Providing More Efficient Access to Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege

JASON R. BARON, University of Maryland, USA
MAHMOUD F. SAYED, University of Maryland, USA
DOUGLAS W. OARD, University of Maryland, USA

At present, the review process for material that is exempt from disclosure under the Freedom of Information Act (FOIA) in the United States of America, and under many similar government transparency regimes worldwide, is entirely manual. Public access to the records of their government is thus inhibited by the long backlogs of material awaiting such reviews. This paper studies one aspect of that problem by first creating a new public test collection with annotations for one class of exempt material subject to the deliberative process privilege, and then by using that test collection to study the ability of current text classification techniques to identify those materials that are exempt from release under that privilege. Results show that when the system is trained and evaluated using annotations from the same reviewer, even difficult cases can often be reliably detected. However, results also show that differences in reviewer interpretations, differences in record custodians, and differences in topics of the records used for training and testing can pose challenges.

CCS Concepts: • **Information systems → Clustering and classification**.

Additional Key Words and Phrases: Sensitivity review; Evaluation; Freedom of Information Act; Deliberative process privilege

ACM Reference Format:
Jason R. Baron, Mahmoud F. Sayed, and Douglas W. Oard. 2021. Providing More Efficient Access to Government Records: A Use Case Involving Application of Machine Learning to Improve FOIA Review for the Deliberative Process Privilege. *ACM J. Comput. Cult. Herit.* 1, 1, Article 1 (January 2021), 19 pages. https://doi.org/10.1145/3481045

1 INTRODUCTION

Early on in the COVID-19 pandemic, UNESCO recognized that the "way the world is responding to this unprecedented global crisis will be part of history books." UNESCO went on to say that "[m]emory institutions, including national archives . . . are already recording the decisions and actions being made which will help future generations to understand the extent of the pandemic and its impact on societies," and that "it is important, now more than ever, for memory institutions to become even more readily accessible to researchers, policymakers . . . and the community at large" [35].

Authors' addresses: Jason R. Baron, University of Maryland, College Park, Maryland, USA, jrbaron@umd.edu; Mahmoud F. Sayed, University of Maryland, College Park, Maryland, USA, mfayoub@cs.umd.edu; Douglas W. Oard, University of Maryland, College Park, Maryland, USA, oard@umd.edu.

XXXX-XXXX/2021/1-ART1 $15.00
https://doi.org/10.1145/3481045

In the United States, the Freedom of Information Act (FOIA)[1] was enacted in 1966 with the goal of fostering government accountability and transparency, by providing citizens with the opportunity to access the documentary heritage of the country. Under the FOIA, all records created or received by federal agencies are presumptively open to public access, subject to nine enumerated statutory exemptions that under law are to be construed narrowly.[2] In recent decades, FOIA has allowed the public to request records in electronic form, including e-mail and word processing documents.[3]

However, as many critics have noted, a systemic weakness of FOIA administration is delays in response time to requestors [11, 41]. In large part, this is due to the near-uniform practice of FOIA officers engaging in manual searches for responsive records, coupled with even more time-consuming manual review of putatively responsive records to withhold (redact) material considered to be covered under a FOIA exemption. A recent example of where delays in the FOIA process have had a measurable negative impact is in finding information on the government's response to COVID-19. A number of federal agencies have been involved in the government's response in managing COVID-19.[4] At the same time, subsequent to the initial outbreak of the pandemic in the United States in March 2020, the U.S. government faced extraordinarily severe criticism regarding their overall efforts to contain the spread of the virus. During this moment in US history, federal agencies experienced a "steep increase" in the number of FOIA requests for information on how the government has been responding to the crisis [23]. Given the fact that the federal government receives on the order of 800,000 FOIA requests on an annual basis [24], it is not surprising that agencies might be overwhelmed, causing substantial delays to requestors. Indeed, during 2020 public interest groups sued agencies not only to receive specific records, but also for delaying the processing of FOIA requests due to COVID-19 [30].

