Sensitivity Review of Large Collections by Identifying and Prioritising Coherent Documents Groups

Hitarth Narvala
University of Glasgow, UK
h.narvala.1@research.gla.ac.uk

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

Iadh Ounis
University of Glasgow, UK
iadh.ounis@glasgow.ac.uk

ABSTRACT

With the massive increase in the volume of digitally produced documents, government departments face a logistical issue when conducting the manual sensitivity review of documents that should be opened to the public. When reviewing a document, sensitivity reviewers often need to quickly access related information from other documents in the collection. For example, documents that mention the same topic or event can provide the reviewers with useful contextual information and assist the reviewers to make consistent sensitivity judgements more quickly. However, it is infeasible to manually identify groups of such related documents in large unstructured collections. In this work, we present a sensitivity review system that automatically identifies groups of related documents to assist reviewers and increase the efficiency of sensitivity review. In particular, our system groups the documents that are to be sensitivity reviewed based on the documents' semantic categories (e.g., criminality). Moreover, the system identifies chronological and coherent information threads to describe the full context of an event, activity or discussion that may be spread across multiple documents. Additionally, the system prioritises the identified semantic categories and information threads for review by leveraging automatic sensitivity classification to maximise the number of documents that can be opened to the public in a limited reviewing time-budget.

CCS CONCEPTS

• **Information systems → Clustering and classification**.

KEYWORDS

sensitivity review, document clustering, information threading

ACM Reference Format:
Hitarth Narvala, Graham McDonald, and Iadh Ounis. 2022. Sensitivity Review of Large Collections by Identifying and Prioritising Coherent Documents Groups. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557182

1 INTRODUCTION

In more than a hundred countries [11], governments are required to make their documents open to the public to comply with Freedom of Information Laws (FOI). Such documents often contain sensitive information, for example personal or confidential information. Therefore, the documents must be manually sensitivity reviewed to identify and protect any sensitive information prior to opening the documents to the public. However, with the massive volume of digitally produced documents, governments face a logistical issue that makes a fully manual sensitivity review infeasible.

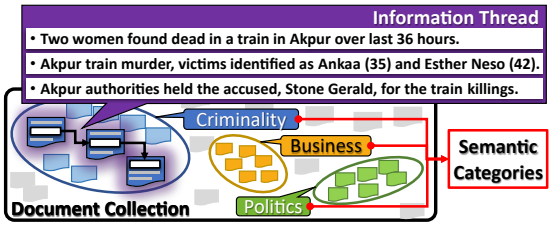

Figure 1: Examples of semantic categories and information threads in a document collection.

Different types of sensitivities can correspond to specific categories of semantically related documents (i.e., *semantic categories*). Moreover, a semantic category can contain multiple documents that discuss similar, or related, sensitivities. For example, multiple documents that discuss criminal incidents may include the personal details of victims. Additionally, sensitive information about an event, activity or discussion can be spread over multiple documents. In such cases, as is illustrated in Figure 1, extracting a chronological thread of coherent information from the documents (i.e., an *information thread*) can describe an event, activity or discussion to explain the full context of a potential sensitivity.

When sensitivity reviewing a document, it is often easier for reviewers to make sensitivity judgements when they are aware of groups of related documents (i.e., semantic categories and information threads). Identifying such groups can help to indicate how likely particular topics are to be sensitive in a particular group. For example, the details of an employee's salary are more likely to be sensitive in documents about business discussions than mentions of salaries in documents about political discussions, since politicians' salaries are usually in the public domain. *Prioritising* particular groups of related documents for review can also help to increase the number of documents that can be opened to the public when there are limited reviewing resources [9] (i.e., *openness* [6]). However, in large unstructured document collections, it is not practical for reviewers to manually identify such groups of related documents.

