A Test Collection for Relevance and Sensitivity

Mahmoud F. Sayed
University of Maryland
mfayoub@cs.umd.edu

William Cox
University of Maryland
wcox@terpmail.umd.edu

Jonah Lynn Rivera
University of Maryland
jriver15@terpmail.umd.edu

Caitlin Christian-Lamb
University of Maryland
cclamb@umd.edu

Modassir Iqbal
University of Maryland
miqbal@terpmail.umd.edu

Douglas W. Oard
University of Maryland
oard@umd.edu

Katie Shilton
University of Maryland
kshilton@umd.edu

ABSTRACT

Recent interest in the design of information retrieval systems that can balance an ability to find relevant content with an ability to protect sensitive content creates a need for test collections that are annotated for both relevance and sensitivity. This paper describes the development of such a test collection that is based on the Avocado Research Email Collection. Four people created search topics as a basis for assessing relevance, and two personas describing the sensitivities of representative (but fictional) content creators were created as a basis for assessing sensitivity. These personas were based on interviews with potential donors of historically significant email collections and with archivists who currently manage access to such collections. Two annotators then created relevance and sensitivity judgments for 65 topics, divided approximately equally between the two personas. Annotator agreement statistics indicate fairly good external reliability for both relevance and sensitivity annotations, and a baseline sensitivity classifier trained and evaluated using cross-validation achieved better than 80% F_1, suggesting that the resulting collection will likely be useful as a basis for comparing alternative retrieval systems that seek to balance relevance and sensitivity.

CCS CONCEPTS

• **Information systems → Test collections**.

KEYWORDS

test collection; relevance; sensitivity

ACM Reference Format:
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 *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401284

SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401284

1 INTRODUCTION

Test collections enable controlled experiments to characterize retrieval effectiveness. Typically, a test collection is thought of as having 3 components: documents, topics (i.e., information need statements), and relevance judgements (which record the degree of relevance of some document to some topic). Our work adds a fourth component: annotations that indicate not just which relevant documents should be found, but also which documents should be protected because they contain content that someone—the content provider or the searcher—would consider inappropriate to show. We refer to what should not be shown—even if relevant—as *sensitive*.

In this paper we describe such a test collection for email, in which the sensitivities to be protected are those of the (modeled) content provider. This is not the first information retrieval test collection to contain sensitivity judgments. Hearst reports on the creation of annotations for about 1,700 messages from the (then newly released) Enron email collection [4]. The messages were annotated as part of a class project for categories such as secrecy, shame, and purely personal.[1] These might be considered sensitivity annotations, which were added in addition to annotations for topics such as political influence, the California energy crisis, or government actions. However, the small scale of the annotated collection limits its utility for evaluation of information retrieval systems. Jabbari et al. took up the challenge at larger scale a couple of years later, annotating about 14,000 Enron email messages as one of six categories of business messages, or as one of three categories of personal messages [5]. Although we might reasonably treat thir "close personal" category as sensitive (as, for example, Hillary Clinton did when removing personal emails before turning over her professional email while serving as Secretary of State to the State Department), the six business categories are considerably broader than is typical of topics in an information retrieval test collection (e.g., core business, or routine admin). One promising result in that paper was that a classifier trained to detect close personal messages achieved 80% F_1, suggesting that sensitivity classification may be a tractable problem.

Perhaps the most ambitious effort to date involving sensitivity annotation for email has been the TREC 2010 Legal Track, which annotated of the Enron email collection for both relevance (to specific requests for "production" of documents on some topic germane to a lawsuit), and sensitivity (in that case, for the legal concept broadly referred to as "privilege" in which documents can be withheld from production because of, for example, attorney-client privilege or the

[1] http://bailando.berkeley.edu/enron/enron_categories.txt

attorney work-product doctrine) [3]. One limitation of the TREC Legal Track test collection, however, is that because the relevance detection and sensitivity detection tasks were modeled separately, different documents were annotated for relevance and for sensitivity. It is not straightforward, therefore, to use that collection to measure the ability of a system to find relevant documents while simultaneously withholding sensitive documents, even if those sensitive documents are relevant. For that we need a test collection of the type we introduce in this paper, in which every annotated document is annotated both for relevance and for sensitivity.

