Towards a Classifier for Digital Sensitivity Review

Graham McDonald, Craig Macdonald, Iadh Ounis, and Timothy Gollins

School of Computing Science
University of Glasgow, G12 8QQ, Glasgow, UK
firstname.lastname@glasgow.ac.uk

Abstract. The sensitivity review of government records is essential before they can be released to the official government archives, to prevent sensitive information (such as personal information, or that which is prejudicial to international relations) from being released. As records are typically reviewed and released after a period of decades, sensitivity review practices are still based on paper records. The transition to digital records brings new challenges, e.g. increased volume of digital records, making current practices impractical to use. In this paper, we describe our current work towards developing a sensitivity review classifier that can identify and prioritise potentially sensitive digital records for review. Using a test collection built from government records with real sensitivities identified by government assessors, we show that considering the entities present in each record can markedly improve upon a text classification baseline.

1 Introduction

Democratic governments are increasingly following policies of openness and transparency by enacting freedom of information legislation that permits anyone to request records from a publicly funded organisation. In the United Kingdom (UK), this is supplemented with regulations that release all government records to the archives after 20 years, subject to some limitations. For example, the UK's Freedom of Information Act 2000 (FOIA)[1] specifies that records containing personal information, or information that might harm international relations, should be withheld for longer periods. In some cases, the requested records may be redacted or *closed / withheld* entirely. To determine such *sensitivities*, all records are reviewed before being released to the archives.

The process of sensitivity review for paper records is a long-established practice[2], however the transition to digital records (including emails) will bring new challenges from several factors: A significant increase in volume (more digital records are created); the structure of digital records is more complex and diffuse than paper; the present resources make the current linear (page-by-page) review procedures infeasible; and there are no commercially available tools to manage digital review procedures. Moreover, the risk associated with the inadvertent release of sensitive records also increases in the digital scenario due to the ubiquitous nature of online search.

We strongly believe that without reliable and efficient sensitivity review tools, policies for transparent government and freedom of information will fail as public bodies will be forced into the precautionary closure of digital records, reducing public scrutiny

[1] http://www.legislation.gov.uk/ukpga/2000/36/contents

[2] http://www.nationalarchives.gov.uk/information-management/our-services/sensitivity-reviews-on-selected-records.htm

and negatively impacting social science research. Therefore, we are developing assistive tools to increase the efficiency of digital sensitivity review, building upon information retrieval (IR) technology.

In general, this paper provides an introduction to the IR problems that we believe must be addressed. The contributions of our paper are twofold: (i) We provide an overview of challenges in sensitivity review (ii) We detail an initial empirical investigation of a classification tool to support digital sensitivity review using real government records with real sensitivities.

The remainder of this paper is structured as follows: Section 2 discusses the sensitivity problem; Section 3 describes the classification features that we employ in this work; Section 4 describes our experimental setup; We describe our classification results in Section 5, while concluding remarks follow in Section 6.

2 Sensitivity in Digital Records

Freedom of information laws are enacted within many countries, providing access to public information. For instance, in the European Union, it is enshrined within Article 42 of the Charter of Fundamental Rights, while the USA and other countries have analogous provisions. In this section we discuss the sensitivities relating to public records.

The assumption behind freedom of information is that public records should be open. A common attribute of freedom of information enactments, however, are some standard exemptions limiting what can be released. For instance, information that may prejudice commercial operations is commonly exempt.

The role of sensitivity review is to ensure that all appropriate exemptions are checked before the records are opened. Hence, it is essential that the sensitivity review is efficient and cost-effective so that it does not stop the timely release of records. In this paper, we present work into developing a classifier for digital sensitivity review, so that reviewers can prioritise the records of highest perceived risk.

Our work is framed within the context of the UK, and of a set of government records and assessors that we have access to[3]. However, the notions encapsulated within the UK exemptions transfer easily to the legislation in many other countries.

Table 1 lists exemptions of the FOIA that apply to historical records. As can be seen, each record must be reviewed within the context of fifteen exemptions. Each exemption is assessed against a detailed set of criteria. For example, in assessing Section 27, which aims to limit damage to international relations, evidence of information being passed in confidence is of particular importance for informing a reviewer's decision.

In this work, we focus on two exemptions, namely Section 27 (International Relations) and Section 40 (Personal Information), as we believe these sections to be both representative of the issues we might expect to find in addressing many other exemptions and sufficiently challenging to test our proposed approach. The closest related work that we are aware of addresses tasks such as anonymisation of unstructured data [1], or data-loss prevention classifiers [2]. However, we believe that no other work has directly addressed the task of sensitivity review. In the next section, we describe our approach to sensitivity review.

[3] Due to the obvious sensitivities involved, the collection is not publicly available.

Table 1. UK Freedom of Information Act 2000: Exemptions that apply to historical records.