There are some existing systems that can support human sensitivity reviewers by providing functionalities to search and navigate a collection of documents, or to assist the reviewers with automatic sensitivity classification predictions (e.g. [7, 8]). However, these systems are not able to present reviewers with groups of chronologically ordered and semantically related documents to help to improve the efficiency and openness of the sensitivity review process.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557182

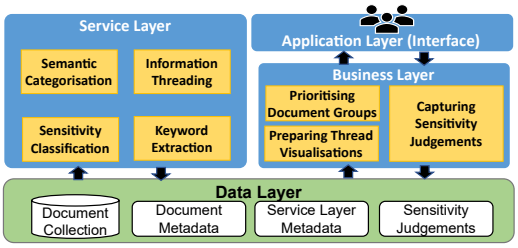

Figure 2: Layered system architecture.

In this work, we present a sensitivity review system that automatically identifies latent semantic categories and information threads within a collection. A video capture of our system is available at https://youtu.be/L4AgwAgKVkQ. The system integrates a sensitivity classifier to prioritise document groups for review. Moreover, the prioritisation criteria can be customised to generate finer-grained categories depending on the needs of reviewers. The system incorporates two user interfaces for reviewing a collection of documents. First, an interface for sequentially reviewing the documents grouped into semantic categories and, second, an interface for reviewing information threads from multiple documents.

2 RELATED WORK

In this section, we discuss previous work on systems for general document review tasks and specific systems for sensitivity review.

For general document review tasks, such as in e-Discovery, Vo et al. [12] presented a system called DISCO that leverages semantic relatedness between documents to assist reviewers with information discovery in a collection. In particular, DISCO enabled the reviewers to perform complex exploratory search tasks using user-defined keywords or automatically extracted keywords by document clustering techniques. Another work by Abualsaud et al. [1] proposed a document review system called HiCAL. HiCAL integrated an active learning classifier, search functionality and a reviewing interface to assist the reviewers in efficiently navigating a collection and assessing the relevance of a document. However, differently from DISCO and HiCAL, where the focus is on retrieving a set of relevant documents for review, in the sensitivity review task, all documents in a collection are required to be reviewed. Therefore in this work, we focus on the *prioritisation* of related document groups (i.e., semantic categories and information threads) to improve the efficiency and openness of the sensitivity review process.

Recently, a few systems have been proposed to assist sensitivity reviewers in exploring sensitivities in a collection, and performing accurate and efficient sensitivity judgements. Narvala et al. [8] proposed Receptor that can assist sensitivity reviewers with gaining insights from latent relations between entities and events that can potentially indicate sensitive information in a collection. Receptor integrated interactive visualisations and advanced search functionalities using graph search to enable reviewers to find documents that are likely to contain sensitive information. Differently from Receptor, our proposed system aims to automatically identify groups of chronologically ordered and semantically related documents to improve the efficiency of sensitivity reviewers by enabling them to review documents that are about a topic, event, activity or discussion, at the same time. Another work by McDonald et al. [7] presented a sensitivity review interface that integrated a sensitivity classifier to assist the reviewers in making sensitivity judgements.

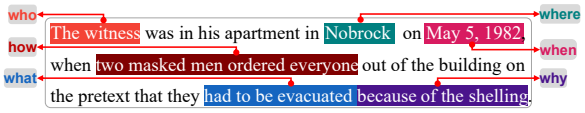

Figure 3: An example of 5W1H extraction.

McDonald et al. [7] showed that the efficiency and accuracy of sensitivity reviewers can be significantly improved when the reviewers are provided with sensitivity predictions from a sensitivity classifier. Differently from the work of McDonald et al. [7], in this work, our system integrates sensitivity classification and prioritises automatically identified semantic categories and information threads based on their amount of predicted sensitivity, to increase openness.