Since our goal is to protect the sensitive information in the collection then the last thing we would want to do is to highlight where to find that sensitive content. Doing so in a public test collection such as Enron thus raises some ethical concerns that do not seem to us to have yet been adequately commented upon. We have therefore chosen to annotate the Avocado Research Email Collection, which is distributed by the Linguistic Data Consortium on a restricted research license that includes content nondisclosure provisions [10]. This license effectively precludes crowdsourcing, so all annotations were performed locally. To further protect specific individuals, we created representative personas for two fictional individuals and we used those personas as a basis for sensitivity annotations (as described in more detail below). Relevance judgments were created for 65 topics, out of which one topic is excluded because it has no relevant documents. This paper describes the process by which the topics were designed and the relevance judgments performed, and it presents some experiment results that serve to illustrate the utility of the test collection for evaluation of information retrieval systems in which the protection of sensitive content is a design goal. The annotations will be distributed by the Linguistic Data Consortium as an addendum to the Avocado Research Email Collection.

2 TEST COLLECTION

This section describes the creation of the test collection.

2.1 Topic Creation

To test systems on relevance, a number of topics had to be created. To test systems on sensitivity some topics must find sensitive content. With these goals in mind, three people created 137 topics that were designed to explore many parts of an individual's life, from business matters like promotions and shareholders, to current events like the Olympics and Columbine, to personal matters, like drug use and vacations, to attitudinal topics like selfish and tired. While creating the topics, the creator searched the test collection to verify that at least some emails were possibly relevant to those topics; the emails noted as possible relevant were recorded for future use. Searching was done using terms from the topics themselves, and also other related terms. For each topic the creator created title, description, and narrative fields, as is often done in TREC topics.

2.2 Personas

'Sensitivity' is an ambiguous concept [8], with both personal [6] and social components [9]. Ideally, individuals would code their own content as sensitive to account for both individual preferences and social norms. But approaching individual Avocado employees would be intrusive, and is not allowed by the license terms. To understand consistent components of sensitivity in professional emails, we therefore conducted interviews with 10 archivists who had worked with email collections and 9 distinguished academics whose email collections are of potential interest to future scholars. Interviews focused on identifying types of content that creators deem sensitive.

Next, we translated the qualitative interview data into a form that coders could use to annotate the test collection by using personas, a concept drawn from the human-centered design literature [2, 11]. Personas are archetypal representations of users that include their goals, attitudes and other relevant design details. We found that respondents differed on both the diversity of information they found sensitive (from very few kinds of information to quite a lot), and how careful they felt they had been in their professional email practices (from uninhibited to circumspect). We grouped our respondents into three types: 1) the cautious writer: circumspect and very sensitive; 2) careful without cause: circumspect despite the fact that they found fewer types of information to be sensitive; and 3) the diarist: uninhibited in their email habits despite finding many kinds of content sensitive. Each category was informed by at least two interview respondents. Finally, we created personas as composite characters for the two categories that together reflected the broadest range of sensitivities, giving the characters names, and backstories loosely based on our interviews.

John Snibert was the cautious writer: motivated to donate his emails to an archive because of their documentation of his career, and relatively assured of his care in writing emails over the years, but worried that he may have overlooked some of the many kinds of information about which he was sensitive. In contrast, Holly Palmer was the diarist: reluctant to donate her emails because she knows how much sensitive information they contain. We gave each persona details about the character's background, how they use email, and lastly, their particular concerns about sensitivity and the email types they consider sensitive. We adapted these personas to a corporate setting to match the nature of the Avocado collection and then gave them to our annotators and asked them to infer sensitivity based on concerns expressed by the persona.

Our use of personas created a methodological challenge because the Avocado emails were authored by hundreds of individuals. Because creating hundreds of personas was untenable, and because the personas were written as composite people from a creator's point of view, we decided that the persona should represent the owner of any email that was being judged for sensitivity.

2.3 Relevance and Sensitivity Judgments

Annotation for relevance and sensitivity was performed by two of the three topic creators; both are co-authors of this paper. We formed pools to be annotated by combining the potentially relevant documents identified during topic creation with search results from 18 automatic ranked retrieval systems. We built these 18 systems by varying the search terms (title, title + description, or title + description + narrative), query expansion using blind relevance feedback (yes or no) and different retrieval models (DPH , BM25, or Cosine), all implemented in Terrier.[2] We merged the top 25 results

[2] http://terrier.org/

from each system, added the manual results from topic creation, and deduplicated. The resulting pools average about 100 documents.

We used Turkle to implement the system, as shown in Figure 1.[3] Each Human Intelligence Task addressed a single (topic, document) pair for a single persona with 3 questions: 1) How relevant is the email to the topic?, 2) Is the email sensitive according to the specified persona?, and, for contrast, 3) Is the email sensitive in the annotator's personal opinion?

