is $\geq 800$ documents reviewed (according to a sign test with $p < 0.01$).

Figure 7.3(b) presents the approaches' recall scores. We note, from Figure 7.3(b), that the Margin and Utility approaches begin to identify sensitivity noticeably quicker than the Entropy and sConf approaches. The Margin and Utility approaches achieve 0.3 recall when only 240 and 380 documents, respectively, are reviewed. However, to achieve the same recall, sConf requires 980 documents to be reviewed and Entropy requires 1240 documents to be reviewed. Margin and Utility both achieve 0.3 recall using statistically significantly less reviewing effort than either sConf or Entropy (sign test, $p < 0.01$). Moreover, Margin and Utility require significantly less reviewing effort to achieve a comparable recall score to sConf or Entropy almost throughout the experiment (Utility shows a notable drop in recall when $\geq 2250$ documents are reviewed.).

The higher recall that is achieved by Margin and Utility has a clear impact on the overall performance of the approaches. From Figure 7.3(e), we can see that Margin and Utility achieve BAC scores of 0.58 when only 250 documents have been reviewed, while sConf achieves 0.525 BAC and Entropy results in a random classifier (0.5 BAC). Moreover, from Figure 7.3(d), we can see that Margin and Utility consistently perform better than Entropy and sConf in terms of $F_2$. Margin and Utility require significantly less reviewing effort to achieve 0.4 $F_2$ (980 documents reviewed) compared to sConf (1480 documents) or Entropy (1760 documents), according to a sign test, $p < 0.01$. However, we note that Margin consistently performs better than Utility in terms of $F_1$, $F_2$ and BAC when the number of documents reviewed $\geq 1000$. Therefore, in response to RQ7.1, we conclude that the Margin document prioritisation active learning strategy

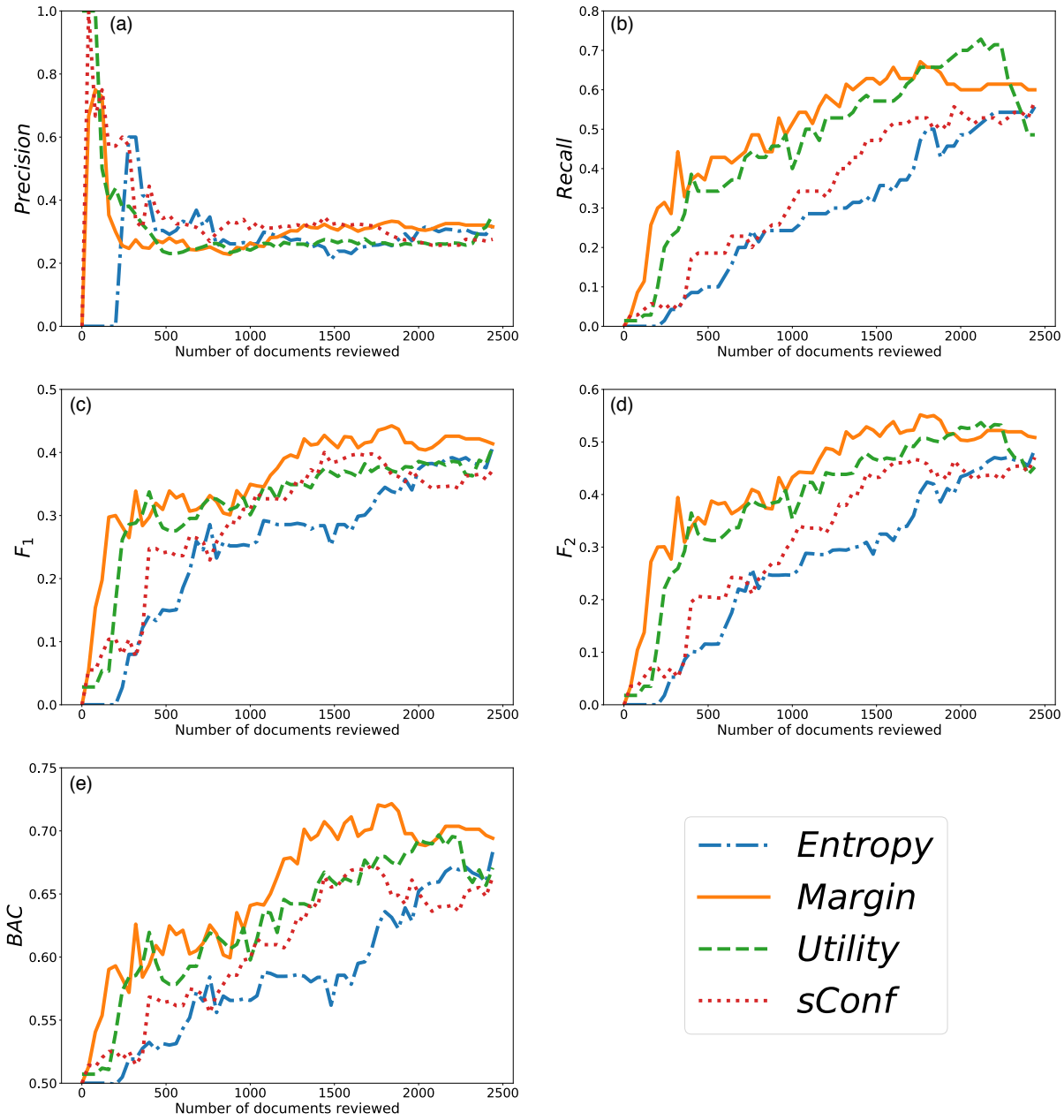



Figure 7.3: Results for the document selection active learning strategies, *Entropy*, *Margin*, *Utility* and the confidence that a document is sensitive *sConf*. The figure shows the Precision, Recall, $F_1$, $F_2$ and Balanced Accuracy (BAC) scores plotted against the number of documents reviewed.

is the most effective document prioritisation strategy for quickly learning an effective sensitivity classifier on our collection.

### 7.7.2 Integrating Reviewer Feedback

In this section, we firstly evaluate each of the approaches for integrating sensitivity annotation features, +*Anno*, +*InfAnno* and +*AnnoPool*, extending each of the raw document prioritisation strategies, before comparing the best performing sensitivity annotations approach with the raw document prioritisation strategies. Figure 7.4 presents the results of each of the sensitivity annotation strategies extending document prioritisation. We are interested in which of the sensitivity annotation strategies, +*Anno*, +*InfAnno* or +*AnnoPool*, requires the least reviewing effort to train an effective sensitivity classifier. Firstly, we note from Figure 7.4 that when $\leq 1000$ documents have been sensitivity reviewed margin consistently performs best in terms of recall, $F_1$, $F_2$ and BAC when the approach is extended with either of the +*Anno* or +*InfAnno* sensitivity annotations strategies. Moreover, from the approaches presented in Figure 7.4, we note that when $\leq 1000$ documents have been reviewed, all of the approaches perform least well in terms of recall, $F_1$, $F_2$ and BAC when the +*AnnoPool* strategy is applied. Therefore, we shall focus our comparison on the performance of +*Anno* and +*InfAnno*.

In comparing the performances of +*Anno* and +*InfAnno*, we note from Figure 7.4 that +*Anno* results in the best recall, $F_1$, $F_2$ and BAC in the initial iterations of the review cycle (i.e., when $\leq 100$ documents are reviewed). Again, the most notable improvements are observed for the Margin document prioritisation strategy. After the first 100 documents are reviewed, Margin+*Anno* achieves 0.6 BAC compared with 0.56 BAC for Margin+*InfAnno*. However, the amount of reviewing effort required to achieve 0.6 BAC with Margin+*Anno* is not statistically significantly less than that required by Margin+*InfAnno* (sign test, $p < 0.01$).

When the number of documents reviewed is $\geq 100$ the +*InfAnno* strategy begins to perform notably better than +*Anno*. Margin+*InfAnno* achieves 0.5 $F_2$ when only 500 documents have been reviewed. This is significantly less reviewing effort than the 1900 documents that are required to be reviewed for Margin+*Anno* to achieve 0.5 $F_2$ (sign test, $p < 0.01$). Moreover, in terms of BAC, Margin+*InfAnno* achieves 0.7 BAC when only 820 documents have been reviewed. This is significantly less reviewing effort than is required for Margin+*Anno* to learn the combination's most effective classifier (0.69 BAC, 1820 documents), (sign test, $p < 0.01$). Therefore, we select +*InfAnno* as the best performing sensitivity annotations strategy to evaluate the effectiveness of extending the raw document prioritisation strategies with sensitivity annotations in the remainder of this section.

Turning our attention to the effectiveness of extending the raw document prioritisation strategies with sensitivity annotations, Figure 7.5 presents the best performing sensitivity annotations strategy (+*InfAnno*) compared with the document prioritisation without additional sensitivity annotation features (Raw (No Anno)). We can see from Figure 7.5 that the addition of the +*InfAnno*

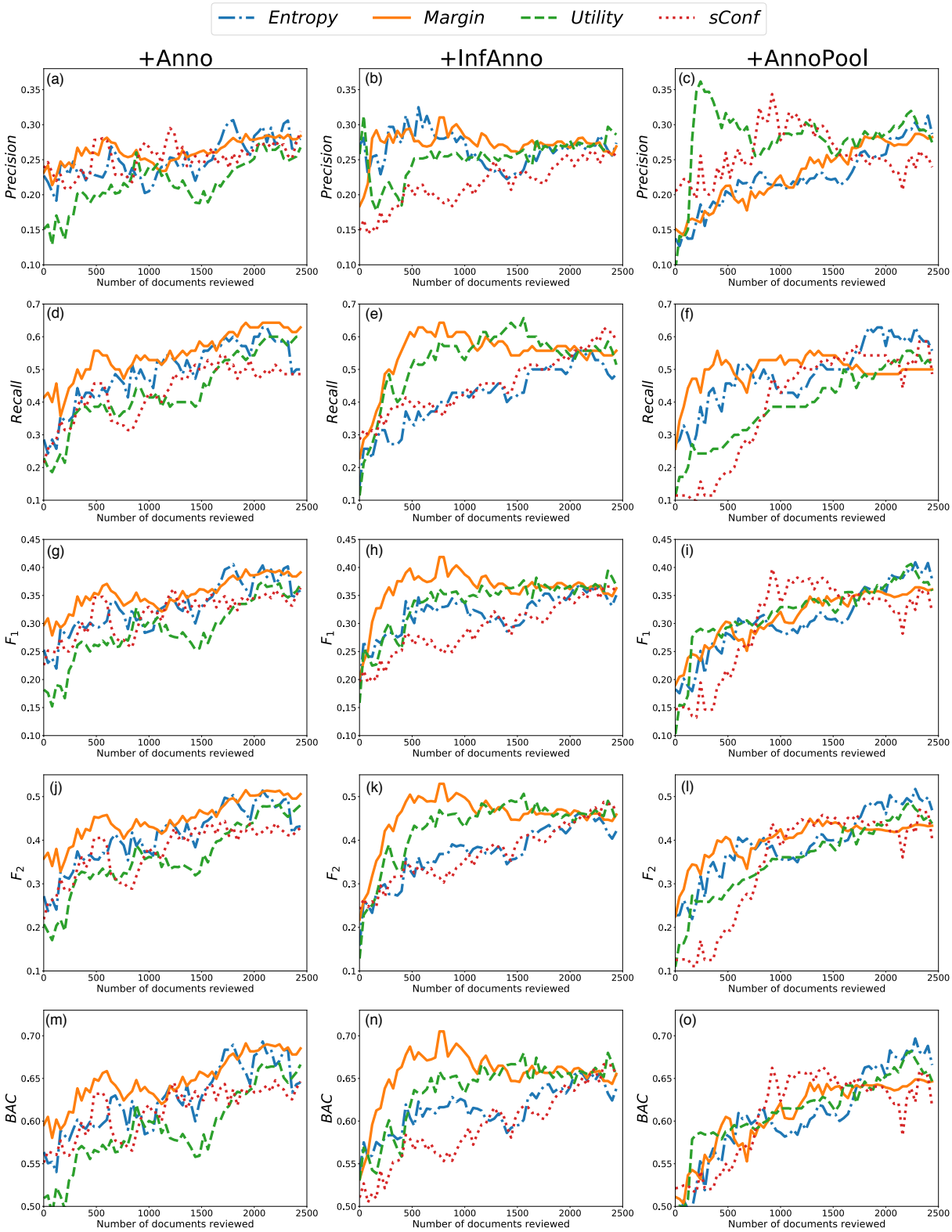



Figure 7.4: Results of the document selection strategies for active learning, *Entropy*, *Margin*, *Utility* and *sConf* for each of the methods for incorporating reviewer feedback *+Anno*, *+InfAnno* and *+AnnoPool*. The figure shows the Precision, Recall, $F_1$, $F_2$ and Balanced Accuracy (BAC) scores plotted against the number of documents reviewed.

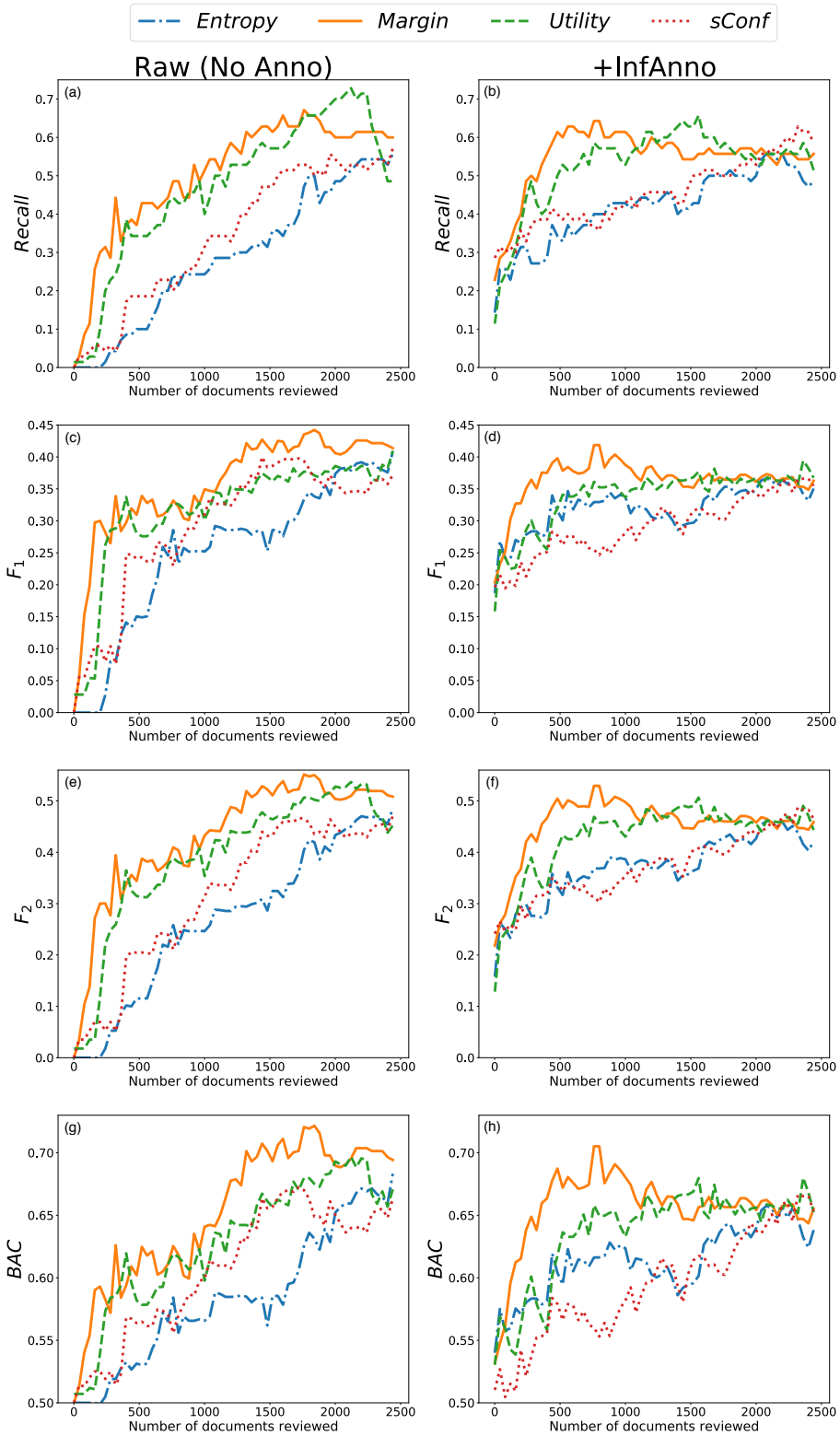



Figure 7.5: Results of the document selection strategies for active learning, *Entropy*, *Margin*, *Utility* and *sConf* without additional sensitivity annotations features, *Raw (No Anno)*, and extended with *+InfAnno* sensitivity annotation features. The figure shows the Recall, $F_1$, $F_2$ and Balanced Accuracy (BAC) scores plotted against the number of documents reviewed.

sensitivity annotation strategy enables all of the document prioritisation approaches to correctly classify sensitive documents using markedly less reviewing effort. Comparing the recall scores from Figures 7.5(a) and 7.5(b), we can see that when 500 documents are reviewed: raw Entropy achieves 0.1 recall, while Entropy+*IntAnno* achieves 0.31 recall; raw sConf achieves 0.18 recall, while sConf+*IntAnno* achieves 0.38 recall; raw Utility achieves 0.33 recall, while Utility+*IntAnno* achieves 0.5 recall; and raw Margin achieves 0.41 recall, while Margin+*IntAnno* achieves 0.61 recall.

In terms of $F_1$, (Figure 7.5(c) vs. Figure 7.5(d)) and $F_2$ (Figure 7.5(e) vs. Figure 7.5(f)), we can see that the addition of sensitivity annotation strategy enables us to learn a more effective sensitivity classifier using notably less reviewing effort. The +*InfAnno* sensitivity annotations results in Margin achieving 0.5 $F_2$ when only 500 documents have been reviewed. This is significantly less than the 1260 documents that are required to be reviewed for raw margin to achieve 0.5 $F_2$ (sign test, $p < 0.01$). Finally, comparing Figures 7.5(g) and (h), we can see that Margin+*InfAnno* sustains initial gains in classification effectiveness and reaches a peak classification performance (0.7 BAC) when significantly less document have been reviewed than when the Margin strategy is deployed without sensitivity annotations features (according to the sign test, $p < 0.01$). Margin+*InfAnno* requires only 820 documents to be reviewed to achieve 0.7 BAC as opposed to 1700 documents when Margin is deployed without sensitivity annotations features. This is a 51% reduction in amount of reviewer effort, in terms of the number of documents reviewed, that is required to learn an effective sensitivity classifier.

In response to RQ7.2, we conclude that extending document prioritisation active learning strategies with our proposed sensitivity annotations strategies is indeed effective for reducing the number of documents that have to be reviewed to learn an effective sensitivity classifier. Moreover, we found that +*InfAnno* is the most effective sensitivity annotations strategy. In response to RQ7.3, we conclude that a combination of the Margin document prioritisation active learning strategy and the +*InfAnno* sensitivity annotations strategy (Margin+*InfAnno*) is the most effective approach for reducing the number of documents that have to be reviewed to learn an effective sensitivity classifier, on our document collection.

### 7.7.3 The Effect of the Batch Size on Learning

In an active learning scenario, the number of documents that are reviewed and labelled in each iteration, i.e., the batch size $k$, can have an impact on the amount of reviewing effort that is required to learn an effective model. For example, selecting smaller values of $k$ can result in more efficient learning but selecting a larger value of $k$ can result in a more stable classifier, i.e., more predictable levels of improvement at each iteration (Brinker, 2003; Schohn & Cohn, 2000). In our experiments that we have presented in this chapter, we set $k = 20$. In this section, we provide a brief analysis of the effect of varying $k$ on the resulting learned sensitivity classifier. Figure 7.6 presents the results of the Margin+*InfAnno* active learning strategy for batch sizes $k \in$

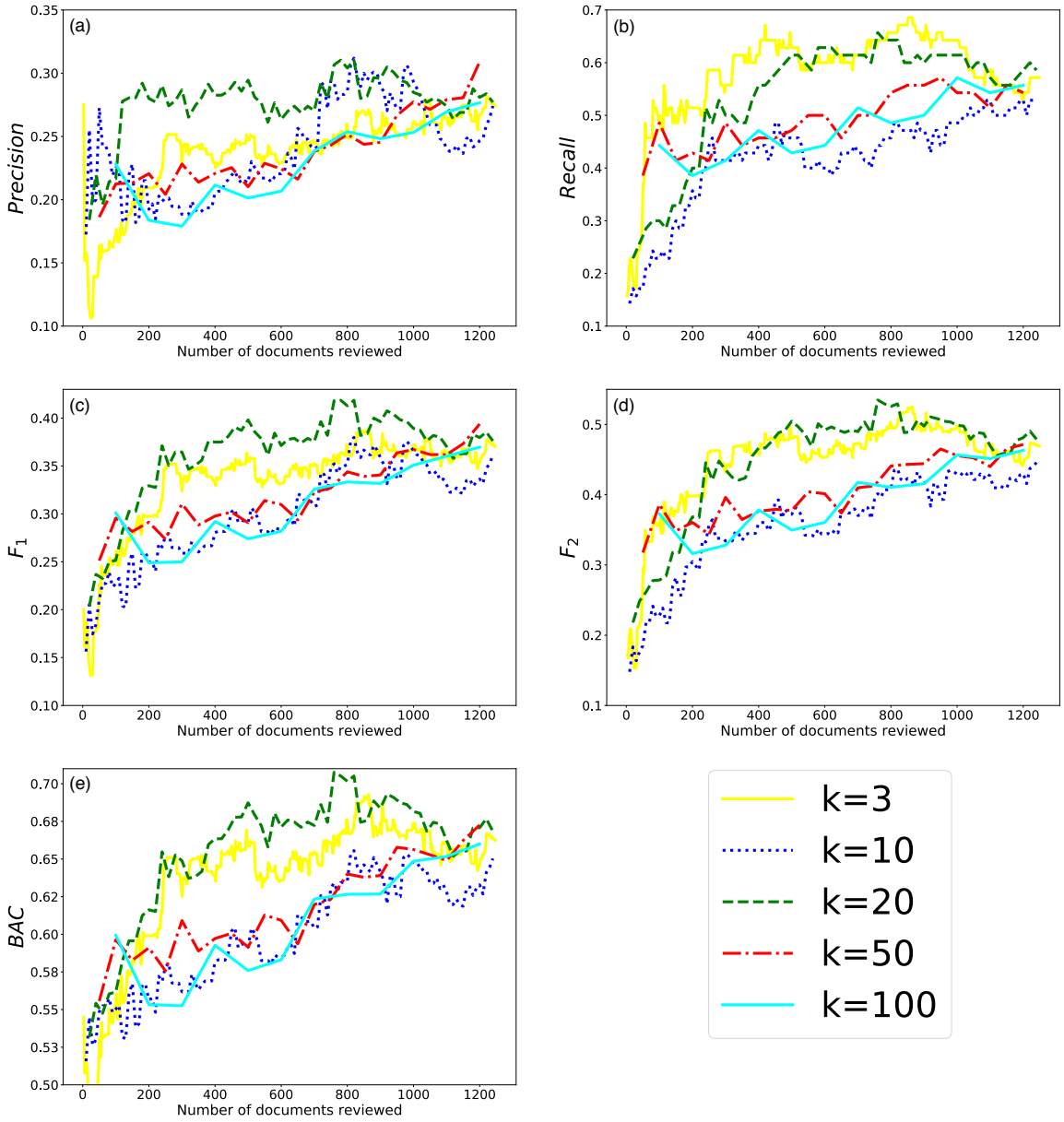



Figure 7.6: The effect of varying the batch size, $k$, of documents that are sensitivity reviewed at each iteration of the review cycle, for Margin extended with +*InfAnno* the sensitivity annotations strategy. The plot shows the Precision, Recall, $F_1$, $F_2$ and Balanced Accuracy (BAC) scores plotted against the number of documents reviewed.