3 SYSTEM ARCHITECTURE

As discussed in Section 1, the goal of our system is to assist sensitivity reviewers in quickly reviewing related documents and to increase the number of documents that can be opened to the public in a fixed reviewing time-budget, i.e., openness. To achieve this, our system incorporates: (1) the identification of latent semantic categories and information threads in a collection, (2) automatic sensitivity classification to compute the probability that a document contains sensitive information, (3) an integrated interface for exploring information threads and semantically related groups of documents, and sensitivity reviewing individual documents. Figure 2 presents the architecture of our system. The system has four distinct layers that we discuss in this section, namely, the data layer, the service layer, the business layer, and the application layer:

• **Data Layer**: The data layer manages the storage of the document collection and the associated document metadata attributes. For this demo, we use the GovSensitivity [9] collection, which comprises government documents that are annotated for FOI sensitivities. The data layer further records the outputs from semantic categorisation, information threading and sensitivity classification along with the reviewers' sensitivity judgements.

• **Service Layer**: In this layer, all documents in the collection are clustered by their semantic categories, identified as part of any information thread, and classified as being sensitive or non-sensitive.

We identify semantic categories using document clustering, based on our previous work [9]. In particular, we deploy DEC [13], which is a popular deep neural clustering approach that simultaneously learns feature representation and clustering assignments.

For information threading, we deploy an approach that extracts mentions of events, activities or discussions in each of the documents to identify groups of coherent information from the documents. Government documents are often long and can mention multiple events. Therefore, we first split the documents into passages and, as shown in Figure 3, extract answers to the questions *who*, *what*, *why*, *where*, *when* and *how* (referred to as 5W1H questions) to describe the event in the passage of text. We deploy Giveme5W1H library [4] for 5W1H extraction. The information threading component clusters the document passages with similar 5W1H questions' answers, and chronologically orders the document passages in the identified clusters using the documents' creation timestamps.

For each identified semantic category and information thread, the system also extracts keywords to describe the particular document group to the reviewers. The system assigns weights to each of the words in the vocabulary of a group using their document

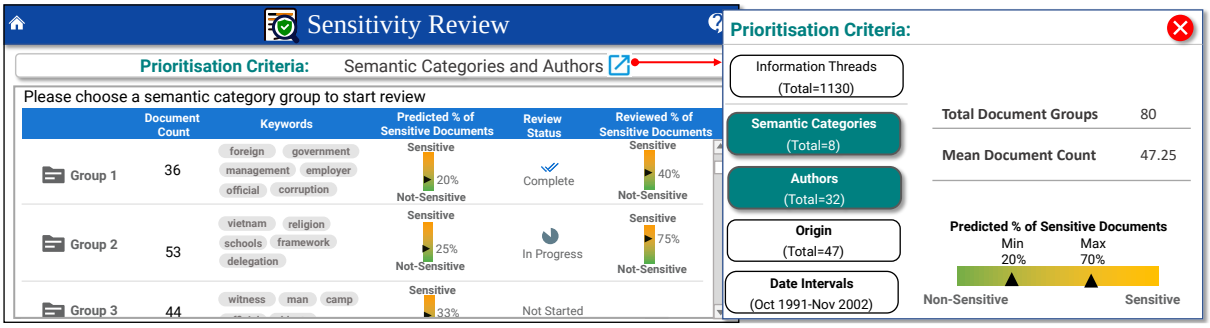

Figure 4: Customisable review prioritisation of the document groups in the increasing order of predicted sensitivities.

frequencies and element-wise mean of TF-IDF document vectors, and outputs the word with the highest weights.

Lastly, for sensitivity classification, we deployed an SVM text classification approach from the literature [5]. To promote a modular architecture, specific methods for the aforementioned service layer components are incorporated as pluggable modules to be able to easily upgrade the system with newly developed components.