FOIA Exemption 5 allows the withholding of "inter-agency or intra-agency memorandums or letters which would not be available by law to a party other than an agency in litigation with the agency." In turn, Exemption 5 allows agencies to justify withholding records through invocation of the "deliberative process privilege."[5] The purpose of this privilege is "to encourage honest and frank communication within the agency without fear of public disclosure."[6] To satisfy being considered within the scope of the "deliberative process privilege" a record must in whole or in part meet two conditions: first, it must be "predecisional," i.e., "antecedent to the adoption of agency policy."[7] Second, the record in whole or in part must be "deliberative," i.e., "when it reflects the give and take of decision making," including "recommendations, draft documents, proposals, suggestions, and other subjective documents "written by subordinate staff for consideration" by final decision makers.[8]

The FOIA statute also requires that "any reasonably segregable portion of a record" is releasable "after deletion of the portions which are exempt."[9] Under the statute, agencies shall "consider whether partial disclosure of information is possible whenever the agency determines that a full disclosure of a requested record is not

[1] Title 5, U.S. Code, Section 552.
[2] 5 U.S.C. 552(b).
[3] See Electronic Freedom of Information Act Amendments of 1996, Pub. L. 104-231.
[4] These include the Centers for Disease Control (CDC) and the National Institute of Health (NIH), as part of the White House Coronavirus Task Force; the State Department, in its dealings with the World Health Organization; and the Federal Emergency Management Agency's efforts in providing emergency public assistance to state, local and tribal government entities.
[5] U.S. Fish and Wildlife Service v. Sierra Club, 141 S. Ct. 777 (2021). FOIA Exemption 5 also covers other forms of material exempt from disclosure, including records covered by the attorney-client privilege or constituting attorney work product. Some exempt material may be exempt from disclosure for multiple reasons. Here, we solely have limited our scope to evaluating textual material within the scope of the deliberative process privilege, irrespective of whether it may also be withholdable on other grounds.
[6] American Center for Law and Justice v. DOJ, 325 F.Supp.3d 162 (D.D.C. 2018).
[7] Ancient Coin Collectors Guild v. U.S. Dep't of State, 641 F.3d 504, 513 (D.C. Cir. 2011) (quoting prior case law).
[8] Judicial Watch v. State Department, 349 F.Supp.3d 1, 7 (D.D.C. 2018) (quoting in part Coastal States, 617 F.2d 854, 866 (D.C. Cir. 1980).
[9] 5 U.S.C. 552(b).

possible."[10] For purposes of the deliberative process privilege, a key consideration in determining whether material is segregable is whether the text of a document is discussing factual matters, as opposed to matters involving opinions, proposals, suggestions, and recommendations.

In acknowledgement of the increasingly archaic aspects of how FOIA searches are conducted throughout the government, a recent report written by an advisory council of FOIA experts recommended that the Archivist of the United States "should work with other governmental components and industry in promoting research into using artificial intelligence, including machine learning technologies, to (i) improve the ability to search through government electronic recordkeeping repositories for responsive records to FOIA requests and (ii) identify sensitive material for potential segregation in government records, including but not limited to material otherwise within the scope of existing FOIA exemptions and exclusions" [31]. The experiments in this paper represent an initial attempt to meet the goal expressed in (ii) for one portion of one FOIA exemption.

This paper makes the following contributions:

- We show that when classifiers are trained and tested under consistent conditions it is possible to design classifiers that achieve F_1 measures between 70% and 83% (i.e., if tuned so that precision and recall were equal, we would expect that between 70% and 83% of the exempt material would be found, and that the same fraction of the content identified by the classifier as exempt would truly be exempt).
- We study the effects of differences between reviewers, between the materials held by different custodians, and within the topical content of the records being classified to identify which differences pose the greatest challenge for current text classifiers.
- We control for the effect of document type and recognizable characteristics of content items to study classifier effectiveness on the content and document types that human reviewers find most difficult.
- We suggest directions for future work, identifying a need to model contextual factors that require access to evidence beyond the boundaries of specific documents.
- We introduce a new freely distributable test collection that is annotated for the deliberative process privilege under Exemption 5 of the FOIA.[11]

