The Role of Latent Semantic Categories and Clustering in Enhancing the Efficiency of Human Sensitivity Review

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

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

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

ABSTRACT

Government documents must be manually sensitivity reviewed to identify and protect any sensitive information (e.g. personal information) in the documents before the documents can be opened to the public. However, due to the large volume of born-digital documents that need to be reviewed, there is a growing need for technologies to assist human reviewers and improve the efficiency of the review process. For example, in sensitivity review, a reviewer needs to be able to quickly find documents that belong to specific latent semantic categories (e.g., documents about criminality that contain the personal details of victims). However, manually identifying such document categories is a challenging task when reviewing digital documents, due to the size of, and lack of structure in the collections. We hypothesise that reviewing documents that are clustered by their latent semantic categories will increase the efficiency of the human reviewers, since the reviewers will be able to review related documents in sequence. In this work, we conduct a user study to evaluate the effectiveness of different clustering techniques, document metadata and automatic sensitivity classification, for grouping and prioritising documents for review, to increase the efficiency of the review process. Our study shows that reviewing documents in semantic clusters can significantly improve the efficiency (i.e., speed) of the sensitivity reviewers (+15.65%, T-Test, $p<0.05$) while maintaining the reviewers' accuracy. Moreover, we propose a novel strategy for prioritising document clusters for review to maximise the number of documents that are opened to the public within a fixed reviewing time budget. Our proposed prioritisation strategy results in a significant increase in the number of documents that are opened to the public (+37.99%, T-Test, $p<0.05$) compared to prioritising documents without clusters.

ACM Reference Format:
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 *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22), March 14–18, 2022, Regensburg, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3498366.3505824

CHIIR '22, March 14–18, 2022, Regensburg, Germany

ACM ISBN 978-1-4503-9186-3/22/03…$15.00
https://doi.org/10.1145/3498366.3505824

1 INTRODUCTION

Many government documents contain sensitive information such as personal or confidential information that is exempt from public release through Freedom of Information (FOI) laws. Therefore, such documents are manually reviewed to identify and protect any sensitive information before opening the documents to the public. However, the volume of born-digital documents that must be sensitivity reviewed is too large to be manually reviewed with the available reviewing resources, and government departments have frequently reported [1, 8, 21] a massive backlog in selecting documents for public release. Consequently, there is a growing need for tools to assist governments in conducting efficient and accurate sensitivity reviews.

In the sensitivity review process, there are primarily two types of users: (1) *Review Organisers*, who prioritise documents that are more relevant (likely to be opened) for review to maximise *openness*, i.e., the number of the documents selected for public release in a fixed review time budget. (2) *Sensitivity Reviewers*, who make judgements about whether documents contain any sensitive information. Both review organisers and sensitivity reviewers often seek information about latent semantic categories to understand the type of content in a collection. Semantic categories can group related documents about a specific subject domain (e.g. criminality), which can assist the sensitivity reviewers to quickly provide consistent review judgements for related documents. Moreover, different semantic categories can indicate how likely are the documents in a category to be sensitive, which can assist the review organisers in prioritising documents for review. For example, documents about "criminal incidents" may contain sensitive personal details of victims, whereas, in the documents about "political events", personal details of individuals are often publicly available and therefore not sensitive.

Document clustering is a popular approach in the literature for identifying semantic document categories in document collections. Moreover, previous studies [17, 22, 25] have shown the importance of document clustering to assist with human tasks in document review systems. In this work, we hypothesise that sensitivity reviewing documents that are clustered by their semantic categories can improve the efficiency of human reviewers since the semantic categories will provide the reviewers with additional useful information about the underlying context between related documents. We further hypothesise that document clustering can assist the review organisers in prioritising document groups for review, to maximise openness.

As shown in Figure 1, we propose a system for sensitivity review that leverages document clustering techniques to facilitate the understanding of the latent semantic categories of documents in a collection. We present a novel strategy of effectively prioritising document clusters for review by leveraging document metadata attributes (e.g. document author) and predictions from a sensitivity

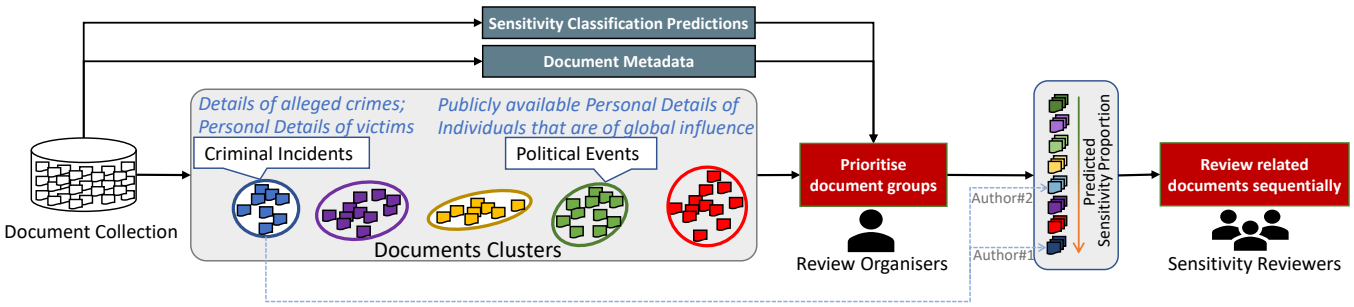


Figure 1: Document clusters for sensitivity review.

classifier [13]. In particular, we split document clusters into smaller semantic groups using document metadata attributes, and deploy a hierarchical ranking approach based on the predictions from the sensitivity classifier to first rank the smaller Cluster+Metadata groups by the proportions of predicted sensitivity and then rank the individual documents within the groups by predicted sensitivity.

We conduct two user studies that evaluate the effectiveness of: (1) different clustering techniques for grouping semantically related documents to increase the efficiency of sensitivity reviewers, and (2) document metadata and sensitivity classification for prioritising document clusters for review to assist the review organisers in increasing the number of documents opened to the public in a fixed time-frame. The contributions in this paper are 3-fold:

(1) We propose a system for sensitivity review that leverages document clustering to assist the human reviewers in understanding the type of content that is included in a collection.
(2) We conduct two user studies to evaluate the impact of reviewing documents clustered by their semantic categories on the efficiency, accuracy and openness of sensitivity review.
(3) We show that sensitivity reviewing documents in semantic clusters can significantly improve the review efficiency and openness while maintaining the accuracy of the review.

On a collection of government documents with real sensitivities (GovSensitivity [16]), our user studies show that reviewing documents in semantic clusters can significantly improve the reviewing speed by 15.65% (T-Test, $p < 0.05$). Furthermore, we show that our proposed review prioritisation strategy that leverages document metadata attributes for ranking clusters with finer grained sensitivity proportions can significantly improve the hourly openness by 37.99% and openness as a function of time (area under the curve) by 23.78% (T-Test, $p < 0.05$).

2 RELATED WORK

In this section, we discuss related work on the applications of clustering in assisting text-based human-computer interaction tasks, and technological solutions for assisting sensitivity reviewers.

Application of Clustering: Clustering is a popular technique used in interactive information retrieval (IR) systems. Previous studies, for example [3] and [2], have shown that clustering can be effectively integrated with IR systems to assist users and analyse user interactions. Bouadjenek *et al.* [3] performed clustering of twitter search results to assist users with coherent groups of related tweets. In particular, Bouadjenek *et al.* proposed a relevance-driven clustering approach to present relevant clusters to the users based on the user queries. Another work by Bogaard *et al.* [2] studied user interests and their search behaviour in a collection by clustering user search session database. Bogaard *et al.* implemented clustering using search metadata and click logs to gain insights from user behaviours such as parts of a collection that are most searched or parts where users spent most/least time. Differently from the works of Bouadjenek *et al.* and Bogaard *et al.*, in this work we perform a series of user studies to evaluate the effectiveness of clustering approaches for assisting human sensitivity reviewers and increasing the efficiency of sensitivity review.