Round	Topics	Rel. κ	Sens. κ
1	#14: Bias, discrimination #133: Cubicles	0.22	0.27
2	#134: Drinking and hangover #51: Shareholder	0.23	0.15
3	#17: Porn #21: Lawsuit	0.76	0.66
4	#14: Bias, discrimination #134: Drinking and hangover	NA	0.53

Table 1: Annotator agreement, training phase, 532 observations. Rounds 1-3: John Snibert, Round 4: Holly Palmer.

Annotators were trained on system use and persona decision-making. There were two topics used for each of 4 training rounds; 3 for John Snibert and then 1 for Holly Palmer. Annotator agreement was measured per round using Cohen's kappa [1], as shown in Table 1. The first round tested the annotation system, and helped to acclimate the annotators to the process. After the second round, annotator disagreements on relevance and sensitivity were discussed. For example, one difference involved a joke, leading to a discussion on whether certain kinds of jokes might be sensitive. The third round tested the level of agreement between the two annotators on John Snibert's persona, again followed by a discussion of disagreements. The last round used the Holly Palmer persona with two recycled topics, both to acclimate one annotator to that persona and to test the level of agreement on that persona.

Excluding the 6 training topics, 65 topics were then randomly selected from the full set of topics and annotated, 35 for John Snibert by one annotator and 35 for Holly Palmer by the other annotator. The last 5 of the 35 topics in each set were duplicated in the other set, permitting recomputation of annotator agreement on relevance at the last stage of the annotation process (see Table 2).

Kappa was computed as follows. We integrated a scale for our nominal variable, relevance, that measured highly relevant, somewhat relevant, and not relevant documents. Table 1 shows little difference between rounds 1 and 2, but a large difference on round 3 after the discussion following round 2. Table 2 shows that kappa on relevance at the end of the process was fairly high, except for Topic 135: Storage Space, which may have been more difficult to interpret consistently.

3 EMPIRICAL ANALYSIS

Both the John Snibert and Holly Palmer personas have sensitivity and relevance judgments for 35 topics. The annotator for John Snibert had a median annotation time of 36 seconds; the annotator for Holly Palmer had a median annotation time of 30 seconds. Over 35 topics, the John Snibert annotator found a mean of 29.6 relevant documents (to any degree) per topic and a mean of 45.2 sensitive documents per topic. Over a different set of 35 topics (5 of which were common with John Snibert), the Holly Palmer annotator also found a mean of 29.6 relevant documents (to any degree) per topic, but a mean of only 16.4 sensitive documents per topic. As Figure 2 shows, some documents have a high fraction of relevant documents that are also sensitive, whereas others do not.

Topics	Rel. κ
#13: Vacation	0.52
#84: Fax	0.70
#100: Chechnya	0.64
#113: Parents	0.71
#135: Storage Space	0.42

Table 2: Annotator agreement, test phase, 585 observations.

3.1 Sensitivity Classification

One use of this collection is to compare techniques for detecting whether a given email is sensitive. Classification results are presented in Table 3 for two classification models: Logistic regression and Support Vector Machine (SVM). We adopt a simple feature representation, using a bag of words from the email and attachment text. Results show that we can predict sensitivity with F_1 above above 80% for John Snibert and above 65% for Holly Palmer. The lower F_1 for Holly Palmer results from class imbalance, with fewer sensitive documents from her perspective.

3.2 Training Search and Protection Engines

Sayed and Oard described three designs for search and protection engines: pre-filtering, postfiltering, or jointly modeling relevance and sensitivity [12]. To illustrate the use of the collection for evaluation, we held out 20% of messages for training a sensitivity classifier and tested on the remaining 80%. Table 4 shows results for pre-filtering and for a no-filtering baseline. As expected, pre-filtering reduces nDCG, which measures relevance but not sensitivity, while improving nCS-DCG (Sayed and Oard's measure balancing relevance and sensitivity); the pattern is consistent over both personas.