• **Business Layer**: The business layer manages the interaction with the application layer to capture user inputs and report the desired outputs. The business layer prioritises the document groups as per user-defined criteria such that documents that are more likely to be opened to the public can be reviewed earlier to increase openness. To prioritise the semantic category document groups, we deploy a hierarchical ranking strategy. This hierarchical strategy first ranks the document groups with the increasing order of mean sensitivity classification probability of documents in the group. The documents within the groups are then ranked using their individual sensitivity probabilities. The information threads are prioritised only at the thread level (using the mean sensitivity probability), since the documents within a thread are already ordered chronologically. Additionally, when a reviewer requests to review a document, the business layer reports the document and options to explore any information threads associated with the reported document. For reviewing an information thread, the business layer prepares the visualisation of the document passages in a thread using the extracted 5W1H information to indicate the related information about an event, activity or discussion. Finally, the business layer stores the sensitivity judgements provided by the reviewers for each document.

• **Application Layer**: This layer comprises a web-based user interface, which enables the reviewers to explore and review the documents in a collection. The interface presents the prioritised semantic categories and information threads to the reviewers, and incorporates the capabilities of reviewing individual documents in a category or passages from multiple documents in an information thread.

The system is implemented in python. Scikit-learn [10] is used for deploying document clustering and sensitivity classification, and the web interface is implemented using Django [3].

4 KEY FUNCTIONALITIES

Our system provides numerous functionalities to the sensitivity reviewers, which we summarise as follows:

• **Sequentially Reviewing Related Documents:** The system provides functionality to sensitivity reviewers to sequentially review documents that are clustered by their semantic categories. As briefly discussed in Section 1, the sequential review of documents that belong to a semantic category can facilitate the understanding of associated sensitivities in the category to reviewers. Therefore, this sequential review of semantically related documents can improve the reviewers' reviewing speed and assist them in providing consistent judgements for related documents [9]. As shown in Figure 4, the reviewers are presented with the identified semantic categories prioritised by their predicted sensitivity probability. The reviewers can then select a category to review the associated documents sequentially in the order of their predicted sensitivity.

• **Collectively Reviewing Coherent Information Threads:** Apart from the document-by-document review, the system leverages information threads to provide the functionality of collectively reviewing coherent information from multiple documents that describe an event, activity or discussion. As shown in Figure 5, the system visualises related information about an event in chronological order and highlights the extracted 5W1H questions' answers to illustrate how the information is related. The collective review of such chronological and coherent information threads can enable reviewers to quickly identify a context of sensitive information that is spread across multiple documents, which is otherwise challenging in a document-by-document review scenario. For example, the thread shown in Figure 5 presents a discussion from multiple documents about the extradition of JTL terrorists from Flavania to Saplos (anonymised names of countries). The thread presents sensitive information, such as the names of individuals being extradited, along with potentially sensitive information about international relations between the governments of Flavania and Saplos.

• **Customised Review Prioritisation:** In large collections, semantic categories can contain many sensitive and non-sensitive documents. Therefore, the mean probability of documents being sensitive in a large semantic category may not be indicative to prioritise the category for review. For example, in a semantic category about politics, documents from internal government agencies are more likely to include sensitive information than political documents from sources such as publicly available media reports. Moreover, in a limited reviewing time budget, reviewers may not have available resources to accommodate large semantic document categories for the sequential document-by-document review. In such cases, the system enables the reviewers to split the large semantic categories into smaller finer-grained semantic groups using document metadata attributes. The system dynamically prioritises the smaller semantic groups for review based on their sensitivity probabilities. As shown in Figure 4, the reviewers can choose the desired criteria for splitting the semantic categories by one or more metadata attributes, namely, authors, origins or intervals of document creation date.