Various studies, for example [17, 22, 25], have highlighted the importance of document clustering in document review systems. Oard *et al.* [17] discussed the importance of clustering in e-Discovery by identifying duplicates and near duplicates along with identifying chains of messages in an email collection. Vo *et al.* [25] presented a system called DISCO that implemented document clustering to assist reviewers by providing cluster keywords to perform complex exploratory search tasks. Trappey *et al.* [22] leveraged document clustering on legal documents to determine clusters of trademark litigation case documents as precedent for a given target case. In particular Trappey *et al.* deployed clustering in a recommendation setting by inferring the terminology associated with a legal case. In contrast to document review in e-Discovery, where the task is to find relevant documents in response to a request for production, sensitivity review is not a query-driven process and requires *all* of the documents in a collection to be reviewed. Therefore differently from the application of clustering in e-Discovery solutions (e.g., [17, 22, 25]) for identifying relevant document groups for review, our work is focused on leveraging the semantic relatedness of documents for the *prioritisation* of document groups, to maximise openness and improve review efficiency.

Sensitivity Review: In recent years, multiple studies have proposed different research directions for assisting sensitivity reviewers. Hutchinson [6] explored using topical clusters to identify documents that are relevant for sensitivity review. The work by Hutchison showed that topical clusters can identify semantic categories of documents, for example *human resource*, that may be likely to require to be sensitivity reviewed to protect personal information. Differently from the work of Hutchison, in this work, we study document clustering for sensitivity review in an interactive setting to evaluate the impact of reviewing documents in semantic clusters, on the efficiency, accuracy, and openness of sensitivity review. McDonald *et al.* [14] proposed a method of predicting the amount of time

that a reviewer would need to review a collection of documents and show that their proposed method can notably improve openness by prioritising the documents that are predicted to take less time to review. In contrast, in this work, we present a method of prioritising *groups* of documents for review to maximise openness. Another work by McDonald *et al.* [15] show 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 to the work of McDonald *et al.* [15], in this work we aim to improve the efficiency of sensitivity reviewers by reviewing documents in semantic clusters.

3 PRIORITISATION OF DOCUMENT CLUSTERS FOR SENSITIVITY REVIEW

As discussed in Section 1, document clusters can present an overview of semantic information in a document collection that we hypothesise will help the sensitivity reviewers to quickly provide sensitivity judgements for related documents, and improve their reviewing speed. However, clustering alone may not be sufficient to effectively improve *openness* since openness is dependent on two factors: (1) the reviewers' reviewing speed, and (2) the order in which documents are presented for review (i.e., review prioritisation). In this section we present our proposed review prioritisation strategy to prioritise document clusters for review using document metadata attributes and sensitivity classification to maximise openness, and illustrate the potential effectiveness of our proposed strategy using an example as shown in Figure 2. The example in Figure 2, compares the effectiveness of three ranking strategies for review prioritisation (that we discuss in this section) on a set of 8 documents to maximise the number of documents opened in the available reviewing time. In Figure 2, the document reviewing speed is controlled as constant to isolate the effect of review prioritisation on openness.

In a sensitivity review system that includes a sensitivity classifier, the review organisers can prioritise the documents in a collection that are more likely to be released, i.e., documents that are predicted to be non-sensitive. For example, in Figure 2(a), documents are ranked according to the increasing order of sensitivity classification probability P_S (least sensitive ranked at top), which leads to the prioritisation of only the predicted non-sensitive documents for review in the available reviewing time. However, ranking documents across the collection is not applicable when documents are semantically clustered to maintain the semantic grouping for quickly reviewing related documents. Moreover, in large collections, the proportion of predicted sensitivity within a cluster may not be an effective criteria for prioritising document clusters since large clusters can often contain a mix of many sensitive and non-sensitive documents. For example, in the "criminal incidents" cluster, documents from Author#1 may contain detailed information about a crime including personal sensitive information of victims, while documents from Author#2 may include general non-sensitive information about how a country is dealing with criminal activities. Figure 2(b) shows one such example of hierarchical ranking of clustered documents by first, the mean sensitivity probability of all documents d_c in a cluster c followed by predicted sensitivity $P_s^{d_c}$ for each document within c. Compared to the approach from Figure 2(a), in Figure 2(b) only 3 out of 4 prioritised documents are non-sensitive since all the documents from cluster C_3 (including the predicted sensitive document d_6) are ranked above the non-sensitive documents in other clusters (C_1 and C_2).

To address this problem of effective review prioritisation of semantic document groups, we propose to leverage document metadata attributes to split large clusters into smaller document groups (Cluster+Metadata) that can have finer-grained sensitivity proportions. This is illustrated in the previous example where the "criminal incidents" cluster can be divided into two groups respectively for each author (metadata) that are more indicative of potentially sensitive/non-sensitive information. The choice of metadata attribute(s) to split large clusters is specific to the document collection. For example, splitting the clusters using the documents' author attribute may not be suitable for a collection where most of the documents are published by different authors since the resulting Cluster+Metadata groups may be very small. In this work, we leverage the documents' author attribute in the GovSensitivity collection for splitting large clusters. However, depending on the collection, other metadata attributes such as a document's origin, year or month of creation can also be potentially useful. We deploy a hierarchical ranking strategy for review prioritisation where first the Cluster+Metadata groups are ranked according to the finer-grained proportions of predicted sensitivity followed by ranking the documents within the Cluster+Metadata groups. Figure 2(c) shows an example of our proposed review prioritisation strategy of hierarchically ranking the documents by the mean sensitivity probability of all documents d_{cm} in a cluster c having metadata attribute m followed by ranking the documents within Cluster+Metadata group cm by $P_s^{d_{cm}}$. In Figure 2, Cluster+Metadata strategy achieves similar openness to prioritising documents without clustering in the available reviewing time (4 documents each). In addition to the semantic grouping of documents by clustering techniques, document metadata attributes such as "Author" can group documents with similar structure and writing style that can help the reviewers to better understand the document structure in a group and reduce the difficulty in making sensitivity judgements. We hypothesise that Cluster+Metadata groups will benefit both: (1) Review Organisers, by opening more documents to the public within a particular reviewing time budget, and (2) Sensitivity Reviewers, by improving their reviewing speed through interpreting underlying context and structure of documents.

In the remainder of the paper, we present two user studies to evaluate the impact of reviewing documents in semantic clusters on the efficiency and accuracy of human sensitivity reviewers, and the effectiveness of the three review prioritisation strategies shown in Figure 2 in improving review openness.

4 PRELIMINARY SETUP

Before discussing the design of our user studies, in this section, we describe the preliminary setup that is required for conducting the user studies. In particular, we first describe: (1) the document collection used in the studies and for training clustering approaches, (2) the specific clustering approaches that we evaluate, (3) selection of the appropriate number of clusters in the collection.

