# SARA: A Collection of Sensitivity-Aware Relevance Assessments

Jack McKechnie j.mckechnie.1@research.gla.ac.uk University of Glasgow Glasgow, Scotland, UK Graham McDonald graham.mcdonald@glasgow.ac.uk University of Glasgow Glasgow, Scotland, UK

### **ABSTRACT**

Large archival collections, such as email or government documents, must be manually reviewed to identify any sensitive information before the collection can be released publicly. Sensitivity classification has received a lot of attention in the literature. However, more recently, there has been increasing interest in developing sensitivity-aware search engines that can provide users with relevant search results, while ensuring that no sensitive documents are returned to the user. Sensitivity-aware search would mitigate the need for a manual sensitivity review prior to collections being made available publicly. To develop such systems, there is a need for test collections that contain relevance assessments for a set of information needs as well as ground-truth labels for a variety of sensitivity categories. The well-known Enron email collection contains a classification ground-truth that can be used to represent sensitive information, e.g., the Purely Personal and Personal but in Professional Context categories can be used to represent sensitive personal information. However, the existing Enron collection does not contain a set of information needs and relevance assessments. In this work, we present a collection of fifty information needs (topics) with crowdsourced query formulations (3 per topic) and relevance assessments (11,471 in total) for the Enron collection (mean number of relevant documents per topic = 11,  $\sigma^2$  = 34.7). The developed information needs, queries and relevance judgements are available on GitHub and will be available along with the existing Enron collection through the popular ir datasets library. Our proposed collection results in the first freely available test collection for developing sensitivity-aware search systems.

### **CCS CONCEPTS**

• Information systems  $\rightarrow$  Test collections.

### **KEYWORDS**

test collection, relevance, sensitivity

### ACM Reference Format:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

# 1 INTRODUCTION

Sensitive information, for example information about a person's medical history, finances or citizenship, is often present in large collections of documents, such as email archives, internal company or government documents, or documents that are requested through e-discovery [10, 23] in litigation trials. Currently, collections that potentially contain sensitive information cannot be made available publicly without first being manually reviewed by experts to ensure that no sensitive documents are released. Sensitivity review is a time-consuming and expensive process that can prohibit many such collections from ever being released. For example, the 2014 US Presidential candidate Hilary Clinton released a collection of emails from when she was working for the government. Clinton sent 30,490 emails to the US State Department to be reviewed and released as quickly as possible. However, the reviewing process took nearly a full year with 25 people working on the review.

Archival document collections can be a valuable resource if they are made available to be searched. For example, email collections can be valuable to historians as they serve as a permanent record of an organisation's or person's activities [8] and medical records can be useful for medical research if they are made available [26]. Indeed, the UK Government has noted that some of the data that would be of the most value cannot be released due to concerns about sensitivity [2]. Most of the previous work on identifying sensitive information has focused on sensitivity classification. Sensitivity classification has been shown to be useful for helping sensitivity reviewers to increase the speed and accuracy of their review [17, 19] and to identify which documents should be prioritised for review to increase the number of documents that can be released with finite reviewing resources [20]. However, recently, there has been an increasing interest in developing search systems that are able to index entire collections (including any potential sensitivities) and deploy a retrieval model that is capable of providing the search engine users with relevant results while ensuring that no sensitive documents are included in the search results, for example in [25]. We refer to this as sensitivity-aware search.

The development of sensitivity-aware search engines requires test collections that contain relevance assessments for a set of information needs along with ground-truth labels for categories of sensitivities. There are a couple of test collections available that have made progress towards this goal. However, each of these collections has disadvantages for some researchers. The TREC 2010 Legal Track provides a collection of documents labelled for both relevance and for legal privilege [7]. Legal privilege is a form of sensitive information. However, it is very specific to the e-discovery context and, therefore, the collection is not suitable for general sensitivity tasks. Moreover, since the collection was assessed separately for relevance and for sensitivity, only a small subset of the documents

have both relevance and sensitivity assessments. The Avocado Research Email Collection [22, 24] contains relevance and sensitivity assessments. However, the Avocado collection is somewhat costly to obtain, which makes its use prohibitive to researchers that are not in a position to purchase it. Moreover, the collection is restricted in what it can be used for due to its licence agreement (for example, the collection cannot be used for crowdsourced experiments).

Another potentially useful test collection is the UC Berkeley version of the Enron email collection. The UC Berkley Enron collection provides classification labels for a rich taxonomy of classification categories [9], some of which are representative of sensitive information. For example, the *Purely Personal* and *Personal but in Professional Context* categories are representative of sensitive personal information. However, the existing collection does not contain a set of information needs and relevance assessments that are necessary for developing sensitivity-aware search systems.

In this work, we present an extension to the UC Berkeley version of the Enron email test collection  $^2$  to make the collection suitable for the development of sensitivity-aware search. Using a topic modelling approach, we identify fifty topics of discussion in the Enron emails and manually create short passages of text to represent each of the identified information needs (topics). Three representative query formulations are crowdsourced for each of the topics and a pooling approach is used to obtain crowdsourced relevance judgements (11,471 in total). The mean number of relevant documents per topic = 11 ( $\sigma^2$  = 34.7). Our extension to the Enron collection provides sensitivity-aware relevance assessments (SARA) that make the labelled Enron collection a valuable resource for the development of sensitivity-aware search.