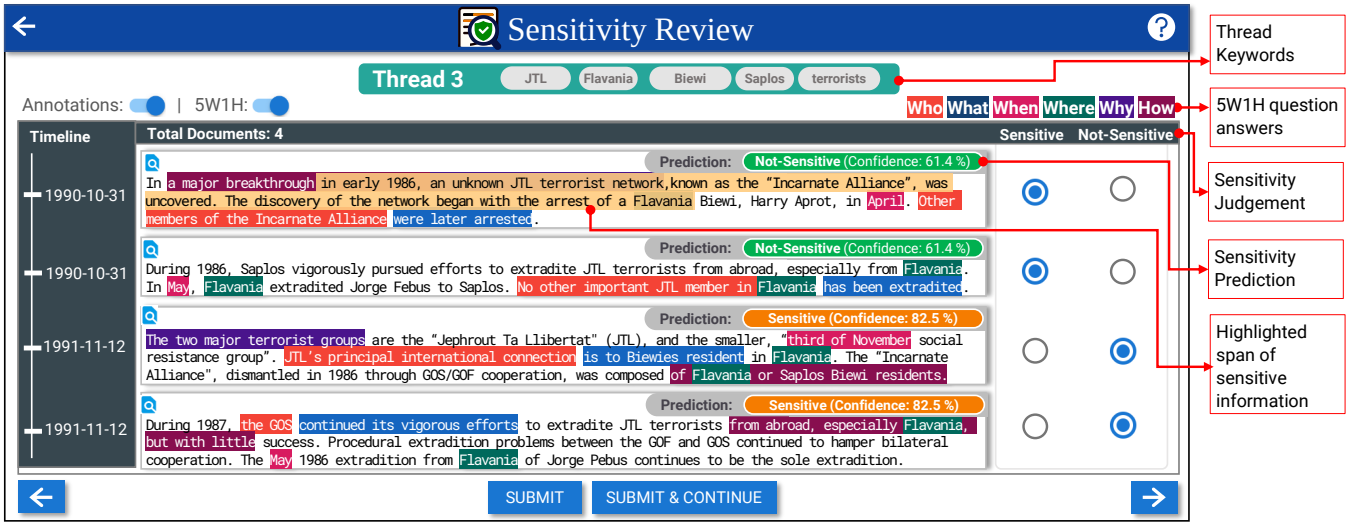

Figure 5: Visualisation of an Information Thread with highlighted 5W1H questions' answers, and options to collectively provide sensitivity judgements to document passages (Sensitive information such as real names of individuals/places are anonymised).

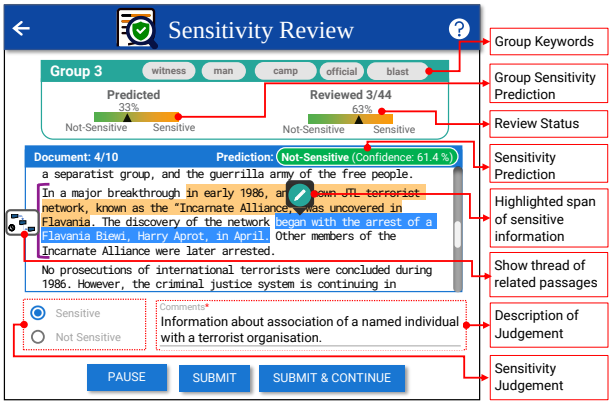

Figure 6: Various options to capture review judgements for each document sequentially in a document group.

• **Comprehensive Sensitive Review Judgements:** In a sensitivity review system, it is essential to record the sensitivity judgements in a manner that can succinctly explain the judgements for future reference. As shown in Figure 6, our system enables the reviewers to capture the sensitivity judgement for each document in three forms: (1) providing an overall sensitive or non-sensitive judgement for a document, (2) highlighting portions of the document text that correspond to a piece of sensitive information, and (3) providing a brief justification to describe why a document is judged as being sensitive. The reviewers are also presented with an option to explore the information threads corresponding to any text passages in the documents being reviewed to provide passage-level judgements.

Additionally, inspired by McDonald et al. [7], the system provides classification confidence scores for documents and document groups to the reviewers to improve their accuracy and efficiency.