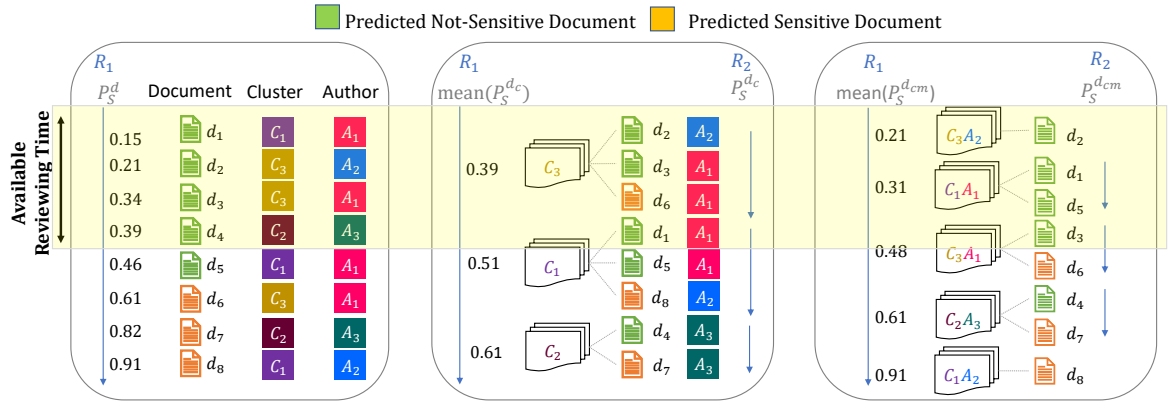


Figure 2: An example of different document ranking strategies for sensitivity review, where R_i is the i^{th} hierarchy level of the ranking, P_S is the predicted sensitivity probability, d_c is a document in cluster c, and d_{cm} is a document with metadata attribute m in cluster c.

Sensitivity Collection: To train the clustering approaches we use a collection (GovSensitivity [16]) of 3801 government documents (502 sensitive) that are annotated at document-level and sentence-level by government sensitivity reviewers for two FOI sensitivities, i.e, "Personal Information" and "International Relations". In the user studies we use passages of the documents instead of the documents itself to reduce the complexity in reviewing large documents. We split the documents into passages by leveraging the textual discourse units (paragraphs) in the documents, and utilise sentence-level annotations to label a passage as sensitive or non-sensitive.

Clustering Approaches: In our user studies, we evaluate three different clustering approaches from the literature.

• K-Means [11, 12]: K-Means is one of the most popular clustering technique in the literature. We deploy the scikit-learn [18] implementation of K-Means. To train K-Means, we construct TF-IDF term feature representation of the GovSensitivity passages and project the sparse TF-IDF vectors to a lower 200-D space using Latent Semantic Analysis (LSA).

• DEC [26]: DEC is a deep neural clustering approach that simultaneously learns feature representations by deploying a deep autoencoder [24], and learns clustering assignments by minimising the Kullback-Leibler (KL) Divergence Loss [9]. We utilise the publicly available implementation of DEC by Kim *et al.* [7], and leverage TF-IDF term features as input for the DEC autoencoder component.

• SCCL [27]: SCCL is a short-text clustering approach that leverages instance-wise contrastive learning and transformer-based [23] contextual word embeddings. We utilise the publicly available implementation of SCCL shared by the authors [27]. As described in [27], to support contrastive learning, we generate a pair of augmented passages for each of the GovSensitivity passages by word substitution using Bertbase and Roberta models of Contextual Augmenter[1]. We then determine the representation of the original and augmented passages using the *distilbert-base-nli-stsb-mean-tokens* model of the Sentence Transformer library [19].

[1]https://github.com/makcedward/nlpaug

Selecting the Number of Clusters: To determine the number of clusters k in the GovSensitivity collection, we first perform stratified sampling to split the collection across 5-folds to perform Cross Validation. We then perform K-Means on TF-IDF vectors of the passages in GovSensitivity for different values of k. We adopt two popular approaches for selecting k, *first*, the elbow method of plotting the within-cluster-sum-of-squares (WCSS) as a function of the number of clusters to identify the point that represents elbow of the curve as the value of k. WCSS is defined as:

$$WCSS = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \left\| x_i - \mu_j \right\|^2 \tag{1}$$

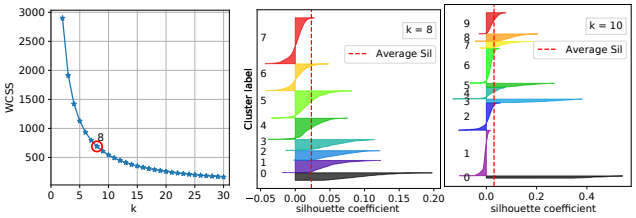
where x_i is the i^{th} input data point, $w_{ij} = 1$ if $x_i \in \text{cluster}_j$ and 0 otherwise, and μ_j is the cluster centroid. For the GovSensitivity collection, the elbow plot in Figure 3(a) shows the elbow around $k = 8$. *Second*, we perform silhouette analysis of the separation distance between the clusters by plotting the silhouette coefficient of each data point defined as:

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \tag{2}$$

where, for the i^{th} data point, a_i is the mean intra-cluster distance and b_i is the mean distance from the nearest-cluster that b is not a part of. From the silhouette plots for $k = 8, 10$ in Figure 3(b), we analyse: (1) clusters with below average silhouette scores, and (2) skewness in cluster sizes. Figure 3(b) shows that all clusters for $k = 8$ are above the average silhouette scores and the data partitions are less skewed as compared to the plot for $k = 10$. Therefore, we keep $k = 8$ to cluster the GovSensitivity collection.

5 USER STUDIES

In this section, we present the two user studies that we conducted to investigate if presenting human sensitivity reviewers with clusters of semantically similar documents to review can increase the efficiency of the sensitivity review process. Section 5.1 presents the reviewing interface that we developed for our user studies. Section 5.2 presents our first user study, which evaluates the impact of semantic clusters on the efficiency and accuracy of sensitivity



(a) Elbow Plot **(b) Silhouette Plots**

Figure 3: Selecting the value of number of clusters.

review. Section 5.3 presents our second user study, which evaluates our proposed review prioritisation strategy that leverages document metadata and sensitivity classification predictions to prioritise document clusters for increasing the overall openness of a review.

For both of the user studies we recruited participants using the MTurk[2] crowdsourcing platform, and restricted the participants to be aged 18+ years and from countries with English as their first language. The study participants were presented with multiple text passages and asked to make a *judgement* as to whether a passage *did* or *did not* contain any sensitive information, and to record a brief textual justification of their judgement. We particularly focused on personal sensitive information in our user studies, i.e., personal details of individuals that are not available in a public domain such as newspapers. For the passages that were presented to the participants, we sanitised all sensitive information, such as names of the individuals, to protect the identities of real persons. Before starting the study, the participants were provided with a detailed description of the sensitivity review task along with examples of sensitive personal information. To ensure that each of the participants understood the task, we validated that the participants achieved at least 50% accuracy on the sensitivity judgements to qualify their assignments for evaluation. We further validated whether the submitted justifications are relevant for the sensitivity review task.

5.1 Reviewing Interface

We developed a web-based reviewing interface for our user studies. The reviewing interface, as is shown in Figure 4, presents a series of documents to the study participants and enables the participants to: (1) highlight any text that they judge as being sensitive information, (2) record an overall judgement about whether a document is sensitive or non-sensitive, and (3) record a brief description of the judgement to justify why they judged the document as being sensitive. Additionally, the interface also records the amount of time that a reviewer takes to review each of the documents. The interface also allows the participants to pause the experimental system at any time, to ensure that the system recorded accurate reviewing times when a study participant took a rest break.