Section 21: Information Accessible by Other Means	Section 34: Parliamentary Privilege
Section 22: Information Intended for Future Publication	Section 37: Certain Aspects Relating to the Royal Family and Honours
Section 23: Bodies Dealing with Security Matters	Section 38: Health and Safety
Section 24: National Security	Section 39: Environmental Information
Section 26: Defence	Section 40: Personal Information
Section 27: International Relations	Section 41: Information Provided in Confidence
Section 29: The Economy	Section 44: Prohibitions on Disclosure
Section 31: Law Enforcement	

3 Features For Sensitivity Review

Our aim in this work is to study appropriate techniques to classify a record's likely sensitivities, focusing upon Section 27 and Section 40. We believe that such automatic classification goes beyond textual/topic classification, as addressed in classical text classification test collections such as 20 Newsgroups[4]. Hence, in the following, we propose several features that we postulate can aid in effective sensitivity classification.

Firstly, a record'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 link to one or more countries. These links can be implicit within a record, which makes the task of identifying sensitivity-entity links very challenging. Moreover, for Section 27, expressed sentiment relating to an entity may help in deciding if the information is sensitive. For these reasons, we chose to focus on the identification of named entities and subjectivity within records.

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[5] knowledge base. We also use a dictionary of 131,232 person names, constructed from the Drupal Name Database[6] and from the lists of unambiguous names supplied with *deid* [3], 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[7] to efficiently match dictionary entries with record instances and for each record r we define the number of extracted named and generic person entities, as the *nEntity* and *pCount* features respectively.

Country entities are significant for certain sensitivities, for example Section 27. Relations between countries are not all on par, therefore, the accidental release of records has varying potential for damaging the international relations between a country producing the records and a referenced country or a third-party. The real nature of these relations is privileged information and in flux. Therefore, we model this fragility using *our* perception of current international relations and supply a country-risk map as a system parameter. We define the country risk score of a record r as follows:

[4] http://qwone.com/∼jason/20Newsgroups/
[5] http://dbpedia.org/
[6] https://drupal.org/project/namedb
[7] http://alias-i.com/lingpipe/

$$cRisk(r) = \sum_{c \in r} countryRisk(c) \quad (1)$$

where *c* is a country occurring in record *r* and *countryRisk* is the risk score from the set {1:None, 2:Moderate, 3:High} associated with country *c*.

Next, we hypothesise that subjective expressions of sentiment might be correlated with sensitivity. For instance, a negatively phrased discussion about another country might be closed under Section 27. For this reason, and inspired by previous work on sentiment analysis (e.g. identifying opinionated content in blog posts [4]), we use the Opinion Finder (OF) sentiment analysis toolkit [5] to detect opinionated sentences within a record and score the record on its subjectivity. Following the work of [4] in finding overall opinionated scores for documents, each record *r* is scored as follows:

$$subjConf(r) = summConf \cdot \frac{\#subjective}{\#sentences} \quad (2)$$

where *#subjective* is the number of subjective sentences, *#sentences* is the total number of sentences in the record and *summConf* is the sum of the confidence score from OF's precision-oriented subjective sentence classifier.

4 Experimental Setup

The research question that we address is the following: Can we improve upon a text classification baseline for identifying sensitive records? In this section, we describe the experimental setup and test collection used to address this research question.

The test collection comprises 1111 government records, sampled from a larger corpus addressing international activities. The sampled records were split between seventeen assessors, twelve of whom are from government departments and experienced in sensitivity review. As discussed in Section 2, we consider only two areas of sensitivity relating to the FOIA, namely Section 27 and Section 40. Each assessor conducted at least 50 initial judgements to gain familiarity with the collection.

Assessors were supplied with a guidance document and were asked to judge whether each record contained sensitive information, that would be withheld by the government, under the sensitivities of interest. Four judgement options were provided, *Not Sensitive*, *Sensitive (Section 27)*, *Sensitive (Section 40)*, or *Sensitive (Both)*. Of the 1111 judged records, 104 were sensitive for Section 27, and 86 for Section 40.

To assess inter-assessor agreement, 150 records were judged by two assessors and 50 records were judged by four assessors. Agreement was found to be 0.5525 measured by Cohen's κ [6] for the double-judged records and a Fleiss' κ [7] score of 0.4414 for records which received four judgements each. While these values indicate a moderate agreement [8], we note that levels of agreement in the current paper-based review process are unknown, as only one assessor will routinely judge each record. Record labels were assigned based on the judgements, using a majority vote where appropriate.

As a baseline, we deploy a text classification approach, where a record is represented by a term frequency vector, over all terms in the collection. This is intuitive as there may exist inherent underlying topical patterns to sensitivities within genres [9] of a collection - e.g. non-sensitive press releases may have various co-occurring terms. Text classification features are extracted and scored using the Terrier IR platform [10]. We then extend the text classification approach by individually applying each of our identified features: *pCount*, *cRisk*, *nEntity* and *subjConf*.

Table 2. Results for sensitivity analysis for Sections 27 & 40.