2 RELATED WORK

There have been numerous work in detecting sensitive content. One line of work is to redact named entities, e.g. person names, organizations, countries, etc. Additionally, it allows to redact special patterns, e.g. social security numbers, date, etc [1]. These techniques could be applied in different domains, e.g. medical records [29, 32]. Document sanitization is one example that masks terms that are believed sensitive [26]. Sánchez et al. detects a sensitive term based on a measure called Information Content. This measure computes the probability of a term to appear in a corpus. A low probability of a term might indicate that the term is sensitive, and hence should be withheld from the containing document before release. In our work, one of the classifiers predicts the sensitivity of a word based on the surrounding words. We think that our approach is more general and can adapt to different kinds of sensitivity.

So far as the authors are aware, to date there has not been research specifically aimed at applying computational methods in the form of machine learning techniques to segregate sensitive material in documents corresponding to any applicable U.S. FOIA exemptions. However, within the U.K. there has been a substantial ongoing project through the SVGC Consortium to use machine learning techniques for assisting with sensitivity review and automated classification for information exempt under the U.K. Freedom of Information (FOI) Act (2000) [17]. McDonald et al. used a test collection of 3,801 government documents, 502 (13.2%) of which contained sensitive information pertaining to the categories of "international relations," "personal information," or both. The authors

[10] 5 U.S.C. 552(a)(8)(A)(ii).

[11] The test collection is available at https://github.com/mfayoub/FOIA-Test-Collection

studied the effect of lexical, syntactic and semantic features on classifier effectiveness, and conduced user studies to determine the utility of the resulting annotations to actual FOIA reviewers [18, 19]. Their test collection can not be redistributed because of the sensitive content that it contains. Unlike this prior work, we chose to work on a paragraph-level as we assumed there to be still non-sensitive paragraphs inside otherwise sensitive documents that could be released. Also, we tried different classifications algorithms. More fundamentally, McDonald's research and our own are aligned in designing computational models [42] to help reviewers identify FOIA exempt materials. These models illustrate a number of computational thinking practices. For example, Underwood and Marciano [34] suggest that Weintrop et al.'s taxonomy of 22 practices [40] can offer useful structure when considering the application of computational thinking to aspects of archival practice. Our work best illustrates Weintrop et al.'s framework for designing computational models, assessing computational models, and using computational models to find and test solutions.

There have also been several other lines of work that have yielded insights that could quite clearly be applied to various FOIA exemptions. The first is research regarding the detection of sensitive content in text [27, 28]. Sayed et al. developed different search systems that try to find relevant, but not sensitive, content. These systems could be applied to protect personally identifiable information (PII), withholdable under FOIA Exemption 6 [3].[12] Complementing that work is research regarding detection of material covered under the attorney client privilege and attorney work product, for withholding in U.S. civil litigation [6, 8, 22, 36]. That research could be extended to detection of equivalent sensitivities under FOIA Exemption 5. Finally, early research at the Georgia Tech Research Institute that explored rule-based techniques for managing the processing of FOIA requests and Presidential Records Act (PRA) materials [15], and for enforcing manually coded access restrictions [12], provide useful context for the types of processing environments now in commercial use[13] within which the machine learning techniques that we describe in this article might ultimately be employed.

We also note that although this paper focuses exclusively on detecting material that is exempt from release under the FOIA, there has been extensive work in the information retrieval field on the prerequisite task of finding relevant material that requires review in response to a request [4, 5, 10, 14, 28]. There is also at least one test collection in which both relevant and sensitive content are annotated, although in that case the sensitivities are personal concerns rather than codified exemptions as in FOIA [27].

3 PROPOSED APPROACH

Our research approach was designed to model the workflow that agency staff follow in carrying out FOIA reviews to determine whether agency records should be disclosed or withheld from the public.