The remainder of this paper is structured as follows: In Section 2, we discuss related work on sensitivity classification and sensitivity-aware search. In Section 3 we present the existing Enron Email collection before, in Section 4, presenting the process for constructing our extension to the collection for sensitivity-aware search. In Section 5, we present some analysis of our extension, before providing some additional discussions in Section 6 and concluding words in Section 7.

# 2 RELATED WORK/BACKGROUND

In this section, we first discuss previous work relating to sensitivity classification before, secondly, discussing work relating to sensitivity-aware search.

Gollins et al. [8] identified some of the challenges that are associated with sensitivity reviewing archival collections of digital documents, and how automatic sensitivity classification approaches may be able to assist with the human review process.

Since then there have been numerous works that have investigated using sensitivity classifiers to assist sensitivity reviewers, for example by highlighting documents which are likely to need to be reviewed and by automatically classifying documents entirely. For example, McDonald et al. [14, 15, 18] proposed several sensitivity

classification approaches in the context of predicting if information is exempt from being publicly released through the United Kingdom Freedom of Information Act 2000<sup>3</sup>. McDonald et al. [18] proposed using classification features which take into account the expected sensitivity-related risk that was related to the countries which were mentioned in the documents' text. Syntactic and semantic document features have also been shown to be useful for improving sensitivity classification [14, 15].

Narvala et al. [19] investigated the identification of latent semantic categories in document collections to improve the efficiency of human sensitivity reviewers and McDonald et al. [16] developed active learning strategies to reduce the amount of reviewing effort that is required to be able to train an effective sensitivity classifier. All of these works ([14–16, 18, 19]) used a test collection of 3801 real government documents that contained real government sensitivities. While it is important for these works to be able to evaluate their proposed approaches on real-life sensitivities, the highly sensitive content in the collection means that the collection is not easily sharable among researchers. Moreover, although the collection has a sensitivity classification ground truth, it does not contain a ground truth of information needs with relevance assessments, and is therefore not suitable for developing sensitivity-aware search approaches.

There has been comparatively little previous work investigating sensitivity-aware search approaches. Saved and Oard [25] proposed a variant of the Discounted Cumulative Gain metric [11], named Cost Sensitive Discounted Cumulative Gain (CS-nDCG), that incorporated both the relevance of the documents in a ranking as well as the preponderance of sensitive document that are in the ranking. The CS-nDCG metric facilitates the evaluation of retrieval models that try to filter out sensitive documents from the list of search results. Sayed and Oard [25] also proposed a novel Learningto-Rank approach that considered the probability of a document being sensitive, as defined by a pre-trained classifier, and integrated a loss function that optimised for normalised CS-nDCG. However, unable to use a collection containing real sensitivities, Sayed and Oard [25] evaluated their proposed approach using the OHSUMED dataset and a subset of PubMed labels (categories of diseases) as a proxy for sensitive categories. This highlighted the lack of, and need for, a freely available and easily accessible test collection that contains sensitivity labels and relevance assessments for a set of queries, to evaluate sensitivity-aware search.

To try to address the lack of an available test collection for sensitivity-aware search, Sayed et al. [24] developed an extension to the Avocado Research Email Collection<sup>4</sup> that included judgments for sensitivity and relevance assessments for a set of information needs. In [24], the sensitivity labels were based on two fictional personas that were developed to take into account the behaviour patterns of people with a range of levels of risk-aversion as to what they consider to be sensitive. The Avocado collection from Sayed et al. [24] provides a useful test collection for evaluating sensitivity-aware search. However, it is somewhat costly to obtain, which makes its use prohibitive to researchers in certain cases, and the collection is also restricted in what it can be used for due to its

<sup>1</sup> https://bailando.berkeley.edu/enron\_email.html

<sup>&</sup>lt;sup>2</sup>There were a number of prior releases of the Enron Collection. The May 7 2015 version is the stable version of the collection and earlier versions of the collection should not be used for research. The UC Berkeley version of the Enron collection is a subset of the May 7 2015 collection. For the sake of brevity, in this work, we refer to the UC Berkeley version of the collection simply as the labelled Enron collection.

<sup>&</sup>lt;sup>3</sup>https://www.legislation.gov.uk/ukpga/2000/36/part/II

<sup>4</sup>https://catalog.ldc.upenn.edu/LDC2015T03



Figure 1: Enron Classification Categories.

licensing arrangements. For example, the collection cannot be used in crowdsourced experiments.

Finally, it is worth noting that, although Sayed et al. [24] introduced a collection that is suitable for sensitivity-aware search, our sensitivity-aware relevance assessments extension to the Enron collection provides a valuable new resource for the field for two main reasons: (1) our sensitivity-aware relevance assessments and the Enron email collection are both made available to the community free of charge and without any prohibitively restrictive licensing arrangements. This makes our extended Enron collection suitable as a first test collection of choice for exploratory researchers that are wanting to evaluate their ideas without having to commit to a substantive financial outlay to do so, or are wanting to perform crowdsourced evaluations of their experiments, and (2) sensitivity classifiers that are developed and evaluated on a narrowly defined set of sensitivities are unlikely to generalise well since sensitivity is often very broadly defined and can be subjective. The Enron text collection provides a range of different classification categories that are representative of different types of sensitivities. This provides opportunities to deploy experimental setups that explicitly evaluate how generalisable, in terms of sensitivity, the developed models are likely to be. Moreover, researchers can evaluate this further by using our extended version of the Enron collection alongside the Avocado collection from Sayed et al. [24].