$\{3, 10, 20, 50, 100\}$. As can be seen from Figure 7.6, when $k$ is large ($k = 50$ or $k = 100$) we do indeed observe a more gradual and steady increase in performance in terms of $F_1$ (Figure 7.6(c)), $F_2$ (Figure 7.6(d)) and BAC (Figure 7.6(e)). Moreover, it can actually be beneficial to set $k$ to a larger value at the beginning of the review process to generate a pool of candidate documents to train the initial classifier. This approach has previously been deployed for technology-assisted review tasks, such as e-discovery (Cormack & Grossman, 2014). In our experiments, when 100 documents have been reviewed the classifiers that are learned when $k$ is set at either $k = 100$ or $k = 50$ achieve the highest $F_1$ (Figure 7.6(c)) and BAC (Figure 7.6(e)) scores. Additionally, in terms of $F_2$, only $k = 1$ results in a comparably effective classifier. However, when the number of documents reviewed $\geq 100$, smaller values of $k$ can be more beneficial. As can be seen from Figure 7.6 the classifiers that are learned when $k = 1$ and $k = 20$ are comparable in terms of $F_2$ (Figure 7.6(d)). However, there is a notable increase in classification effectiveness in terms of BAC (Figure 7.6(e)) when $k = 20$ and, on our collection, $k = 20$ results in the most effective active learning approach for reducing the amount of reviewing effort that is required to learn an effective classifier.

## 7.8 Conclusions

In this chapter, we argued that the vocabulary that is associated with sensitive information is likely to vary between different collections. Some of the terms that are used often in sensitive information in one collection might be used often but not associated with sensitive information in another collection. Therefore, a sensitivity classifier must be able to quickly identify what the sensitivities in a specific collection *look like* to be able to assist reviewers, for example by providing the reviewers with useful information about which documents in the collection contain sensitive information. We proposed to address the problem of quickly learning to identify, i.e., classify, the sensitive information in a specific collection as an active learning task. However, most active learning approaches focus only on trying to select the most informative documents to have reviewed and integrate the reviewer's feedback at the document-level, i.e., they only integrate the class labels for the newly reviewed documents. We argued that this is not an optimal strategy for sensitivity classification since, it is often only a small portion of a document that is sensitive and document-level labels can not identify the specific vocabulary that is sensitive when only a small number of documents have been reviewed. To address this, we proposed to have a sensitivity reviewer annotate, or redact, the sensitive text in a document that they judge to be sensitive, while they are reviewing the document. Moreover, we proposed to identify the most informative terms in the reviewer's annotations to construct a representation of what the sensitivities in a collection look like from the reviewer's feedback. Moreover, we argued that extending active learning (document prioritisation) strategies with the informative annotation features would result in fewer documents be required to be reviewed to learn an effective sensi-

tivity classifier.

In particular, in this chapter, we evaluated four active learning strategies for prioritising specific documents to have reviewed, so that the documents that are the most *informative* for the sensitivity classifier are reviewed before the less informative documents (see Section 7.3). We showed that, on our collection, the margin active learning strategy is the most effective document prioritisation strategy, when the strategies are not extended with sensitivity annotations features (see Figure 7.3). Moreover, we evaluated three strategies for integrating a reviewer's sensitivity annotations into our framework for technology-assisted sensitivity review to construct a representation of the sensitivities in a collection (see Section 7.4). We showed that our proposed sensitivity annotations strategy further reduced the amount of reviewing effort required to develop an effective sensitivity classifier, in terms of the number of document that are reviewed. Deploying the margin active learning document prioritisation strategy and extending it with high Information Gain (+*InfAnno*) sensitivity annotation (term) features, enables our framework to learn an effective sensitivity classifier (0.7 BAC) using significantly less reviewing effort (according to a sign test, $p < 0.01$) (see Figure 7.5).

In the following chapter, we investigate how our proposed framework, and sensitivity classification, can assist the sensitivity review process in the first of the two realistic digital sensitivity review scenarios that we identified in Chapter 1, namely the *limited review* user model.

# Chapter 8

# Maximising Openness in the Limited Review User Model

## 8.1 Introduction

In the previous chapter, we proposed to reduce the amount of reviewing effort that is required to learn an effective sensitivity classifier by integrating explicit reviewer feedback about the terms in a collection's vocabulary that are most associated to sensitive information. In particular, we proposed to have a sensitivity reviewer annotate any sensitive information within a document as the reviewer sensitivity reviews a document. Moreover, we proposed to integrate the most informative terms into the sensitivity classification model through an iterative active learning process. This is analogous to integrating the redaction process into the sensitivity review process and using the redacted information to construct a representation of the sensitivities within a collection that can provide the classifier with additional evidence about what the sensitivities *look like*. We empirically showed, in Section 7.7.2, that integrating explicit reviewer feedback from sensitive information annotations can significantly reduce the amount of reviewing effort that is required to learn an effective sensitivity classifier (see Figure 7.5).

Sensitivity classification provides our framework for technology-assisted review with a mechanism to assist sensitivity reviewers, and government departments, to perform the digital sensitivity review task in a number of scenarios. For example, sensitivity classification can provide reviewers with assistance by showing reviewers which documents in a collection are most likely to contain sensitive information. Sensitivity classification can also be used within our framework to assist government departments in making strategical decisions when they are planning which documents to review.

In this chapter, we investigate how our proposed framework can assist the sensitivity review process in the first of two realistic digital sensitivity review user models that we investigate in this thesis, namely the *limited review* user model. The limited review user model addresses a scenario in which there are not enough reviewing resources available to review all of the

documents in a collection that is to be transferred to the archive. This user model is motivated by the expectation in government that government departments will not be able to recruit enough sensitivity reviewing resources to review all of the digital documents that are to be archived (The National Archives, 2016*a*).

The Public Records Act 1958 (c. 51) imposes a strict time-to-transfer obligation that, in effect, states which documents are due to be publicly archived each year. The time-to-transfer obligation is currently transitioning from transferring 30 years after a document's creation to 20 years after creation (Constitutional Reform and Governance Act 2010, c. 25). This transition means that there are twice as many documents to be reviewed each year. Government departments each have a different number of documents to review and varying amounts of sensitive information in their documents. Moreover, the amount of available reviewing resources varies between departments and some government departments are not currently meeting their time-to-transfer obligations (Allan, 2014). Furthermore, government departments are not expected to be able to recruit enough reviewing resources to review all of the digital documents that are to be transferred (The National Archives, 2016*a*). Therefore, government departments that do not have enough reviewing resources will need to make strategic decisions to allocate reviewing resources effectively.

This thesis argues that sensitivity classification can be deployed along with other techniques, such as predicting the amount of reviewing time that is likely to be required to review a document, to increase the number of documents that can be reviewed and released to the public when there are not enough reviewing resources available. We argue that the productivity of the available reviewing resources can be increased by focusing the reviewers' effort on reviewing the documents that are most likely to be released, i.e., documents that are not sensitive. Additionally, by prioritising for review the documents that are expected to take less time to review, more documents can be reviewed and released to the public while using the same amount of reviewing resources. This will, in turn, mean that government departments should be able to meet the Public Records Act 1958 (c. 51) time-to-transfer obligations for a larger percentage of the documents that are awaiting review.

In this chapter, we propose an approach for prioritising specific documents to be reviewed so that the total number of documents that are reviewed and released to the public with the available reviewing resources is increased. We refer to this approach as *Maximising Openness*. Our proposed approach models aspects of the sensitivity reviewers' reviewing process, such as whether the document is predicted to be judged as sensitive or not, and the collection that is to be reviewed, to prioritise for review the documents that are predicted to be most likely to be released and reviewed quickly. We conduct a user study to analyse reviewers' behaviour, such as the time taken to review documents, and use the log data from the user study to develop and evaluate our proposed approach for prioritising documents for review. Moreover, we evaluate how the distribution of sensitive information within the collection that is being reviewed, and

the effectiveness of the deployed sensitivity classifier, affect the effectiveness of our proposed document prioritisation approach. The remainder of this chapter is structured as follows:

- In Section 8.2, we formally define the *limited review* user model and the effectiveness measures that we use to evaluate the *openness* of our proposed document prioritisation approach.
- In Section 8.3, we present details of a sensitivity review user study that we perform to observe the reviewing behaviour of sensitivity reviewers, e.g., the time taken to review documents, and construct a test collection for developing and evaluating our proposed approach. We present details of the study design, the documents that were used and the participants that took part in the study in Section 8.3.1. In Section 8.3.2, we present details of the test collection that we generate from the user study.
- In Section 8.4, we present our proposed approach for predicting the amount of time that a reviewer is likely to need to review a document. We present the features that we use for our reviewing times predictions in Section 8.4.1 and our experimental methodology in Section 8.4.2. We present the results of our reviewing time experiments in Section 8.4.3.
- In Section 8.5, we present our proposed approach for prioritising documents for review to maximise the number of documents that can be opened to the public with the available reviewing resources. Our proposed approach prioritises for review the documents in a collection that are predicted to require less time to review. Thereby, prioritising non-sensitive documents. We present the experimental methodology that we use to evaluate our approach in Section 8.5.1 and the results of our experiments in Section 8.5.2. In particular, in Section 8.5.2, we evaluate how the distribution of sensitive information in a collection affects our proposed approach. Moreover, since our proposed document prioritisation approach relies on sensitivity classification to predict a document's reviewing time, we evaluate whether our baseline sensitivity classifier and our enhanced sensitivity classifier that we proposed in Chapters 5 and 6 respectfully, are sufficiently effective to be deployed for prioritising documents for review.
- Finally, in Section 8.6, we summarise our conclusions from this chapter.

## 8.2 Limited Review User Model

In the limited review user model, we assume that there is a collection of documents, $D$, that would require a total of $T_D$ hours to sensitivity review. Moreover, there is an available reviewing time budget, $T_b$, of $b$ hours that defines the amount of reviewing resources that are available. Furthermore, we assume that $T_b < T_D$. Therefore, to assist sensitivity review in the limited review

user model, the task is to generate a ranking of documents, $r_i = d_1 .. d_{|D|}$, such that if a sensitivity reviewer starts by reviewing $d_1$ and continues reviewing each or the documents in $r$ sequentially, i.e., in the order that they are ranked, until the reviewing time budget $T_b$ has expired, then a predetermined measure of sensitivity reviewing effectiveness is increased.

We argue that it is reasonable to measure the effectiveness of a document prioritisation approach for the limited review user model by the number of documents that are opened to the public with the available reviewing time budget. As we previously discussed in Section 8.1, given the time-to-transfer obligation imposed by the Public Records Act 1958 (c. 51), the backlog of documents that are awaiting review (Allan, 2014) and the expectation that government departments will not be able to recruit enough reviewing resources to review all of the documents that are due for transfer (The National Archives, 2016*a*), increasing the number of documents that are released will enable government departments to meet the time-to-transfer obligations for a larger percentage of the documents that are awaiting review. Therefore, we define the following two metrics to evaluate this task:

Mean Hourly Openness (Absolute Openness) for a ranking strategy $r_i$ is the average number of documents that are sensitivity reviewed and released to the public in 1 hour of reviewing time:

$$O_A(r_i) = \frac{o}{T_b} \tag{8.1}$$

where $o$ is the total number of documents that are reviewed and *released* to the public (i.e., documents that are not sensitive) and $T_b$ is the total number of reviewing hours.

Mean Hourly Openness Ratio (Openness Ratio) is the ratio of reviewed documents that are actually released to the public (i.e., documents that are not sensitive):

$$O_R(r_i) = \frac{o}{SR_b} \tag{8.2}$$

where $o$ is the total number of documents that are reviewed and *released* to the public (i.e., documents that are not sensitive) and $SR_b$ is the total number of documents that are sensitivity reviewed within the reviewing time budget $T_b$.

## 8.3 Reviewing Times User Study

To investigate how sensitivity classification predictions can be used to improve the effective allocation of sensitivity reviewing resources we, firstly, conducted a sensitivity review user study. The aim of this study is three-fold:

1. To gather evidence about how sensitivity reviewers perform the reviewing task, such as the length of time that a reviewer takes to review a document and whether reviewers revisit their decisions for previously judged documents.

2. To provide insights about features of the reviewing process and the reviewers' behaviour that can be useful for predicting the time that is required to review a specific document and develop an approach for increasing the number of documents that can be opened with the available reviewing resources.

3. To construct a test collection for developing and evaluating our proposed approach for improving the allocation of sensitivity reviewing resources by prioritising specific documents to be reviewed.

We provide details of the design of the study and the study participants in Section 8.3.1 before, in Section 8.3.2, presenting details of the test collection that is constructed from the user study.

### 8.3.1 Study Design and Participants

We recruited 16 volunteers from the official UK government archive (The National Archives[1]) to sensitivity review a random sample of documents from the collection of 251,287 formal government communications that we previously introduced in Chapter 3[2]. The sensitivity review task adhered to the same structure and objectives as the reviewing task that we presented in Chapter 3. Reviewers were provided access to the same web based reviewing interface that we presented in Figure 3.3. The interface enables reviewers to record a document level classification stating if a document is not sensitive, or contains Sections 27 international relations sensitivities, or contains Sections 40 personal information sensitivities, or contains both Section 27 and Section 40 sensitivities. As per the reviewing procedure that we previously presented in Chapter 3, the reviewers were also asked to annotate any sensitive passages of text in the documents that they judged to be sensitive.

The reviewers were familiar with sensitivity review. However, they were provided the same detailed guidance, regarding (1) the scope of the task that they were being asked to perform and (2) the reviewing interface, that was provided to the reviewers for generating the test collection presented in Chapter 3. Moreover, the reviewers attended a half-day workshop prior to the start of the study where they received a presentation about the task and reviewed a batch of practice documents to familiarise themselves with the interface and have an opportunity to raise any questions that they had. In line with current sensitivity review practices, the reviewers were allowed to perform the reviewing task at times suitable to themselves over a period of two months. Since the reviewers in the study were volunteers with no obligation or financial incentive to complete the task, they were initially assigned twenty documents to review. Whenever a reviewer had reviewed half of their assigned documents they were assigned a further batch of twenty documents to review. Since we were not able to estimate the number of documents that

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

[2] The documents are sampled from the same collection as is used to generate the test collection in Chapter 3. However, the samples of documents that are reviewed in this chapter and in Chapter 3 are disjoint sets.

Table 8.1: The generated reviewing times test collection. Document length is measured by number of words. The average reviewing time is measured in seconds.

| | docs | %sensitive | Avg. Length | Avg. Review Time |
|---|---|---|---|---|
| Training Data | 184 | 9.63 | 824.6 | 321.05 |
| Test Data | 181 | 17.4 | 710.3 | 385.77 |

we would be able to have reviewed in this study, and in the interests of obtaining as many example reviews as possible, each document was only reviewed by a single reviewer.

To ascertain the duration taken to review a document, we logged the time when a document was loaded to view, $t_0$, and when a judgement was saved, $t_1$. The reviewing time, $t_r$, for a document, $d$, is then calculated as $t_r(d) = t_1 - t_0$. Previously judged documents could also be revisited. For revisited documents, we calculate reviewing time as:

$$t_r(d) = \sum_{i=1}^{n} t_{1i} - t_{0i} \tag{8.3}$$

where $n$ is the number of times the document was viewed and judged.

461 documents were reviewed in total by the 16 reviewers. 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 $\sigma = 45.4$. We use the reviewed documents and their associated reviewing log data to construct a test collection for developing and evaluating our proposed approach for maximising openness in the limited review user model. We present details of the test collection in the following section.

### 8.3.2 Test Collection Constructed from the User Study

We use the judgements and the log data that we collected from the study to generate a test collection for developing and evaluating our proposed approach for maximising the number of documents that can be opened to the public with the available reviewing resources. To ensure that the test collection only contains data from the reviewers who committed to the task, we include reviews from the reviewers who 1) made at least 10 judgements, and 2) recorded sensitivity annotations. This resulted in eleven reviewers contributing to the test collection. Additionally, since we could not control for reviewers taking breaks, we do not include in the test collection documents that took longer than two hours to review.

Each reviewer's reviews were ordered by the order that they were judged. We then split the reviews from each reviewer 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 8.1 provides an overview of the training and test data for the generated test collection.

## 8.4 Predicting Reviewing Times

For the limited review user model, we aim to generate a ranking of documents that can increase the number of documents that are released to the public with the available reviewing resources. Our proposed approach prioritises documents that are predicted to require less time to sensitivity review and is based on a three-step process:

**Step 1** Automatically classify the documents in the collection by whether they do or do not contain any sensitive information.

**Step 2** Predict the length of time a reviewer is likely to require to review a specific document.

**Step 3** Use the reviewing time predictions to generate a ranking of documents that prioritises documents that do not contain any sensitive information and that are quickest to review.

The intuition behind our approach is that to maximise the number of documents that are released to the public, the available reviewing resources should be focused on reviewing non-sensitive documents. Moreover, the non-sensitive documents that will require the least time to review should be prioritised over the non-sensitive documents that will require more time to review. There is an additional reviewing time cost associated to reviewing sensitive documents, since the reviewer has to record any identified sensitivities (this additional time cost is accounted for in our user study of Section 8.3 by reviewers having to annotate the sensitive text in a document). Therefore, by integrating sensitivity classification into our approach for predicting the amount of time that a reviewer will require to review a document, we postulate that our approach will prioritise documents that are both non-sensitive and require less time to review. Thereby, increasing the number of documents that are released to the public.

In remainder of this section, we present our approach for the 2nd step of our proposed approach: predicting the length of time a reviewer will require to review a specific document. In particular, we present our approach and the features that we use (including the output from Step 1 of our proposed approach) for predicting reviewing times in Section 8.4.1, our experimental methodology in Section 8.4.2 and our results in Section 8.4.3. We will present our analysis of Step 3, i.e., using the reviewing times predictions to prioritise documents for review, in Section 8.5.

### 8.4.1 Predictions Reviewing Times Approach and Features

Step 2 of our proposed process requires that we predict the length of time a reviewer is likely to require to review a specific document. Predicting a document's reviewing time is a complex task since there are many variables that can lead to large variations in reviewing times, such as a document's length, the complexity of the document or a specific reviewer's reading speed.

Jethani & Smucker (2010) modelled the average time to judge relevance as a function of document length. In that work, the authors learned a linear model to predict reviewing times for two user models. Jethani & Smucker (2010) used the adjusted $R^2$ ($R^2_{Adj}$) metric as a measure of the amount of variance accounted for by their model. We will provide more details about the $R^2_{Adj}$ metric in Section 8.4.2. The authors found that when reviewers have to review an entire document to make a decision of relevance (as is the case for sensitivity review) their model could account for 26% of the variance in the time taken to review a document (compared to 45% of the reviewing time variance when reviewers only reviewed query-biased summaries). The work of Jethani & Smucker (2010) predicted the reviewing times for eight topics from the TREC Robust track[3], which used the AQUAINT[4] collection of newswire documents. This task is not the same as reviewing for sensitivity. However, we also use a linear model to predict a document's reviewing time and the observed 26% variance accounted for by Jethani & Smucker (2010) provides us with a *ball park* figure for identifying a model that could be effective enough to deploy in our ranking strategy for Step 3 of our proposed approach[5].

As we previously mentioned, there are many variables that contribute to the amount of time that a reviewer will require to review a document. One of these variables is the reading speed of a reviewer. When prioritising document to be sensitivity reviewed, we do not want to have to learn a separate reviewing time prediction model for each reviewer since this would require additional resources. Moreover, we argue that identifying the priority documents before the documents are assigned to a specific reviewer will be less restrictive for departments, since they can assign whichever reviewers are available to the priority documents. Therefore, we aim to predict the time that an *average* reviewer would take to review a document.

Damessie *et al.* (2016) used a reviewer's dwell time, i.e., the time from a reviewer first viewing a document until the reviewer records a relevance judgement, 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, the authors proposed *normalised dwell time* (NDT) to measure the reviewing time of an average reviewer. The NDT for a document, $d$, is defined as

$$NDT = exp^{(\log(time)+\mu-\mu_\alpha)} \tag{8.4}$$

where $\log(time)$ is the log of the time taken to review $d$, $\mu$ is the global mean reviewing time calculated over all documents for all reviewers, and $\mu_\alpha$ is the mean reviewing time for the reviewer who reviewed $d$.

We also use NDT as a measure of the time that an average reviewer would require to review

[3]https://trec.nist.gov/data/robust.html

[4]https://catalog.ldc.upenn.edu/LDC2002T31

[5]To the best of our knowledge this is the most closely aligned task from the literature that we can use to compare our performance against.

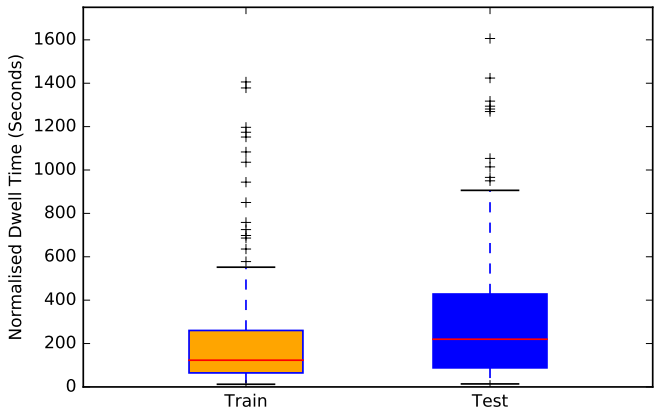



Figure 8.1: Normalised Dwell Time (NDT) distributions in seconds for the training and test data of our test collection constructed from the reviewing times user study.

Table 8.2: The Feature groups that we evaluate for predicting a document's reviewing time.

| **Feature Group** | **Features** |
|---|---|
| Decision | (1) Number of previously reviewed documents (2) Sensitivity classification prediction |
| Surface | (3) Number of sentences (4) Total prepositions (5) Total syllables (6) The ratio of unique words |
| Complexity | (7) Simple Measure of Gobbledygook (SMOG) (8) Automated Readability Index (ARI) (9) Coleman-Liau Index (10) Gunning Fog Index |

a document. However, differently from Damessie *et al.* (2016), since calculating NDT relies on the means $\mu$ and $\mu_\alpha$ which we do not have before the documents are reviewed, we learn a linear regression model to predict a document's NDT. Figure 8.1 presents the *actual* NDT distributions in the training and test data of our test collection that we presented in Section 8.3.2. In Step 3 of our approach, we use the *predicted* NDT of a document (from the test data) to prioritise (rank) the documents that are to be reviewed.

**Features for Reviewing Times:** To predict a document's NDT, we use ten features that are separated into three sets of features, namely Decision, Surface and Complexity features. Table 8.2 lists the features in each of the feature groups. We now discuss each of the feature groups and their features:

- **Decision**: The first set of features that we use represent the main aspects of a reviewer's decision process when making a sensitivity judgement that we expect to affect the length of time that they will require to review a document:
  - (1) The number of documents that a reviewer has reviewed prior to the current document. As the reviewer reviews more documents, they should become more familiar with the collection and the sensitivities. Therefore, they may get quicker at reviewing as they review more documents.

  - (2) Whether the document is predicted to be sensitive or not sensitive. In sensitivity review, there is an inherent additional reviewing time cost for sensitive documents compared with not-sensitive documents. This additional reviewing time is due to the fact that the sensitivities in a document must be recorded so that the reviewing department can apply to the Advisory Council[6] to have the information closed. As previously mentioned in Section 8.4, this additional reviewing time cost is accounted for in our user study by having a reviewer annotate any sensitive information in a document. Therefore, whether a document is sensitive or not has a direct impact on the amount of time that a reviewer will need to review a document. With this in mind, we include a document's sensitivity classification prediction as a feature when predicting a document's reviewing time.

- **Surface**: The second set of features that we use are basic statistics about the grammatical content, or surface features, of a document that we expect to affect the length of time that is required to review a document:

  - (3) The number of sentences in a document.
  - (4) The total number of prepositions, such as *at*, *with* or *from*, in a document.
  - (5) The total number of syllables in a document.
  - (6) The ratio of unique words / total words in a document.