	Section 27 International Relations				Section 40 Personal Information			
	Precision	Recall	F-measure	BAC	Precision	Recall	F-measure	BAC
Text Classification	0.3360	0.2972	0.3125	0.6197	0.3756	0.5595	0.4020	**0.7372**
+ *pCount*	0.2224	0.2205	0.2154	0.5689	0.3698	0.3961	0.3188	0.6352
+ *cRisk*	0.3157	0.3067	0.3089	0.6201	0.3549	0.5719	0.3844	0.7088
+ *nEntity*	**0.3605**	**0.3072**	**0.3282**	**0.6255**	**0.3901**	0.5595	**0.4107**	0.7186
+ *subjConf*	0.3123	0.2872	0.2938	0.6125	0.2684	**0.5941**	0.3150	0.6868

As a classifier, we use SVMLight [11] with a linear kernel. We measure our results using several classification measures, namely: Precision, Recall, and F-measure, as well as Balanced Accuracy (BAC), which is the arithmetic mean of true positive and true negative rates with a BAC of 0.5 indicating a random prediction. All measures are reported over a 5-fold cross validation. Moreover, there is a bias in the distribution of sensitive and non-sensitive records throughout the collection, with over 80% of records being non-sensitive. Hence, due to SVM's sensitivities to imbalanced training data, we up-sample the training sets in each fold by repeating sensitive records until the number of sensitive and non-sensitive records match. The test sets in each fold retain their observed distribution of sensitivities.

5 Results

Table 2 reports the results of the Section 27 and Section 40 classification tasks. Focusing on BAC, as it accounts for imbalanced test sets, we observe that the text classification baseline achieves a BAC of 0.6197 and 0.7372 for Sections 27 and 40, respectively, which are markedly above random. Next, we find that the *cRisk* and *nEntity* features improve the classifier's performance for Section 27 to 0.6201 and 0.6255 respectively. Identifying the presence of entities of interest and countries risk factors, intrinsic to the notion of International Relations, appear to be promising future research directions.

Conversely, *nEntity* and *cRisk* are not fundamental to the notion of Personal Information and we see that these features are detrimental to the baseline classifier's BAC. The detrimental effects of applying the selected feature sets across Sections 27 and 40, illustrate the need for individual feature sets for different aspects of sensitivities.

The *pCount* feature, a simple count of person names occurring in a record, performs poorly for both areas of sensitivity, reducing the performance of the text classification baseline. This is likely due to an over-aggressive selection process, as not all mentions of people's names are in fact personal information.

Finally, it is surprising that *subjConf* leads to the classifier's degradation for Section 27. Opinions within the records have been tagged by the judges as indicative of that exemption, and we believe that this feature is worthy of future investigation. Overall, in addressing our research question, we find that our proposed domain-specific features for sensitivity review, *cRisk* and *nEntity*, can provide benefit in enhancing the accuracy of a text classification approach for digital sensitivity review, especially for Section 27.

6 Conclusions and Future Work

We have provided an overview of the challenges faced by government departments from the imminent switch to digital records sensitivity review. Moreover, we presented some

work to develop a classification tool to assist the review process. The challenges discussed in this paper will inevitably be of increasing importance for governments obliged to be transparent and open, while securing the safety of individuals and countries.

We found that two features, namely the number of people in specific roles of interest and a risk score for countries identified within a record, can help to identify sensitive records that risk damaging international relations, by improving on a text classification baseline. We further found that these features did not help to improve BAC for personal information sensitivities. This illustrates the need for individual feature sets to identify different aspects of sensitivity.

As future work, we intend to conduct a study of assessor disagreement, having assessors revisit the disputed judgements to construct a gold standard test collection. We also intend to develop our feature usage. For example, determining term specificity using the Z-score statistical measure [12], investigating feature co-occurrence such as subjective sentences containing named entities and applying automatic features selection.

Acknowledgements. The authors would like to thank Michael Moss, Norman Gray and James Girdwood for their valuable comments, as well as, the assessors for their help in constructing the test collection.

References

1. Nguyen-Son, H.Q., Nguyen, Q.B., Tran, M.T., Nguyen, D.T., Yoshiura, H., Echizen, I.: Automatic anonymization of natural languages texts posted on social networking services and automatic detection of disclosure. In: Proc. ARES. (2012)
2. Hart, M., Manadhata, P., Johnson, R.: Text classification for data loss prevention. In Fischer-Hübner, S., Hopper, N., eds.: Privacy Enhancing Technologies. Volume 6794. (2011) 18–37
3. 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.: Automated de-identification of free-text medical records. BMC medical informatics and decision making **8**(1) (2008) 32
4. He, B., Macdonald, C., Ounis, I.: Ranking opinionated blog posts using opinionfinder. In: Proc. SIGIR. (2008)
5. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., an d Siddharth Patwardhan, E.R.: Opinionfinder: a system for subjectivity analysis. In: Proc. HLT/EMNLP. (2005)
6. Cohen, J., et al.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1) (1960) 37–46
7. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5) (1971) 378
8. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics **33**(1) (1977) 159–174
9. Orlikowski, W.J., Yates, J.: Genre repertoire: The structuring of communicative practices in organizations. Administrative science quarterly **39**(4) (1994) 541–574
10. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proc. OSIR. (2006)
11. Joachims, T.: Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms. Kluwer/Springer (2002)
12. Savoy, J.: Authorship attribution based on specific vocabulary. ACM Transactions on Information Systems (TOIS) **30**(2) (2012) 12