5.2 User Study#1: Review Efficiency

For our first user study, referred to as our Review Efficiency study, we deployed the clustering techniques that we presented in Section 4 (K-Means, DEC and SCCL) to evaluate the impact of reviewing documents in semantic clusters on the efficiency and accuracy of the sensitivity reviewers (i.e., the study participants). We expect that reviewing documents in semantic clusters will improve the efficiency (reviewing speed) of reviewers without affecting their accuracy. The Review Efficiency study is implemented as a *mixed* experimental design, i.e., we evaluate the impact of reviewing documents with or without clustering in a within-subject design, and we evaluate the effectiveness of the aforementioned clustering techniques in a between-subject design. Our Review Efficiency study aims to answer the following two research questions:

- **RQ1** Does sensitivity reviewing documents in semantic categories improve the reviewers' efficiency without affecting their accuracy?
- **RQ2** Which of the evaluated clustering techniques results in the greatest improvements in the reviewers' efficiency and accuracy?

Dataset: To select passages to use in the user study, we sampled 40 passages from the GovSensitivity collection in two different sets *A* & *B*, i.e., 20 passages per set. We use the cluster assignments from K-Means, DEC and SCCL for each of the sampled passages in Set A and Set B. Table 1 shows the average length of passages in each set along with the number of clusters assigned by each of the clustering methods. We controlled the number of sensitive documents in each set as 5, i.e., 25% of the sample size.

Table 1: Sampled passages for the Review Efficiency study.

	Length (words)		**Number of Clusters**		
	mean	**std**	**K-Means**	**DEC**	**SCCL**
Set A	72.5	7.71	4	6	4
Set B	73.3	10.39	4	6	3

Experiment Design: For RQ1, we evaluate the impact of reviewing documents that are clustered by each of the clustering techniques (K-Means, DEC and SCCL) compared to reviewing documents without clustering. We evaluate RQ1 in a within-subject experiment design, i.e., all participants in our experiment were presented with the same two tasks corresponding to the test conditions of RQ1, i.e, reviewing documents without clustering (No-Cluster) vs reviewing documents that are semantically clustered (Cluster). In the control condition (No-Cluster), the passages were presented randomly in a single batch, while in the treatment condition (Cluster) the passages were presented in semantic clusters. We used a different set of passages (A & B) for each of the conditions. Overall, every participant was required to review 40 passages (20 in each condition). For RQ2, we compare the effectiveness of the aforementioned three clustering techniques in improving the efficiency and/or accuracy of the reviewers. In RQ2, we chose to deploy a between-subject experimental design for the three test conditions (K-Means, DEC and SCCL) , since a within-subject design would require the participants to review $(1+3)*20 = 80$ passages. Having the participants sensitivity review 80 passages of text would have markedly increased the cognitive load for the participants and resulted in a high risk of participant fatigue. Therefore, to investigate RQ2, each participant was asked to review passages that are clustered by a single clustering method. As per the aforementioned mixed experiment design, we created 12 participant groups and counterbalanced the allocation of document sets and clustering approaches, as shown in Table 2. The study participants were also asked to complete a follow-up questionnaire at the end of each of the two tasks and

[2]https://www.mturk.com/

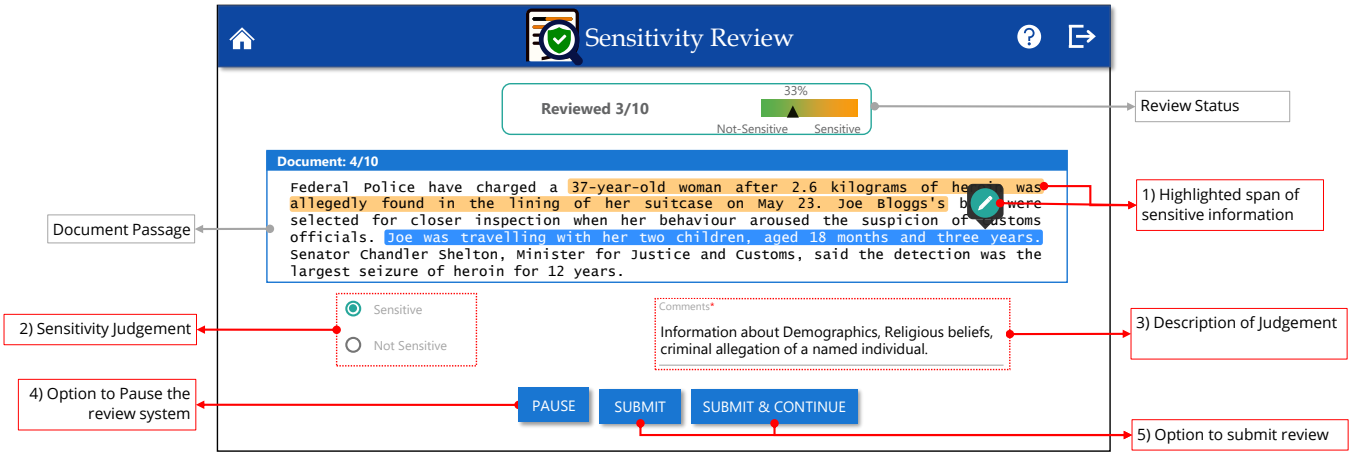


Figure 4: Components of the review interface.

a final questionnaire to analyse user preference between the control and treatment conditions along with ratings on task difficulty, cluster-interpretability and usefulness of cluster keywords.

Participant Recruitment: We recruited 42 participants for this study. The participants were remunerated $7.00 USD for completing the study. The mean completion time for the study was 40 minutes.

Evaluation Criteria: To evaluate the performance of participants, we deploy two criteria: (1) Balanced Accuracy (BAC) to evaluate the accuracy of participant reviews compared to the sensitivity labels in the GovSensitivity collection discussed in Section 4, and (2) Normalised Processing Speed (NPS) [4] to evaluate the efficiency of participants i.e., the reviewing speed in words per minute. NPS is defined as follows:

$$NPS = \frac{|d|}{\exp\left(\log\left(time\right) + \mu - \mu_\alpha\right)} \tag{3}$$

where $|d|$ is the length of a document d in words, $\log(time)$ is the natural logarithm of time taken by a reviewer to review document d, μ_α is the mean $\log(time)$ of the reviewer and μ is the global mean $\log(time)$ for all reviewers. The NPS calculation controls for variations in the documents' lengths and the reviewers' reading speeds.

We measure statistical significance in our mixed experiment design using two-way mixed ANOVA to test the interaction between our within-subject factors (i.e., No-Cluster and Cluster) and the between-subject factors (i.e., K-Means, DEC and SCCL). We report the observed power and the Partial Eta Squared (η^2) effect size for the ANOVA tests, and follow it with post-hoc tests using paired samples T-Test for the within-subject factor and one-way ANOVA for the between-subject factor. We select $p < 0.05$ as our significance threshold.

5.3 User Study#2: Review Openness

In our second user study, referred to as our Review Openness study, we evaluate the effectiveness of our proposed review prioritisation strategy (Cluster+Metadata) in terms of the number of documents that are released (opened) to the public in a fixed time frame, i.e., the review *openness*, compared to document prioritised without clusters (No-Cluster) and prioritisation of document clusters (Cluster) discussed in Section 3. The Review Openness study is implemented as

Table 2: Participant groups for the Review Efficiency Study.