- **Complexity**: The last set of features that we test are standard readability metrics that represent the complexity, or reading difficulty, of a document. We postulate that more complex documents are more difficult to read and, therefore, will require more time to review:

  - (7) Simple Measure of Gobbledygook (SMOG) (Mc Laughlin, 1969) is a simple readability metric based on the number of polysyllabic words per sentence within a 30-sentence sample from a document.
  - (8) The Automated Readability Index (ARI) (Smith & Senter, 1967) is a weighted sum of the mean words per sentence and the mean number of characters per word.
  - (9) The Coleman-Liau Index (Coleman & Liau, 1975) is a weighted sum of the average number of characters per 100 words and the average number of sentences per 100 words.
  - (10) the Gunning Fog Index (Gunning, 1952) is a weighted sum of the average sentence length and the percentage of *complex* words, i.e., words with three or more syllables.

[6] http://www.nationalarchives.gov.uk/about/our-role/advisory-council/

In this section, we have presented our proposed approach for predicting the time that is required to review a document and the features that we use in our proposed approach. In the following section, we present our experimental methodology that we use to evaluate our three sets of features for predicting reviewing times, before presenting the results of our experiments in Section 8.4.3.

### 8.4.2 Experimental Methodology

We will evaluate the effectiveness of our predicted reviewing times approach for prioritising documents for review in Section 8.5. The research question that we wish to answer in this section is as follows:

- RQ8.1: Which feature set(s) are most effective for predicting a documents normalised dwell time?

To answer this research question, we learn a linear regression model using the feature sets that we presented in Section 8.4.1 (Table 8.2). Feature number (2) requires us to predict if a document is sensitive or not. The effectiveness of the sensitivity classifier that we use for this feature is likely to have a direct impact on the accuracy of the reviewing times predictions. Moreover, it may be that the effectiveness of the sensitivity classifier affects which of the feature sets should be used. Therefore, we evaluate the feature sets individually for three different levels of sensitivity classification effectiveness: *Perfect*, *Good* and *Baseline*. For our perfect classifier, we use the actual judgements of the reviewers in the user study. As our good classifier we use predictions from the best performing sensitivity classifier from the approaches that we presented in Chapter 6, i.e., text classification plus language and semantic features (denoted as Text+$TN_7$+$WE_{wp}$+$WE_{gn}$(concat) in Table 6.9). As our baseline classifier, we use the predictions from the best performing sensitivity classifier from the approaches that we presented in Chapter 5, i.e., text classification (denoted as SENSITIVE in Table 5.7).

We use the test collection that we presented in Section 8.3.2 to train and evaluate the regression model. We evaluate the effectiveness of each of the feature groups individually, in pairs and all of the feature groups combined. We select root mean squared error (RMSE) as our main metric as it provides an absolute measure of variance, in seconds, for our predictions (smaller RMSE values mean better prediction effectiveness). Additionally, we report $R^2$, defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{8.5}$$

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 $R^2$:

$$R^2_{Adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \tag{8.6}$$

where $n$ is the number of documents and $k$ is the number of features. $R^2_{\text{Adj}}$ enables a fair comparison between models with different numbers of features, i.e., when a new feature is added to a model $R^2_{\text{Adj}}$ increases only if the model improves more than would be expected by chance. Additionally, as we previously mentioned in Section 8.4.1, $R^2_{\text{Adj}}$ provides us with an indication of whether our model's effectiveness is in line with the effectiveness achieved by Jethani & Smucker (2010).

### 8.4.3 Results

Table 8.3 presents the results for our approach for predicting a document's NDT. The table shows the root mean squared error (RMSE), $R^2$ and adjusted $R^2$ ($R^2_{\text{Adj}}$) achieved when either of the Perfect, Good or Baseline classifiers are deployed to supply the sensitivity classification prediction feature for the linear regression model.

As can be seen from Table 8.3, for each level of classification effectiveness (Perfect, Good and Baseline), our NDT prediction model performs best when it is deployed with all three feature sets (All Features). Deploying the model with All Features results in a RMSE of 261.23 (4 minutes 21 seconds) when the Perfect sensitivity classifier is deployed, a RMSE of 275.82 (4 minutes 35 seconds) when the Good sensitivity classifier is deployed and a RMSE of 278.20 (4 minutes 38 seconds) when the Baseline sensitivity classifier is deployed. We argue that a RMSE in the range of 261.23 to 278.20 provides relatively good predictions since, as can be seen from the actual NDT distributions that we presented in Figure 8.1, 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 8.3 also shows that the $R^2_{\text{Adj}}$ for our NDT prediction model is 0.23, i.e., 23% of the variance in NDT in the test data is explained by the model, when All Features and a Perfect sensitivity classifier are deployed. This is in line with the 0.26 $R^2_{\text{Adj}}$ observed by Jethani & Smucker (2010) when reviewers were required to read an entire document to make a relevance judgement. This gives us additional confidence that our model provides relatively good predictions for this configuration. We note that the $R^2_{\text{Adj}}$ is reduced to 0.14 and 0.12 when all features are used with the Good and Baseline sensitivity classifiers respectively. Therefore, although there is a relatively small difference in RMSE when any of the levels of sensitivity classifier effectiveness are deployed ($\sim$17 seconds difference between Perfect Classification Predictions (261.23) and Baseline Classification Predictions (278.20), from Table 8.3), there is more variation in the accuracy of the NDT predictions as the level of sensitivity classification effectiveness is reduced.

Table 8.3: Reviewing Time Predictions. The root mean squared error (RMSE) in seconds, $R^2$ and adjusted $R^2$ ($R^2_{Adj}$) achieved by our linear regression model for predicting a document's Normalised Dwell Time (NDT). The table shows the results achieved when either of a Perfect, Good or Baseline sensitivity classifier is deployed to predict a document's sensitivity.

| Feature Set | $R^2$ | $R^2_{Adj}$ | RMSE (sec) |
|---|---|---|---|
| *Perfect Classification Predictions* | | | |
| 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** |
| *Good Classification Predictions* | | | |
| Decision | -0.0294 | -0.0352 | 310.52 |
| Surface | 0.1095 | 0.0941 | 288.81 |
| Complexity | -0.0639 | -0.0822 | 315.68 |
| Decision+Surface | 0.1858 | 0.1622 | 276.16 |
| Decision+Complexity | 0.0050 | -0.0236 | 305.26 |
| Surface+Complexity | 0.1087 | 0.0721 | 288.93 |
| All Features | 0.1877 | 0.1444 | 275.82 |
| *Baseline Classification Predictions* | | | |
| Decision | -0.0397 | -0.0456 | 312.06 |
| Surface | 0.1095 | 0.0941 | 288.81 |
| Complexity | -0.0639 | -0.0822 | 315.68 |
| Decision+Surface | 0.1728 | 0.1488 | 278.34 |
| Decision+Complexity | -0.0117 | -0.0409 | 307.83 |
| Surface+Complexity | 0.1087 | 0.0721 | 288.93 |
| All Features | 0.1737 | 0.1296 | 278.20 |

Therefore, in response to RQ8.1, we conclude that using all of the feature sets together is most effective for predicting a document's NDT. Therefore, we select to use all three feature sets when prioritising documents by their predicted reviewing times, in Section 8.5, to maximise the number of documents that can be reviewed and released to the public.

## 8.5 Maximising Openness

The sensitivity review of (paper) government documents is currently performed on a file-by-file basis (Moss & Gollins, 2017). This can result in documents and files essentially being reviewed chronologically. However, reviewing documents chronologically may not be the most effective approach for our limited review user model, i.e., when there are not enough reviewing resources to review all of the documents in a collection that is due to be publicly archived. Moreover, the fact that digital documents are not stored in a logically structured file-plan, such as is the case for paper government documents, raises the question of how to identify an effective order in which to select documents to be reviewed. We hypothesise that prioritising documents that take less time to review will result in more documents being released to the public for the same amount of reviewing resources. We refer to this as increasing the *openness* of the sensitivity review.

In this section, we use the reviewing time prediction models that we presented in Section 8.4 to prioritise for review the documents which are predicted to require less time to review. Our proposed approach ranks documents in ascending order by the documents' predicted Normalised Dwell Time. We refer to our approach as *shortest predicted reviewing time* and we denote it as SPR. We evaluate our SPR approach against three other approaches, namely: Shortest Document First (SDF), this strategy naively assumes that shorter documents take less time to review; Chronological (CHR), a strategy that is currently deployed for sensitivity review by some government departments; and Random selection (RND) as a baseline approach.

### 8.5.1 Experimental Methodology

The research questions that we wish to answer in this section are as follows:

- RQ8.2: Does prioritising documents that are predicted to take less time to review result in more documents being released when there are not enough reviewing resources to review all of a collection?
- RQ8.3: Is our proposed enhanced sensitivity classification approach effective enough to deploy when prioritising documents for review by their predicted reviewing times?

In Section 8.5.2, we deploy a perfect sensitivity classifier to evaluate how effective our model can be on collections that contain different amounts of sensitive documents, before evaluating how the effectiveness of the sensitivity classifier affects our proposed approach for prioritising

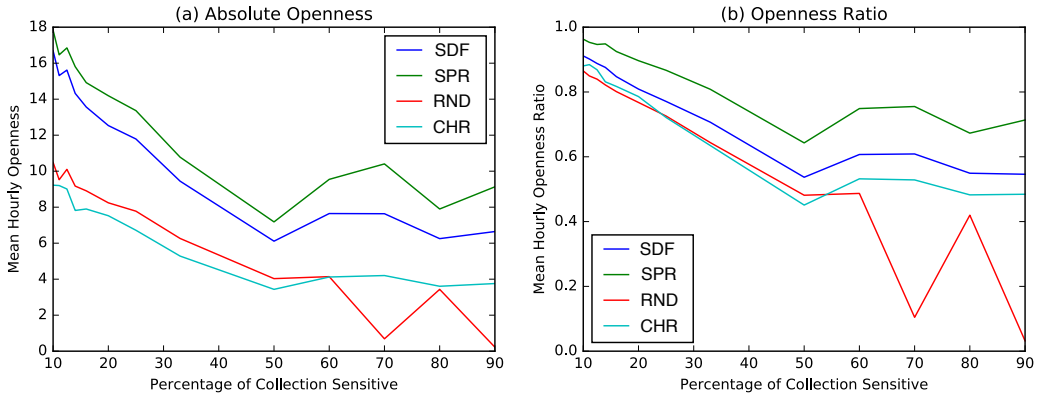



Figure 8.2: (a) Number of documents opened per hour. (b) Ratio of reviewed documents opened.

documents. For our evaluation, we simulate collections with varying distributions of sensitivity by sampling with replacement from the test data to fit the desired sensitivity distribution. We simulate nine separate collections, ranging from 10% - 90% sensitive data, where for each collection, $C$, we sample documents such that the total *actual* NDT ($NDT_A$) for the collection is 2 hours:

$$\sum_{i=0} NDT_A(d_i) = 2 \text{ hours}, d_i \in C \tag{8.7}$$

We then rank the documents in $C$ by a prioritisation strategy (SPR, SDF, CHR or RND) and set the reviewing time budget $T_b$ at one hour. 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 sampled collections for each distribution of sensitivity. Therefore, in Section 8.5.2, we report mean values over 100 * 1 hour ($NDT_A$) samples of the test data in the test collection that we presented in Section 8.3.2.

As our metrics, we report the metrics that we previously defined in Section 8.2, namely: Mean Hourly Openness (Absolute Openness) which is the average number of documents that are sensitivity reviewed and released to the public in 1 hour of reviewing time; and Mean Hourly Openness Ratio (Openness Ratio), which is the ratio of reviewed documents that are actually released to the public (i.e., documents that are not sensitive). We test for statistical significance using the sign test ($p < 0.05$).

### 8.5.2 Results

In this section, we present the results of our proposed shortest predicted reviewing time approach (SPR) for prioritising documents for review when there are not enough resources to review the whole collection. We, firstly, evaluate how effective our proposed approach can be when it is deployed with perfect sensitivity classification on simulated collections of varying sensitivity distributions. Secondly, we evaluate how our proposed approach is affected by the effectiveness of the available sensitivity classifier.

Table 8.4: The achieved openness for our proposed approach (SPR) with perfect sensitivity classification, compared against the shortest document first (SDF), chronological (CHR) and random (RND) approaches. The table presents the Absolute Openness ($O_A$) and the Openness Ratio ($O_R$) of each approach on simulated collections in which 10%, 30%, 50% or 70% of the documents are sensitive. Approaches that perform statistically significantly better than random for all sensitivity distributions are denoted as † (Sign test, $p < 0.05$).

| | 10% | | 30% | | 50% | | 70% | |
|---|---|---|---|---|---|---|---|---|
| | $O_A$ | $O_R$ | $O_A$ | $O_R$ | $O_A$ | $O_R$ | $O_A$ | $O_R$ |
| SPR † | 18 | 0.98 | 11.9 | 0.82 | 7.5 | 0.63 | 10.4 | 0.79 |
| SDF † | 16.5 | 0.90 | 10.2 | 0.72 | 6.1 | 0.55 | 8 | 0.61 |
| CHR | 9.1 | 0.86 | 5.9 | 0.68 | 3.8 | 0.45 | 4.2 | 0.52 |
| RND | 10.2 | 0.85 | 6.9 | 0.69 | 4.1 | 0.48 | 0.9 | 0.10 |

Figure 8.2 presents the effectiveness of our proposed SPR approach compared to the shortest document first (SDF) and chronological (CHR) approaches. Figure 8.2 also shows the performance of the random (RND) baseline approach. Firstly, from Figure 8.2(a), we note that ordering documents by their expected time to review (SPR), results in more documents being released to the public than the next best approach, i.e., shortest document first (SDF), for all of the sensitivity distributions that we tested. Secondly, we note that the improvements in openness are fairly consistent when $< 50\%$ of the collection is sensitive[7]. However, when the collection has high levels of sensitivity, SPR can result in higher relative gains in openness. Figure 8.2(b) presents the ratio of reviewed documents that were released. As can be seen from Figure 8.2(b), our proposed approach also consistently results in more of the documents that are reviewed being released, i.e., more of the reviewed documents are not sensitive. Moreover, in our experiments, for collections that are 60%-70% sensitive, SPR results in a 30% increase in the ratio of reviewed documents that are actually opened, e.g., on our simulated collection in which 70% of documents contain some portion of sensitive information our SPR ranking strategy results in an extra 200 documents being released for 100 hours of reviewing time.

Table 8.4 presents the corresponding Absolute Openness ($O_A$) and Openness Ratio ($O_R$) scores for the collections containing 10%, 30%, 50% and 70% sensitive documents. We note, from Table 8.4, that our proposed SPR approach and the shortest document first (SDF) approach results in significantly more documents being released to the public than random selection for all sensitivity distributions (sign test, $p < 0.05$). Lastly, we note that reviewing documents chronologically actually resulted in the fewest number of documents being released on simulated collections where $\leq$50% of the documents are sensitive (which we argue is the most likely case). This suggests that this prioritisation strategy, which is currently used by some government departments, is not a suitable strategy when the objective is to review and open as many documents as possible.

[7]We note that we would expect that in most collections that are to be sensitivity reviewed $< 50\%$ of the documents in the collection will contain sensitive information.

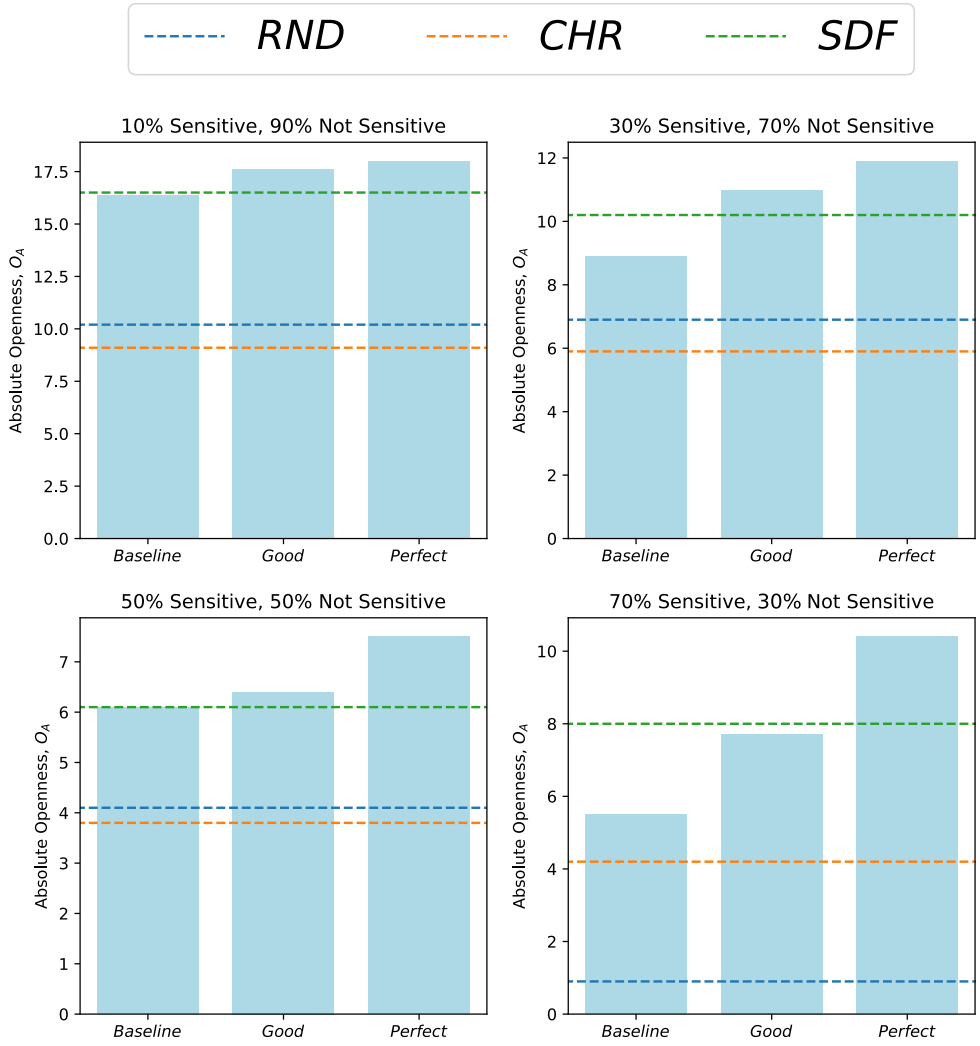



Figure 8.3: The resulting Absolute openness achieved by our proposed SPR document prioritisation approach when either the Baseline, Good or Perfect sensitivity classifier is deployed.

We now turn our attention to how the effectiveness of the available sensitivity classifier (Baseline vs Good vs Perfect) affects the performance of our proposed shortest predicted reviewing times approach. As previously noted in Section 8.4.2, as our baseline classifier, we use the predictions from the best performing sensitivity classifier from the approaches that we presented in Chapter 5, i.e., text classification (denoted as SENSITIVE in Table 5.7). As our good classifier we use predictions from the best performing sensitivity classifier from the approaches that we presented in Chapter 6, i.e., text classification plus language and semantic features (denoted as Text+$TN_7$+$WE_{wp}$+$WE_{gn}$(concat) in Table 6.9).

Figure 8.3 presents the achieved Absolute Openness when the available sensitivity classifier has a baseline or good level of effectiveness, compared to the perfect classification effectiveness (SPR) and the SDF, CHR and RND approaches for four sensitivity distributions. From Figure 8.3, firstly, we note that the effectiveness of our proposed SPR approach does decrease as the effectiveness of the sensitivity classifier decreases. However, the achieved openness when

the good classifier is deployed is still significantly better than the shortest document first (SDF) approach when $\leq$50% of the documents in the collection contain sensitive information (Sign test, $p < 0.05$). Moreover, the SPR approach performs significantly better than either of the CHR and RND approaches for all sensitivity distributions. Therefore, in response to RQ8.2, we conclude that prioritising documents that are predicted to take less time to review does result in more documents being released to the public when there are not enough reviewing resources to review all of the documents in a collection. Moreover, in response to RQ8.3, we conclude that our proposed enhanced sensitivity classification approach, that we presented in Chapter 6 is sufficiently effective to be deployed for prioritising documents for review by their predicted reviewing times. As can be seen from Figure 8.3, our proposed SPR document prioritisation approach with the predictions from our classifier from Chapter 6 results in more documents being released to the public than the three other approaches that we evaluated (SDF, CHR and RND) when either 10%, 30% or 50% of the collection is sensitive.

## 8.6 Conclusions

In this Chapter, we investigated how our proposed framework can assist the sensitivity review process in the first of two realistic sensitivity review user models that we investigate in this thesis, namely the limited review user model. Limited review addresses a scenario where there are not enough reviewing resources available to sensitivity review all of the documents in a collection that is due to be publicly archived. We proposed an approach for prioritising documents for review by predicting the amount of reviewing time that an average reviewer would require to review them. We argued that, when there are not enough reviewing resources to review all of the documents that are to be transferred, the productivity of the sensitivity review can be increased by focusing reviewing resources on documents that are (1) more likely to be released, i.e., documents that are not sensitive, and (2) quicker to review. We showed that our proposed approach that predicts the amount of time that an average reviewer would require to review each of the documents in a collection and prioritises the documents with the shortest predicted reviewing times can result in significantly more documents being released compared to a random selection strategy (Table 8.4). Moreover, we showed that our proposed enhanced sensitivity classification approach (that we presented in Chapter 6) is effective enough to be deployed to predict a document's sensitivity as an input to our proposed SPR document prioritisation strategy (Figure 8.3). In conclusion, we argue that prioritising documents for review by our shortest predicted reviewing times strategy (SPR) can help to enable government departments to meet the Public Records Act 1958 (c. 51) time-to-transfer obligations for a larger percentage of the digital documents that are to be reviewed, when there are not enough reviewing resources to review all of the documents. In the following chapter, we investigate how sensitivity classification can assist sensitivity reviewers in the second of our user models, namely *exhaustive review*.

# Chapter 9

# Assisting Sensitivity Reviewers in the Exhaustive Review User Model

## 9.1 Introduction

In the previous chapter, we presented our proposed approach for prioritising documents for review to maximise the number of documents that can be opened to the public in the limited review user model, i.e., when there are not enough reviewing resources to review all of the documents in a collection that is to be publicly archived. Our proposed approach uses sensitivity classification and features of the reviewing process to predict the amount of time that an average reviewer would require to review a specific document and prioritises the documents that are predicted to require the least time to review. We showed that prioritising documents for review by our proposed shortest predicted reviewing time strategy can result in an increase in the number of documents that can be reviewed and released to the public while using the same amount of reviewing resources (see Section 8.5.2). Moreover, we showed that our proposed enhanced sensitivity classification approach, that we presented in Chapter 6, is sufficiently effective to be deployed to provide sensitivity classification predictions for our proposed shortest predicted reviewing time document prioritisation strategy (see Figure 8.3).

In this chapter, we investigate how sensitivity classification can assist sensitivity reviewers in the second, and final, of our realistic sensitivity review scenarios that we investigate in this thesis, namely the *exhaustive review* user model. The exhaustive review user model addresses the scenario in which there are sufficient reviewing resources available to sensitivity review *all* of the documents in a collection that is to be publicly archived. In an exhaustive review, the order in which the documents are reviewed does not directly impact the number of documents that are released to the public[1]. Moreover, it is generally accepted that the sensitivity review of

[1]We say that the order of review does not *directly* impact the number of documents that are released since it is possible that the order in which the documents are reviewed could have an effect on a reviewing decision. However, investigating these questions is outside the scope of this thesis.

government documents will continue to be done by human sensitivity reviewers at least until automatic technologies have matured enough for governments and reviewers to develop trust in the technology (The National Archives, 2016*a*).

As we previously stated in Chapter 8, government departments are not expected to be able to recruit enough reviewing resources to review all of the digital documents that are to be transferred (The National Archives, 2016*a*). Currently, for many central UK government departments, such as the FCO, sensitivity review is conducted by expert reviewers who have a good knowledge of the expected sensitivities within the department (The National Archives, 2017). One potential approach to address the shortfall in the available reviewing resources would be to employ a greater number of reviewers by recruiting less experienced reviewers, at less expense that experienced reviewers, and assisting them with sensitivity classification predictions. This strategy would enable government departments to focus the experienced resources on reviewing high-risk documents or documents that are more difficult to review (e.g., re-reviewing sensitivity judgements that the less experienced reviewers disagree about.)