Upon receipt of a FOIA request, staff in a FOIA office decide which components of an agency are most likely to possess records responsive to the request. Sometimes the request needs to be clarified with the requestor to determine where in the agency to conduct a reasonable search. A FOIA officer, working with both records managers and with individual custodians of records (e.g., employees), conducts a search that may include both hard-copy records in traditional files and electronic records. In cases involving modest amounts of potentially responsive records, searches are carried out manually (and often inconsistently) by these individuals. However, due to recent policy changes encouraging digital government, the volume of repositories of electronic records, and particularly e-mail and attachments, many agencies either are already or soon will be approaching in the millions of discrete records to be searched for FOIA requests [21, 25].

The authors are unaware of any federal agency that employs machine learning either for the purpose of conducting initial searches for responsive records in electronic repositories, or for finding exempt records or

[12]Exemption 6 allows for withholding information in certain records which "would review foconstitute a clearly unwarranted invasion of personal privacy." 5 U.S.C. 552(b)(6).
[13]See, for example, https://www.ains.com/foiaxpress

partially exempt records for purposes of determining what can be released pursuant to FOIA. The initial search process is performed using some combination of manual searching and automated searching based on metadata (e.g., date or original custodians) and keywords, while a second round of sensitivity review for exempt material is conducted solely by manual means.

With respect to finding textual material subject to the Exemption 5 deliberative process privilege, the manual review process consists of filtering using the following protocol:

(1) Is the record one where both the creator and all recipients are employed within the Executive branch? Subject to minor exceptions, if the answer is "Yes" the record satisfies the Exemption 5 threshhold test for being "inter-agency" or "intra-agency" in nature.
(2) Are the entire contents of a particular record considered to be both "pre-decisional" and "deliberative"?
(3) If the answer to (2) is no, is there any portion (or portions) of the record that is considered to be "pre-decisional" and "deliberative"? Ideally, determinations should be made based on a line-by-line review, but as a practical matter due to time and resources reviews often end up defaulting to decisions being made on a document-by-document basis at the first level of FOIA review, followed by a "paragraph by paragraph" basis on appeal.

Initial FOIA review is conducted by FOIA office staff. In cases where FOIA determinations end up being subject to initial appeal, a different reviewer (often an attorney) will re-review records that have been withheld in whole or in part, for possible release. Requestors have the right to file a lawsuit if they continue to disagree with any non-disclosures of information.

Several aspects of the above protocol are of special importance to note. First, the protocol described above consists of both easy and difficult tasks for a human reviewer. Assume a newspaper article attached to an email has been identified in an initial search as potentially responsive to a given FOIA request due to a keyword appearing in the article. A human reviewer would, however, immediately recognize the fact that a newspaper article is not an intra- or inter-agency document, and therefore would not be considered exempt. For example, if the article were attached to an otherwise responsive and nonexempt email, both would be released to the requestor. Optimally, a classifier must be trained to distinguish inter- and intra-agency records, from those arising and solely communicated within the Executive branch. For our purposes here, we refer to categorically exempt documents as "Easy."

Second, human reviewers have little or no difficulty in recognizing that some limited portions of agency records are "factual" in nature, outside the scope of what could possibly be deemed to be "deliberative." This would be true for "Date," "To, "From," and "Subject" lines in a traditional document, as well as agency letterheads or other metadata. For purposes of the experiments run here, we have chosen to designate this type of information in a document as "trivial." For the same reason, we annotate header information and signature blocks in individual e-mail records as trivial since, with very limited exceptions, such information can not be withheld under Exemption 5 (although it may be exempt from release under some other exemption).

Third, a human reviewer would have little difficulty segregating out "final" versions of records representing public releases of agency positions or testimony or other statements that on the face are not "pre-decisional" or "deliberative" in nature. Ideally, a classifier needs to be trained to recognize final versions of records, as opposed to drafts or other pre-decisional deliberations. For this research exercise, however, "final" versions of documents have been annotated as if they were drafts, since draft detection is not a research topic that we address here.