### 3 EXISTING ENRON EMAIL COLLECTION

The original version of the Enron Email Collection was made available by the Federal Energy Regulatory Commission<sup>5</sup> (FERC) as a public archive of the commission's investigation into the Western Energy Crisis of 2000/2001 [1]. Originally, the collection contained ~1.5 million emails. However, there were also many duplicate emails and unused folder structures. An updated version of the Enron email collection with approximately 500,000 emails was made publicly available after the collection was acquired by researchers at the Massachusetts Institute of Technology<sup>6</sup> (MIT) [12]. The updates to the MIT version of the collection included converting invalid email addresses to valid placeholders, removing any attached files and removing a number of emails that Enron employees had requested to be removed. Subsequently, several updated versions of the collection were released with further filtering of the emails that were contained in the collection. The current version of the Enron

email collection [12] was released in 2015 and is available from the Carnegie Mellon University (CMU) website. $^7$ 

In order to aid the development of sensitivity-aware search engines, we aim to create a test collection which includes relevance assessments for topics, as well as sensitivity labels. For this to be possible, a collection with a number of different characteristics is needed. A variety of classes which can reasonably be considered as sensitive are essential as the collection needs to provide labels for sensitivity which reflect true sensitive information. The Hearst [9] labelled version of the Enron Email Collection is a subset of the CMU collection that contains 1702 emails that were annotated as part of a class project at UC Berkley. Students in the Natural Language Processing course were tasked with annotating the emails as relevant or not relevant to 53 different categories. Therefore, the Hearst [9] labelled version of the Enron email collection provides a rich taxonomy of labels which can be used for multiple definitions of sensitivity such as the Purely Personal and Personal but in a Professional Context. Figure 1 presents the categories that the emails were annotated for.

The collection also needs to have a diverse enough set of topics that are discussed in the documents, in order to develop information needs based on the topics. Previous works have identified plentiful and rich topics in the collection using topic modelling [4], further motivating the choice of the labelled Enron collection. The context of the creation of the emails in this collection also influences our decision to extend it. The emails were created in a real-world context, with employees in an actual company interacting with each other. This is reflective of personal and corporate email collections which are currently not available to be released due to their sensitive information. Finally, the collection must not be held under a restrictive license and must be available for little to no cost so that other researchers are able to use the collection to develop sensitivity-aware search solutions. The labelled Enron collection is freely available and not held under a restricting licence. As can be seen from Figure 1, there are three high-level categories in the collection: Coarse Genre, Included/forwarded information, and Emotional Tone. Each of the high-level categories has an associated set of up to nineteen subcategories that an email can be associated with.8 The categories and their subcategories were hand crafted by Hearst [9] and refined after discussions with the class. Each email was read and annotated by at least two students.

When using the labelled Enron collection as a collection for identifying sensitive information, there are a number of the subcategories labels that could reasonably be chosen to represent sensitive categories. For example, the *Shame* and *Worry / Anxiety* subcategories of the *Emotional Tone* category. However, 1699 of the emails are categorised for their course genre (e.g., *Company Business, Strategy, etc, Logistic Arrangement or Purely Personal*), whereas only 310 emails are annotated for their emotional tone. Therefore, in this work, we consider the *Coarse Genre* subcategories of *Purely Personal* and *Personal but in a Professional Context* to be sensitivity labels and emails with these labels are referred to as sensitive emails. The full distribution of classification labels of the *Coarse Genre* categories

<sup>5</sup>https://www.ferc.gov/

<sup>6</sup>https://www.mit.edu/

 $<sup>^7</sup> http://www.cs.cmu.edu/~enron/$ 

<sup>&</sup>lt;sup>8</sup>The subcategories of *Primary topics* are only applied to emails that are labelled as the subcategory 1.1 *Company Business, Strategy,* and the *Emotional Tone* subcategories are only applied to emails that are not neutral in tone.

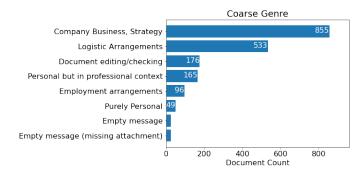


Figure 2: Distribution of classification labels across the subcategories of the Coarse Genre category

can be seen in Figure 2. There are 211 emails that have the *Purely Personal* and *Personal but in a Professional Context* labels, more than other potentially sensitive categories, and because personal sensitivities have been researched in the literature significantly, e.g., in [3, 15, 18]. The *Purely Personal* category includes invitations to weddings and events and conversations between family members. These are emails that are personal but do not include any relation to work being done at Enron. Conversely, the *Personal but in a Professional Context* category contains emails which are personal, but are related to work that was being done at Enron. This includes comments about the quality of people's work and expressions of feelings about employee treatment.