Therefore, we wish to know how sensitivity classification can benefit sensitivity reviewers, who have less reviewing experience than expert reviewers, when *all* of the documents in a collection must be reviewed by a human reviewer. One way that sensitivity classification can potentially benefit sensitivity reviewers is by providing the reviewers with useful information about the documents they are to review, to *assist* them in making their reviewing decisions. In particular, informing a reviewer about any sensitivities that are in a document, before the reviewer reads the document, has the potential to enable the reviewers to make accurate sensitivity judgements more quickly. Moreover, sensitivity classification predictions can potentially lead to increased consistency in the level of accuracy of sensitivity reviewers, since the classifier could bring to the attention of a reviewer sensitivities that they may have missed by themselves. Furthermore, sensitivity classification has the potential to increase the agreement between reviewers, which in turn could result in less time being required for discussing disputed reviewing decisions and fewer closure applications being challenged by the Advisory Council[2].

To investigate if sensitivity classification predictions can benefit sensitivity review in the exhaustive review user model, we present the results of a controlled sensitivity review user study. The study evaluates how two aspects of sensitivity classification predictions affect the sensitivity judgements of human sensitivity reviewers, namely: (1) the accuracy of sensitivity classification predictions, i.e., the classifier's effectiveness; and (2) the level of (simulated) confidence that the classifier has in its individual predictions. Participants in the study review a collection of documents while being assisted by sensitivity predictions for three levels of classification effectiveness treatments, namely: *None* (i.e., no classification predictions); *Medium* accuracy; and *Perfect* accuracy. Moreover, each of the predictions has an associated score that indicates if the classifier had *Low*, *Medium* or *High* confidence in its prediction. The study investigates how

[2]http://www.nationalarchives.gov.uk/about/our-role/advisory-council/

the two aspects of sensitivity classification (effectiveness and confidence) affect three aspects of the reviewer's performance, namely (1) the number of documents that a reviewer correctly judges to contain, or to not contain, sensitive information (reviewer accuracy); (2) the length of time that it takes for a reviewer to sensitivity review a document (reviewing speed); and (3) the amount of agreement between the reviewers' judgements and the classifier's predictions (reviewer-classifier agreement). The remainder of this chapter is structured as follows:

- In Section 9.2, we introduce the *exhaustive review* user model and provide details of the motivation for the study.
- Section 9.3 presents our two hypotheses that we investigate in this chapter. Our first hypothesis looks at how the classifier's effectiveness (accuracy) affects the participant reviewers' accuracy, reviewing speed and the reviewer-classifier agreement. Our second hypothesis looks at how the level of (simulated) confidence that the classifier has when it makes a prediction affects the reviewers, in terms of reviewing speed and reviewer-classifier agreement.
- In Section 9.4, we present details of a controlled within-subject user study that we conduct under laboratory conditions to evaluate the benefits of providing non-expert sensitivity reviewers with automatic sensitivity classification predictions for the documents that they are to review. The study investigates how the level of effectiveness of the sensitivity classifier's predictions, *none*, *medium* or *perfect*, affects the reviewing decisions that a reviewer makes. Moreover, the study investigates how the reviewers are affected by being informed that the classifier has *low*, *medium* or *high* confidence in its prediction for a specific document. In particular, we provide a short discussion about the ground truth that we use to assess the sensitivity judgements that are made by the reviewers in our study in Section 9.4.1, before presenting details of the reviewing interface and how we log the reviewers interactions in Section 9.4.2. We present details of the study design and the documents that we use for the study in Section 9.4.3, before providing details about the study participants, incentives and instructions in Section 9.4.4 and the evaluation metrics that we use in Section 9.4.5.
- Section 9.5 presents the findings of our sensitivity review user study. In particular we evaluate the impact of classification effectiveness on the reviewers' performance in Section 9.5.1 and the impact that the classifier's simulated confidence has on the reviewers' performance in Section 9.5.2. We conclude the findings of our study in Section 9.5.3.
- Finaly, in Section 9.6, we summarise our conclusions from this chapter.

## 9.2 Exhaustive Review User Model

In the exhaustive review user model, there are sufficient available reviewing resources to review all of the documents that are to be transferred to the public archive. Moreover, when conducting a sensitivity review, the sensitivity reviewer must identify all of the passages of sensitive text in a document, so that the reviewer can provide comprehensive information for the redaction process and to the Advisory Council. Therefore, sensitivity reviewers must read all of the text in each of the documents that are being reviewed[3]. Since all of the documents in the collection will be manually sensitivity reviewed, in the exhaustive review user model the role of sensitivity classification, and our proposed framework, is to provide reviewers with useful information that can potentially assist the reviewer to perform the reviewing task. We postulate that, by providing a reviewer with an automatic classification prediction about whether a document is sensitive or not before the reviewer reads the document, the reviewer will be able to use this prior knowledge to guide their reviewing actions. For example, if the classifier is very confident that a document is sensitive, then the reviewer may be able to read the document in less detail, i.e., to quickly *scan* the document, to identify passages that are clearly sensitive. In the following section, we present our hypotheses and discuss why we expect automatic sensitivity classification predictions to benefit sensitivity reviewers and assist them in making reviewing decisions.

## 9.3 Hypotheses

As previously mentioned in Section 9.1, in this chapter, we investigate if providing sensitivity reviewers with sensitivity classification predictions, to assist their reviewing task, affects the speed, accuracy and agreement of reviewers. In this section, we present our hypotheses that we investigate in the user study that we present in Section 9.4 and our analysis in Section 9.5.

We, firstly, wish to know how the effectiveness of a sensitivity classifier affects how a reviewer performs the sensitivity review task. If we assume that a perfect classifier does indeed benefit a reviewer and, therefore, enables the reviewers to review more quickly and/or more consistently, then we would intuitively expect that as the level of classification effectiveness decreases towards random predictions, the benefit provided to the reviewer by the classification predictions would also decrease. Therefore, we state our first hypothesis as:

**H1:** As the effectiveness of the classifier increases from no classification predictions to perfect classification predictions, the classifier will be of more benefit to reviewers and, therefore, reviewers will:

(a) Make more correct and less incorrect judgements.

[3]We note that a reviewer may not have to read all of a document if they discover early in the process that the document is sensitive enough that the entire document has to be closed.

(b) Make quicker reviewing decisions (i.e., review documents faster) on average.

(c) Agree with the classifier's predictions more often.

Secondly, we wish to know how the level of confidence that a classifier has in its predictions affects the reviewers' decisions. In our discussions with sensitivity review professionals from UK government departments, it has often been suggested to us that sensitivity reviewers would likely benefit from being provided with information about how confident the classifier is in its predictions. In general, it was suggested that supplying reviewers with a confidence score about the classifier's decisions would provide a level of transparency for reviewers and would also help the reviewers to build trust in the technology. We postulate that the level of confidence that a classifier has in its predictions will have a direct influence on how much trust reviewers have in the classifier's predictions and how quickly reviewers make reviewing decisions. Our second hypothesis is stated as follows:

**H2:** When the classifier is confident about its predictions, 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 compared to when they disagree with the classifier. Moreover, the *difference* in reviewing speeds when reviewers agree or disagree with the classifier will increase as the classifier's confidence increases.

In the following section, we present the user study that we conduct to investigate these hypotheses.

## 9.4 Assisted Sensitivity Review: User Study

In this section, we present details of the controlled sensitivity review user study that we conduct to test the two hypotheses that we stated in Section 9.3. The study is designed to evaluate the effects on non-expert sensitivity reviewers' actions and decisions when the reviewers are provided with sensitivity classification predictions for the documents that they are to review. Participants in the study, i.e., *reviewers*, are provided with a reviewing interface and asked to sensitivity review three *batches* of twenty documents. The study evaluates the effects of two variables of sensitivity classification that have the potential to influence the amount of benefit that sensitivity reviewers can get from the predictions. The first variable that we test is the accuracy of the classification predictions. In the study each batch of documents that a reviewer reviewed has an associated level of classification prediction accuracy, *none*, *medium*, or *perfect*. The second variable of sensitivity classification that we test is the (simulated) confidence that the classifier has in its individual predictions. Each sensitivity classification prediction has an associated

level of simulated confidence, *Low*, *Medium* or *High*. We log the reviewers actions and sensitivity judgements to evaluate how the two variables of sensitivity classification (effectiveness and confidence) affects the number of documents that a reviewer correctly judges to contain, or to not contain, sensitive information (reviewer accuracy), the length of time that it takes for a reviewer to sensitivity review a document (reviewing speed) and the amount of agreement between the reviewers' judgements and the classifier's predictions (reviewer-classifier agreement). We discuss the ground truth that we use for evaluating the participants accuracy in Section 9.4.1, before providing details of the reviewing interface in Section 9.4.2 and the experimental design in Section 9.4.3. We present details of the participants, incentives and instructions in Section 9.4.4 and the metrics that we use to test our hypotheses in Section 9.4.5, before presenting a discussion of our findings from the study in Section 9.5.

### 9.4.1 Ground Truth

In this study, we use a sample of documents from our test collection of Chapter 3. We will provide a full description of how we sampled the documents for our user study in Section 9.4.3. We use the expert reviewers' sensitivity judgements from our test collection of Chapter 3 as a ground truth when evaluating the study participants' reviewing accuracy. In effect, we are assuming that it is reasonable to expect the availability of gold standard judgements that reliably identify what is or is not sensitive within a collection. We argue that this is a reasonable assumption to make, since it is reasonable to expect that the effectiveness of sensitivity classifiers that are trained on the judgements of many sensitivity reviewers will continue to improve as more data becomes available to learn from. However, as is also the case when assessing relevance in other fields, for example when evaluating the performance of Information Retrieval systems in the TREC evaluation campaigns (Voorhees, 1998), there is always a level of disagreement, even among expert reviewers. The levels of inter-assessor agreement in the paper-based sensitivity review are unknown since, although some departments have senior reviewers and committees to check samples of reviews (Allan, 2014), historically each document was typically only assessed by a single reviewer. When constructing our test collection that we presented in Chapter 3, we conducted a small study to assess the inter-assessor agreement of expert reviewers in the sensitivity review of born-digital documents (McDonald *et al.*, 2014). We found that there was moderate agreement between expert reviewers, with a Cohen's $\kappa$ of 0.55 for 150 documents assessed by two reviewers and a Fleiss' $\kappa$ of 0.44 for 50 documents assessed by four reviewers.

### 9.4.2 Reviewing Interface and Logging Interactions

As previously mentioned in the introduction to Section 9.4, reviewers are provided with a (web-based) interface to navigate the collection and record any sensitivities in the documents. The interface is mostly identical to the interface that we presented in Chapter 3 (Figure 3.3) and


Figure 9.1: Reviewing Interface Information Panel: The panel displays the classification prediction (Sensitive or Not Sensitive), and the classifier's prediction simulated confidence score. The panel also enables participants to record their sensitivity judgements and provide comments.

Chapter 8. However, the interface that the study participants are provided has an additional information panel to inform the reviewer about the sensitivity classifier's prediction for the displayed document and the classifier's simulated confidence about the prediction. Figure 9.1 presents the reviewing functionalities that the interface provides to reviewers. As can be seen from Figure 9.1, the information panel at the top of the screen shows participants the current document's classification prediction (Sensitive or Not Sensitive) and a simulated prediction confidence score. The document to be reviewed is displayed to participants below the panel in Figure 9.1. Therefore, when reviewing a document, the reviewer is presented with the classification prediction and simulated confidence score before actually reviewing the document.

As can be seen from Figure 9.1, differently from the interface of Chapters 3 and 8, the participants are provided with a button to pause the system. In the study, participants can use this button at any time, for example to have a comfort break or ask a question. This functionality helps to improve the accuracy of the recorded timings of when participants are focused on the reviewing task.

Figure 9.1 also shows how reviewers record their judgements as to whether a document contains sensitive information. This functionality is identical to the reviewing interface that we presented in Chapters 3 and 8. Firstly, participants record a sensitivity judgement by selecting one of the four radio button options at the left of the panel. Participants are also asked to provide a short explanatory comment about their decision, in the text box at the centre of the panel, for any documents that they judge as being sensitive. In addition to providing this comment, for documents that are judged to be sensitive, the participants are asked to highlight any sensitive text within the document. A simple mouse-click and drag functionality facilitates the highlighting of sensitive text. Additionally, participants can indicate if a particular judgement is particularly hard to make. In addition to logging the participants' sensitivity judgements, the interface also logs a timestamped record of when a participant loads a document, saves a judgement, pauses or restarts the system.

Table 9.1: The distribution of automatic classification predictions for documents in batches representing different classification effectiveness treatments along with the resulting $F_2$ and Balanced Accuracy (BAC) scores.

| Classification | TP | FN | FP | TN | Sensitive | Not Sensitive | Total | $F_2$ | BAC |
|---|---|---|---|---|---|---|---|---|---|
| None | - | - | - | - | 5 | 15 | 20 | - | - |
| Medium | 3 | 2 | 3 | 12 | 5 | 15 | 20 | 0.5769 | 0.7 |
| Perfect | 5 | 0 | 0 | 15 | 5 | 15 | 20 | 1.0 | 1.0 |

### 9.4.3 Study Design

The user study is a within-subject design, where each participant is exposed to all of the conditions being evaluated. Participants are asked to review batches of 20 documents and, for each document, record a sensitivity judgement as to whether the document is "*Not Sensitive*" or contains either "*Section 27*" (international relations), "*Section 40*" (personal information) or "*Both: Section 27 & Section 40*" sensitive information.

Participants are asked to review batches in a prescribed order. Documents within a batch are presented in random order, consistently between reviewers (i.e all of the reviewers are presented the documents in the same order). In the study, participants are advised to proceed linearly through a batch. However, they are able to (re)select documents within a batch in any order to re-visit documents and change any previously made judgements, if they so wish.

We deploy a simulated classifier in our user study. We use the expert sensitivity reviewers ground truth as gold standard judgements, and the classification predictions from our enhanced sensitivity classification approach that we presented in Chapter 6, i.e., text classification plus language and semantic features (denoted as Text+$TN_7$+$WE_{wp}$+$WE_{gn}$(concat) in Table 6.9), we randomly sample documents from our test collection of Chapter 3 (Table 3.3) to fit the distributions of sensitive and not-sensitive documents presented in Table 9.1. As can be seen from Table 9.1, each batch of 20 documents has an associated *treatment* of sensitivity classification predictions, where each treatment has a certain overall level of classifier accuracy, either *None* (i.e., no classification predictions were provided)[4], *Medium* (i.e., the accuracy of the classification predictions is 0.7 BAC) or *Perfect* (i.e., the classification predictions agree with the expert reviewers gold standard and, therefore, has an accuracy of 1.0 BAC).

Table 9.1 also presents the distributions of correct and incorrect sensitivity classification predictions that are associated to each of the classification treatments. As can be seen from Table 9.1, each batch of documents contains 5 *sensitive* documents and 15 *not sensitive* documents, resulting in 25% of documents in each batch containing sensitive information. This is slightly higher than in our test collection of Chapter 3 (Table 3.3), in which 13.2% of documents contain sensitive information. Also, we note that it would have been desirable to have had more treatments in the study design, with classification effectiveness levels between *Medium* and *Perfect*.

[4]When participants are reviewing documents from batches with classification effectiveness *None*, the information panel, presented in Figure 9.1, says "Classification Prediction :: Off".

Table 9.2: Distributions of *Low*, *Medium* and *High* simulated confidence scores for each classification effectiveness.

| Classification | Low | Medium | High |
|---|---|---|---|
| Medium | 7 | 6 | 7 |
| Perfect | 7 | 6 | 7 |

However, we developed the experimental design containing three classification accuracy levels and 20% 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 our proposed enhanced sensitivity classification approach that we presented in Chapter 6 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 have associated True Positive (TP) predictions, 2 documents have False Negative (FN) predictions, 3 documents have False Positive (FP) predictions and 12 documents have associated True Negative (TN) predictions (see Table 9.1), where *sensitive* is the positive class and *Not Sensitive* is the negative class.

In treatment batches with either *Medium* or *Perfect* classification effectiveness, each prediction has an associated score in the range $(0, 1)$ that represents the level of simulated confidence the classifier has about its prediction. Each assigned simulated confidence score represents either *Low*, *Medium* or *High* confidence, where $Low < 0.35 < Medium < 0.7 < High$. Simulated confidence scores are assigned to classifier predictions randomly to fit the distribution presented in Table 9.2. However, the assignments are manually checked to ensure that the allocations are credible. This manual check helps to control for the possibility of participants learning to distrust the simulated confidence scores due to their random allocation.

Participants review 3 batches of documents each (1**None*, 1**Medium*, 1**Perfect* classification effectiveness), i.e., 60 documents each. To control for learning effects and fatigue, we counterbalance the allocation of batches, i.e., we permute the order in which batches are reviewed by different reviewers.

As we previously discussed in Chapter 8, in sensitivity review, there is an additional processing cost associated to sensitive documents, since any identified sensitivities must be recorded to provide evidence of why the information is being withheld. This additional processing cost is accounted for in the study design by having reviewers (1) provide a short explanation of their decision if they judged a document to be sensitive and (2) highlight the specific text that they judge to be sensitive. Importantly, having reviewers highlight the sensitive text also ensures that participants read the documents and do not solely rely on the provided classification predictions to make their decisions.

### 9.4.4 Participants, Incentives and Instructions

We recruited eight students from the University of Glasgow as participants for the study. In this study, since participants were being asked to identify information relating to international relations and personal information sensitivities, we ensured that participants had a good knowledge of the Freedom of Information Act 2000 (c. 36) (FOIA) and were familiar with the concepts that they were being asked to review, by limiting participants to those whose main subject of education was politics or international law. Additionally, we limited subjects to those who had been speaking English for at least 10 years. Full ethical approval for the study was obtained from our university's ethics IRB (Application Number 300170056).

At the beginning of the study, there was a 1 hour training session where participants were provided with training on the reviewing interface and the sensitivities that they were being asked to identify. Participants were provided with the same reviewing guidelines and interface user manual as was provided to the expert sensitivity reviewers when constructing the test collection in Chapter 3 (Table 3.3), i.e the ground truth for this study. Moreover, the participants were given a presentation of the information in the reviewing guidelines. The 1 hour training session included a discussion of examples of sensitive and not-sensitive documents. Moreover, as part of the training session, participants were given time to review a batch of 8 practice documents, and discuss their reviewing decisions with the study coordinator, before the study began.

Participants were remunerated £7.50 per hour for taking part in the study. In total, including training times, each participant took between 10-12 hours, split over 2 separate sessions, to complete the study. There was a 30 minute refresher training session on the task and sensitivities at the beginning of the second session. Reviewing for sensitivity over a period of 10-12 hours requires a reasonable amount of effort from the participants. In line with the findings of McLean *et al.* (2001), participants were advised to take regular and frequent short breaks. As previously stated in Section 9.4.2, the reviewing interface was set up to not include time spent on breaks as part of the reviewing times.

### 9.4.5 Evaluation and Metrics

As previously mentioned in Section 9.4.1, 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); the length of time that it takes for a reviewer to sensitivity review a document (reviewing speed); and the amount of agreement between the reviewers' judgements and the classifier's predictions (reviewer-classifier agreement).

To evaluate the impact that the accuracy of sensitivity classification predictions has on the reviewers' performance, we compare how well reviewers perform on average for each of the clas-

sification treatments, *None*, *Medium* and *Perfect*. Therefore, we report the mean performance (calculated over all reviewers) for each of the classification treatments, in terms of reviewer accuracy, reviewing speed and reviewer-classifier agreement. However, when evaluating the effects of (simulated) classifier confidence, we evaluate the effects of the *Low*, *Medium* and *High* confidence levels over the *Medium* and *Perfect* classification batches combined (i.e., 14 documents with *Low* confidence predictions, 12 documents with *Medium* confidence predictions and 14 documents with *High* confidence predictions). We do this since (1) the distributions of *Low*, *Medium* and *High* simulated confidences are the same in the *Medium* and *Perfect* effectiveness batches and (2) it does not make sense to evaluate the classifier confidence for the classification effectiveness *None*, i.e., when there are no predictions provided. We evaluate how the classifier's simulated confidence affects the participants' performance in terms of reviewing speed and reviewer-classifier agreement.

When evaluating the reviewer accuracy, we select BAC and $F_2$ as our metrics. We select BAC since it provides a reliable accuracy score for performance on both classes (i.e., sensitive and not sensitive) when the distribution of classes is heavily skewed, as is the case in this study with only 20% of the documents being sensitive. Moreover, BAC is easily interpretable since 0.5 BAC indicates randomness. We also select $F_2$ since, as previously explained in Chapter 7, it puts more emphasis on correctly identifying sensitive documents and reflects the fact that, in sensitivity review, there are more severe consequences from not identifying a sensitive document than there are from falsely judging a document to be sensitive.

When evaluating the participants' reviewing speeds, we use Normalised Processing Speed (Damessie *et al.*, 2016) (NPS). NPS is a measure of reviewing speed that controls for the effects of varying reading speeds between reviewers and document lengths. NPS is related to the Normalised Dwell Time (Damessie *et al.*, 2016) (NDT) metric that we used in Chapter 8 since NPS is defined as:

$$\frac{|d|}{exp(\log(time) + \mu - \mu_\alpha)} \tag{9.1}$$

where $|d|$ is the document length, measured in number of words, and $exp(\log(time) + \mu - \mu_\alpha)$ is the NDT metric. As such, $\log(time)$ is the natural logarithm of the time taken to review $d$, $\mu_\alpha$ is the mean $\log(time)$ for the reviewer who reviewed $d$, calculated over a particular treatment condition, and $\mu$ is the global mean $\log(time)$ calculated for all reviewers over all documents.

When presenting our findings from the study in Section 9.5, we plot the participants' performance to show the mean participant score (e.g., in terms of BAC or NPS) and 95% confidence intervals. To calculate confidence intervals, we use the Cousineau (2005) update of the Loftus & Masson (1994) method, with the Morey (2008) correction. This method, commonly known as the Cousineau and Morey method, is specifically suited to within-subject study designs. Using the Cousineau and Morey method, we would expect that 5 out of 6 participants would be included in this interval in a replication study (Cumming & Maillardet, 2006). Importantly, the Cousineau and Morey method enables the reader to use the *rule of eye* to evaluate the signifi-

cance of the results from the plots, i.e., we can expect $p < 0.01$ for non-overlapping intervals and $p < 0.05$ when two intervals overlap by <50%. To calculate statistical significance, we use a one-way repeated measures univariate ANOVA in pair-wise comparisons between treatment conditions, e.g., *Medium* classification effectiveness vs. *Perfect* classification effectiveness. We select $p < 0.05$ as our significance threshold.

## 9.5 Results and Evaluation

In this section, we present our findings from the user study that we presented in Section 9.4. The study evaluates how sensitivity classification can assist sensitivity reviewers when there are sufficient reviewing resources to sensitivity review all of the documents in a collection that is to be publicly archived, i.e., in the scenario of an exhaustive review. Firstly, in Section 9.5.1 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; (b) agree with other reviewers' judgements more often; and (c) agree with the classifier's predictions more often. Secondly, in Section 9.5.2, 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 simulated confidence increases; and (b) make quicker reviewing decisions when they agree with the classifier.

### 9.5.1 The Impact of Classification Effectiveness on Reviewer Performance

To evaluate the impact of the classification effectiveness on the reviewers' performance, we compare the mean reviewer performance, in terms of reviewer accuracy, reviewing speed and reviewer-classifier agreement, for each of the classification effectiveness levels *None*, *Medium* and *Perfect*. Firstly, 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 9.2(a) presents the mean participant performance in terms of their Balanced Accuracy (BAC) for each of the levels of classification effectiveness, while Figure 9.2(b) presents the analogous participant performance in terms of $F_2$.