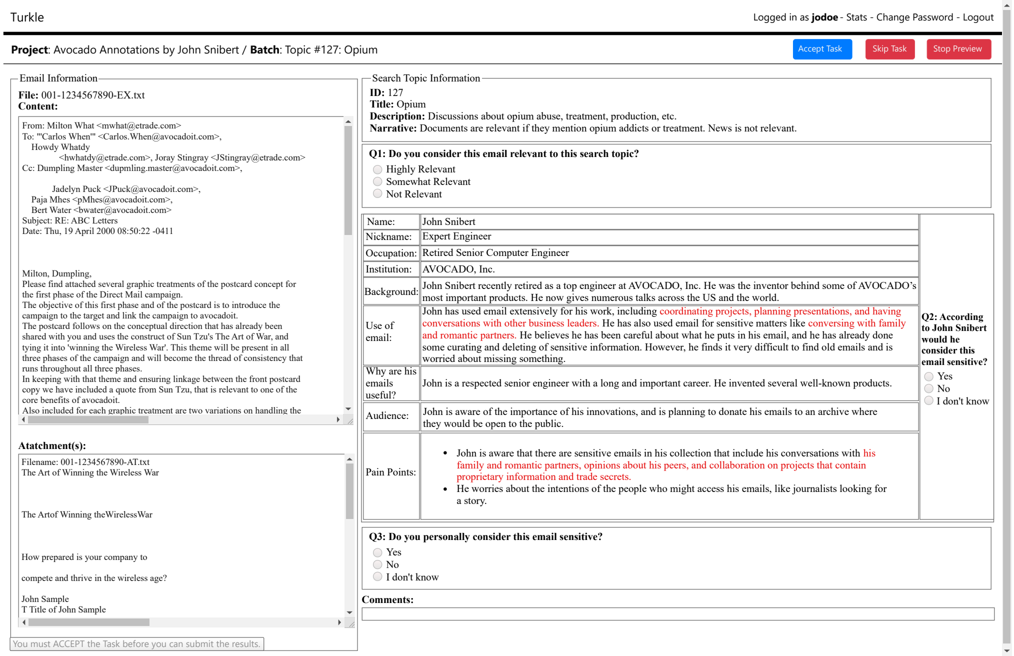
4 CONCLUSION AND FUTURE WORK

We have created a test collection to support experimentation with a search and protection engine, and we have illustrated the use of that collection both for training a sensitivity classifier and for evaluating a simple search and protection engine. Of course, much remains to be done. We already have a larger topic set, so we can extend our collection using the same procedures. We have sensitivity judgments both for personas and for the annotators themselves, so

Model	Precision	Recall	F1	Accuracy
Logistic Regression	78.34	83.33	80.70	79.85
SVM	80.15	82.09	81.05	80.60
Logistic Regression	69.41	63.80	66.38	87.97
SVM	74.09	58.17	65.03	88.37

Table 3: Sensitivity classifier results (upper John Snibert, lower Holly Palmer), 10-run mean, 20%/80% train/test split.

[3] https://github.com/hltcoe/turkle



Project: Avocado Annotations by John Snibert / **Batch**: Topic #127: Opium

Email Information

File: 001-1234567890-EX.txt
Content:

From: Milton What <mwhat@etrade.com>
To: '"Carlos When'" <Carlos.When@avocadoit.com>,
Howdy Whatdy
<hwhatdy@etrade.com>, Joray Stingray <JStingray@etrade.com>
Cc: Dumpling Master <dupmling.master@avocadoit.com>,

Jadelyn Puck <JPuck@avocadoit.com>,
Paja Mhes <pMhes@avocadoit.com>,
Bert Water <bwater@avocadoit.com>
Subject: RE: ABC Letters
Date: Thu, 19 April 2000 08:50:22 -0411

Milton, Dumpling,
Please find attached several graphic treatments of the postcard concept for the first phase of the Direct Mail campaign.
The objective of this first phase and of the postcard is to introduce the campaign to the target and link the campaign to avocadoit.
The postcard follows on the conceptual direction that has already been shared with you and uses the construct of Sun Tzu's The Art of War, and tying it into 'winning the Wireless War'. This theme will be present in all three phases of the campaign and will become the thread of consistency that runs throughout all three phases.
In keeping with that theme and ensuring linkage between the front postcard copy we have included a quote from Sun Tzu, that is relevant to one of the core benefits of avocadoit.
Also included for each graphic treatment are two variations on handling the

Atatchment(s):

Filename: 001-1234567890-AT.txt
The Art of Winning the Wireless War

The Artof Winning theWirelessWar

How prepared is your company to

compete and thrive in the wireless age?

John Sample
T Title of John Sample

You must ACCEPT the Task before you can submit the results.

Search Topic Information

ID: 127
Title: Opium
Description: Discussions about opium abuse, treatment, production, etc.
Narrative: Documents are relevant if they mention opium addicts or treatment. News is not relevant.

Q1: Do you consider this email relevant to this search topic?