Finally, with respect to textual material of a substantive nature at the heart of the sensitivity review, the review process places a priority on ensuring that all exempt material be properly identified. This consideration significantly outweighs considerations of the volume of records to be reviewed. In other words, FOIA reviewers place the value of high recall over high precision, to ensure that textual material potentially subject to the

deliberative process privilege is flagged. This also partially explains and contributes to delays in the overall FOIA process.

Aside from any difficulties posed for a classifier by the above considerations, the learning task involved here is made more difficult for two primary reasons. First, this is due to interpretive nuances that have evolved over time in court decisions interpreting the scope of what constitutes "predecisional" and "deliberative" discussions,[14] as well as allowance for ad hoc determinations in restricted cases on what constitutes "reasonable segregability."[15] Second is that the degree of discretion that FOIA reviewers have in considering whether textual material falls in or outside of the scope of the exemption based on contextual considerations outside the four corners of a given document (e.g., see the discussion in Section 3.2 of the "foreseeable harm" test). Both considerations lead to differences in application amongst FOIA subject matter expert reviewers, and introduce some measure of difficulty in achieving "gold standard" certainty in training a classifier.

3.1 A Document Collection

An initial difficulty encountered with this research was finding a suitable test collection. Ideally, such a collection would be comprised of two version sets of documents (1) an "original" set of documents previously withheld in whole or in part under Exemption 5, and (2) the same documents now released or accessible to the public without the withheld portions. This was the approach taken by the Declassification Engine.[16] It proved difficult, however, to identify an adequately large collection of documents that either had been withheld under Exemption 5 but later released in litigation containing Exemption 5 withholdings, or a collection that had been withheld from public release due to its containing Exemption 5 material, but later reviewed and released after Exemption 5 concerns no longer applied due to the passage of time.[17]

An alternative strategy would have been to access selected archival records held at the National Archives and Records Administration (NARA), identifying through existing finding aids records once in senior agency officials' files. Such files would have a high probability of including policy option papers and other forms of consultation and recommendations, as part of the particular decisionmaking processes. This would normally have been possible with some scanning and OCR, but the closure of public access to NARA's archival collections due to COVID-19 eliminated the possibility of pursuing this alternative.

A third option, however, presented itself in the form of searching through online record collections maintained on Presidential library websites. Records covered by the PRA are accessible through FOIA five years after the end of a President's term in office, subject to restrictions authorized by the statute that authorize a President to exempt designated categories of records from disclosure for up to 12 years. Under one of these restrictions (colloquially referred to as "P5"), "confidential communications requesting or submitting advice, between the President and the President's advisers, or between such advisers," may be restricted for up to 12 years.[18] Moreover, after the restriction period ends, such records cannot be withheld on the basis of Exemption 5.[19] In other words, as a surrogate test collection, formerly restricted P5 records that have been opened in connection with FOIA requests would be excellent candidates for inclusion within this research, subject to understanding that there is

[14] See Judicial Watch v. Dep't of State, 349 F.Supp.3d 1, 7 (D.D.C. 2018) (an ex post communication may still be predecisional if it discusses recommendations not expressly adopted).

[15] For example, a court may decline to order an agency to commit significant time and resources to the separation of disjointed words, phrases, or even sentences which taken separately or together have minimal or no information content. See Mead Data Center v. Department of the Air Force, 566 F.2d 242, 261 n.55 (D.C. Cir. 1977). In addition, when nonexempt information is "inextricably intertwined with exempt information, reasonable segregation is not possible." Id.

[16] https://web.archive.org/web/20130607100457/http://www.declassification-engine.org/redactions/#/

[17] The deliberative process privilege ceases to be a basis for withholding after 25 years from creation date of an agency record. 5 U.S.C. 552(b)(5).

[18] 44 U.S.C. 2204(a)(5).

[19] Id., 2204(c).

not perfect congruence as a matter of law as between the P5 restriction and the deliberative process privilege. As it turned out, we found many records where the P5 restriction had not been invoked, that nevertheless contained passages arguably withholdable under Exemption 5 had those records had been created in a federal agency subject to FOIA.

