DOI: http://dx.doi.org/10.14236/ewic/FDIA2015.15

A Framework for Enhanced Text Classification in Sensitivity and Reputation Management

Graham McDonald
School of Computing Science
University of Glasgow
Glasgow
G12 8QQ
g.mcdonald.1@research.gla.ac.uk

Freedom of Information (FOI) laws state that government documents should be open to the public. However, many government documents contain *sensitive* information that is exempt from release. In this PhD programme, we aim to develop a framework that can automatically classify sensitive information in documents. However, automatic classification of sensitive information is a complex task that requires a relative judgement on the *effect* of a combination of factors. In this paper, we present an overview of the features of sensitivity that we can use to automatically classify documents containing FOI exemptions, such as *International Relations*. Moreover, we argue that current Named Entity Recognition (NER) approaches to classifying sensitive information are not appropriate for classifying FOI exemptions and, therefore, we need classification models that consider the document's *content* and *context* at the *time* of classification.

Keywords: Information retrieval, Text classification, Sensitive information

1. INTRODUCTION

Democratic governments are increasingly following policies of openness and transparency. Moreover, Freedom of Information (FOI)[1 2] laws state that government documents should be open to the public. However, many government documents contain information that is of a *sensitive* nature, such as *personal* or *confidential* information. Therefore, FOI laws make provisions that exempt sensitive information from being released into the public domain. It is essential that all such sensitivities are identified in government documents prior to transfer to the archives. Therefore, the governments of the United Kingdom (UK) and America (USA) have recently recognised that there is a timely need for new algorithms that can detect sensitive information in documents to avoid accidental disclosure (D.A.R.P.A. (2010); Allan (2014)).

In this PhD programme, we aim to develop a framework that can automatically classify sensitive information in documents. However, assessing the sensitivity of information and, moreover, automatically classifying sensitive information is a complex task. For example, in our initial work, we focus on a particular UK FOI exemption, namely *International Relations*, that protects the interests of the UK abroad. This exemption can apply to a document if it contains inappropriate language or content that is potentially reputationally damaging. Therefore, assessing the sensitivity of information requires a relative judgement on the *effect* of a combination of factors.

In the remainder of this paper, we argue that to be able to automatically classify FOI sensitivities, such as International Relations, we need to identify features of sensitive information that relate to three key attributes of sensitivity, namely, the document's *content*, the *context* in which the document was created and the *time* at which we are classifying the document.

2. RELATED WORK

Most research into automatically classifying sensitive information in documents has focused on personal data. Early approaches to document anonymisation came from within the domain of clinical records (Tveit *et al.* (2004)) and used medical dictionaries for term-matching or regular expressions for pattern identification (Sweeney (1996)). However, these approaches were costly, fragile and restricted in their application generalisability. Therefore, recent research into classifying sensitive information has tended to focus on more generalisable approaches.

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

Named Entity Recognition (NER) is a popular general approach for detecting sensitive information in documents. For example, Abril *et al.* (2011) adapt approaches from Statistical Disclosure Control (Willenborg and De Waal (2001)) and Privacy-Preserving Data Mining (Agrawal and Srikant (2000)) to mask named entities in documents. However, in these approaches all named entities are considered sensitive and, therefore, applying NER masking can reduce a document's utility. With this in mind there has been a shift in the focus of sensitive information classification from simple NER redaction to document sanitisation.

Document sanitisation aims to produce a privacy-preserved version of a document that retains the original document's utility. Sánchez *et al.* (2012) presented a document sanitisation approach that assumed sensitive text is more specific than non-sensitive text. Using the Information Content (IC) of noun phrases as a measure of the phrase's sensitivity they classified phrases with an IC score above an empirically set threshold β as sensitive. This approach focused on identifying personal information. However, they also identified potentially confidential phrases and showed that their approach has the potential for identifying a broader range of sensitivities than NER approaches.

In our previous work (McDonald *et al.* (2014)), we deployed a text classification approach to classify Personal Information and International Relations FOI exemptions. In that work, we extended the text classification with additional features such as the entities in the document, a country risk score and a subjective sentences count. We achieved promising results for a proof-of-concept, however, to fully address the problem of automatic classification of sensitive information we must consider the three key attributes of sensitivity outlined in Section 1.

3. FEATURES OF SENSITIVITY

Sensitive information in documents can arise from three key attributes of sensitivity. Firstly, a document can contain sensitive content, such as inappropriate language. Secondly, often the sensitive nature of a document is a result of the context in which the document was created and, thirdly, sensitivities emergence and decay over time.

Content: A document's content has many potential sources of sensitivity. Firstly, the topic could be sensitive in its own right. However, sensitivity relating to a topic usually arises from what is said about the topic or how it is said. For example, in a report claiming that Croatia are suspected of having violated an international treaty, discussion of the treaty could be sensitive. However, the discussion of the violation is sensitive since the information could be disputed and potentially damage relations with Croatia. Knowledge of the existence of the treaty can help to highlight the potential sensitivity. However, to detect this sensitivity we need to look more closely at the language used and the structure of the text.

A second posible source of sensitivity is the tone of the language used in reference to an entity, for example commenting that a foreign government is "lazy" or "corrupt". Moreover, culturally inappropriate or politically incorrect references about significant figures can be deemed sensitive.