From Figure 9.2(a), 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 effectiveness (+38%) and 0.8 BAC when the classification predictions agree perfectly with the ground truth (+16% compared to *Medium*, +60% compared to *None*). Importantly, 0.5 BAC indicates that, on average, without classification predictions the participants' judgements were effectively random[5]. This is

[5]Although BAC 0.5 shows that on average the non-expert reviewers' judgements were effectively random, in terms of $F_2$, 0.4166 indicates random on the collection used in the study. Figure 9.2(b) shows that reviewers achieved

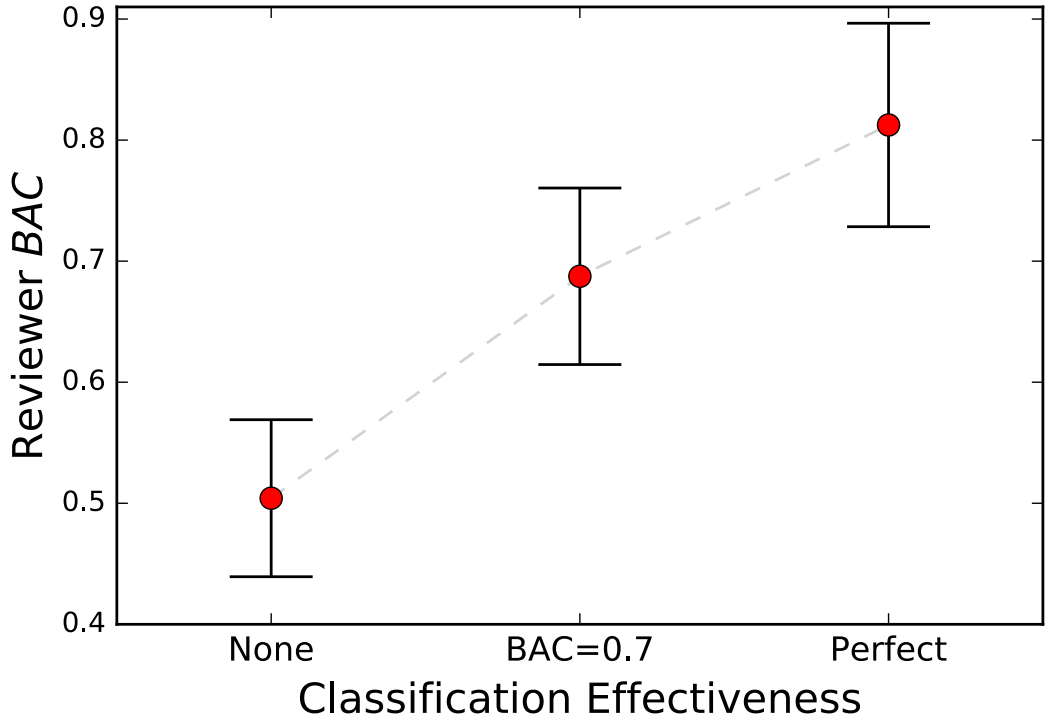
indicative of the difficulty of the sensitivity reviewing task, and underlines why government departments have typically employed expert reviewers for the task (The National Archives, 2017).

From Figure 9.2(a), we also note that for the *Medium* classification effectiveness treatment, the mean participant performance is almost equivalent to the level of classification effectiveness (participants = 0.69 BAC, classifier = 0.7 BAC). However, the participants only achieved an accuracy of 0.8 BAC when they were provided with perfect classification predictions. This shows that even when the classifier is perfect (i.e., its predictions are the same as the ground truth), reviewers still disagree with the classifier. We will discuss the reviewer-classifier agreement in more detail at the end of this section. However, none of the participants completely agreed with the classifier when its predictions were the same as the expert reviewers' judgements.

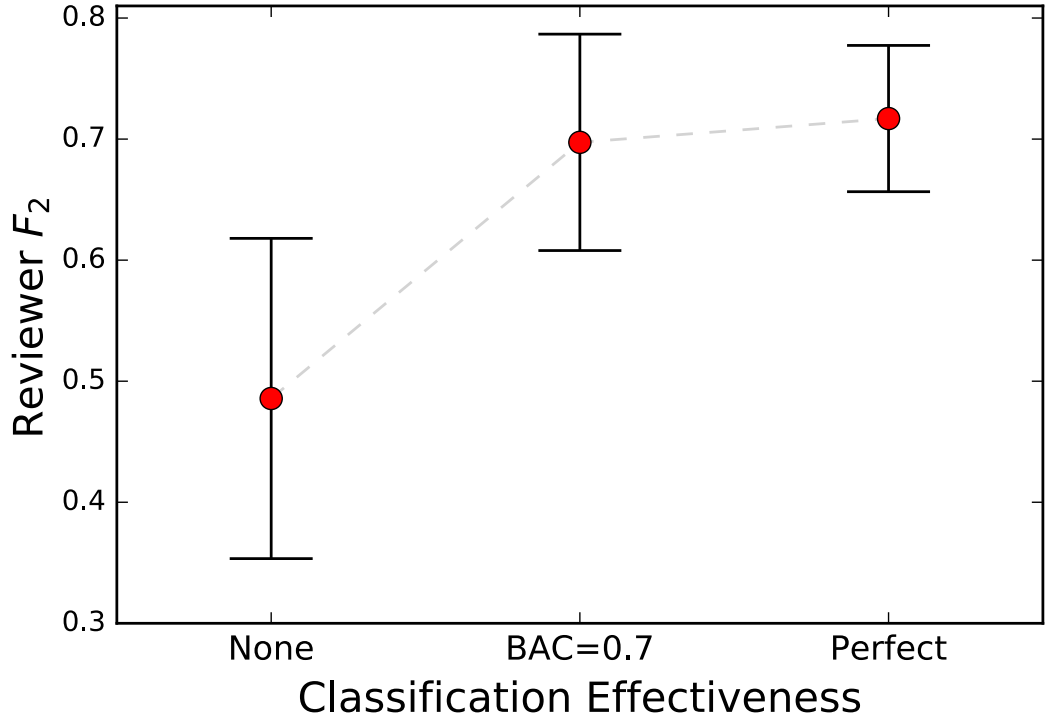
Turning our attention to Figure 9.2(b), which presents the participant performance in terms of $F_2$, we note that the relative mean participant performance increase is much greater between the no classification and the *Medium* classification effectiveness than between the *Medium* and *Perfect* classification. The ANOVA test of mean participant performance between no classification and the *Medium* classifier effectiveness shows significant improvements, both in terms of Balanced Accuracy (BAC) $[F(1,7) = 23.528, p = 0.002]$ and $F_2$ $[F(1,7) = 7.936, p = 0.026]$. However, comparing the participants' performance improvements between the *Medium* and *High* classifier prediction accuracy, the ANOVA test shows significant improvements in terms of BAC $[F(1,7) = 6.377, p = 0.040]$ but not in terms of $F_2$ $[F(1,7) = 0.560, p = 0.479]$. This finding shows that the main increase in participant performance between the *Medium* and *Perfect* classification effectiveness came as a result of participants making more True Negative judgements, since 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.

In response to hypothesis **H1**(a), which states that improved classifier effectiveness will lead to reviewers making more correct judgements, we conclude that improved classification effectiveness does indeed lead to a significantly improved performance of the participant reviewers. We observed significant improvements in the reviewers' accuracy in terms of BAC when the reviewers are provided with predictions from a classifier with an effectiveness of 0.7 BAC and further significant improvements when the classifier's predictions are perfect. In terms of $F_2$, we observed significant improvements in the reviewers' accuracy when the reviewers are provided with predictions from a classifier with 0.7 BAC effectiveness, compared to when the reviewers were not provided classification predictions. Importantly, we note that providing reviewers with sensitivity predictions from a classifier with an accuracy of 0.7 BAC, which is in line with the 0.7149 BAC achieved by our enhanced sensitivity classification approach that we presented in Chapter 6 (Table 6.9), led to significant improvements in reviewer accuracy in terms of BAC $[F(1,7) = 23.528, p = 0.002]$ and $F_2$ $[F(1,7) = 7.936, p = 0.026]$. This shows that our pro-

0.49 $F_2$ without classification predictions. The $F_2$ score, along with the BAC score, shows that, overall, the reviewers were more accurate when reviewing sensitive documents than non-sensitive ones. However, they over-predicted the amount of sensitivity in the collection.




(9.2(a)) Mean reviewer Balanced Accuracy (BAC).




(9.2(b)) Mean reviewer $F_2$.

Figure 9.2: Mean reviewer accuracy (in terms of BAC and $F_2$), with 95% confidence intervals, for each classification treatment: *None*, *Medium* (0.7 BAC) and *Perfect*.

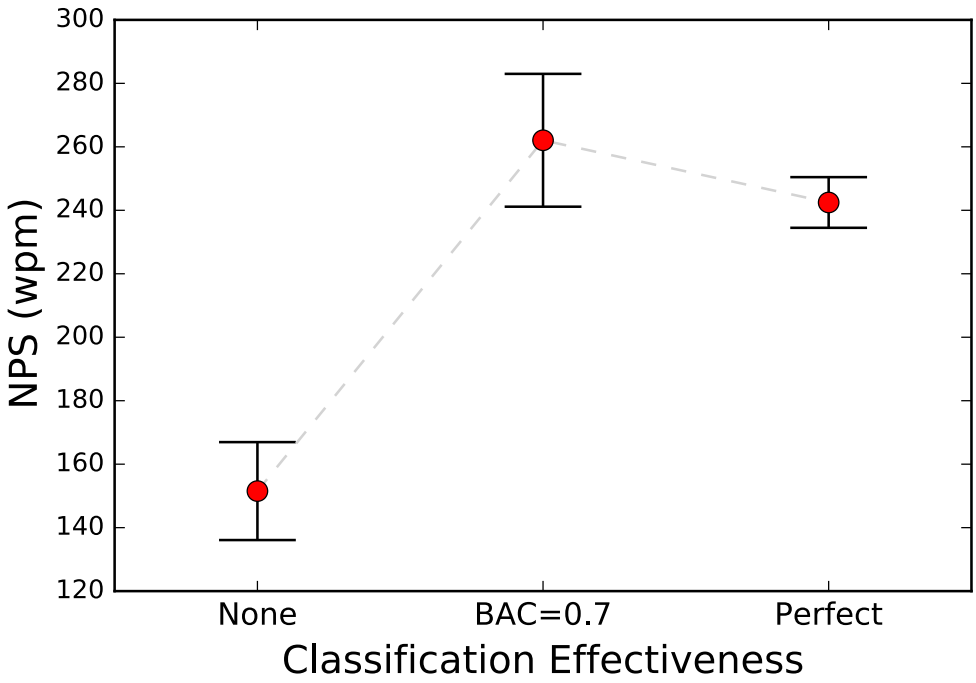



Figure 9.3: Normalised Processing Speed (NPS) (words per minute), with 95% confidence intervals, for each classification treatment: *None*, *Medium* (0.7 BAC) and *Perfect*.

posed sensitivity classifier can assist reviewers to make more accurate reviewing decisions. We note, however, that in this study there appears to be diminishing gains in reviewer performance improvements as the classification effectiveness increases. We argue that the identification of a threshold above which the classification effectiveness does not further enhance the reviewers' accuracy would be a valuable future research direction for digital sensitivity review research.

Turning our attention to hypothesis **H1**(b), which tests if more effective classification predictions will result in the reviewers processing document faster on average. Figure 9.3 presents the participants mean Normalised Processing speed (NPS) (defined in Eq. 9.1), in words per minute (wpm), for each of the levels of classification effectiveness. As can be seen from Figure 9.3, 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% to 260 wpm, when the classifier predictions have an accuracy of 0.7 BAC. The one-way ANOVA between *None* and *Medium* classification shows that this is a significant result, $[F(1,7) = 79.549, p = 0.0001]$.

Interestingly, we note from Figure 9.3 that the mean reviewing speed is slightly less when reviewers are provided with classification predictions that agree perfectly with the ground truth, 260 wpm (0.7 BAC) vs 244 wpm (Perfect). The one-way ANOVA between *Medium* and *Perfect* classification shows that this decrease is not significant ($[F(1,7) = 4.210, p = 0.079]$). The significant gains in reviewing speeds from providing classification predictions are sustained over both levels of classification predictions accuracy. Therefore, in response to **H1**(b), we conclude that, providing reviewers with classification predictions leads to significant increases in review-

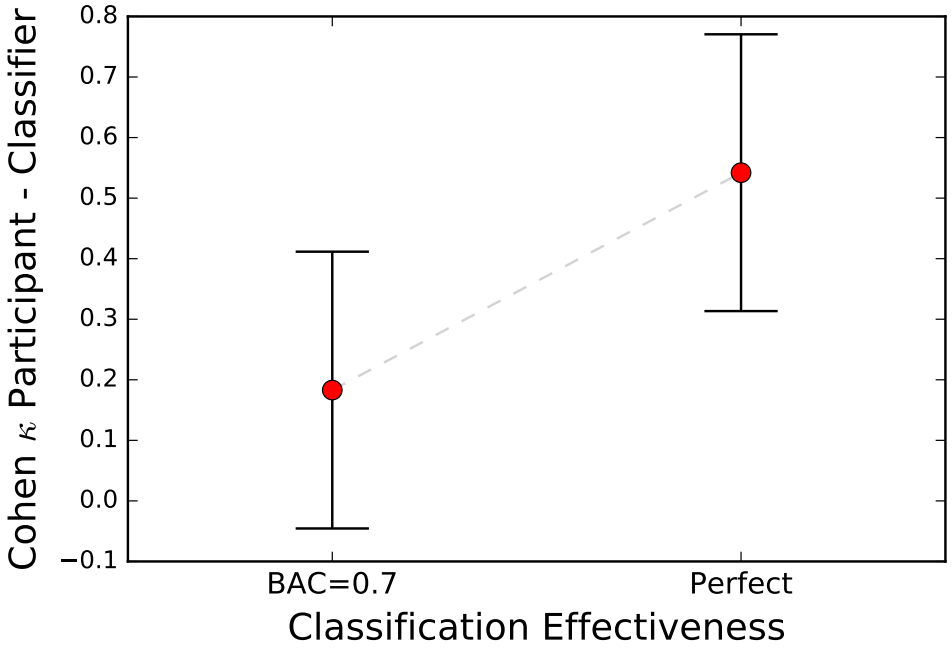



Figure 9.4: Cohen $\kappa$ participant classifier agreement, with 95% confidence intervals, for each classification treatment: *None*, *Medium* (0.7 BAC) and *Perfect*.

ing speeds. However, the observed increased reviewing speeds do not continue to increase when the classifier predictions agree perfectly with the ground truth. This result is somewhat counter-intuitive since, as we will discuss in the following paragraph, there is more agreement between the reviewers' judgements and the classifier's predictions when the classifier is perfect. The participants in the study were not informed of the accuracy of the classifier's predictions. We hypothesise that the reviewers expect that a classifier will not be 100% accurate and when the classifier is perfect the reviewers spend some additional time looking for the classifiers mistakes. However, it will require a further user study to investigate this hypothesis.

Finally for **H1**, we turn our attention to **H1**(c). Figure 9.4 presents the mean Cohen $\kappa$ agreement between participants and the classifier's predictions for each of the classification effectiveness levels. As can be seen from the figure, on average, participants do indeed agree with the classifier more when the classifier is more effective. Moreover, from performing a one-way ANOVA, we can see that this increase in agreement is significant $[F(1,7) = 6.897, p = 0.034]$. Therefore, in response to hypothesis **H1**(c), we conclude that the participant-classifier agreement does indeed increase with an improved classification effectiveness. Viewed on its own, this result might seem somewhat expected. However, we believe that the level of agreement between a reviewer and the classifier will have important implications for the speed of review (since disagreeing with a classification prediction could lead to a reviewer doubting their decision) and, therefore, for **H2**, which we examine in the next section.

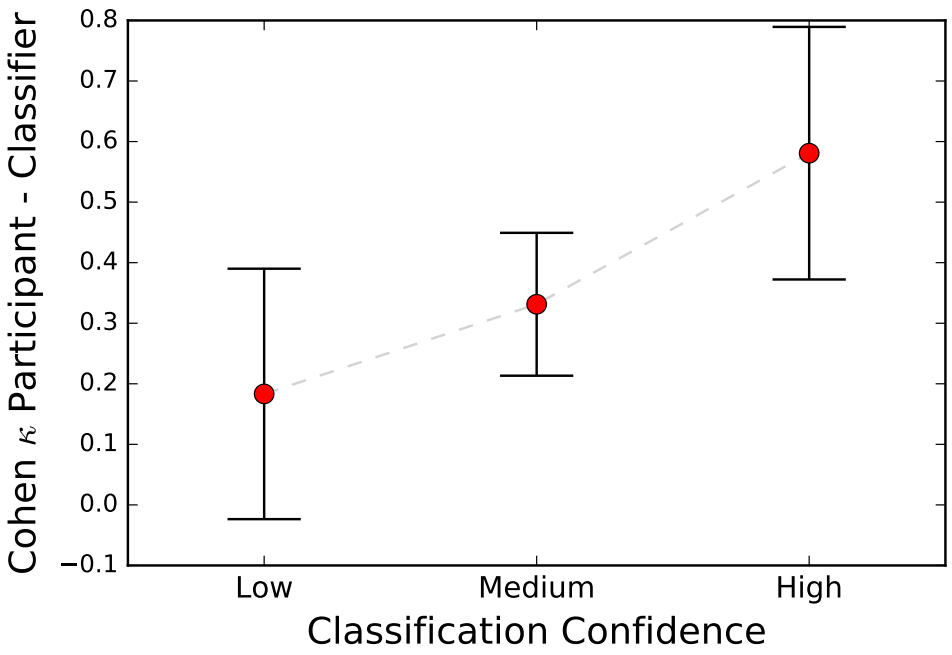



Figure 9.5: Cohen's $\kappa$, and 95% confidence intervals, for participant and classifier agreement for each simulated confidence level; Low, Medium and High.

### 9.5.2 The Impact of Classifier Confidence on Reviewer Performance

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

Firstly, addressing **H2**(a), Figure 9.5 presents the mean Cohen's $\kappa$ scores for the agreement between participants and the classification predictions for each of the classifier's simulated 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 simulated confidence level increases. The observed increase in agreement between the *Low* and *Medium* levels of classifier simulated confidence is not significant ($[F(1,7) = 2.631, p = 0.149]$). However, the ANOVA test between the *Medium* and *High* classifier simulated confidences shows that this relative increase in agreement is significant $[F(1,7) = 7.247, p = 0.031]$. Therefore, in response to hypothesis **H2**(a), we conclude that we do indeed observe significantly increased agreement with the classifier when the classifier claims to be confident about its predictions.

Turning our attention to hypothesis **H2**(b), which states that reviewers will review faster when they agree with the classifier and the difference in the reviewing speeds when the reviewer does or does not agree with the classifier will increase as the classifier's (simulated) confidence increases. Figure 9.6 presents the mean participant Normalised Processing Speed (NPS) for each

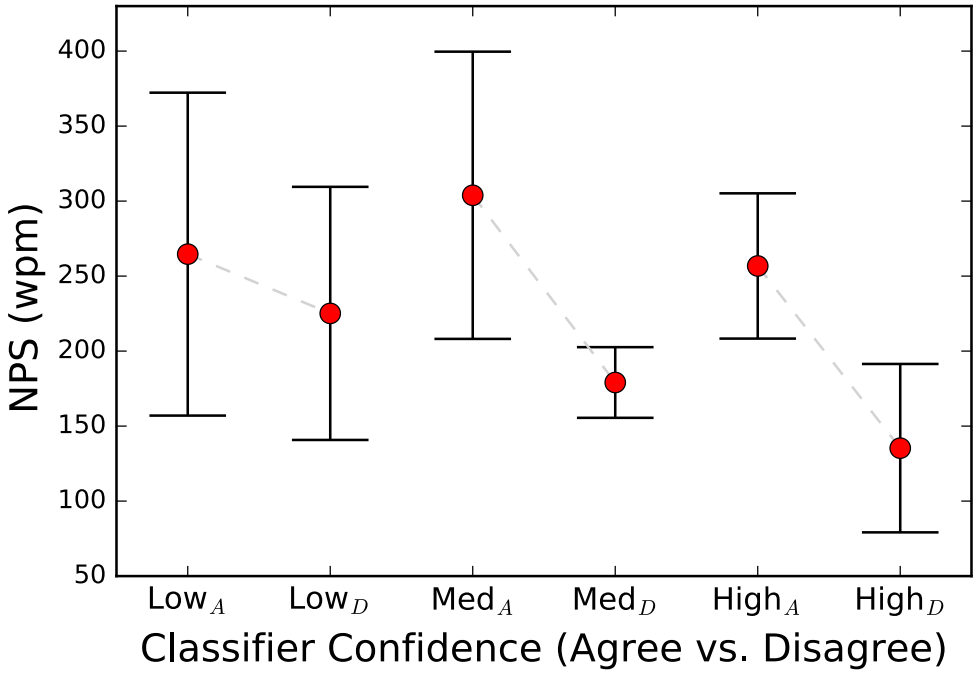



Figure 9.6: Normalised Processing Speed (words per minute), and 95% confidence intervals, for when participants agree (subscript $A$) or disagree (subscript $D$) with the classifier, for each of the classifier's simulated confidence levels: Low, Medium and High.

of the classifier simulated confidence levels *Low*, *Medium* and *High*, when participants either agree (subscript $A$) or disagree (subscript $D$) with the classifier's predictions. We note, from Figure 9.6, that there is a clear trend that reviewers do indeed review documents faster when they agree with the classifier's predictions. This trend holds over all levels of classifier simulated confidence, as is illustrated in Figure 9.7. An ANOVA shows that the trend is statistically significant when the classifier's simulated confidence is either *Medium*, $[F(1,7) = 12.662, p = 0.009]$, or *High* $[F(1,7) = 16.507, p = 0.005]$. The dashed lines in Figure 9.6 show the difference (or change) in reviewing speeds when the reviewers either agree or disagree with the classifier, for each of the classifier simulated confidence levels *Low*, *Medium* and *High*. As can be seen from the direction (or gradient) of the lines, the difference in reviewing speeds does indeed increase between the *Low* and *Medium* simulated confidence levels and is sustained when the classifier has *High* simulated confidence.

These observations validate hypothesis **H2**(b). However, surprisingly, Figure 9.6 shows that overall, in our study, there is a slight decrease in reviewing speeds when the classifier's simulated confidence level is *High*. This decrease in reviewing speeds is not significant, as confirmed by an ANOVA, which reports $[F(1,7) = 3.513, p = 0.103]$ (*Medium* vs. *High* simulated confidence). However, this observation shows that, for a reviewer, disagreeing with a classification prediction when the classifier has a high level of (simulated) 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. Therefore, in summary, reviewing speeds are indeed

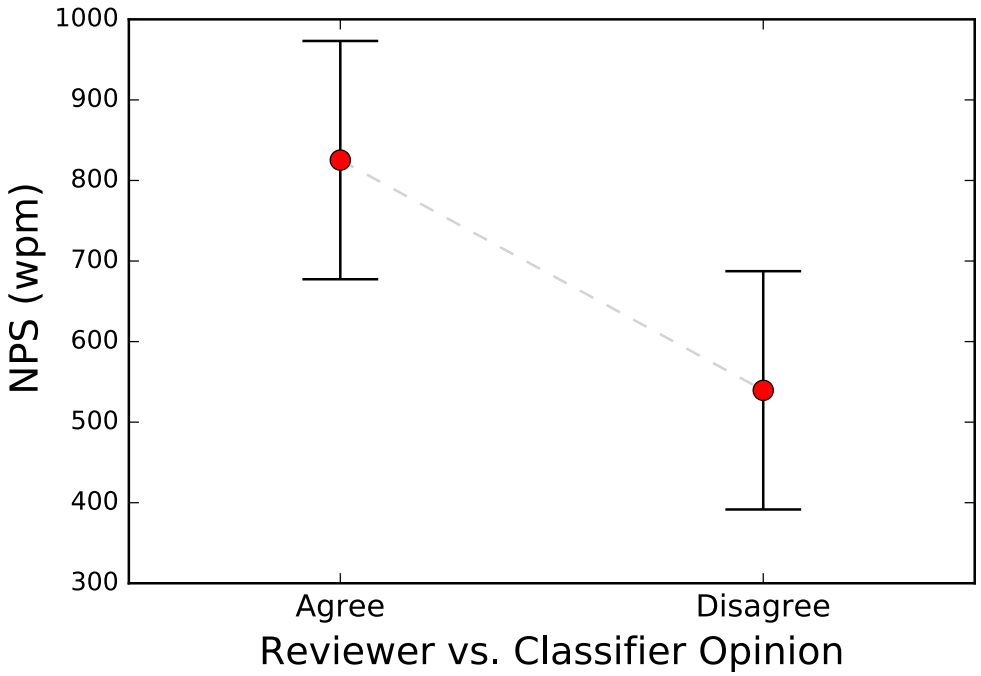



Figure 9.7: Normalised Processing Speed (words per minute) and 95% confidence intervals when participants either agree or disagree with the classifier predictions, for all judgements made on documents with associated classifier predictions.