# 4 SENSITIVITY-AWARE SEARCH TEST COLLECTION

In this section, we present our sensitivity-aware relevance assessments (SARA) extension to the labelled Enron email collection. We deploy a topic modelling approach to identify topical themes in the labelled Enron collection that serve as a basis for our information needs and relevance assessments. Two separate crowdsourcing tasks are carried out in the development of SARA. Firstly, query formulations are crowdsourced to represent the information needs and, secondly, relevance assessments are crowdsourced for a pooled set of documents from the labelled Enron collection for each of the information needs. We discuss each of these processes in-turn before providing an overview of the resulting Enron collection sensitivity-aware relevance assessments extension.

**Topic Identification:** To create our set of sensitivity-aware relevance assessments for the labelled Enron email collection, we first identify a set of topical subjects that reflect the contents of the emails in the collection. Our identified topics are used to form descriptions of information needs that the crowdworkers are asked to provide query formulations for, i.e., example queries that a crowdworker would enter into a search engine to find relevant information to satisfy the information need. We use a topic modelling approach to identify the information needs. When identifying topics to be used as information needs, we are interested in identifying general themes that relate to the topics of discussion that might likely be covered in the contents (i.e., the body) of the emails in the collection. The topics are chosen to be broad enough to be able to reasonably expect that there would be relevant documents in the

**Information Need 14:** You are interested in flights that Enron employees took. You want to know more about how they were booked, where the employees went and what kind of employees were taking flights. You would also like to know more about the airlines that they flew on.

Query 1: Enron flights and airlines

**Query 2:** Flights booking enron employees method of booking destinations airlines employee status

Query 3: Which flights for enron employees on which airlines

**Information Need 49**: You are interested in taxes. You want to know what types of taxes Enron paid as an energy supplier, for example; taxes paid on income or electricity prices. You would also like to know more about how changes in tax rates affected the business.

Query 1: What type of taxes did enron pay

Query 2: What were Enrons business tax rates as an energy supplier

Query 3: History breakdown of Enrons tax activities

Figure 3: Two example information needs and queries gathered from the crowdsourcing tasks.

collection, and not so specific that it would require specialist knowledge to make a judgement of relevance on the subject. We choose to use a topic modelling approach since topic modelling is an efficient way of identifying topics within a collection of documents that is based on statistical modelling of the documents, rather than a blind search through documents attempting to identify topics manually.

We use the Gensim<sup>9</sup> implementation of Latent Dirichlet Allocation [5] topic modelling to generate fifty topics. As a sanity check to ensure that there are relevant and sensitive documents for each of the identified topics, we select the top 10 terms for each topic and use these as search terms to retrieve documents from the labelled Enron collection. We manually review the top  $\sim$  20 retrieved documents to ensure that at least one relevant document and one sensitive document is retrieved for each of the identified topics. We use PyTerrier [13] retrieval model to perform this sanity check.

Subsequently, satisfied that there are at least one relevant and one sensitive emails that are relatively easily retrievable for each of the identified topics, we manually construct short passages of text to serve as descriptions of the information needs that are to be searched for in the collection by the crowdworkers. Figure 3 presents two illustrative examples of the generated information needs. All of the information needs are of the same structure and are of similar size to the examples in Figure 3. The mean number of words in an information need passage is 41.48, with  $\sigma^2 = 27.37$  words.

Crowdsourced Query Formulations: In order to collect relevance assessments for pairs of emails and information needs, different query formulations are first needed to generate pools of documents. Query formulations for each topic are collected from crowdworkers from the Prolific crowdwork platform. Ten information needs are shown to each crowdworker and they are asked to provide a query formulation that they would use to get relevant documents to satisfy the information need they are presented with. Ten crowdworkers are recruited for each batch of ten information needs. Attention

<sup>9</sup>https://radimrehurek.com/gensim/

check questions are used to ensure that the crowdworker is paying attention. These questions are aligned with the guidelines of the Prolific Platform  $^{10}$ . Crowdworkers are presented with a short piece of text which contains the correct - randomised - answer to the question "What is your favourite colour?". The crowdworkers must select the correct option in a multiple choice question. Those who do not answer the attention check correctly are rejected on the Prolific platform. The crowdworkers are paid at a rate of £7/Hour. Therefore, the output of this crowdworking task is ten different queries for each of the fifty information needs are collected from crowdworkers, examples of which can be seen in Figure 3.

Pooling Documents for Relevance Assessments: Given that there are 1702 documents and 50 information needs, yielding 85,100 information need/query pairs, gathering relevance judgments for every information need, email pair is financially infeasible. Therefore, a select number of documents that are likely to be relevant to the given information need are chosen to be judged [27]. This is a pooling approach. Eighteen different retrieval models are used to generate the pools and the top documents in these pools are judged. The eighteen retrieval models were made up of each combination of 3 different queries, three different retrieval models (DPH, BM25, and PL2) and using Bo1 query expansion or not. The three different queries that are used to make the pools are chosen as they provide a diverse set of documents when used with the retrieval models and, after being read by the authors, they are deemed to be well formulated. The top twenty documents in the generated pools are chosen to be judged by crowdworkers.