Thirdly, information that can give a competitor a strategic advantage can also be sensitive, such as reporting that a country is inadequately prepared for a terrorist attack. This sensitivity is more difficult to detect than it might first appear. The reporting of a terrorism incident, or a government's reaction to terrorism, is not by itself sensitive information. It is the appraisal of the government's ability that causes the sensitivity.

Lastly, the source of information is significant in deciding if the information is sensitive. For example, information that has been supplied in confidence is sensitive, however, reporting information from a press conference is not.

To automatically classify the content of sensitivities such as International Relations, we need to identify sensitivity-specific language constructs, such as sequences of terms or parts-of-speech, that are indicative of the sensitivity and use the identified vocabularies to train sensitivity-specific classifiers.

Context: The context in which a document is created is important for sensitivity for two main reasons. Firstly, documents created in a particular context, such as by the *same author* or in a particular *date range*, can discuss related or similar content. Therefore, clusters of related sensitivities can exist. For example, documents produced by a particular government department within a certain date range are likely to produce many documents on a topic. The sensitivities associated to a particular topic are likely to share certain attributes and features. Therefore, we should be able to better identify the sensitivities relating to certain batches of documents if they are viewed within the context that they were created.

Secondly, sensitivities can span multiple documents. Moreover, the sensitive nature of one document might not be apparent without viewing other related documents. To address this we can classify documents from within the same context and propagate

(potential) sensitivities to related documents to see if inter-document sensitivities exist. Moreover, by training context-specific classification models we expect to better identify context-dependent sensitivities.

Time: Sensitivities evolve and decay at varying rates. Moreover, the duration of existence for some types of sensitivity are not well defined. For example, documents accounting the sinking of the Argentinian war ship ARA General Belgrano in the Falklands War of 1982 were considered highly sensitive for many years after the event. However, many of details in these documents are now freely available. Therefore, to effectively classify sensitivity in documents over time, the classification models must be able to adapt to the changing vocabulary of currently sensitive content. External resources can help to identify information that is in the public domain. However, as previously outlined, sensitivity arises from the specific aspects of the topics being discussed and this increases the complexity of the task.

4. IMPLEMENTATION

In our initial work (McDonald *et al.* (2015)) we have looked at statistical methods for automatically identifying sensitive content that relates to information supplied in confidence. More specifically, we found that by identifying part-of-speech n-grams that are specific to this sensitivity, we can use the identified n-grams to train sensitivity-specific classifiers. Moreover, we found that this approach can achieve markedly improved recall of this sensitivity compared to a recent approach from the literature, that has been shown to achieve high levels of recall of sensitive text in other domains.

To further develop this work we aim to answer three main research questions: RQ1 What are the most effective methods for automatically identifying sensitivity-specific language constructs? RQ2 How can sensitivity-specific vocabularies be constructed and maintained to be effective for sensitivity classification over time? and RQ3 What is the impact of training context-specific classification models on sensitivity classification?

5. CONCLUSIONS

In this paper, we have presented an overview of sensitivity relating to FOI exemptions such as International Relations. We have argued that to successfully classify these sensitivities we need to go beyond the current NER based approaches. Moreover, we need to develop classification models that can identify features of sensitivity relating to the document's content, the context in which the document was created and the current sensitivities at the time of classification. More specifically, we argue that effective classification of sensitive documents can be achieved by constructing sensitivity-specific vocabularies from language constructs, such as sequences of parts-of-speech, that are specific to individual sensitivities. Moreover, we can use the identified vocabularies to train effective classifiers that can identify passages of sensitive text in documents. Furthermore, by training context-specific classification models we will be able to better identify inter-document sensitivities.

6. ACKNOWLEDGEMENTS

We thank Dr Iadh Ounis and Dr Craig Macdonald for supervising this PhD programme and the ELIAS ESF Research Networking Programme for the scholarship award to participate in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Abril, D and NA, G and Torra, V (2011) *On the De-classification of Confidential Documents* Modeling Decision for Artificial Intelligence. Springer.

Agrawal, R and Srikant, R (2000) *Privacy-preserving data mining*. ACM Sigmod Record. Vol 29.

Allen, A (2014) *Records Review*. UK Government. https://www.gov.uk/government/publications/records-review-by-sir-alex-allan.

Defense Advanced Research Projects Agency (2010) *DARPA, New technologies to support declassification*. Request for Information (RFI).

McDonald, G and Macdonald, C and Ounis, I and Gollins, T (2014) *Towards a Classifier for Digital Sensitivity Review*. In *Proc of ECIR*.

McDonald, G and Macdonald, C and Ounis, I (2015) *Using Part-of-Speech N-grams for Sensitive-text Classification*. In *Proc of ICTIR*.

Sánchez, D and Batet, M and Viejo, A (2012) *Detecting sensitive information from textual documents: an information-theoretic approach*. Modeling Decisions for Artificial Intelligence. 173-184.

Sweeney, L (1996) *Replacing personally-identifying information in medical records, the Scrub system.* In *Proc. of AMIA*.

Tveit, A and Edsberg, O and Rost, TB and Faxvaag, A and Nytro, O and Nordgard, MT and Ranang, MT and Grimsmo, A *Anonymization of general practitioner medical records*. In *Proc of HelsIT*.

Willenborg, L and de Waal, T (2001) *Elements of Statistical Disclosure Control*. LNCS. Vol 155. Springer-Verlag New York.