After further review, several presidential records collections accessible on the website of the William J. Clinton Presidential Library appeared to meet the requisite requirements as outlined above. The Clinton Library maintains its collections in designated "files," each of which contains individual documents.[20] A "file" consists of documents on a particular subject, similar to how a traditional file cabinet would be organized by subject. Both files and documents are indexed by the White House component associated with their creation or receipt, as well as by the individual custodians who would have held the file. In particular, a review of the records of now-Justice Elena Kagan, who in the 1990s worked as a lawyer on the Domestic Policy Council in the Clinton White House, contained numerous documents with formerly P5 restricted content that subsequently had been opened through FOIA requests or as part of the ongoing processing work of NARA archivists. An additional search for records associated with Cynthia Rice, a second lawyer on the Domestic Policy Council, were made part of the review. A keyword search for "Elena Kagan" produced 2,945 PDF files that contained varying numbers of documents. From this we selected 32 files, a total of 432 documents, that on inspection were found to have a reasonable number of documents containing textual material arguably within the scope of the deliberative process privilege. A keyword search for "Cynthia Rice" produced 1,072 files, from which we selected 5 files (a total of 77 documents) that met the same criterion.

3.2 Annotation Protocols

Table 1 describes the collection of files used in the research. "Batches" consist of an arbitrary number of subject matter files that were selected because they contained documents comprising a large number of pages. The five batches of documents were reviewed sequentially in the order that they are numbered. Four Batches (K1, K2, K3 and K5) were from files associated with Elena Kagan, consisting of a combined total of 32 files. An additional Batch R4 consisted of three Cynthia Rice files. For the first four batches, documents that the first reviewer felt were unlikely to contain material within the scope of the deliberative process privilege were skipped. The entire fifth batch was annotated, however, including some documents that would easily be categorically excluded by a human reviewer. Documents from the fifth batch that the first reviewer felt could be categorically excluded from review were excluded from batch K5 and were instead used to used to form batch E5 (where E indicates an "Easy" decision).

Individual documents in each batch were annotated by the lead author of this paper (Reviewer A), an attorney with subject matter expertise in FOIA law. As detailed below, each document was divided into "paragraphs" for purposes of designation as either within or outside the scope of the exemption. A second attorney with FOIA subject matter expertise (Reviewer B) also reviewed documents in Batch E2. In the case of Batch K2, the two reviewers independently annotated documents in all 10 files, and then met together to review their annotations to see in cases of disagreement whether they could agree on a "consensus" annotation for training purposes.

Files in each batch ranged across a wide collection of subjects within the purview of the Domestic Policy Council, as indicated by the listed file names in Table 1. To support analysis of classifier performance on specific topics, Reviewer A assigned a single topic label to each document. The list of topic labels assigned by Reviewer A is shown in Table 11.

A number of artificial constraints were placed on the attorney review process in order to simplify the core goal of finding pre-decisional and deliberative material within individual documents. For the most part, these

[20] In keeping with the vocabulary of text classification, we refer to the items as documents.

Batch	Custodian	Files	Paragraphs	File Names	Reviewer(s)
K1	Elena Kagan	9	523	Superfund, Welfare Budget, Welfare-Blair Visit, Service Summit Policy, Service General, Veterans Affairs/Filipinos, Drugs Coerced Abstinence, Drugs Heroin Chic	A
K2	Elena Kagan	10	447	Education/ TIMSS Meeting, Education/Troops to Teachers, Education/Vouchers, Environment/Climate Change, Kids Executive Order, Family Child Care Policy, Social Security/Nazis, Social Security/Prisoners, Drugs/Drug Testing	A & B
K3	Elena Kagan	10	670	Emails Received, Health/Radiation Experiments, Health/ Organ Transplants, Health/ Nursing Homes, Health/Medicaid Cap, Health/Immunization, Health/Genetic Screening, Drugs/Southwest Border, Environment/Port Dredging	A
R4	Cynthia Rice	5	466	Child Support/Gambling, Child Support/License, Fathers/Bayh Bill, Budget 2001 FY New Ideas, Disability-Kennedy-Jeffords 1999	A
K5	Elena Kagan	3	631	Tax Proposals; Drugs/Media Campaign, Drugs/ Meth Report	A
E5	Elena Kagan	3	286	Tax Proposals; Drugs/Media Campaign, Drugs/ Meth Report	A

Table 1. Document batches in the test collection.

constraints reflect the fact some of the relevant criteria for making a real-world classification decision as to material within the scope of the deliberative process privilege are "contextual" in nature.