**Crowdsourced Relevance Assessments:** In order to create a test collection useful for sensitivity-aware search, relevance labels for pairs of information needs and emails are required. Crowdworkers are shown an information need and an email and asked to rate the document as being either *Highly Relevant*, *Partially Relevant*, or *Not Relevant* to the information need.

If a crowdworker judges a document as being relevant, i.e., either *Highly Relevant* or *Not Relevant*, then they are asked to copy and paste the section of the email text that they thought made the email relevant. This is used as an attention and quality check in addition to the crowdworker being asked to complete an attention check question in the same style as the attention check that is used in the first crowdworking task. The system used for this crowdworking task can be seen in Figure 4. The crowdworkers recruited are paid at a rate of £7/Hour.

Each information need/document pair is judged by three crowdworkers and a majority vote is used to generate a ground truth label. Since each email-information need pair is judged by three crowdworkers and there are three possible labels, *Highly Relevant*, *Partially Relevant*, and *Not Relevant*, it is possible for each of the labels to be selected by one crowdworker. In practice, this only happened for 134 pairs. In such cases, ties are broken by having one of the authors read the document and make an additional judgement.

As the relevance labels are being crowdsourced, checks are done to ensure that a number of documents judged as relevant could be returned by baseline retrieval models. In order to ensure that

1. Information Need: You are interested in Enron's contacts in companies and organisations. You want to know more about the roles of these people and who they worked for. You are also interested in the ways in which Enron leveraged these contacts in different situations. Email: Message-ID: <28790569.1075852654964.JavaMail.evans@thyme Date: Mon. 18 Jun 2001 13:47:54 -0700 (PDT) From: kevinscott@onlinemailbox.net To: jeff.skilling@enron.com Subject: More Than Words Cc: sherri.sera@enron.com Mime-Version: 1.0 Content-Type: text/plain; charset=ANSI\_X3.4-1968 Content-Transfer-Encoding: 7bit Bcc; sherri.sera@enron.com X-From: Kevin Scott @ENRON X-To: Skilling, Jeff X-cc: Sera, Sherri X-bcc: X-Folder: \JSKILLIN (Non-Privileged)\Deleted Items X-Origin: Skilling-J X-FileName JSKILLIN (Non-Privileged).pst There aren't any "corporate speak" words that can express what I feel, so I will use plain English: I am thrilled that I will be working for you at Enron. I look forward to helping you and your team continue to transform Enron into the number one company in the world and I have sent my bio to Steve Kean and will send him my list of my preferred California contacts. By mid-week, I will send you a cleaned-up version of the suggestions I outlined in our meeting. I look forward to the next steps. It is an honor to be working with you again. I pledge to you my very best. Sincerely. [IMAGE] Contact Information E-mail kevinscott@onlinemailbox.net (213) 926-2626 Fax (707) 516-0019 Traditional Mail PO Box 21074 ?Los Angeles, CA 90021 - image001.png - image002.gif 9205 Highly Relevant 2. Copy and paste the sections of the email that make it relevant to the query. If the email is not relevant to the query then please enter N/A

Figure 4: The system used for the relevance assessments crowdworking task

sensitive documents definitely have relevance labels they were also judged by one of the authors for each of the information needs. The crowdsourcing tasks were given ethical approval by the University of Glasgow College of Science and Engineering Ethics Committee. Overview of the Enron Collection Sensitivity-Aware Relevance Assessments Extension:

The Enron Collection Sensitivity-Aware Relevance Assessments extension consists of fifty different information needs, three different queries for each information need, and 11,471 relevance judgments on a three-point scale of *Highly Relevant, Partially Relevant*, and *Not Relevant*. On average, each information need has 11.38 ( $\sigma^2 = 34.73$ ) relevant documents, and on average 6.02 ( $\sigma^2 = 26.72$ ) relevant documents that are sensitive. Figure 5 shows how many relevant and relevant sensitive documents each information

 $<sup>^{10}\</sup>mbox{https://researcher-help.prolific.co/hc/en-gb/articles/360009223553-Prolific-s-Attention-and-Comprehension-Check-Policy$ 

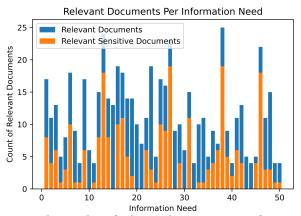


Figure 5: The number of relevant documents per information need.

need has. Information needs have the same numbering in Figure 5 as in the collection.

The Enron Collection Sensitivity-Aware Relevance Assessments extension is available at https://github.com/JackMcKechnie/SARA-A-Collection-of-Sensitivity-Aware-Relevance-Assessments. TSV files that include the queries, relevance assessments, and information needs are made available for download, along with a README file containing information about the collection. The sections in the emails that influenced the crowdworkers decisions on which label to give an email will also be released through the GitHub page. The files which contain the emails can be downloaded from the UC Berkley website<sup>11</sup>. The Enron Collection Sensitivity-Aware Relevance Assessments extension will also be made available via the popular ir\_datasets library. The dataset is held under an Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) licence which allows for it to be adapted, transformed and built upon.