○ Highly Relevant
○ Somewhat Relevant
○ Not Relevant

Name:	John Snibert	**Q2: According to John Snibert would he consider this email sensitive?** ○ Yes ○ No ○ I don't know
Nickname:	Expert Engineer	
Occupation:	Retired Senior Computer Engineer	
Institution:	AVOCADO, Inc.	
Background:	John Snibert recently retired as a top engineer at AVOCADO, Inc. He was the inventor behind some of AVOCADO's most important products. He now gives numerous talks across the US and the world.	
Use of email:	John has used email extensively for his work, including coordinating projects, planning presentations, and having conversations with other business leaders. He has also used email for sensitive matters like conversing with family and romantic partners. He believes he has been careful about what he puts in his email, and he has already done some curating and deleting of sensitive information. However, he finds it very difficult to find old emails and is worried about missing something.	
Why are his emails useful?	John is a respected senior engineer with a long and important career. He invented several well-known products.	
Audience:	John is aware of the importance of his innovations, and is planning to donate his emails to an archive where they would be open to the public.	
Pain Points:	• John is aware that there are sensitive emails in his collection that include his conversations with his family and romantic partners, opinions about his peers, and collaboration on projects that contain proprietary information and trade secrets. • He worries about the intentions of the people who might access his emails, like journalists looking for a story.	

Q3: Do you personally consider this email sensitive?

○ Yes
○ No
○ I don't know

Comments:

Figure 1: Sample Annotation Task (synthetic message and attachment created for public dissemination).

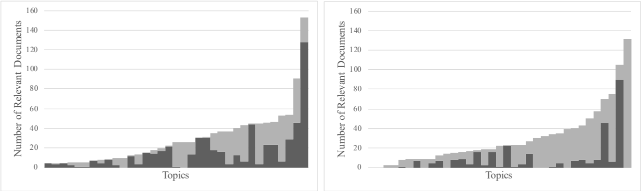


Figure 2: Relevant docs to any degree (top) and sensitive relevant docs (bottom), by topic. Snibert (left), Palmer (right).

Model	John Snibert		Holly Palmer	
	nDCG	nCS-DCG	nDCG	nCS-DCG
BM25/no filter	0.33	0.69	0.27	0.75
BM25/prefilter	0.25	0.84	0.24	0.78

Table 4: Results@10 for retrieval (nDCG) and for balancing retrieval and protection (nCS-DCG), 35 topics per persona.

we can study the degree to which agreement on sensitivity varies by topic, by annotator, and by whether that annotator is making personal or inferred judgements. We have shown that by using words as features we can obtain potentially useful levels of classification accuracy for sensitivity, but considerably better sensitivity classifiers might result from using semantic features [7], or the addition of contextual features such as times at which messages were sent, burst patterns in those times, people to whom or from whom those messages were sent, the roles of those people in the organization whose email we are working with, social networks among those people, and the temporal dynamics of the message flows in those networks. We have shown that a simple protect-then-search engine can achieve a degree of protection using one evaluation measure, but we might do better by jointly modeling those two goals, and we might productively consider a broader range of evaluation measures. Perhaps the greatest limitation of our work is that we now have only one test collection, and thus it is difficult to separate what works well in general from what works well on this specific collection. Nonetheless, by moving from no collections to one we have gained insights to support a robust evaluation infrastructure for the critical task of providing privacy-sensitive access to important text collections.

ACKNOWLEDGMENTS

The authors are grateful to Benjamin Mao for his help with topic creation. This work was supported in part by NSF grant IIS-1618695.

REFERENCES

[1] R Artstein and M Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.
[2] A Cooper. 2004. The origin of personas. *Innovation* 23, 1 (2004), 26–29.
[3] G Cormack et al. 2010. Overview of the TREC 2010 Legal Track. In *TREC*.
[4] M Hearst. 2005. Teaching applied NLP: Triumphs and tribulations. In *ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL*.
[5] S Jabbari et al. 2006. Towards the Orwellian Nightmare: Separation of Business and Personal Emails. In *COLING/ACL poster sessions*.
[6] K Martin and H Nissenbaum. 2016. Measuring Privacy: An Empirical Test Using Context to Expose Confounding Variables. *Colum. Sci. & Tech. L. Rev.* 18 (2016).
[7] G McDonald et al. 2017. Enhancing Sensitivity Classification with Semantic Features Using Word Embeddings. In *ECIR*.
[8] D Mulligan et al. 2016. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Phil. Trans. Royal Soc. A* 374 (2016).
[9] H Nissenbaum. 2009. *Privacy in context: Technology, policy, and the integrity of social life.* Stanford University Press.
[10] D Oard et al. 2015. Avocado Research Email Collection. LDC2015T03.
[11] J Pruitt and T Adlin. 2010. *The persona lifecycle: keeping people in mind throughout product design.* Elsevier.
[12] M Sayed et al. 2019. Jointly Modeling Relevance and Sensitivity for Search Among Sensitive Content. In *SIGIR*.