Group	Task#1 Set	Task#2 Config	Task#2 Set
1	A	K-Means	B
2	B	K-Means	A
5	A	DEC	B
6	B	DEC	A
9	A	SCCL	B
10	B	SCCL	A

Group	Task#1 Config	Task#1 Set	Task#2 Set
3	K-Means	B	A
4	K-Means	A	B
7	DEC	B	A
8	DEC	A	B
11	SCCL	B	A
12	SCCL	A	B

a between-subject experimental design . The study aims to answer the following two research questions:

- **RQ3** Can the Cluster+Metadata review prioritisation strategy increase the number of opened documents in a specific time-frame?
- **RQ4** Does reviewing documents in Cluster+Metadata groups offer similar or improved review efficiency and accuracy compared to reviewing documents in clusters?

Dataset: We sampled 20 passages (mean length 95.05 words) from the GovSensitivity collection. We restricted the number of sensitive documents to 25% of the sampled passages (the same as for our Review Efficiency study). For this study, we chose only DEC clustering as it was found to be the best approach for improving reviewers' efficiency from the results of the Review Efficiency user study that we discuss in Section 6. We chose document author as the metadata attribute for splitting the DEC clusters into Cluster+Metadata groups. DEC assigned 3 cluster labels to the sampled passages that were further divided into 7 Cluster+Metadata groups.

Sensitivity Classification: We performed sensitivity classification on the documents in the GovSensitivity collection to obtain the sensitivity probabilities of the documents for review prioritisation as discussed in Section 3. We deployed an SVM text classification approach as described in [13] to classify the documents as either sensitive or non-sensitive. To train the classifier, we used a 5-fold cross validation with stratified samples of the GovSensitivity collection as described in [16]. The effectiveness of the learned classifier was 0.733 BAC.

Review Prioritisation: To prioritise the document passages for review as per the three strategies described in Section 3 (No-Cluster, Cluster and Cluster+Metadata), we performed hierarchical ranking based on the increasing order of following two scores: (1) the classification probability P_s of a document being sensitive, and (2) length L (in words) of a passage. For the sampled passages, we used the same value of P_s as for the document that contained the passage. The length parameter L provides more control on the ranking such that if two passages have same classification probability, then the shortest passage will be reviewed first. We ranked the passages for each of the three review prioritisation strategies as follows:

• No-Cluster: $argsort_{d_i \in D}(P_s^{d_i}, L^{d_i})$, i.e., ranking a passage d_i in a set D by the sensitivity probability $P_s^{d_i}$ followed by the length L^{d_i}.

• Cluster: $argsort_{d_i \in D}(\mu_c, P_s^{d_i}, L^{d_i})$, i.e., ranking d_i by the mean sensitivity probability μ_c of all the passages in cluster c that contains d_i followed by $P_s^{d_i}$ and L^{d_i}.

• Cluster+Metadata: $argsort_{d_i \in D}(\mu_m, P_s^{d_i}, L^{d_i})$, i.e., ranking d_i by the mean sensitivity probability μ_m of all the passages in a Cluster+Metadata group m that contains d_i followed by $P_s^{d_i}$ and L^{d_i}.

We note that the output of our proposed prioritisation strategies that rank clusters by their sensitivity classification probabilities can be dependent on the effectiveness of the deployed sensitivity classifier. Moreover, reviewing shorter documents first may not always lead to a faster review of documents, as is described by McDonald *et al.* [14]. In this study, we control these variables as constant across all the aforementioned review prioritisation strategies to isolate the effectiveness of our proposed Cluster+Metadata approach.

Experiment Design: We evaluate RQ3 and RQ4 in a between-subject design, i.e., each participant in our experiment was assigned to one of three review prioritisation configurations. The participants were each required to review 20 passages in a specific order, as was defined by their assigned prioritisation configuration. The participants were asked to complete a follow-up questionnaire at the end of the experiment to analyse the task difficulty and cluster-interpretability among the three configurations.

Participant Recruitment: For this study, we recruited 36 participants (12 in each group). The participants were remunerated $4.00 USD for completing the task. The mean time taken to complete the study was 25 minutes.

Evaluation Criteria: To evaluate openness of documents in our experiment under the three configurations (No-Cluster, Cluster and Cluster+Metadata), we deploy two metrics: *first*, Absolute Openness (O_{Abs}) to measure the number of documents selected for release per unit time defined as:

$$O_{Abs} = \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{n} t_i}, \qquad \lambda_i = \begin{cases} 1, & \text{if } d_i \text{ is non-sensitive} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where n is the number of documents that are to be reviewed, d_i is the document passage at the i^{th} position of the document ranking, and t_i is the time taken to review d_i. To account for the difference in reading speeds of the participants in our experiment, we use Normalised Dwell Time (NDT) [4] as the measure of reviewing time t_i, which is the denominator part of the NPS measure that we previously presented in Equation 3. *Second*, we deploy Openness AUC (Area Under the Curve) (O_{AUC}) to measure the number of documents selected for release as a function of time. We calculate O_{AUC} by determining the area under the curve for the plot between the cumulative count of non-sensitive documents in the particular review ranking and the cumulative sum of review time (NDT). We report the openness metrics using the sensitivity labels from the ground truth (True Labels, O^T_{Abs} & O^T_{AUC}) and from the reviewers' predictions (Predicted Labels, O^P_{Abs} & O^P_{AUC}). The O^P_{AUC} metric closely models the real-life sensitivity review scenario, where openness is measured as the interaction between the number of documents selected by the sensitivity reviewers for public release and the total time taken by the reviewers to achieve this number of selected documents. Therefore, we consider O^P_{AUC} as our main metric to measure openness. We also report BAC and NPS for the reviewers similar to the Review Efficiency study to evaluate RQ4 and to compare the consistency of results between the two user studies.

We measure statistical significance for our between-subject factor under three review prioritisation configurations using one-way ANOVA test. We report the observed power and the Partial Eta Squared (η^2) effect size for the ANOVA tests, and follow it with post-hoc tests using independent samples T-Test. We select $p < 0.05$ as our significance threshold.

Table 3: BAC and NPS of participants in different configurations of the Review Efficiency Study.

Group	Configuration	**mean BAC** (±95% CI)	**mean NPS** (wpm) (±95% CI)
1-4	No Cluster	**0.755** (±0.060)	132.22 (±8.77)
	K-Means	0.727 (±0.076)	**149.09** (±10.32)*
5-8	No Cluster	**0.790** (±0.067)	140.43 (±13.04)
	DEC	0.786 (±0.065)	**162.41** (±13.55)*
9-12	No Cluster	0.823 (±0.054)	138.99 (±5.69)
	SCCL	**0.846** (±0.069)	**151.99** (±9.85)*

6 RESULTS

In this section, we present the results of our user studies. We first discuss the results for the Review Efficiency study that we previously presented in Section 5.2, before discussing the results of our Review Openness study that we presented in Section 5.3.

In the Review Efficiency study, for the two-way mixed ANOVA significance test comparing the overall interaction between No-Cluster/Cluster (RQ1) and the different clustering techniques (RQ2: K-Means, DEC and SCCL), the data samples for NPS & BAC meet the assumptions of homogeneity of variance for the between-group factor (clustering techniques) as assessed by Levene's test ($p > 0.05$), and homogeneity of covariance as assessed by Box's test ($p > 0.05$). From the two-way mixed ANOVA tests we found that, for NPS (participants' reviewing speed), there is a statistically significant interaction between No-Cluster and Cluster conditions ($F(1, 35) = 56.158$, $p < 0.0005$, $\eta^2 = 0.616$, observed power = 1.0). However, there is no statistically significant interaction between No-Cluster/Cluster and the different clustering techniques for NPS ($F(2, 35) = 1.372$, $p = 0.267$, $\eta^2 = 0.07$, observed power = 0.275). For BAC (participants' reviewing accuracy), the results were found not significant as per the two-way mixed ANOVA ($p < 0.05$).