# 5 COLLECTION ANALYSIS

The development of sensitivity-aware search engines requires test collections that have classification labels for sensitivity categories and relevance assessments for a set of information needs. Extending the existing Enron test collection with our sensitivity-aware relevance assessments results in such a collection. We now present some preliminary experiments to demonstrate the utility of our extended Enron test collection for developing sensitivity-aware search systems.

Previous work, e.g., Sayed and Oard [25], has proposed approaches for developing sensitivity-aware retrieval models that integrate sensitivity predictions directly into the retrieval model. However, when developing a sensitivity-aware search engine, a reasonable baseline approach is to train a sensitivity classifier separately from the retrieval model. The sensitivity predictions can then be applied to the search engine's results to filter out any documents from the results list that are predicted to be sensitive. Sayed and Oard [25] refer to this baseline approach as a *post-filtering* approach. In this section, we also deploy post-filtering sensitivity-aware search to illustrate the baseline properties of our extension to the Enron test collection.

Table 1: Classification scores on the Enron Email Collection.

	Precision	Recall	$F_1$	BAC
SVM	0.4856	0.5976	0.5358	0.7540
LR	0.3493	0.6923	0.4643	0.7548

In the remainder of this section, we first discuss the sensitivity classifiers that we deploy, and their effectiveness, before presenting our sensitivity-aware search experiments.

Sensitivity Classification: For our experiments, we train two different sensitivity classifiers, namely Support Vector Machine (SVM) and Logistic Regression (LR), to be deployed as two different postfiltering approaches in our sensitivity-aware search experiments. For both of the classifiers, documents (i.e., emails in the labelled Enron collection) are represented by their TF-IDF features. We use a stratified 20%/80% train/test split of the existing Enron test collection, resulting in ~ 12% of the emails in each of the splits being sensitive. The training set is downsampled to account for the unbalanced nature of the sensitivity distribution in the collection. We use scikit-learn<sup>13</sup> to train the classifiers (using the default parameters) and to generate the stratified train/test splits. As our classification metrics, we report Precision, Recall, F1 and Balanced Accuracy (BAC). We report BAC since it provides a generalisable measure of the performance of a classifier when the distribution of the class labels is unbalanced, as is the case in our collection. A BAC score of 0.5 denotes random predictions.

Table 1 presents the effectiveness of the sensitivity classifiers. As can be seen from the table, the SVM and LR classifiers correctly classify sim60% and sim69% of the sensitive emails respectively. The SVM classifier makes slightly fewer incorrect predictions than the LR classifier (0.4856 precision SVM vs 0.3493 precision LR), resulting in a higher  $F_1$  score for SVM (0.5358  $F_1$  SVM vs 0.4643  $F_1$  LR). The LR classifier achieves the highest BAC score with 0.4578 compared to 0.7540 for the SVM classifier.

Our deployed classifiers are only intended to illustrate the performance of a baseline sensitivity classification approach in post-filtering sensitivity-aware search. As such, we note that these classification results could be improved upon by deploying more sophisticated classification approaches, e.g., [15].

**Sensitivity-Aware Search:** In the remainder of this section, we deploy three post-filtering sensitive-aware search approaches to illustrate their performance on our extended Enron test collection. The post-filtering approaches first deploy an off-the-shelf retrieval model to retrieve an initial ranking of documents, before deploying a classifier to predict if each of the documents in the initial ranking are either sensitive or non-sensitive. Documents that are predicted to be sensitive are then removed from the initial ranking to form the final ranked list of results.

For our retrieval model, we deploy the PyTerrier [13] implementation of BM25 (with default parameters) to retrieve documents from the same 80% of the documents that is used for evaluating the performance of the classifiers (Table 1). We then deploy each of our trained classifiers, SVM and LR, separately and remove any documents from the ranking that are predicted to be sensitive. This results

 $<sup>^{11}</sup> https://bailando.berkeley.edu/enron\_email.html\\$ 

<sup>12</sup> https://ir-datasets.com/

<sup>&</sup>lt;sup>13</sup>https://scikit-learn.org/stable/

Table 2: Sensitivity-aware search scores using SARA

	P@10	R@10	nDCG@10	CS-nDCG@10
BM25-NoFilter	0.1893	0.2355	0.2009	0.7111
BM25-PostFilter_Oracle	0.1601	0.1841	0.1693	0.8623
BM25-PostFilter_SVM	0.1773	0.2230	0.1945	0.7236
BM25-PostFilter LR	0.1773	0.2292	0.2044	0.7280

in two post-filtering approaches, denoted as BM25- $PostFilter_{SVM}$  and BM25- $PostFilter_{LR}$  in the remainder of this section. Additionally, we deploy a post-filtering approach that uses the ground truth sensitivity classification labels to make perfect classification predictions and, therefore, removes all of the sensitive documents from the initial ranking, denoted as BM25- $PostFilter_{Oracle}$ . Lastly, we deploy a baseline approach that does not apply any post-filtering to remove sensitive documents, denoted as BM25-NoFilter. When indexing the documents, we remove stopwords and applying Porter stemming using PyTerrier.