Table 9.3: Summary table of hypothesis conclusions.

| | Treatment | Metric | Validated? | Source |
|---|---|---|---|---|
| H1(a) | Effectiveness | Increased judgement accuracy | ✔ | Figures 9.2(a) & 9.2(b) |
| H1(b) | Effectiveness | Increased reviewing speed | ✔ | Figure 9.3 |
| H1(c) | Effectiveness | Increased reviewer-classifier agreement | ✔ | Figure 9.4 |
| H2(a) | Confidence | Increased reviewer-classifier agreement | ✔ | Figure 9.5 |
| H2(b) | Confidence | Agreement=Increased reviewing speed | ✔ | Figures 9.6 & 9.7 |

significantly faster if the reviewer agrees with the sensitivity classification predictions when the classifier is confident. We note however, that, although we observe a clear trend that reviewers review faster when they agree with the classifier, in our study, this does not lead to an *overall* increase in reviewing speed when the classifier is confident. We hypothesise that when the classifier's (simulated) confidence is high reviewers appear to have taken more time to ensure that they fully reviewed the documents, so as not to rely solely on the classifier prediction, thereby reducing their reviewing speed. However, again, it will require a further user study to test this additional hypothesis.

### 9.5.3 Study Conclusions

As shown by the results reported in Sections 9.5.1 and 9.5.2, both of our stated hypotheses hold. Table 9.3 provides a summary of our hypothesis conclusions, and the sources of supporting evidence from Sections 9.5.1 and 9.5.2. In short, providing classification predictions to non-expert sensitivity reviewers increases the accuracy and the speed of the reviewers, as well as increasing their agreement with the classifier (**H1**). For **H2**, we found that reviewers agree with the classi-

fier more when it is confident and, moreover, their reviewing speed is enhanced when they agree with the classifier. We argue that, since both of our hypotheses hold, our findings in this study demonstrate that sensitivity classification predictions are a viable technology to effectively provide non-expert 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. Importantly, our findings suggest that governments may be able to increase the volume of digital documents that can be reviewed, while maintaining high levels of reviewing accuracy, if they increase the number of reviewers by recruiting less experienced reviewers (at less expense than expert reviewers) (as shown by this study) and assisting them with automatic sensitivity classification predictions. This, in turn, would enable the expert reviewers to focus on reviewing the more *high risk* documents, or the more difficult sensitivities where there is a higher level of disagreement.

We note though, that the results of this study are limited to the bennifits of sensitivity classification for *non-expert* sensitivity reviewers and that further work is required to evaluate the bennifits of sensitivity classification for *expert* reviewers when *all* of the documents in a collection will be manually sensitivity reviewed. In essence, this work demonstrates to governments and other stakeholders the importance of conducting further research and development in sensitivity classification, both to gain further insights about the additional benefits of improving the effectiveness of sensitivity classifiers and, also, to identify additional ways in which sensitivity classification can assist sensitivity review. Moreover, we suggest that more work needs to be done in evaluating different methods for portraying the classifier's decisions (i.e., transparency), and explaining how each decision was obtained. For example, there is a potential for passage classification approaches to highlight sensitivities, such as we presented separately in (McDonald *et al.*, 2015), and thereby further improve the accuracy and/or speed of sensitivity reviewers.

## 9.6 Conclusions

In this chapter, we investigated how sensitivity classification can assist sensitivity review in the second, and final, of our realistic sensitivity review scenarios that we investigate in this thesis, namely the *exhaustive review* user model (see Section 9.2). The exhaustive review user model addresses the scenario in which there are sufficient reviewing resources available to sensitivity review *all* of the documents in a collection that is to be publicly archived. In particular, we conducted a within-subject digital sensitivity review user study under laboratory conditions, to evaluate the benefits, and effects, of automatic sensitivity classification predictions for non-expert sensitivity reviewers (see Section 9.4). We found that providing the reviewers with sensitivity classification predictions resulted in significant improvements in the number of correct sensitivity judgements made by the participants in our study (see Figure 9.2), the speed that reviewer review documents (+72%) (see Figure 9.3), and the amount of agreement between the reviewer's

and the classifier's predictions (see Figure 9.4) (repeated measures ANOVA, $p < 0.05$). Our findings provide strong evidence showing that sensitivity classification is a viable technology to assist sensitivity review and can also enable governments to increase the number of digital documents that can be sensitivity reviewed, while maintaining high levels of reviewing accuracy, with non-expert sensitivity reviewers. Both of our hypotheses hold (see Section 9.3 and Table 9.3), and hence we argue that the results of our study demonstrate that the sensitivity classification is a viable technology to effectively provide reviewers with valuable information about the sensitivities within a collection. Moreover, sensitivity classification enabled the study participants to increase both their accuracy and their speed when conducting the sensitivity review task.

At this point, we argue that through our experiments and user studies that we have presented thus far we have validated the statement of this thesis. In particular, in Chapter 4, we presented our proposed framework for assisting digital sensitivity review. In Chapter 5, we showed that text classification is a viable approach for automatically identifying sensitive documents and, in Chapter 6 we showed that sensitivity classification can be enhanced by automatically identifying latent vocabulary, syntactic and semantic features of sensitive and non-sensitive information. In Chapter 7, we showed that we can reduce the amount of reviewing effort that is required to train an effective sensitivity classifier by having reviewers annotate, or redact, the sensitive information as they review, and identifying the most informative terms in the reviewers annotations. In Chapter 8, we showed how our proposed framework can be deployed to increase the amount of documents that can be opened to the public with the available reviewing resources. In this chapter, we have shown that sensitivity classification can assist sensitivity reviewers to make accurate reviewing decisions more quickly. Next, in Chapter 10, we will close this thesis by summarizing the conclusions and contributions from each of the chapters in the thesis and show how they validate our thesis statement. Moreover, we will discuss possible directions for future research uncovered by this work.

# Chapter 10

# Conclusions and Future Work

## 10.1 Contributions and Conclusions

The amount of digital information that is created has far outstripped the volume of information that was previously created in the paper-based world. Moreover, the collections of digital information that are being created are often an unorganised mix of banal, important and sensitive information (Oard *et al.*, 2016). Governments, and other public bodies, that are subject to freedom of information laws have a legal obligation to make their documents publicly available. However, the volumes and lack of structure in digital document collections makes the challenge of identifying the digital documents that should or should not be publicly archived currently unmanageable (Moss *et al.*, 2018). Moreover, identifying, and protecting, the sensitive information in these large digital collections is a challenging task. Therefore, the need for new technologies to assist with identifying sensitive information in digital government documents has been highlighted by governments, such as in the UK (The National Archives, 2016*a*) and the USA (Defense Advanced Research Projects Agency, 2010). In a 2015 report to former UK Prime Minister David Cameron, Sir Alex Allen (the Prime Minister's Advisor on Ministerial Standards) highlighted the need for such technologies, stating that "*The risk if this issue is not addressed satisfactorily is either that material will be released to TNA without proper review, leading to embarrassment when sensitive material is found to be in the public domain; or, perhaps as a reaction to the discovery of such releases, departments become risk averse and apply for blanket closures of records*". Moreover, this risk is increased in the digital world, since the likelihood that accidentally released sensitive information will be discovered by search technologies is much greater than that of paper documents being discovered on the shelves of the archive (Redwine *et al.*, 2013).

In this thesis, we addressed the problem of assisting human reviewers to sensitivity review born-digital government documents, to identify sensitive information that is exempt from public release through the Freedom of Information Act 2000 (c. 36) (digital sensitivity review). In tackling this problem, we proposed to address the challenge of automatically identifying sensi-

tive information as a document classification task, we refer to this as sensitivity classification. However, sensitivity classification on its own is not enough to effectively assist digital sensitivity review. To this end, we proposed a framework that builds upon sensitivity classification to assist digital sensitivity review by discovering the sensitivities within a collection, prioritising documents for review to increase the number of documents that can be released to the public and by providing sensitivity reviewers with useful information that can help the reviewers to accurately sensitivity review born-digital government documents faster.

In particular, we proposed to address sensitivity classification as a text classification task (see Chapter 5). Moreover, we empirically showed that the effectiveness of a baseline (text classification) sensitivity classifier can be improved by identifying the latent vocabulary, syntactic and semantic similarities that are in sensitive, or non-sensitive, documents within a collection (see Chapter 6). Furthermore, we proposed to deploy sensitivity classification within our framework for technology-assisted sensitivity review, to prioritise specific documents for review at different stages of the sensitivity review process. We showed that by prioritising specific documents to have reviewed we can identify the sensitivities in a collection and learn an effective classifier more quickly (see Chapter 7) or increase the number of documents that can be reviewed and released to the public with the available reviewing resources (see Chapter 8). Moreover, we have shown that sensitivity classification can assist reviewers to make accurate judgements more quickly (see Chapter 9). Through experiments and user studies addressing two realistic sensitivity review scenarios, we have drawn insights and concluded that digital sensitivity review could be assisted by automatic sensitivity classification to increase the accuracy, speed and agreement of sensitivity reviewers. Moreover, we concluded that our proposed framework could increase the number of digital documents that can be sensitivity reviewed and released to the public while using the same amount of reviewing resources.

In the remainder of this chapter, we firstly present the main contributions of this thesis in Section 10.1.1, before discussing the limitations of the work in Section 10.1.2. We summarise our main achievements and our conclusions that validate our thesis statement in Section 10.1.3 before, in Section 10.2, discussing some future research directions for IR, machine learning, the humanities and further afield, that arise from the digital sensitivity review task. Finally, we present our closing remarks in Section 10.3.

### 10.1.1 Contributions

In this section, we outline the main contributions of this thesis. We begin by listing the main contributions of this thesis to technology-assisted sensitivity review, before concluding the section by outlining our main contributions to the field of Information Retrieval.

The main contributions of this thesis to digital sensitivity review are as follows:

- In Chapter 4, we proposed a novel framework to assist government departments and human reviewers to sensitivity review digital government documents. The framework defines four components that work together and can be instantiated to provide different functionalities at different stages of the iterative sensitivity review process. In particular, the Document Representation component (see Section 4.4) identifies latent vocabulary, syntactic and semantic structures or relationships that can be reliable features for identifying documents that contain sensitive information. The Document Prioritisation component (see Section 4.5) identifies the documents that should be prioritised for review at any particular stage of the reviewing process. In the earlier iterations of the sensitivity review process, when the sensitivities in the collection are not known, the Document Prioritisation component can prioritise the documents that will provide the most information to the classifier so that an effective sensitivity classifier can be learned more quickly. When an effective sensitivity classifier has been learned, the Document Prioritisation component can prioritise documents that are not sensitive and that will require less time to review, to focus the reviewing resources on documents that will be released to the public and increase the number of documents that are opened to the public with the available reviewing resources. The Feedback Integration component (see Section 4.6) integrates explicit feedback from a reviewer about (1) the sensitivity judgements that a reviewer makes, to construct a representation of the sensitivities in a collection that is being sensitivity reviewed or (2) aspects about the reviewers interactions, such as the length of time that is required to review a document. Finally, the Learned Predictions component (see Section 4.7) combines the information that has been generated by the other three components. As the final step in each iteration of the sensitivity review process, the Learned Predictions component is responsible for making accurate sensitivity classification and expected reviewing time predictions for each of the documents that have not yet been sensitivity reviewed, thereby helping reviewers to accurately review documents more quickly and release more documents to the public. Our proposed framework is a key contribution of this thesis in that it defines the processes that are required to learn from sensitivity reviewers and to assist with digital sensitivity review.

- In Chapter 5, we showed that a state-of-the-art document sanitisation technique is not a viable approach for classifying sensitive information, as defined by the Freedom of Information Act 2000 (c. 36) (see Section 5.2). Instead, we proposed to address the task of automatically identifying documents that contain sensitive information as a text classification task. Moreover, we presented a thorough analysis of document representation and feature reduction techniques for a baseline sensitivity classifier that classifies sensitivity as a single category of information (see Section 5.3). Furthermore, we evaluated the effectiveness of classifying individual Freedom of Information Act 2000 (c. 36) exemptions using hand-crafted features of sensitivity (see Section 5.4) and an ensemble classifica-

tion approach for combining sensitivity classifiers (see Section 5.5.1). We showed that, on our collection, classifying sensitive information as a single category of information is an effective approach for sensitivity classification.

- In Chapter 6, we proposed an enhanced sensitivity classification approach that integrates *automatically* generated features of sensitive information. Our proposed sensitivity features identify latent structures or patterns, in the vocabulary (see Section 6.2), the syntax (see Section 6.3) and the semantics (see Section 6.4) of sensitive and not sensitive documents, that can be a reliable indication of the presence or absence of sensitive information. Moreover, and importantly, our proposed sensitivity classification features rely solely on distributional statistics of words within a collection and, therefore, the proposed approach is not constrained to being deployed within a specific government department (or for the specific sensitivities that we have investigated in this thesis). Furthermore, the terms within a document, or the collection of documents, that led to the classifier's sensitivity prediction can be identified using a simple heuristic, which is important for providing a level of transparency in the classifier's predictions (since governments will be held accountable for the reviewing decisions). We also evaluated a sequence classification technique for sensitivity classification using part-of-speech (POS) sequences. In particular, we evaluated kernel functions (see Section 6.3.1) sensitivity classification using POS sequence and ensemble classification approaches for combining sequence classification and text classification techniques for sensitivity classification (see Section 6.3.2). We empirically showed that POS sequence classification can be combined with text classification, by deploying a simple weighted majority vote ensemble and a linear (SVM) kernel for sequence classification, to improve the overall effectiveness of sensitivity classification.

- In Chapter 7, we proposed to reduce the amount of reviewing effort that is required to develop an effective sensitivity classifier by having a reviewer *annotate*, or *redact*, any passages of sensitive text in the documents that the reviewer judges to be sensitive. Moreover, we proposed to construct a representation of the sensitivities within the collection from the most informative terms in the reviewer's annotations, i.e., the vocabulary that is most informative about what that sensitivities in the collection *look like*. We evaluated four active learning approaches from the literature for selecting the most informative documents to have reviewed in each iteration of the sensitivity review process (see Section 7.3). Moreover, we evaluated three approaches for extending the active learning strategies to incorporate the sensitivity annotation features (i.e., the informative terms) from the reviewer's annotations (see Section 7.4). We showed that we can reduce the number of documents that need to be sensitivity reviewed to learn an effective sensitivity classifier by extending the *margin* active learning strategy to identify the most informative annotated (redacted) terms and giving them greater importance in the classifier.

- In Chapter 8, we investigated how our proposed framework can assist the sensitivity review process in the first of two realistic digital sensitivity review scenarios that we investigate in this thesis, namely the *limited review* user model (see Section 8.2). The limited review user model addresses a scenario in which there are insufficient reviewing resources available to review all of the documents in a collection that has been selected for transfer to the public archive. We proposed an approach for prioritising documents for review to maximise the number of documents that can be opened to the public with the available reviewing resources (see Section 8.4.1). In particular, we proposed an approach that uses sensitivity classification and the log data from reviewers to predict the amount of time that an average reviewer is likely to require to review a specific document (see Section 8.4). Moreover, our proposed approach increases the number of non-sensitive documents that are reviewed by prioritising the documents that are predicted to require the least amount of time to sensitivity review (see Section 8.5).

- In Chapter 9, we investigated how sensitivity classification can assist sensitivity reviewers in the second, and final, of our realistic sensitivity review scenarios that we investigated in this thesis, namely the *exhaustive review* user model. The exhaustive review user model addresses the scenario in which there are sufficient reviewing resources available to sensitivity review *all* of the documents in a collection that is to be publicly archived. We presented a controlled sensitivity review user study (see Section 9.4) that evaluated how two aspects of sensitivity classification predictions, namely: (1) the accuracy of the sensitivity classifier's predictions; and (2) the level of confidence that the classifier has in its individual predictions, affect the sensitivity judgements of (non-expert) human sensitivity reviewers. We evaluated how the two aspects of sensitivity classification affects three aspect of the reviewer's performance, namely (1) the number of documents that a reviewer correctly judges to contain, or to not contain, sensitive information (reviewer accuracy); (2) the length of time that it takes for a reviewer to sensitivity review a document (reviewing speed); and (3) the amount of agreement between the reviewers' judgements and the classifier's predictions (reviewer-classifier agreement) (see Section 9.5).

This thesis is focused on developing and evaluating approaches for assisting the task of digital sensitivity review. So far, in this section, we have discussed our main contributions to digital sensitivity review. In the remainder of this section, we discuss the four main computing science contributions of this thesis to the field of Information Retrieval (IR).

Firstly, in Chapter 6, we proposed to represent documents as sequences of parts-of-speech (POS) and frame the document classification task as a sequence classification (Xing *et al.*, 2010) task. Moreover, we showed that combining text classification with POS sequence classification, as a simple weighted majority vote ensemble approach, can be an effective approach for document classification (see Table 6.5). To the best of our knowledge this is a novel approach to

document classification that, we argue, is potentially appropriate for some other document classification tasks in which the grammatical or syntactic structure of the documents is a useful classification feature.

Secondly, in Chapter 6, we introduced a novel approach for identifying the terms that are most frequently associated with important classification features derived from word embedding document representations (Balikas & Amini, 2016; Collobert *et al.*, 2011; Mikolov, Chen, Corrado & Dean, 2013). Our proposed approach identifies the relative importance of terms in a collection by a four step process. Firstly, we rank the features of a linear classification model by the features' weights to identify the most important feature. Secondly, we identify the feature's value in the document representations of each of the documents that the model was trained on before, thirdly, tracing back to the word embedding representations of each of the terms in a document, to identify the term that is responsible for the value in the document representation. Lastly, we count the number of times that a term is responsible for a value in the important dimension of a document representation. Each time that a term is responsible for such a value we count it as a vote for the term's importance, or contribution to, the classification feature. We argue that our proposed approach is a useful technique to evaluate the relative importance of terms, in a document collection, when training a linear classification model from semantic document representations derived from word embeddings.

Thirdly, in Chapter 7, we showed that semi-automated text classification (Berardi *et al.*, 2012) (SATC) can be deployed as an active learning strategy. SATC assumes that there is an available classifier that is optimal with the available data but is not effective enough to meet a strict effectiveness threshold imposed by the task (e.g., due to the operational constraints of a company). The goal of SACT is to generate a ranking of documents so that, if a reviewer starts at the top of the ranking and proceeds down the ranking correcting miss-classified documents until the available reviewing time has expired, the overall accuracy of the classification predictions is maximised. We showed that by simply utilising the corrected classification predictions as explicit feedback for the classifier SACT can perform competitively as an active learning strategy, where the goal is to select for review the documents that are most informative for developing an effective classifier, to use minimal reviewing resources.

Lastly, in Chapter 8, we introduced a novel approach for predicting the amount of time that a reviewer will require to review and judge a document. In our proposed approach, we showed that predicting a reviewers Normalised Dwell Time (Damessie *et al.*, 2016) (NDT) can be an effective approach for predicting reviewing times. The NDT metric relies on the calculation of mean reviewer times which means that the metric cannot be *calculated* until after the documents have been reviewed. However, we showed that combining document complexity features with features of reviewing behaviour to train a linear regression model we can *predict* the NDT of a reviewer for a specific document. Moreover, we showed that our approach can result in effective (i.e. sufficiently accurate to be useful) reviewing time predictions.

### 10.1.2 Limitations of this Work

In this section, we discuss some of the limitations of our work that we present in this thesis.

Firstly, this thesis addresses the task of assisting human reviewers to sensitivity review digital government documents. However, we have only evaluated how to classify two categories of Freedom of Information Act 2000 (c. 36) sensitive information, i.e., Section 27: International relations and Section 40: Personal Information. These sensitivities are representative of the most frequent sensitivities within a UK government context (The National Archives, 2016*b*) but there remains a need for further experimentation to evaluate our proposed framework on other categories of Freedom of Information Act 2000 (c. 36) sensitivities. Moreover, our work is framed wholly within a UK government context. Many government departments, public bodies and organisations out-with of this context will be tasked with the challenge of reviewing digital documents to identify sensitive information. Moreover, within each context, the definitions and categories of sensitive information will vary. With this in mind, there is also a need for experimentation on sensitivities out-with the Freedom of Information Act 2000 (c. 36) context.

Secondly, the size of our test collection is relatively small compared to other benchmark text classification datasets. The collection is large enough that statistical significance tests can provide a good degree of confidence in the validity of the experiments that we have performed. However, the size of the test collection limits the breadth of failure analysis that can be performed on some of the individual subcategories of the sensitivities, e.g., the S27 subcategory Treaty (see Table 3.2). Moreover, the size of the test collection influenced our choice to present the results in Chapters 5 and 6 as the best performing settings from a parameter sweep of the variables that we evaluated. This choice is appropriate for this stage of our research. However, future work will have to investigate learning suitable and robust parameter settings for sensitivity classification.

Finally, in Chapter 9, the results of our Exhaustive Review user study are limited to evaluating the benefits of sensitivity classification predictions for non-expert reviewers. For the effectiveness of our framework to be fully evaluated, there is a need for further experimentation to evaluate the impact and benefits of providing sensitivity classification predictions to expert reviewers when all of the documents in a collection will be manually sensitivity reviewed by those expert sensitivity reviewers.

### 10.1.3 Conclusions

In this section, we summarise the main conclusions and achievements of this work. In particular, these conclusions validate the statement of this thesis proposed in Section 1.4 using the newly created test collection with expert sensitivity review judgements that we presented in Section 3.4.

***Effectiveness of Text Classification for Automatically Classifying Sensitive Documents***: In Chapter 5 we proposed to address the task of automatically identifying documents that contain sensitive information that is exempt from public release through the Freedom of Information Act 2000 (c. 36) as a text classification task. Moreover, we performed a thorough evaluation of document representation and feature reduction techniques to develop a strong baseline (text classification) sensitivity classification approach. Table 5.2 presented the document representation and feature reduction combinations that we evaluated. We showed that representing documents as TDF-IDF vectors resulted in statistically significant (McNemar's test $p < 0.05$) improvements compared to *tf* and BIN for eight combinations of feature reduction techniques (see Table 5.3). We also showed that, on our test collection, the advanced feature reduction techniques that we evaluated, Information gain (IG) and Chi-Squared ($\chi^2$), did not result in improved sensitivity classification effectiveness (see Table 5.3). Moreover, we showed that the most effective document representation and feature reduction combination for sensitivity classification (on our collection) is a TF-IDF document representation with retained stopwords and no stemming applied (denoted as $\text{TF-IDF}_{\text{stopNoSm}}$ in Table 5.3). We showed that the $\text{TF-IDF}_{\text{stopNoSm}}$ combination performed statistically significantly better than all the other document representations and feature reduction combinations that we evaluated (McNemar's test $p < 0.05$). Moreover, $\text{TF-IDF}_{\text{stopNoSm}}$ resulted in an +4.12% increase in $F_2$ and +2.2% increase in BAC compared to the next best document representation approach ($\text{BIN}_{\text{stopNoSm}}$) and a +4.8% increase in $F_2$ and +3% increase in BAC compared to the next best performing *basic* feature reduction combination ($\text{TF-IDF}_{\text{noSpNoSm}}$) (see Table 5.2). We also evaluated the effectiveness of classifying individual Freedom of Information Act 2000 (c. 36) exemptions, (as opposed to classifying sensitivity as a single category of information) and ensemble approaches for combining sensitivity classifiers. We showed that, on our collection, classifying sensitive information as a single category of information resulted in the most effective sensitivity classifier in terms of $F_2$ and BAC, but classifying individual Freedom of Information Act 2000 (c. 36) sensitivities ($\text{Individual}_{27/40}$) may be more appropriate in risk-averse situations, e.g., in specific government departments or for specific collections, since it results in the highest recall score (0.7490) (see Table 5.7).