Table 4: Hourly Openness achieved by participants in different configurations of the Review Openness Study.

Configuration	True Labels		Predicted Labels	
	mean O^T_{Abs} (±95% CI)	mean O^T_{AUC} (±95% CI)	mean O^P_{Abs} (±95% CI)	mean O^P_{AUC} (±95% CI)
No-Cluster	37.324 (±2.651)	4.924 (±0.375)	27.095 (±3.729)	3.819 (±0.578)
Cluster	49.706 (±2.505)★	5.506 (±0.162)★	37.366 (±7.423)★	4.622 (±0.596)
Cluster + Metadata	**49.813** (±2.077)★	**5.671** (±0.209)★	**37.391** (±6.027)★	**4.727** (±0.579)★

To address RQ1, we measure statistical significance between the No-Cluster and Cluster conditions respectively for each of the Cluster configurations (K-Means, DEC and SCCL). We follow the two-way mixed ANOVA test with post-hoc tests using paired samples T-Test comparing difference in NPS between No-Cluster and Cluster conditions (denoted as "★" in Table 3, $p < 0.05$). From Table 3, we observe that our treatment condition (Cluster) shows significant improvements in participants' NPS compared to the control condition (No-Cluster) consistently for all clustering methods, with the best improvement in DEC (+15.65% wpm) followed by K-Means (+12.86% wpm) and SCCL (+9.35% wpm). We also observe that BAC of the participants slightly improves for SCCL clustering compared to No-Cluster (0.846 vs 0.823), although we observe slightly lower BAC in DEC (0.786 vs 0.790) and noticeable lower BAC in K-Means (0.727 vs 0.755). However, the differences in BAC between No-Cluster and Cluster conditions are not statistically significant as discussed in the results of the Two-Way Mixed ANOVA test. Therefore, in response to RQ1, we conclude that reviewing documents in semantic clusters can significantly improve (paired samples T-Test, $p < 0.05$) the efficiency (NPS) of the sensitivity reviewers without significantly affecting the reviewers' accuracy (BAC).

Moving on to RQ2, from Table 3 we observe that DEC achieved the best NPS, followed by SCCL and K-Means (162.41 vs 151.99 vs 149.09 wpm). However, in terms of BAC, SCCL achieved the highest BAC (0.846), followed by DEC (0.786) and K-Means (0.727). Therefore, the results show that the clustering methods DEC and SCCL are more effective in improving BAC and NPS of reviewers compared to K-Means clustering. We follow the two-way mixed ANOVA test with post-hoc test using one-way ANOVA comparing differences in NPS between the three clustering methods. We found that there are no significant differences in NPS for the different clustering techniques ($F(2, 35) = 1.308$, $p = 0.283$, $\eta^2 = 0.070$, observed power = 0.264). Therefore, in response to RQ2, we conclude that the improvement in reviewers' efficiency (NPS) by reviewing documents in semantic clusters is not significantly affected by specific clustering techniques.

Overall, in the Review Efficiency study, we found that semantic clustering of documents can indeed significantly improve the efficiency of human reviewers regardless of different types of clustering methods. We now discuss the effectiveness of prioritising semantic document clusters for review in improving openness based on the results of our Review Openness study that we presented in Section 5.3. We report the mean absolute openness (O_{Abs}) and openness AUC (O_{AUC}) along with 95% confidence intervals (CI) in Table 4. We also report the mean BAC and NPS along with 95% CI in Table 5. For the one-way ANOVA significance test comparing the three different review prioritisation approaches (No-Cluster, Cluster and Cluster+Metadata), the data samples for O^T_{Abs}, O^T_{AUC}, O^P_{Abs}, O^P_{AUC}, NPS and BAC meet the assumption of homogeneity of variance as assessed by Levene's test ($p > 0.05$).

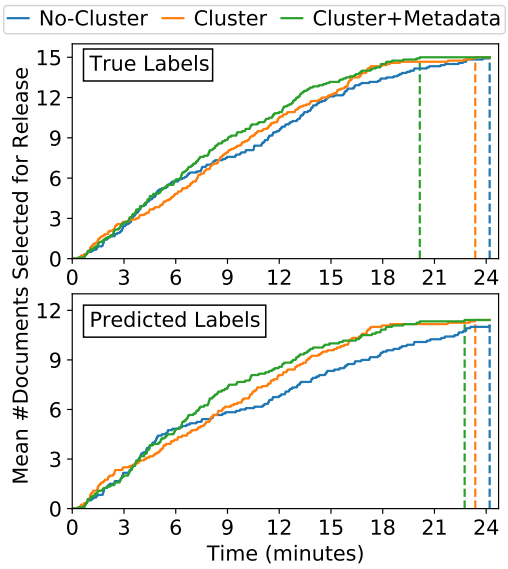


Figure 5: Number of Documents selected for release as a function of time in the Review Openness study

To address RQ3, we evaluate the openness of documents in the aforementioned three review prioritisation strategies. From one-way ANOVA tests we found that the interactions between the three review prioritisation strategies are significant for, O^T_{Abs} ($F(2, 33) = 30.910$, $p < 0.0005$, $\eta^2 = 0.652$, observed power = 1.0), O^T_{AUC} ($F(2, 33) = 7.464$, $p = 0.002$, $\eta^2 = 0.311$, observed power = 0.921) and O^P_{Abs} ($F(2, 33) = 3.536$, $p = 0.041$, $\eta^2 = 0.176$, observed power = 0.617), while not significant for O^P_{AUC} ($F(2, 33) = 2.547$, $p = 0.094$, $\eta^2 = 0.134$, observed power = 0.474). We follow the one-way ANOVA tests with post-hoc tests using independent samples T-Test comparing the pairs of the different review prioritisation strategies. In Table 4, statistically significant improvements compared to No-Cluster are represented as "★" (independent samples T-Test, $p < 0.05$). From Table 4, we observe that our proposed approach Cluster+Metadata achieves the best openness consistently across all four metrics. We also observe that both Cluster and Cluster+Metadata configurations significantly improve O^T_{Abs}, O^P_{Abs} and O^T_{AUC} compared to No-Cluster, however, for O^P_{AUC}, only the Cluster+Metadata improvements are statistically significant compared to No-Cluster (4.727 vs 3.819, independent samples T-Test, $p < 0.05$). Figure 5, shows the plot of the mean number of non-sensitive documents reviewed as a function of review time (NDT). The results from Table 4 are found to be consistent with Figure 5, where both Cluster

Table 5: BAC and NPS of participants in different configurations of the Review Openness study.