We report Precision@10 (*P*@10), Recall@10 (*R*@10), Normalised Discounted Cumulative Gain@10 (*nDCG*@10) [11] and Cost Sensitive Normalised Discounted Cumulative Gain@10 (*CS-nDCG*@10) [25]. *CS-nDCG*@10 is an extension of Normalised Discounted Cumulative Gain which adds a penalty for returning sensitive documents. In other words, *CS-nDCG*@10 penalises a retrieval model if it returns a sensitive document in the results list but it allows for the model's recovery if the model returns many relevant (but non-sensitive) documents.

Table 2 presents the results of our sensitivity-aware search approaches and demonstrates the utility of our sensitivity-aware relevance assessments extension to the Enron Collection. Our expectation in terms of CS-nDCG@10 is that BM25- $PostFilter_{Oracle}$  should perform best since it has access to the true classification labels and, therefore, is able to remove all of the sensitive documents from the results list. The trained classifier post-filtering approaches (BM25- $PostFilter_{LR}$  and BM25- $PostFilter_{SVM}$ ) should achieve the next best performances, with the BM25-NoFilter approach achieving the worst performance in terms of CS-nDCG@10, since it is not able to filter out any sensitive documents. As can be seen from Table 2, our expectation holds with the approaches achieving 0.8623, 0.7280, 0.7236 and 0.7111 respectively.

Interestingly, we note that the  $BM25-PostFilter_{LR}$  approach results in a higher nDCG@10 score compared to the BM25-NoFilter approach. This shows that  $BM25-PostFilter_{LR}$  manages to correctly remove non-relevant sensitive documents from the ranking. Overall, these results illustrate the usefulness of the collection for evaluating sensitivity-aware search. Additionally, we note that there is room for improvement in terms of the deployed approaches. These results could be improved upon by deploying more sophisticated sensitivity-aware approaches, such the Opt.CS-nDCG approach from Sayed and Oard [25].

### 6 DISCUSSION

This section provides additional discussion points on the collection. We discuss the previous concerns about the contents of the Enron Email Collection and the size of the collection.

### 6.1 Contents of the Enron Email Collection

Previous work has had concerns about using the Enron Email Collection for the creation of a dataset for sensitivity-aware search [24]. We address these concerns in this section. Sayed et al. [24] argued that the ultimate goal of works looking at sensitivity-aware search is to protect sensitive information. Therefore, in their opinion, using crowdworkers to annotate a corpus which contains sensitive information goes against the goal of protecting sensitive information. However, the Enron Email Collection has been publicly available for over a decade and has undergone numerous redaction efforts to remove emails upon request by the authors. Therefore, those ex-Enron employees who felt that they did not want their emails to be read by others have an opportunity over the last 10 years to request the removal of their emails. Additionally, the emails that are shown to the crowdworkers to create our relevance assessments are publicly available and widely used. Consequently, we argue that the use of the Enron Email Collection to build a sensitivity-aware search test collection is justified and provides a valuable resource for the community.

# 6.2 Collection Size

Modern-day information retrieval test collections often contain hundreds of thousands, if not millions, of documents with relevance judgements [21, 28]. However, the SARA extension of the Enron Email collection contains only 1702 documents that have been judged for relevance. This is in the nature of the task of sensitivity-aware search - sensitive documents are not often publicly accessible, for the very reason that they are sensitive. Collections that only contain judgements for sensitivity that have been used previously for sensitivity classification are not publicly available and also contain a small number of documents [15].

In the task of sensitivity-aware search, having small collections that contain one specific genre of sensitivity is desirable. We want to be able to test retrieval systems on a number of test collections, each containing one type of sensitivity, so that we can see how well the system performs on each kind of sensitivity. This is due to the nature of the task and the situations that sensitivity-aware search systems would be deployable to. For example, if we desire to build a sensitivity-aware search engine for searching among the emails of authors donated to an archive, how well that system performs on national security sensitivities is not as important as how it performs on personally identifiable information sensitivities. Consequently, having multiple smaller test collections, that concentrate on one variety of sensitivity is more beneficial, and a more accurate representation of the quality of the system, than a larger collection containing many types of sensitivity.

Table 3: Retrieval over all the Enron documents

Name	P@10	R@10	Bpref	nDCG@10
BM25	0.002667	0.004611	0.126165	0.002759
BM25 » Filter (Oracle)	0.110667	0.106246	0.096130	0.127596
BM25 » Filter (Logistic Regression)	0.000667	0.000444	0.103885	0.000290
BM25 » Filter (Support Vector Machine)	0.000667	0.000444	0.108616	0.000290

If desired, it is possible to retrieve over all (500k) documents in the original Enron collection, not just the 1702 that have been

judged for relevance and sensitivity. This provides a larger collection to retrieve over, with sparse judgements for relevance. We present experiments demonstrating this, using the same classifiers as defined in Section 5, the results of which are shown in 3. In addition to the metrics introduced in Section 5, we also report Binary Preference (BPref) [6] as this measure is more resilient against significant deviations from the completeness assumption of the Cranfield Evaluation Methodology. However, as previously described, within the task of sensitivity-aware search, multiple, smaller test collections containing specific genres of sensitivities is a more relevant measure of the quality of a sensitivity-aware search engine.