5 EVALUATION

In our previous work [9], we conducted two user studies using the system that we present in this work. In the user studies, we evaluated the functionalities of our system for efficient sensitivity reviews. In our first user study, we evaluated the effectiveness of sequentially reviewing related documents. Our user study showed that sequentially reviewing related documents can significantly improve the reviewers' efficiency (15.65% Normalised Processing Speed [2], T-Test $p < 0.05$) without significantly affecting their accuracy compared to reviewing documents in a randomised sequence. In our second user study, we evaluated the effectiveness of review prioritisation based on the predicted sensitivity classification probabilities of document groups. The study showed that review prioritisation of the document groups can significantly improve openness (+23.8% Openness AUC [9], T-Test $p < 0.05$) compared to prioritisation of individual documents (i.e., without semantic categorisation). As future work, we plan to evaluate the effectiveness of information threads in assisting reviewers to identify sensitivities that are not apparent in a document-by-document review scenario.

6 CONCLUSIONS

In this paper, we have presented a system that can assist sensitivity reviewers in efficiently reviewing large document collections by identifying latent semantic categories and information threads. The system further aims to maximise the number of documents opened to the public in a limited reviewing time-budget by prioritising the documents for review. The system architecture seamlessly integrates information extraction and various machine learning modules. The web-based interface incorporates the functionality of sequentially reviewing semantically related documents or collectively reviewing coherent information threads from multiple documents. The system can be deployed in government departments to expedite the release of documents to the public to comply with FOI laws. Moreover, with the identification of groups of related document combined with the ability to organise them into finer-grained groups, the system can provide a customisable structure to large unstructured collections. As future work, we plan to investigate additional approaches for identifying relations between document groups, such as network communities of information threads.

REFERENCES

[1] Mustafa Abualsaud, Nimesh Ghelani, Haotian Zhang, Mark D. Smucker, Gordon V. Cormack, and Maura R. Grossman. 2018. A system for efficient high-recall retrieval. In *Proc. of SIGIR*. https://doi.org/10.1145/3209978.3210176

[2] Tadele T. Damessie, Falk Scholer, and J. Shane Culpepper. 2016. The influence of topic difficulty, relevance level, and document ordering on relevance judging. In *Proc. of ADCS*. https://doi.org/10.1145/3015022.3015033

[3] Django Software Foundation. 2021. Django. https://djangoproject.com

[4] Felix Hamborg, Corinna Breitinger, and Bela Gipp. 2019. Giveme5W1H: A universal system for extracting main events from news articles. In *Proc. of RecSys, INRA*. http://ceur-ws.org/Vol-2554/paper_06.pdf

[5] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2017. Enhancing sensitivity classification with semantic features using word embeddings. In *Proc. of ECIR*. https://doi.org/10.1007/978-3-319-56608-5_35

[6] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2018. Towards maximising openness in digital sensitivity review using reviewing time predictions. In *Proc. of ECIR*. https://doi.org/10.1007/978-3-319-76941-7_65

[7] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2020. How the accuracy and confidence of sensitivity classification affects digital sensitivity review. *ACM Transactions on Information Systems* 39, 1 (2020). https://doi.org/10.1145/3417334

[8] Hitarth Narvala, Graham McDonald, and Iadh Ounis. 2020. Receptor: A platform for exploring latent relations in sensitive documents. In *Proc. of SIGIR*. https://doi.org/10.1145/3397271.3401407

[9] Hitarth Narvala, Graham Mcdonald, and Iadh Ounis. 2022. The Role of Latent Semantic Categories and Clustering in Enhancing the Efficiency of Human Sensitivity Review. In *Proc. of CHIIR*. https://doi.org/10.1145/3498366.3505824

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011). http://jmlr.org/papers/v12/pedregosa11a.html

[11] UNESCO. 2019. Powering sustainable development with access to information: highlights from the 2019 UNESCO monitoring and reporting of SDG indicator 16.10.2. https://unesdoc.unesco.org/ark:/48223/pf0000369160

[12] Ngoc Phuoc An Vo, Fabien Guillot, and Caroline Privault. 2016. DISCO: A system leveraging semantic search in document review. In *Proc. of ICCL*. https://aclanthology.org/C16-2014

[13] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proc. of ICML*. https://doi.org/10.48550/arXiv.1511.06335