***Effectiveness of Enhanced Sensitivity Classification with Vocabulary, Syntax and Semantic features***: In Chapter 6, we proposed to enhance sensitivity classification with automatically generated vocabulary, syntactic and semantic features of sensitivity. Moreover, we evaluated combining text classification and sequence classification techniques for sensitivity classification. We showed that combining POS sequence classification with text classification resulted in significant improvements in classification performance according to McNemar's test, $p < 0.05$, (1.5% $F_2$) (see Table 6.5). Moreover, we showed that extending sensitivity classification with additional vocabulary and semantic features ($\text{Text+TN}_7\text{+WE}_{\text{wp}}\text{+WE}_{\text{gn}}$(concat)) resulted in our sensitivity classifier achieving a 5% increase in $F_2$, correctly classifying ~6% more sensitive

documents than the text classification baseline (see Table 6.9).

***Effectiveness of Constructing a Representation of Sensitive Information with Reviewer Annotations***: In Chapter 7, we investigated active learning strategies for constructing a representation of the sensitivities in a collection to learn a sensitivity classifier while using less reviewing resources. Moreover, we investigated extending that active learning strategies with sensitivity annotation features. We showed that the addition of (+InfAnno) sensitivity annotation features enabled all of the document prioritisation strategies that we tested to correctly classify sensitive documents using markedly less reviewing effort (Figure 7.5). Moreover, we showed that the Margin document prioritisation strategy extended with +InfAnno annotation features reaches its peak classification performance (0.7 BAC) using significantly less reviewer effort than the Margin strategy without annotation features (according to the sign test, $p < 0.01$), requiring only 820 documents to be reviewed as opposed to 1700 when Margin is deployed without annotation features (see Figures 7.5(g) and (h)). This is a 51% reduction in amount of reviewer effort that is required to learn an effective sensitivity classifier.

***Effectiveness of Shortest Predicted Reviewing Time Prioritisation for Maximising Openness***: In Chapter 8, we proposed an approach for prioritising documents for review to increase the number of documents that can be opened to the public with the available reviewing resources. We showed that, in our experiments, for collections that are 60%-70% sensitive, our proposed document prioritisation strategy resulted in a 30% increase in the ratio of reviewed documents that are actually opened to the public, e.g., for a collection in which 70% of documents contain some portion of sensitive information our *shortest predicted reviewing time* ranking strategy resulted in an extra 200 documents being released for 100 hours of reviewing time on our simulated collection (see Figure 8.2).

***Effectiveness of Assisting Reviewers with Sensitivity Classification Predictions***: In Chapter 9, we presented our controlled sensitivity review user study, that was conducted under laboratory conditions. We found that providing non-expert sensitivity reviewers with sensitivity classification predictions resulted in a significant improvement in mean participant BAC scores as the effectiveness of the classifier increased, from 0.5 BAC when there are no classification predictions to 0.69 BAC (+38%) for medium classification effectiveness and 0.8 BAC (+16%) for perfect classification (Figure 9.2(a)). Moreover, we observed significant improvements in reviewer accuracy in terms of $F_2$ from a Medium effectiveness sensitivity classifier (Figure 9.2(b)). Providing non-expert reviewers with classification predictions also resulted in a 72% increase in reviewing speeds from 151 wpm to 260 wpm when the classifier predictions have an accuracy of 0.7 BAC (Figure 9.3).

***Validating our Thesis Statement***: The main claim of our thesis statement is that automatic sensitivity classification can be effective for assisting human reviewers with the sensitivity review of digital government documents. We argue that we have validated this claim in Chapters 8 and 9, where we showed that (1) sensitivity classification can be used within our proposed framework to increase the number of documents that can be reviewed and released to the public when there are insufficient reviewing resources to review all of the documents that are due to be reviewed (Chapter 8) and (2) providing the reviewers with sensitivity classification predictions can enable the reviewers to sensitivity review born-digital documents more accurately and more quickly (Chapter 9). Our thesis statement claims that an effective sensitivity classifier can be learned by identifying the latent vocabulary, syntax and semantic language features of the sensitive information in a corpus. We argue the we have validated this in Chapter 6, where we showed that enhancing text classification with vocabulary, syntax and semantic language features led to significant improvements in sensitivity classification effectiveness (according to McNemar's test, $p < 0.05$), either through combining text classification an sequence classification techniques (see Table 6.5) or by extending text classification with the additional features (see Table 6.9). Furthermore, our thesis statement claims we can reduce the number of documents that are required to be reviewed to learn an effective sensitivity classifier by deploying an active learning strategy to select specific documents to be reviewed and having a reviewer annotate, or *redact*, any passages of sensitive text in a document as they perform the review. We argue that we have validated this in Chapter 7, where we showed that identifying the most informative annotated terms and assigning them more importance in the classifier led to an effective sensitivity classifier being learned with significantly fewer documents being reviewed (sign test, $p < 0.01$)(see Figures 7.5(g) and (h)).

With respect to how our proposed framework can be deployed to assist with the sensitivity review of born-digital government documents, our thesis statement states that automatic sensitivity classification predictions can be used to prioritise specific documents for review to increase the number of non-sensitive documents that can be reviewed and released to the public within the available reviewing time budget. We argue that we validated this claim in Chapter 8, where we used sensitivity classification predictions as a feature of modelling (and predicting) the amount of time that a reviewer would require to sensitivity review a specific document (see Section 8.4) as input to our proposed shortest predicted reviewing time document prioritisation strategy to maximise the number of documents that can be opened to the public when there are insufficient reviewing resources available (see Section 8.5). Moreover, our thesis statement states that providing the reviewers with sensitivity classification predictions for the documents that are to be reviewed will enable the reviewers to accurately sensitivity review documents more quickly and the reviewer's agreement will be increased. We argue that we have validated this claim in Chapter 9 where we showed that providing reviewer sensitivity predictions did indeed lead to increased reviewer accuracy (see Figure 9.2), increased reviewing speeds (see Figure 9.3) and

increased agreement with the expert sensitivity reviewers judgements (see Figure 9.4). Therefore, we argue that we have validated all points of our thesis statement.

## 10.2 Directions for Future Work

In this section, we discusses possible directions for future research into sensitivity classification and assisting digital sensitivity review. We split our proposed future research directions into three categories. Firstly, we discuss future research directions that have become apparent as a direct result of the work that we have presented in this thesis. Secondly, we present future research directions for identifying sensitive information that we argue will be of interest to, and are related to, research from the wider IR community. Lastly, we present future sensitivity research directions that will be of interest to the IR and wider scientific community, such as the humanities.

**Research Directions Arising from This Work**:
*Limit of sensitivity predictions for Assisting Reviewers*: In Chapter 9, we saw that, once the sensitivity classifier reaches a certain level of effectiveness, there appears to be diminishing returns in the amount of benefit that sensitivity reviewers get from providing them with predictions about which of the documents in the collection contain sensitive information (we observed that none of the reviewers in our study agreed with the classifier's predictions 100% of the time, even when the classifier's predictions are the same as the expert reviewers gold standard). We argue that studying this relationship further to identify a threshold value, above which it is likely that improving the effectiveness of the classifier may not result in a similar increase in reviewing speed, accuracy and/or agreement would be a valuable future research direction for digital sensitivity review research. We argue that this would be of benefit to government departments and researchers for evaluating where to focus their efforts in the development of approaches for assisting digital sensitivity review.

*Identify Documents That Should be Re-reviewed*: Currently, each central government department has a different policy and procedure for double-checking the sensitivity judgements of reviewers (Allan, 2014). For example, some government departments employ senior reviewers to double-check the reviews from a sample of the reviewed documents. However, many paper documents that are sensitivity reviewed are only assessed by a single reviewer. As we have previously discussed in Chapter 9, when assessing the inter-assessor agreement of expert reviewers in the sensitivity review of born-digital documents we observed only moderate agreement between reviewers (Cohen's $\kappa$ of 0.55 for 150 documents assessed by two reviewers and a Fleiss' $\kappa$ of 0.44 for 50 documents assessed by four reviewers). This suggests that, historically, there

have been documents being released to the public that would have been judged as being sensitive if they had been reviewed by a different sensitivity reviewer. When paper documents are released to the public they are most likely to be accessible only to people who physically go to the archive to look at them. However, the digital documents that are released will be indexed and made available on-line, where anyone with access to a computer will be able to find them. Therefore, with digital documents, it is more likely that sensitive documents that are inadvertently released to the public will be discovered. However, the transition to digital documents also potentially makes it easier to address this problem. We argue that analysing the sensitivity judgements from many reviewers and many government departments will make it easier to identify documents that should be re-reviewed to safe guard against the inadvertent release of sensitive information that is missed by the initial review, and this would be a valuable direction for future technology-assisted sensitivity review research.

*Referring documents to other departments for review*: Currently, sensitivity reviewers initiate the referral of a file to other departments or agencies whenever the reviewer judges that a view on potential sensitivities should be sought from another department. Similarly to the previous point, we argue that this process could potentially be assisted by analysing sensitivities from multiple government departments to automatically identify inter-department sensitivities.

**Research Directions for the IR Community**:
*Transparency and Accountability*: The 3rd Strategic Workshop on Information Retrieval (SWIRL) (Culpepper *et al.*, 2018) was held recently to explore the long-range issues of the Information Retrieval field and to build consensus on some of the key challenges that face the IR community. The SWIRL 2018 participants identified Fairness, Accountability, Confidentiality, and Transparency (FACT) to be one of the three most important topics for discussion at the workshop. The problems and tasks that are associated with sensitive information, and assisting governments to adhere to freedom of information laws, are a great application of future research in making machine learning, search and classification technologies more fair, accountable, confidential and transparent. Algorithms that make decisions for governments need to have a good level of transparency since governments will be held accountable for their decisions. Moreover, it is widely thought within governments that the algorithms that governments use to make decisions will themselves become part of the public record. Therefore, future research on the FACT of machine learning algorithms will be of crucial to assisting digital sensitivity review.

*Mosaic Sensitivities*: In our discussions with professional sensitivity reviewers, we have often been informed that many sensitivities only become apparent when the information in two or more documents is combined, these sensitivities are known as *mosaic* sensitivities. The probability ranking principle in IR assumes that a document's relevance is independent of the other

documents in a collection (Maron & Kuhns, 1960). This assumption has been challenged and shown to be incorrect in certain circumstances (Robertson, 1977). Indeed, in some task-specific scenarios, such as in search results diversification Santos (2013), there has been progress made in relaxing this independence assumption. However, this has mostly been from that view that if you have already identified a relevant document for a specific topic then the relevance of another document covering the same topic might be changed. Future research into automatically identifying sensitive information will need to address the independence assumption to tackle the fact that a document may not be relevant (i.e., sensitive) on its own but becomes relevant only when the information in the document is combined with (an)other document(s). For example, Moss & Endicott-Popovsky (2015) discusses the fact that email threads often splinter into multiple disconnected threads and it is only when all of the information from each of the email threads are combined that the sensitive nature of the information becomes apparent. Therefore, sensitivity identification research could benefit from a test collection being constructed with a ground truth of sensitivities that span multiple documents to identify mosaic sensitivities.

*Reviewing order*: Following from the previous point, the order in which documents are sensitivity reviewed can potentially affect a reviewers judgements. For example, a sensitivity may be contained within a single document but it may not become apparent until the reviewer has information from another document. This problem also challenges the independence assumption of the probability ranking principle (Maron & Kuhns, 1960). However, in this problem, the interrelated relevance of documents is not explicit since the sensitivity is contained within a single document. This also could be a useful direction for future research.

**Research Directions for the Wider Scientific Community**:
*Integrating Policy Changes into Assistive Technologies*: Government policies and a country's laws change and evolve over time. It will be important that any predictive technology for assisting digital sensitivity review, e.g., sensitivity classification, can also easily change and adapt to newly implemented policies and laws. Solving this problem will, however, clearly require more communities to be involved than just the IR / machine learning community, such as the political sciences community.

*Adapting to Forget what has Been Learned*: Sensitivity evolves over time. Information that is considered to be sensitive today will most likely not be sensitive at some point in the future. Algorithms that automatically predict sensitivity will need to be able to adapt to current and future sensitivities. There has already been a lot of research done on this within the legal community (e.g., (Rosen, 2011; Walker, 2012)) initially due to the *right to be forgotten* legal case of Google Spain SL, Google Inc v Agencia Española de Protección de Datos, Mario Costeja González (2014) (Kalis, 2014), and more recently due to the resulting ruling being enshrined

in the Regulation (EU) (2016/679)(Art. 17(2)). We argue that researching best practices for addressing this phenomenon from a computing science perspective will be a valuable direction for future research for assisting digital sensitivity review and other IR tasks that are concerned with handling sensitive data.

## 10.3 Closing Remarks

In this thesis we have addressed a new and challenging task, namely automatically classifying documents that contain context-dependent sensitive information to assist with the sensitivity review of digital government documents. Classifying context-dependent sensitive information is a challenging task for a number of reasons. For example, sensitivity is not usually topic-oriented. Sensitivity does not usually arise from the fact that a document it about a particular topic or subject. More often, it is what is said about the topic or subject, or by whom it is said, that makes the information sensitive (Moss & Gollins, 2017). Moreover, sensitivity is broadly and vaguely defined. For example, sensitive information relating to the Freedom of Information Act 2000 (c. 36) Section 27: *international relations* (Ministry of Justice, 2008*a*) defines information to be sensitive "if its disclosure would, or would be likely to prejudice: relations between the United Kingdom and any other state; relations between the United Kingdom and any other international organisation or international court; the interests of the United Kingdom abroad; or the promotion or protection by the United Kingdom of its interests abroad". International relations sensitivities can, therefore, relate to personal, institutional, political or security matters. Furthermore, potential sensitivity of information, as defined by the Freedom of Information Act 2000 (c. 36), is dependent upon the likely effect of the information being released to the public.

We have argued that sensitivity classification can be deployed to assist human reviewers perform the sensitivity review of digital government documents. We proposed a novel framework for technology-assisted sensitivity review. The basis of our proposed framework is an effective sensitivity classifier that identifies latent vocabulary, syntactic and semantic patterns or relations in sensitive and non-sensitive documents to reliably classify documents that contain sensitive information. We showed that sensitivity classification can be of benefit for assisting human sensitivity reviewers with the sensitivity review of digital government documents, by increasing the speed, accuracy and agreement of sensitivity reviewers and increasing the number of documents that can be released to the public with a limited amount of reviewing resources.

We have made progress in addressing some of the main challenges in assisting government departments perform digital sensitivity review. However, there are many remaining, interesting and challenging, tasks that need to be addressed in the area of identifying sensitive information in large digital collections (some of which we highlighted in Section 10.2). In our various discussions with different stake holders throughout the course of this work, it has become apparent that

the problem of reviewing large collections of born-digital documents to identify context-specific sensitive information is a problem that is being faced not only by government departments but also by various public bodies, such as the police, and intergovernmental organisations, such as the North Atlantic Treaty Organisation (NATO). We argue that this will continue to be an increasingly important field of future research.

# Bibliography


Abril, D., Navarro-Arribas, G. & Torra, V. (2011). On the declassification of confidential documents. *In* 'Proceedings of the 8th International Conference on Modeling Decisions for Artificial Intelligence'. Springer. pp. 235–246. 3.3

Ackerman, J. M. & Sandoval-Ballesteros, I. E. (2006). The global explosion of freedom of information laws. *Admin. L. Rev.* **58**, 85. 1.2

Aggarwal, C. C. & Zhai, C. (2012). A survey of text classification algorithms. *In* 'Mining text data'. Springer. pp. 163–222. 2.2.2, 2.2.3

Agrawal, R. & Srikant, R. (2000). Privacy-preserving data mining. *In* 'ACM Sigmod Record'. Vol. 29. ACM. pp. 439–450. 3.3

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv:1707.02919 [cs.CL]*. 5.3

Allan, S. (2015). Government digital records updated and archives review. Cabinet Office, UK Government. https://www.gov.uk/government/publications/government-digital-records-and-archives-review-by-sir-alex-allan. Last accessed on 12-02-2018. 1.1, 1.2, 2, 4.2

Allan, S. A. (2014). Records review. Cabinet Office and The National Archives. https://www.gov.uk/government/publications/records-review-by-sir-alex-allan. Last accessed on 12-02-2018. 1.2, 4.3, 8.1, 8.2, 9.4.1, 10.2

Aphinyanaphongs, Y., Fu, L. D., Li, Z., Peskin, E. R., Efstathiadis, E., Aliferis, C. F. & Statnikov, A. (2014). A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. *Journal of the Association for Information Science and Technology* **65**(10), 1964–1987. 2.2.2, 2.2.3

Balikas, G. & Amini, M.-R. (2016). An empirical study on large scale text classification with skip-gram embeddings. *arXiv:1606.06623 [cs.CL cs.IR]*. 6.1, 6.4, 6.4, 6.5.2, 10.1.1

Baron, J. R., Lewis, D. D. & Oard, D. W. (2006). TREC 2006 legal track overview. *In* 'Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA,

November 14-17, 2006'. Vol. Special Publication 500-272. National Institute of Standards and Technology (NIST). 2.4

Berardi, G., Esuli, A. & Sebastiani, F. (2012). A utility-theoretic ranking method for semi-automated text classification. *In* 'Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 961–970. 7.1, 7.3, 7.3.2, 7.6, 10.1.1

Berardi, G., Esuli, A., Macdonald, C., Ounis, I. & Sebastiani, F. (2015). Semi-automated text classification for sensitivity identification. *In* 'Proceedings of the 24th ACM International Conference on Information and Knowledge Management'. ACM. pp. 1711–1714. 7.1, 7.3, 7.3.2

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. *In* 'Proceedings of the 20th International Conference on Machine Learning'. AAAI Press. pp. 59–66. 7.7.3

Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. *In* 'Proceedings of the 20th International Conference on Pattern Recognition'. IEEE Computer Society. pp. 3121–3124. 2.2.5

Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E. & Li, H. (2008). Context-aware query suggestion by mining click-through and session data. *In* 'Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining'. ACM. pp. 875–883. 6.3

Cavnar, W. B. & Trenkle, J. M. (1994). N-gram-based text categorization. *In* 'Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval'. Vol. 161175. Citeseer. 1

Chilcot, J (2016). 'The report of the iraq inquiry'. Cabinet Office. https://www.gov.uk/government/publications/the-report-of-the-iraq-inquiry. Last accessed on 18-07-2018. 1.2

Cleverdon, C. (1984). Optimizing convenient online access to bibliographic databases. *Information Services and Use* **4**(1-2), 37–47. 2.2.1

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1), 37–46. 2.2.1, 3.4

Coleman, M. & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology* **60**(2), 283. 8.4.1

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research* **12**, 2493–2537. 6.4, 6.6.1, 10.1.1

Constitutional Reform and Governance Act 2010 (c. 25). http://www.legislation.gov.uk/ukpga/2010/25/section/45. HMSO. Last accessed on 12-02-2018. 1.1, 8.1

Cormack, G. V. & Grossman, M. R. (2014). Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. *In* 'Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 153–162. 1.1, 2.4, 2.4, 7.6, 7.7.3

Cormack, G. V. & Grossman, M. R. (2015). Autonomy and reliability of continuous active learning for technology-assisted review. *arXiv:1504.06868 [cs.IR]*. 2.4

Cormack, G. V. & Mojdeh, M. (2009). Machine learning for information retrieval: Trec 2009 web, relevance feedback and legal tracks.. *In* 'Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009'. Vol. Special Publication 500-278. National Institute of Standards and Technology (NIST). 2.4

Cormack, G. V., Grossman, M. R., Hedin, B. & Oard, D. W. (2010). Overview of the TREC 2010 legal track. *In* 'Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010'. Vol. Special Publication 500-294. National Institute of Standards and Technology (NIST). 2.4

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in Quantitative Methods for Psychology* **1**(1), 42–45. 9.4.5

Culpepper, J. S., Diaz, F. & Smucker, M. D. (2018). Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (SWIRL 2018). *In* 'SIGIR Forum'. Vol. 52. ACM. pp. 34–90. 10.2

Cumby, C. & Ghani, R. (2011). A machine learning based system for semi-automatically redacting documents. *In* 'Proceedings of the 23rd Innovative Applications of Artificial Intelligence Conference'. AAAI. 3.3

Cumming, G. & Maillardet, R. (2006). Confidence intervals and replication: where will the next mean fall?. *Psychological Methods* **11**(3), 217. 9.4.5

Damessie, T. T., Scholer, F. & Culpepper, J. S. (2016). The influence of topic difficulty, relevance level, and document ordering on relevance judging. *In* 'Proceedings of the 21st Australasian Document Computing Symposium'. ACM. pp. 41–48. 8.4.1, 8.4.1, 9.4.5, 10.1.1

Data Protection Act 1998 (c. 29). https://www.legislation.gov.uk/ukpga/1998/29/contents. HMSO. Last accessed on 07-05-2018. 3.2.2

Defense Advanced Research Projects Agency (2010). DARPA, New technologies to support declassification. DARPA-SN-10-73. http://fas.org/sgp/news/2010/09/darpa-declass.pdf. Last accessed on 12-02-2018. 1.1, 10.1

Dernoncourt, F., Lee, J. Y., Uzuner, O. & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* **24**(3), 596–606. 3.3

Deshpande, M. & Karypis, G. (2002). Evaluation of techniques for classifying biological sequences. *In* 'Proceedings of the 6th Pacific-Asia Conference Advances in Knowledge Discovery and Data Mining'. Springer. pp. 417–431. 6.3

Ditterrich, T. (1997). Machine learning research: four current direction. *Artificial Intelligence Magzine* **4**, 97–136. 5.5.1, 6.3.2

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning* **29**(2-3), 103–130. 2.2.4

Duda, R. O. & Hart, P. E. (1973). *Pattern recognition and scene analysis*. Wiley, New York. 2.2.4, 7.5

Dumais, S. (1998). Using svms for text categorization. *IEEE Intelligent Systems* **13**(4), 21–23. 2.2

Eskin, E., Weston, J., Noble, W. S. & Leslie, C. S. (2002). Mismatch string kernels for SVM protein classification. *In* 'Proceedings of the 15th Conference on Advances in Neural Information Processing Systems'. MIT Press. pp. 1417–1424. 6.3.1

Esuli, A., Fagni, T. & Fernández, A. M. (2017). Jatecs an open-source java text categorization system. *arXiv:1706.06802 [cs.CL]*. 7.6

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters* **27**(8), 861–874. 5.3.1

Fetterly, D., Manasse, M., Najork, M. & Wiener, J. (2003). A large-scale study of the evolution of web pages. *In* 'Proceedings of the 12th International World Wide Web Conference'. ACM. pp. 669–678. 5.2.1

Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* **33**(3), 613–619. 2.2.1, 3.4

Forbes, T. (2014). 'The report of the al sweady inquiry'. https://www.gov.uk/government/publications/al-sweady-inquiry-report. Last accessed on 18-07-2018. 1.2

Freedom of Information Act 2000 (c. 36). https://www.legislation.gov.uk/ukpga/2000/36/contents. HMSO. Last accessed on 12-02-2018. (document), 1.1, 1.2, 1.5, 1.7, 2.5, 3.1, 3.2, 3.2.1, 3.2.2, 3.5, 4.1, 5.1, 9.4.4, 10.1, 10.1.1, 10.1.2, 10.1.3, 10.3

Freedom of Information (Scotland) Act 2002 (asp. 13). QPS. http://www.legislation.gov.uk/asp/2002/13/contents. Last accessed on 12-02-2018. 1.1

Freund, Y., Seung, H. S., Shamir, E. & Tishby, N. (1992). Information, prediction, and query by committee. *In* 'Proceedings of the 5th Advances in Neural Information Processing Systems'. Morgan Kaufmann. pp. 483–490. 4.5