Configuration	**mean BAC** (±95% CI)	**mean NPS** (wpm) (±95% CI)
No-Cluster	0.781 (±0.064)	121.83 (±6.07)
Cluster	0.781 (±0.057)	**138.32** (±5.26)*
Cluster + Metadata	**0.847** (±0.069)	136.02 (±4.71)*

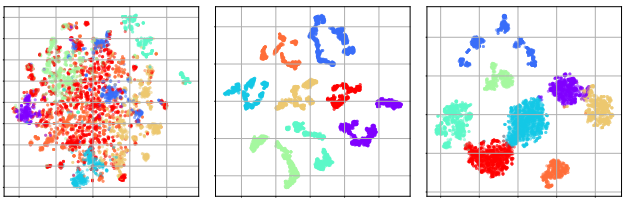
and Cluster+Metadata configurations show a higher number of non-sensitive documents compared to No-Cluster at any point in time, and the Cluster+Metadata configuration achieves the maximum number of non-sensitive documents earlier than the Cluster and No-Cluster configurations. Therefore, in response to RQ3, we conclude that our proposed approach of review prioritisation can significantly improve mean absolute openness and openness AUC calculated using both true labels (+33.4% O^T_{Abs} & +15.2% O^T_{AUC}) and predicted labels (+38.0% O^P_{Abs} & +23.8% O^P_{AUC}) compared to baseline No-Cluster. Moreover, even though none of the metrics shows significant improvement between Cluster and Cluster+Metadata approaches, only Cluster+Metadata approach shows a statistically significant improvement compared to No-Cluster in our main metric, i.e., openness AUC calculated using predicted labels (O^P_{AUC}).

Finally addressing RQ4, we evaluate the BAC and NPS for the participants between the three review prioritisation strategies. From one-way ANOVA tests for BAC and NPS, we found that the interactions between the three review prioritisation strategies are significant for NPS ($F(2, 33) = 9.720$, $p < 0.0005$, $\eta^2 = 0.371$, observed power = 0.972) and not significant for BAC, ($F(2, 33) = 1.298$, $p < 0.287$, $\eta^2 = 0.073$, observed power = 0.261). The post-hoc tests using independent samples T-Tests ($p < 0.05$) are represented by "$\star$" in Table 5 for Cluster and Cluster+Metadata conditions compared to No-Cluster condition. From Table 5 we found that the results for NPS and BAC between No-Cluster and Cluster conditions are consistent with the Review Efficiency study as presented in Table 3. Moreover, the NPS for the participants in both the Cluster+Metadata and Cluster conditions is comparable (136.02 wpm vs 138.32 wpm), and is significantly higher compared to the No-Cluster condition (121.83 wpm). Participants in the Cluster+Metadata condition show a noticeably higher BAC (0.847) compared to Cluster (0.781) and No-Cluster (0.781), however the improvements in BAC are not statistically significant as discussed in the results of one-way ANOVA test. Therefore for RQ4, we conclude that the Cluster+Metadata approach provides similar improvements in the reviewing speed as provided by reviewing documents in semantic clusters compared to No-Cluster.

7 ANALYSIS

In this section, we provide an offline cluster quality analysis and compare it with the qualitative analysis of the participants' responses to the follow-up questionnaires in our user studies.

Cluster Quality: We first analyse the quality of clusters identified by the three clustering methods (K-Means, DEC and SCCL) using two unsupervised metrics: *First*, Hopkins Statistics [5, 10] that measures cluster tendency of the representation of input data points. The value of Hopkins Statistics ranges between (0, 1), where the values closer to 1 means a representation of input data points is highly clusterable. *Second*, Silhouette Score [20] that measures cohesion, i.e. how similar a data point is to its own cluster compared to other clusters. Silhouette score as defined in Equation 2 ranges between [-1,1], where the values towards +1 indicate cohesive clusters.



(a) K-Means **(b) DEC** **(c) SCCL**

Figure 6: t-SNE 2-D visualisations of resulting clusters.

Table 6: Results for Cluster Quality.

	K-Means	**DEC**	**SCCL**
Hopkins Statistics	0.7766	**0.9897**	0.9080
Silhouette Score	0.0233	**0.9286**	0.5637

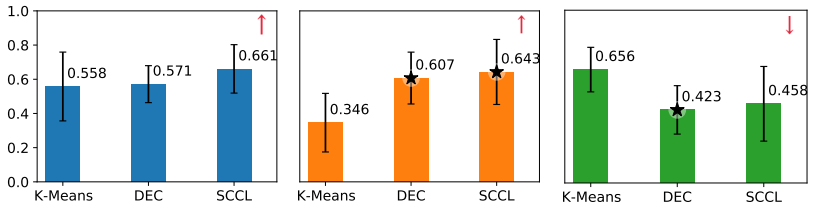
Table 6 presents the results of cluster quality evaluation of K-Means, DEC, and SCCL. In Table 6, the Hopkins statistics shows that all three input data representations, i.e., TF-IDF+LSA for K-Means, DEC's latent embeddings, and SCCL's contextual embeddings are clusterable ($H > 0.5$). Moreover the silhouette scores in Table 6 for the resulting clusters show that DEC forms very tight clusters followed by SCCL, while K-Means leads to overlapping clusters. These results are consistent with the 2-D t-SNE visualisations of the resulting clusters as shown in Figure 6. Overall, we observe that both DEC and SCCL are effective methods for producing quality clusters compared to K-Means for the GovSensitivity collection.

We also analyse the top keywords for each of the resulting clusters from the different clustering methods. Table 7 shows the top-5 keywords for two of the resulting clusters from each of the three evaluated clustering methods. In Table 7, even though the top keywords are often different for the resulting clusters in K-Means, DEC and SCCL, they represent the same high-level semantic categories, i.e. Middle-East and Commercial. We observed that the top keywords from the remaining clusters across the three clustering methods also represented the same semantic categories, i.e., Asia/Far-East, Politics, Medical, Education, Criminality and Legal-Trials.

Table 7: Top-5 Keywords for two clusters

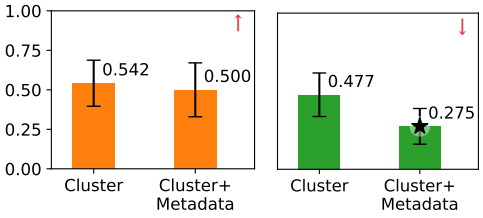
Cluster#1 (Middle-East)			**Cluster#2** (Commercial)		
K-Means	**DEC**	**SCCL**	**K-Means**	**DEC**	**SCCL**
iraq	turkey	president	percent	company	percent
turkey	eu	palestinian	company	law	market
israel	us	israel	investment	foreign	local
us	iraq	ha'aretz	bank	ipr	capital
israeli	turkish	turkey	market	investment	rate

Qualitative Analysis: We now present the analysis of participants' responses to the follow-up questionnaires in our user studies.



(a) Keyword Usefulness (b) Cluster Interpretability (c) Decision Difficulty

Figure 7: Normalised participants' ratings for the Review Efficiency Study. "⋆" denotes statistical significance compared to K-Means (T-Test, $p < 0.05$).



(a) Cluster Interpretability (b) Decision Difficulty

Figure 8: Normalised participants' ratings for the Review Openness Study. "⋆" denotes statistical significance (T-Test, $p < 0.05$).

In our Review Efficiency study (Section 5.2), 85.37% of the participants rated reviewing documents in semantic clusters as their preferred way compared to reviewing documents in a single large group. To compare the effectiveness of the three clustering methods (K-Means, DEC and SCCL), we normalise the participants' ratings (using L_2 norm) on: (1) how useful were the keywords to understand the context of each cluster, as shown in Figure 7(a). (2) how meaningful or interpretable the clusters were, i.e, whether the clusters contain semantically similar documents, as shown in Figure 7(b). (3) how difficult it was to make decisions about sensitivity of documents in semantic clusters, as shown in Figure 7(c). In Figure 7, "⋆" represents statistical significance compared to K-Means clustering as per independent samples T-Test ($p < 0.05$). From Figure 7(a), we observe that participant ratings for the usefulness of cluster keywords are comparable for the three clustering configurations. However from Figure 7(b), clusters from both DEC and SCCL were found to be significantly more interpretable than K-Means. The human interpretability of clusters is found to be consistent with the analysis of cluster quality, where both DEC and SCCL are found to be effective in producing quality clusters compared to K-Means. Interestingly, even though DEC's cluster quality was found to be better than SCCL (Silhouette Score 0.9286 vs 0.5637, Table 6), human interpretability for both methods is found to be comparable and even slightly higher for SCCL. Lastly, as shown in Figure 7(c), the participants rated lower difficulty in making sensitivity decisions for documents in both DEC and SCCL clusters compared to K-Means, where the decision difficulty for DEC is found to be significantly lower than K-Means. Considering the cluster quality results from Table 6 and decision difficultly ratings from Figure 7(c) we extend our response to RQ2 that DEC is indeed an effective document clustering approach for the GovSensitivity collection compared to K-Means and SCCL.