# **CONCLUSIONS**

In this work, we have identified the need for test collections that are suitable for developing sensitivity-aware search systems. We also presented our sensitivity-aware relevance assessments (SARA) extension to the Enron email collection. LDA topic modelling was performed on the labelled Enron collection to identify topics that are present in the emails. These topics were then used to create a set of 50 information needs. Two crowdsourcing tasks were subsequently carried out. The first task involved sourcing query formulations for the information needs. We crowdsourced three queries per information need (150 queries in total). After carrying out a pooling approach to identify documents that were likely to be relevant to information needs, the second crowdsourcing task was carried out. This involved gathering relevance judgments for information needs. Preliminary experiments using the extended Enron test collection were also performed. Post-filtering sensitivity-aware search approaches were deployed using trained sensitivity classifiers. Our experiments illustrate the usefulness of our sensitivity-aware relevance assessments extension to the Enron email collection for evaluating sensitivity-aware search systems.

### REFERENCES

- [1] 2003. Final report on price manipulation in western markets: Fact-finding investigation of potential manipulation of electric and natural gas prices. FERC (Mar 2003), http://elibrary.ferc.gov/idmws/common/opennat.asp?fileID=9666688
- [2] 2019. AI Sector Deal. UK Government (2019). https://www.gov.uk/government/ publications/artificial-intelligence-sector-deal/ai-sector-deal
  [3] Sabah Al-Fedaghi and Abdul Aziz Rashid Al-Azmi. 2012. Experimentation with
- personal identifiable information. Intelligent Information Management 4 (2012).
- [4] Michael W Berry and Murray Browne. 2010. The 2001 annotated (by topic) Enron email data set. URL http://cis. jhu. edu/parky/Enron/Anno\_ Topic\_exp\_LDC. pdf (2010).

- [5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3 (2003), 993–1022.
- Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 25-32
- Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. 2010. Overview of the TREC 2010 Legal Track.. In TREC.
- Timothy Gollins, Graham McDonald, Craig Macdonald, and Iadh Ounis. 2014. On Using Information Retrieval for the Selection and Sensitivity Review of Digital Public Records.. In PIR@ SIGIR. 39-40.
- [9] Marti A Hearst. 2005. Teaching applied natural language processing: Triumphs and tribulations. In Proc. of Workshop on Effective Tools and Methodologies for Teaching NLP and CL.
- [10] Faith M Heikkila. 2008. E-discovery: Identifying and mitigating security risks during litigation. IT Professional 10, 4 (2008), 20-25.
- Kalervo Järvelin and Jaana Kekäläinen. 2003. Cumulated gain-based evaluation
- of IR techniques. *Transactions on Information Systems* 20, 4 (2003), 422–446. [12] Bryan Klimt and Yiming Yang. 2004. Introducing the Enron corpus.. In *Proc. of* CEAS.
- [13] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation in Information Retrieval using PyTerrier. In Proc. of ICTIR.
- [14] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2015. Using part-of-speech n-grams for sensitive-text classification. In Proc. of ICTIR.
- [15] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2017. Enhancing sensitivity classification with semantic features using word embeddings. In Proc. of ECIR.
- Graham McDonald, Craig Macdonald, and Iadh Ounis. 2018. Active learning strategies for technology assisted sensitivity review. In Proc. of ECIR.
- [17] 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 (TOIS) 39, 1 (2020), 1-34.
- Graham McDonald, Craig Macdonald, Iadh Ounis, and Timothy Gollins. 2014. Towards a classifier for digital sensitivity review. In Proc of ECIR.
- [19] 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 CHIIR @ SIGIR.
- [20] Hitarth Narvala, Graham McDonald, and Iadh Ounis. 2022. Sensitivity Review of Large Collections by Identifying and Prioritising Coherent Documents Groups. In Proc. of CIKM.
- [21] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. choice 2640 (2016), 660.
- [22] Douglas Oard, William Webber, David Kirsch, and Sergey Golitsynskiy. 2015. Avocado research email collection. LDC2015T03 (2015).
- [23] Douglas W Oard, Jason R Baron, Bruce Hedin, David D Lewis, and Stephen Tomlinson, 2010. Evaluation of information retrieval for E-discovery. Artificial Intelligence and Law 18 (2010), 347-386.
- Mahmoud F Sayed, William Cox, Jonah Lynn Rivera, Caitlin Christian-Lamb, Modassir Iqbal, Douglas W Oard, and Katie Shilton. 2020. A test collection for relevance and sensitivity. In Proc. of SIGIR.
- [25] Mahmoud F Sayed and Douglas W Oard. 2019. Jointly modeling relevance and sensitivity for search among sensitive content. In Proc of SIGIR.
- [26] Nemanja Vaci, Qiang Liu, Andrey Kormilitzin, Franco De Crescenzo, Ayse Kurtulmus, Jade Harvey, Bessie O'Dell, Simeon Innocent, Anneka Tomlinson, Andrea Cipriani, et al. 2020. Natural language processing for structuring clinical text data on depression using UK-CRIS. BMJ Ment Health 23, 1 (2020), 21-26.
- [27] Ellen M Voorhees. 2002. The philosophy of information retrieval evaluation. In Proc. of CLEF.
- [28] Ellen M Voorhees, Nick Craswell, Bhaskar Mitra, Daniel Campos, and Emine Yilmaz. 2020. Overview of the TREC 2019 Deep Learning Track. (2020).