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1**(1), 55–77. 2.2.4

Fürnkranz, J. (1998). A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence* **3**(1998), 1–10. 6.2

Gabriel, M., Paskach, C. & Sharpe, D. (2013). The challenge and promise of predictive coding for privilege. *In* 'Proceedings of the ICAIL 2013 Workshop on Standards for Using Predictive Coding, Machine Learning and Other Advanced Search and Review Methods in E-Discovery'. DESI V. 5

Gardner, J. & Xiong, L. (2008). Hide: an integrated system for health information de-identification. *In* 'Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems'. IEEE. pp. 254–259. 3.3

Ghosh, D., Guo, W. & Muresan, S. (2015). Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. *In* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing'. ACL. pp. 1003–1012. 6.4

Gollins, T., McDonald, G., Macdonald, C. & Ounis, I. (2014). On using information retrieval for the selection and sensitivity review of digital public records. *In* 'Proceeding of the 1st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security, co-located with 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. PIR. pp. 39–40. 1.6

Government of Wales Act 2006 (c. 32). https://www.legislation.gov.uk/ukpga/2006/32/contents. HMSO. Last accessed on 12-02-2018. 1

Grossman, M. R. & Cormack, G. V. (2010). Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology* **17**(3), 1–48. 2.2.1, 2.4

Grossman, M. R., Cormack, G. V. & Roegiest, A. (2016). Trec 2016 total recall track overview.. *In* 'Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016'. Vol. Special Publication 500-321. National Institute of Standards and Technology (NIST). 2.4

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill, New York. 8.4.1

Guo, Y., Gaizauskas, R., Roberts, I., Demetriou, G. & Hepple, M. (2006). Identifying personal health information using support vector machines. *In* 'Proceedings of the i2b2 workshop on challenges in natural language processing for clinical data'. AMIA. pp. 10–11. 3.3

Gupta, D., Saul, M. & Gilbertson, J. (2004). Evaluation of a deidentification (de-id) software engine to share pathology reports and clinical documents for research. *American Journal of Clinical Pathology* **121**(2), 176–186. 3.3

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* **46**(1-3), 389–422. 5.3.1

Harris, Z. S. (1954). Distributional structure. *Word* **10**(2-3), 146–162. 6.4

Hedin, B., Tomlinson, S., Baron, J. R. & Oard, D. W. (2009). Overview of the TREC 2009 legal track. *In* 'Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009'. Vol. Special Publication 500-278. National Institute of Standards and Technology (NIST). 2.4

Information Commissioner's Office (1998). 'Data protection principles'. https://ico.org.uk/for-organisations/guide-to-data-protection/data-protection-principles. Last accessed on 07-05-2018. 3.2.2

Information Commissioner's Office (2014). 'Personal information (section 40 and regulation 13)'. https://ico.org.uk/media/for-organisations/documents/1213/personal-information-section-40-and-regulation-13-foia-and-eir-guidance.pdf. Last accessed on 07-05-2018. 3.2.2

Inquiries Act 2005 (c. 12). https://www.legislation.gov.uk/ukpga/2005/12/contents. Last accessed on 18-07-2018. 1.2, 4.2

Jethani, C. P. & Smucker, M. D. (2010). Modeling the time to judge document relevance. *In* 'Proceedings of The Simulation of Interaction Workshop at The 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. p. 11. 8.4.1, 8.4.2, 8.4.3

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *In* 'Proceedings of the 10th European Conference on Machine Learning'. Springer. pp. 137–142. 2.2.2, 2.2.3, 2.2.4, 2.2.4, 5.3.1, 6.2, 6.3, 6.3.1, 6.5.1

Johnson, M. (2009). How the statistical revolution changes (computational) linguistics. *In* 'Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?'. ACL. pp. 3–11. 4.4, 6.3

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv:1607.01759 [cs.CL]*. 6.4

Kalis, S. M. (2014). Google spain sl, google inc. v. agencia espanola de proteccion de datos, mario costeja gonzalez: An entitlement to erasure and its endlenss effects. *Tul. J. Int'l & Comp. L.* **23**, 589. 10.2

Kalt, T. & Croft, W. (1996). A new probabilistic model of text classification and retrieval. Technical report. Technical Report IR-78, University of Massachusetts Center for Intelligent Information Retrieval. 2.2.4

Keerthi, S. S. & Lin, C. (2003). Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation* **15**(7), 1667–1689. 6.3.1

Kent, A., Berry, M. M., Luehrs, F. U. & Perry, J. W. (1955). Machine literature searching viii. operational criteria for designing information retrieval systems. *Journal of the Association for Information Science and Technology* **6**(2), 93–101. 2.2.5

Kerr, B. (2003). THREAD ARCS: an email thread visualization. *In* 'Proceedings of the 9th IEEE Symposium on Information Visualization'. IEEE. pp. 211–218. 1

Kibriya, A. M., Frank, E., Pfahringer, B. & Holmes, G. (2004). Multinomial naive bayes for text categorization revisited. *In* 'Proceedings of the 17th Australasian Joint Conference on Artificial Intelligence'. Springer. pp. 488–499. 2.2.4

Kuncheva, L. I. & Rodríguez, J. J. (2014). A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems* **38**(2), 259–275. 5.5.1, 6.3.2

Lain, V. (2013). 'Digital records sensitivity review'. https://blog.nationalarchives.gov.uk/blog/digital-records-sensitivity-review. Last accessed on 12-02-2018. 1.1

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174. 3.4

Lane, T. & Brodley, C. E. (1999). Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security* **2**(3), 295–331. 6.3

Larkey, L. S. & Croft, W. B. (1996). Combining classifiers in text categorization. *In* 'Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 289–297. 2.2.4, 2.4

Lefebvre, C., Manheimer, E. & Glanville, J. (2008). Searching for studies. *Cochrane Handbook for Systematic Reviews of Interventions* **5.1.0**, 95–150. 2.4

Leslie, C. S., Eskin, E. & Noble, W. S. (2002). The spectrum kernel: A string kernel for SVM protein classification. *In* 'Proceedings of the 7th Pacific Symposium on Biocomputing'. PBS. pp. 566–575. 6.3, 6.3.1

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. *In* 'Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 37–50. 2.2.4

Lewis, D. D. & Catlett, J. (1994). Heterogeneous uncertainty sampling for supervised learning. *In* 'Machine Learning Proceedings 1994'. Morgan Kaufmann. pp. 148–156. 2.3.1, 7.3.1

Lewis, D. D. & Gale, W. A. (1994). A sequential algorithm for training text classifiers. *In* 'Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 3–12. 2.2.4, 2.2.5, 2.3, 2.3.1, 2.3.1, 4.5, 7.3, 7.3.1, 7.3.1, 7.5

Lioma, C. & Ounis, I. (2006). Examining the content load of part of speech blocks for information retrieval. *In* 'Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics'. COLING/ACL. pp. 531–538. 6.3, 6.3.1.1, 6.5.1

Loftus, G. R. & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review* **1**(4), 476–490. 9.4.5

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* **11**(1-2), 22–31. 2.2.3

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development* **2**(2), 159–165. 2.2.3

Maron, M. E. (1961). Automatic indexing: an experimental inquiry. *Journal of the ACM* **8**(3), 404–417. 2.2

Maron, M. E. & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the ACM* **7**(3), 216–244. 10.2

Mc Laughlin, G. H. (1969). Smog grading - a new readability formula. *Journal of Reading* **12**(8), 639–646. 8.4.1

McCallum, A., Nigam, K. *et al.* (1998). A comparison of event models for naive bayes text classification. *In* 'Proceedings of the AAAI-98 workshop on learning for text categorization'. AAAI. pp. 41–48. 2.2.4, 7.5

McCallumzy, A. K. & Nigamy, K. (1998). Employing em and pool-based active learning for text classification. *In* 'Proceedings of the 15th International Conference on Machine Learning'. Morgan Kaufmann. pp. 350–358. 2.2.4, 7.3, 7.5

McDonald, G. (2015). A framework for enhanced text classification in sensitivity and reputation management. *In* 'Proceeding of the 6th BCS-IRSG Symposium on Future Directions in Information Access'. BCS. 1.6

Mcdonald, G., García-Pedrajas, N., Macdonald, C. & Ounis, I. (2017). A study of svm kernel functions for sensitivity classification ensembles with pos sequences. *In* 'Proceedings of The 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 1097–1100. 1.6

McDonald, G., Macdonald, C. & Ounis, I. (2015). Using part-of-speech n-grams for sensitive-text classification. *In* 'Proceedings of the International ACM SIGIR Conference on The Theory of Information Retrieval'. ACM. pp. 381–384. 1.6, 9.5.3

McDonald, G., Macdonald, C. & Ounis, I. (2017). Enhancing sensitivity classification with semantic features using word embeddings. *In* 'Proceedings of The 39th European Conference on Information Retrieval'. Springer. pp. 450–463. 1.6

McDonald, G., Macdonald, C. & Ounis, I. (2018*a*). Active learning strategies for technology assisted sensitivity review. *In* 'Proceedings of The 40th European Conference on Information Retrieval'. Springer. pp. 439–453. 1.6

McDonald, G., Macdonald, C. & Ounis, I. (2018*b*). Towards maximising openness in digital sensitivity review using reviewing time predictions. *In* 'Proceedings of The 40th European Conference on Information Retrieval'. Springer. pp. 699–706. 1.6

McDonald, G., Macdonald, C., Ounis, I. & Gollins, T. (2014). Towards a classifier for digital sensitivity review. *In* 'Proceedings of 36th European Conference on Information Retrieval'. Springer. pp. 500–506. 1.2, 1.6, 7.3.2, 9.4.1

McLean, L., Tingley, M., Scott, R. N. & Rickards, J. (2001). Computer terminal work and the benefit of microbreaks. *Applied Ergonomics* **32**(3), 225–237. 9.4.4

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**(2), 153–157. 5.2.1, 5.3.1, 5.4.2, 6.5.1

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs.CL]*. 6.1, 6.4, 10.1.1

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In* 'Proceedings of the 27th Annual Conference on Neural Information Processing Systems'. NIPS. pp. 3111–3119. 6.4

Ministry of Justice (2008*a*). 'Section 27: International relations'. http://www.justice.gov.uk/downloads/information-access-rights/foi/foi-exemption-s27.pdf. Last accessed on 07-05-2018. 3.2.1, 3.2.1, 10.3

Ministry of Justice (2008*b*). 'Section 40: Personal information'. http://www.justice.gov.uk/downloads/information-access-rights/foi/foi-exemption-s40.pdf. Last accessed on 07-05-2018. 3.2.2

Mitchell, J. & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science* **34**(8), 1388–1429. 6.4

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to cousineau (2005). *Reason* **4**(2), 61–64. 9.4.5

Moss, M. & Endicott-Popovsky, B. (2015). *Is Digital Different?: How information creation, capture, preservation and discovery are being transformed*. Facet Publishing. 10.2

Moss, M. S. & Gollins, T. J. (2017). Our digital legacy: an archival perspective. *Journal of Contemporary Archival Studies* **4**(2), 3. 1.2, 1, 2.4, 3.2, 1, 3.3, 6.1, 8.5, 10.3

Moss, M. S., Thomas, D. L. & Gollins, T. (2018). Artificial fibres-the implications of the digital for archival access.. *Frontiers in Digital Humanities* **5**, 20. 10.1

Neamatullah, I., Douglass, M. M., Li-wei, H. L., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G. & Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making* **8**(1), 32. 3.3, 5.4.2

Nettleton, D. F. & Abril, D. (2012). Document sanitization: Measuring search engine information loss and risk of disclosure for the wikileaks cables. *In* 'Proceedings of the International Conference on Privacy in Statistical Databases'. Springer. pp. 308–321. 3.3

Nigam, K., McCallum, A. K., Thrun, S. & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning* **39**(2-3), 103–134. 5.3.2

Nigam, K., McCallum, A., Thrun, S. & Mitchell, T. M. (1998). Learning to classify text from labeled and unlabeled documents. *In* 'Proceedings of the 15th National Conference on Artificial Intelligence and the 10th Innovative Applications of Artificial Intelligence Conference'. AAAI Press / The MIT Press. pp. 792–799. 2.2.4, 7.5

Ntoulas, A., Cho, J. & Olston, C. (2004). What's new on the web?: the evolution of the web from a search engine perspective. *In* 'Proceedings of the 13th international conference on World Wide Web'. ACM. pp. 1–12. 5.2.1

Oard, D. W., Baron, J. R., Hedin, B., Lewis, D. D. & Tomlinson, S. (2010). Evaluation of information retrieval for e-discovery. *Artificial Intelligence and Law* **18**(4), 347–386. 1.2, 2.4

Oard, D. W., Hedin, B., Tomlinson, S. & Baron, J. R. (2008). Overview of the TREC 2008 legal track. *In* 'Proceedings of The Seventeenth Text REtrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, November 18-21, 2008'. Vol. Special Publication 500-277. National Institute of Standards and Technology (NIST). 2.4

Oard, D. W., Shilton, K. & Lin, J. J. (2016). Evaluating search among secrets. *In* 'Proceedings of the 7th International Workshop on Evaluating Information Access @ NTCIR-12'. NII. 10.1

Pauli, M. (2015). 'The changing face of official record keeping in uk, ireland and the netherlands'. http://www.itutility.ac.uk/2015/12/07/threats-to-openness-in-the-digital-world-conference/. Last accessed on 26-02-2018. 3.4

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B. & Callison-Burch, C. (2015). Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. *In* 'Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing'. Vol. 2. ACL. pp. 425–430. 6.4

Porter, M. F. (1980). An algorithm for suffix stripping. *Program* **14**(3), 130–137. 2.2.3

Public Records Act 1958 (c. 51). HMSO. http://www.legislation.gov.uk/ukpga/Eliz2/6-7/51. Last accessed 12-02-2018. 1.1, 1.2, 3.2, 8.1, 8.2, 8.6

Public Records Act (Northern Ireland) 1923 (c. 20). HMSO. https://www.legislation.gov.uk/apni/1923/20/contents. Last accessed 12-02-2018. 1

Public Records (Scotland) Act 2011 (asp. 12). QPS. https://www.legislation.gov.uk/asp/2011/12/contents. Last accessed 12-02-2018. 1

Redwine, G., Barnard, M., Donovan, K. M., Farr, E., Forstrom, M., Hansen, W. M., John, J. L., Kuhl, N., Shaw, S. & Thomas, S. E. (2013). *Born digital: Guidance for donors, dealers, and archival repositories*. Council on Library and Information Resources Washington, DC. 10.1

Regulation (EU) (2016/679). 'General Data Protection Regulation'. OJ. https://eur-lex.europa.eu/eli/reg/2016/679/oj. Last accessed on 07-05-2018. 10.2

Rennie, J. D. M., Shih, L., Teevan, J. & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. *In* 'Proceedings of the 20th International Conference on Machine Learning'. AAAI Press. pp. 616–623. 2.2.4, 7.5

Right2INFO.org (2016). 'Constitutional protections of the right to information'. http://www.right2info.org/constitutional-protections. Last accessed on 12-02-2018. 1.1

Riloff, E. (1995). Little words can make a big difference for text classification. *In* 'Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 130–136. 5.3.2

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation* **33**(4), 294–304. 10.2

Robertson, S. E. & Jones, K. S. (1976). Relevance weighting of search terms. *JASIS* **27**(3), 129–146. 2.2.4

Roegiest, A., Cormack, G. V., Grossman, M. R. & Clarke, C. (2015). Trec 2015 total recall track overview. *In* 'Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015'. Vol. Special Publication 500-319. National Institute of Standards and Technology (NIST). 2.4

Roffman, H. (1975). Freedom of information: Judical review of executive security classifications. *U. Fla. L. Rev.* **28**, 551–567. 3.3

Rosen, J. (2011). The right to be forgotten. *Stan. L. Rev. Online* **64**, 88–92. 10.2

Salton, G. (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc. 2.2.2

Salton, G., Wong, A. & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM* **18**(11), 613–620. 2.2.2

Sánchez, D., Batet, M. & Viejo, A. (2012). Detecting sensitive information from textual documents: An information-theoretic approach. *In* 'Proceedings of the 9th International Conference on Modeling Decisions for Artificial Intelligence'. Springer. pp. 173–184. 3.3, 5.1, 5.2, 5.2.1, 5.2.1, 5.6

Sanderson, M. & Joho, H. (2004). Forming test collections with no system pooling. *In* 'Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 33–40. 2.4

Santos, R. L. T. (2013). Explicit web search result diversification. PhD thesis. University of Glasgow. 10.2

Scheffer, T., Decomain, C. & Wrobel, S. (2001). Active hidden markov models for information extraction. *In* 'Proceedings of the 10th International Symposium on Intelligent Data Analysis'. Springer. pp. 309–318. 2.3.1, 7.3.1

Schohn, G. & Cohn, D. (2000). Less is more: Active learning with support vector machines. *In* 'Proceedings of the 17th International Conference on Machine Learning'. Morgan Kaufmann. pp. 839–846. 4.5, 7.7.3

Scholer, F., Turpin, A. & Sanderson, M. (2011). Quantifying test collection quality based on the consistency of relevance judgements. *In* 'Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 1063–1072. 1.2

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* **34**(1), 1–47. 2.2, 2.2.4, 4.4, 5.3, 5.3.1, 5.3.2, 6.2, 6.5.1

Sebastiani, F. (2015). An axiomatically derived measure for the evaluation of classification algorithms. *In* 'Proceedings of the ACM SIGIR International Conference on The Theory of Information Retrieval'. ACM. pp. 11–20. 2.2.5, 2.2.5

Settles, B. (1995). Active learning literature survey. *Science* **10**(3), 237–304. 2.3

Settles, B. (2011). Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. *In* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing'. ACL. pp. 1467–1478. 2.2.4, 7.4, 7.5, 7.6

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **6**(1), 1–114. 2.3, 2.3.1, 7.3, 7.3.1

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**(3), 379–423. 2.2.3, 2.3.1, 7.3.1

Silva, C. & Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. *In* 'Proceedings of the International Joint Conference on Neural Networks'. IEEE. pp. 1661–1666. 5.3.2

Sloyan, V. (2016). Born-digital archives at the wellcome library: appraisal and sensitivity review of two hard drives. *Archives and Records* **37**(1), 20–36. 4.2

Smith, E. A. & Senter, R. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)* pp. 1–14. 8.4.1

Socher, R., Huang, E. H., Pennin, J., Manning, C. D. & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *In* 'Proceedings of the 25th Annual Conference on Neural Information Processing Systems'. NIPS. pp. 801–809. 6.4, 6.6.1

Song, F., Liu, S. & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern Analysis and Applications* **8**(1), 199–209. 2.2.2, 2.2.2, 4.4, 5.3.2

Souza, R. R., Coelho, F. C., Shah, R. & Connelly, M. (2016). Using artificial intelligence to identify state secrets. *arXiv:1611.00356 [cs.CY]*. 3.3

Suykens, J. A. & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters* **9**(3), 293–300. 1

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the scrub system.. *In* 'Proceedings of the American Medical Informatics Association Annual Fall Symposium'. AMIA. 3.3

The Advisory Council (2016). 'Advisory council on national records and archives 13th annual report 2015-16'. http://www.nationalarchives.gov.uk/documents/advisory-council-annual-report-2015-16.pdf. Last accessed on 16-07-2018. 3.4

The Advisory Council (2017). 'Advisory council on national records and archives 14th annual report 2016-17'. http://www.nationalarchives.gov.uk/documents/advisory-council-annual-report-2016-17.pdf. Last accessed on 16-07-2018. 3.3, 3.4

The Centre for Law and Democracy (2016). 'Global right to information rating'. http://www.rti-rating.org/country-data. Last accessed on 12-02-2018. 1.1

The National Archive (2007). 'Code of practice for archivists and records managers under section 51(4) of the data protection act 1998'. http://www.nationalarchives.gov.uk/documents/information-management/dp-code-of-practice.pdf. Last accessed on 07-05-2018. 3.2.2

The National Archives (2016*a*). 'The application of technology-assisted review to born-digital records transfer, inquiries and beyond'. http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf. Last accessed on 26-02-2018. 1.1, 2, 1.2, 2.4, 4.2, 6.6.3, 8.1, 8.2, 9.1, 10.1

The National Archives (2016*b*). 'The digital landscape in government 2014-2015'. http://www.nationalarchives.gov.uk/documents/digital-landscape-in-government-2014-15.pdf. Last accessed on 26-02-2018. 1.2, 2, 3.2, 4.4, 10.1.2

The National Archives (2016*c*). 'Retention'. http://www.nationalarchives.gov.uk/documents/information-management/retention.pdf. Last accessed on 16-07-2018. 3.4

The National Archives (2017). 'Digital strategy'. http://www.nationalarchives.gov.uk/documents/the-national-archives-digital-strategy-2017-19.pdf. Last accessed on 26-02-2018. 1.2, 2.2.1, 9.1, 9.5.1

The National Archives (2018). 'Information management assessment'. http://www.nationalarchives.gov.uk/documents/information-management/dcms-ima-reassessment-report-2017.pdf. Last accessed on 18-07-2018. 1.2

The National Archives (n.d.). 'Sensitivities and review'. http://www.nationalarchives.gov.uk/information-management/manage-information/public-inquiry-guidance/sensitivities-review/. Last accessed on 18-07-2018. 1.2

Tveit, A., Edsberg, O., Rost, T., Faxvaag, A., Nytro, O., Nordgard, T., Ranang, M. T. & Grimsmo, A. (2004). Anonymization of general practioner medical records. *In* 'Proceedings of the 2nd HelsIT Conference'. 3.3

Uzuner, Ö., Sibanda, T. C., Luo, Y. & Szolovits, P. (2008). A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine* **42**(1), 13–35. 3.3

van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edition*. Butterworths, London. 2.2.5

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York. 2.2.4, 5.3.1

Vinjumur, J. K., Oard, D. W. & Paik, J. H. (2014). Assessing the reliability and reusability of an e-discovery privilege test collection. *In* 'Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 1047–1050. 5

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. *In* 'Proceedings of The 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 315–323. 1.2, 9.4.1

Walker, R. K. (2012). The right to be forgotten. *Hastings LJ* **64**, 257. 10.2

Wang, S. & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. *In* 'Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics'. Vol. 2. ACL. pp. 90–94. 6.2

Webber, W. (2011). Re-examining the effectiveness of manual review. *In* 'Proceedings of the Information Retrieval for E-Discovery Workshop at the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. p. 2. 1.2

Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J. & Hirschman, L. (2007). Rapidly retargetable approaches to de-identification in medical records. *Journal of the American Medical Informatics Association* **14**(5), 564–573. 3.3

Willenborg, L. & De Waal, T. (2012). *Elements of statistical disclosure control*. Vol. 155. Springer Science & Business Media. 3.3

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks* **5**(2), 241–259. 6.3.2

Wu, Y. & Liu, Y. (2007). Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association* **102**(479), 974–983. 1

Xing, Z., Pei, J. & Keogh, E. (2010). A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter* **12**(1), 40–48. 6.1, 6.3, 10.1.1

Yang, B., Sun, J.-T., Wang, T. & Chen, Z. (2009). Effective multi-label active learning for text classification. *In* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining'. ACM. pp. 917–926. 7.3

Yang, X., Macdonald, C. & Ounis, I. (2018). Using word embeddings in twitter election classification. *Information Retrieval Journal* **21**(2-3), 183–207. 6.4

Yang, Y. (1995). Noise reduction in a statistical approach to text categorization. *In* 'Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 256–263. 2.2.3

Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods. *In* 'Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 42–49. 2.2.4

Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *In* 'Proceedings of the 14th International Conference on Machine Learning'. Morgan Kaufmann. pp. 412–420. 2.2.3, 2.2.3

Zheng, G. & Callan, J. (2015). Learning to reweight terms with distributed representations. *In* 'Proceedings of the 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval'. ACM. pp. 575–584. 6.4

Zuccon, G., Koopman, B., Bruza, P. & Azzopardi, L. (2015). Integrating and evaluating neural word embeddings in information retrieval. *In* 'Proceedings of the 20th Australasian document computing symposium'. ACM. pp. 12:1–12:8. 6.4

# Declaration

This thesis has been composed by the author and all work presented therein was carried out by the author unless otherwise explicitly stated or cited.