In the Review Openness study (Section 5.3), we analyse the normalised participants' ratings on cluster interpretability and decision difficulty between Cluster and Cluster+Metadata configuration. Figure 8(a) shows that the human-interpretability of the Cluster+Metadata groups is comparable to the original DEC Clusters. However, as shown in Figure 8(b) the participants rated significantly lower decision difficulty in making sensitivity decisions for the documents in Cluster+Metadata groups compared to the original DEC clusters. This analysis of decision difficulty supports our argument from Section 3 that reviewing documents in the Cluster+Metadata groups can indeed significantly reduce the difficulty for the sensitivity reviewers in making sensitivity judgments.

8 CONCLUSIONS

We presented a system for sensitivity review that leverages document clustering to assist human sensitivity reviewers by allowing them to quickly review related documents in semantic clusters. In addition, we proposed a review prioritisation strategy for effectively prioritising semantic clusters to assist the review organisers in increasing the number of documents opened to the public in a fixed reviewing time budget (i.e., openness). To investigate the impact of reviewing documents that are semantically clustered on the efficiency and openness of human sensitivity review, we developed a new tailored reviewing interface and conducted two user studies that evaluated the effectiveness of different clustering techniques, document metadata and automatic sensitivity classification, for grouping and prioritising documents for review. Results from our conducted user studies showed that reviewing documents into semantic clusters can significantly increase the efficiency (i.e., reviewing speed, +15.65% NPS, paired samples T-Test, p<0.05) of the reviewers without affecting their accuracy. We also showed that our proposed review prioritisation strategy that leverages document metadata and automatic sensitivity classification to prioritise semantic document clusters for review, can significantly improve openness (+23.8% O^P_{AUC}, independent samples T-Test (p<0.05) compared to prioritising documents by predicted sensitivity without clustering. We evaluated three different clustering methods (K-Means, DEC and SCCL) and found that the improvement in reviewing speed by reviewing documents in clusters is not significantly affected by the used clustering methods. However, from the qualitative analysis of the participants' feedback, we found that the neural clustering methods that we evaluated (DEC & SCCL) produced significantly more interpretable clusters compared to K-Means clustering on document term features (independent samples T-Test, $p < 0.05$). Our analysis of the cluster interpretability produced by the three clustering methods was found to be consistent with our offline analysis of the cluster quality where both DEC and SCCL were found to be more effective than K-Means clustering. As future work, we plan to analyse and investigate the underlying factors that help the sensitivity reviewers improve their efficiency while reviewing documents in semantic clusters.

REFERENCES

[1] Sir Alex Allan. 2014. Records review. *Cabinet Office and The National Archives* (2014). https://www.gov.uk/government/publications/records-review-by-sir-alex-allan

[2] Tessel Bogaard, Laura Hollink, Jan Wielemaker, Lynda Hardman, and Jacco van Ossenbruggen. 2019. Searching for old news: User interests and behavior within a national collection. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 113–121. https://doi.org/10.1145/3295750.3298925

[3] Mohamed Reda Bouadjenek and Scott Sanner. 2019. Relevance-Driven clustering for visual information retrieval on twitter. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 349–353. https://doi.org/10.1145/3295750.3298914

[4] Tadele T. Damessie, Falk Scholer, and J. Shane Culpepper. 2016. The influence of topic difficulty, relevance level, and document ordering on relevance judging. In *Proceedings of the 21st Australasian Document Computing Symposium*. 41–48. https://doi.org/10.1145/3015022.3015033

[5] Brian Hopkins and J. G. Skellam. 1954. A new method for determining the type of distribution of plant individuals. *Annals of Botany* 18, 2 (1954), 213–227. https://doi.org/10.1093/oxfordjournals.aob.a083391

[6] Tim Hutchinson. 2017. Protecting privacy in the archives: Preliminary explorations of topic modeling for born-digital collections. In *Proceedings of the 2017 IEEE International Conference on Big Data*. 2251–2255. https://doi.org/10.1109/BigData.2017.8258177

[7] Jaeyoung Kim, Janghyeok Yoon, Eunjeong Park, and Sungchul Choi. 2020. Patent document clustering with deep embeddings. *Scientometrics* 123, 2 (2020), 563–577. https://doi.org/10.1007/s11192-020-03396-7

[8] Jane E. Kirtley. 2006. Transparency and accountability in a time of terror: The Bush administration's assault on freedom of information. *Communication Law and Policy* 11, 4 (2006), 479–509. https://doi.org/10.1207/s15326926clp1104_2

[9] S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79 – 86. https://doi.org/10.1214/aoms/1177729694

[10] Richard G. Lawson and Peter C. Jurs. 1990. New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences* 30, 1 (1990), 36–41. https://doi.org/10.1021/ci00065a010

[11] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[12] James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*. 281–297.

[13] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2017. Enhancing sensitivity classification with semantic features using word embeddings. In *Proceedings of the 39th European Conference on Information Retrieval*. 450–463. https://doi.org/10.1007/978-3-319-56608-5_35

[14] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2018. Towards maximising openness in digital sensitivity review using reviewing time predictions. In *Proceedings of the 40th European Conference on Information Retrieval*. 699–706. https://doi.org/10.1007/978-3-319-76941-7_65

[15] 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

[16] Hitarth Narvala, Graham McDonald, and Iadh Ounis. 2021. RelDiff: Enriching knowledge graph relation representations for sensitivity classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 3671–3681. https://aclanthology.org/2021.findings-emnlp.311

[17] Douglas W. Oard and William Webber. 2013. Information retrieval for e-discovery. *Foundations and Trends® in Information Retrieval* 7, 2-3 (2013), 99–237. https://doi.org/10.1561/1500000025

[18] 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), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[19] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992. https://doi.org/10.18653/v1/D19-1410

[20] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[21] Derigan Silver. 2016. The news media and the FOIA. *Communication Law and Policy* 21, 4 (2016), 493–514. https://doi.org/10.1080/10811680.2016.1216686

[22] Charles V. Trappey, Amy J.C. Trappey, and Bo-Hung Liu. 2020. Identify trademark legal case precedents - Using machine learning to enable semantic analysis of judgments. *World Patent Information* 62 (2020), 101980. https://doi.org/10.1016/j.wpi.2020.101980

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[24] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research* 11, 12 (2010), 3371–3408.

[25] Ngoc Phuoc An Vo, Fabien Guillot, and Caroline Privault. 2016. DISCO: A system leveraging semantic search in document review. In *Proceedings of 26th International Conference on Computational Linguistics: System Demonstrations*. 64–68. https://aclanthology.org/C16-2014

[26] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*. 478–487. http://proceedings.mlr.press/v48/xieb16.html

[27] Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5419–5430. https://doi.org/10.18653/v1/2021.naacl-main.427