First and foremost, as of 2016 the FOIA was amended to codify a "foreseeable harm" test, requiring agencies to withhold information "only if the agency reasonably foresees that disclosure would harm an interest protected by an exemption." Thus, passages that otherwise meet the test for what is considered "pre-decisional" and "deliberative" may nevertheless be required to be disclosed based on human judgment bringing a range of additional information about agency policies and current events to bear. For purposes of this exercise, we did not annotate using a "foreseeable harm" test.

Second, some initial scoping decisions were made so as not to muddle or make unduly difficult the classifier's work. Among these were: (a) parsing individual documents into paragraphs, for annotation purposes; (b) treating header information and signature block information (including in emails) as separate paragraphs from the main text; (c) annotating paragraphs as a whole as either exempt or not, rather than performing a more granular annotation on a sentence-by-sentence basis; (d) ignoring the presence or absence of "DRAFT" headers, instead treating all documents as "drafts" and thus per se within scope; and (e) in Batches K1, K2, K3 and R4, excluding documents that Reviewer A found to be categorically nonexempt. Examples of this nature included records originating outside the Executive branch, including from Congress or from outside lobbying groups; final legal briefs; final reports and white papers of various kinds; and newspaper and magazine articles. We relaxed this last

```
  T0//
Sandra Thurman 01/12/98 10:35:44 AM Record Type: Record
To: Richard Socarides/WHO/EOP cc: Maria Echaveste/WHO/EOP
Subject: Re: Needles/Embryos/Abortion and Other Selected L/HHS General Provisions SPEAK NOW OR...
  D1//
We did comment on the proposed language on needle exchange after consulting with both Chris
Jennings and Kevin Thurm. I will forward a copy of the memo to you.
  D1//
I had a lengthy discussion with Kevin last week regarding this issue. HHS does not plan to do
anything on needle exchange until Satcher is confirmed, assuming that will happen in February.
If indeed the confirmation is held up for some reason,we will have to revisit the
timing of any action.
  D1//
Contrary to what Scott Hitt may have told you, the AIDS community is still vehemently opposed
to any law enforcement component in any compromise we might propose. So are General Mccaffrey
and I. In fact, it may well be the only point upon which we agree on this issue.
  0//
I am meeting again this week with the national AIDS groups to discuss where we are on needle
exchange. I'll keep you posted.
0//
Sandy
```

Fig. 1. An example of an annotated document.

constraint (e) in reviewing the fifth Batch, where the reviewer annotated all documents found in three Kagan files.

Annotations were coded as follows:

D1// Paragraphs that fall within the scope of the deliberative process privilege
E0// "Easy" non-exempt paragraphs—those found in documents easily excluded from review due to not being "inter-agency" or "intra-agency" in nature (only annotated in batch E5)
T0// "Trivial" non-exempt paragraphs (e.g., header information and signature blocks)
D0// "Decided" non-exempt paragraphs (i.e., paragraphs containing only factual content)

Figure 1 shows an example of an annotated e-mail message.

3.3 Annotator Agreement

We measured inconsistencies in annotation as between the two reviewers on batch K2, the only batch that was annotated by both reviewers (see Table 1). For this analysis, we ignored T0 annotations (which were made only by Reviewer A). Table 2 shows the number agreements (on-diagonal when both reviewers annotate a paragraph similarly) and disagreements (off-diagonal). We used Cohen's Kappa coefficient as a measure of the chance-corrected agreement between annotators. According to Landis and Koch [16], the resulting Cohen's Kappa, 0.67, indicates substantial agreement.

During the consultation phase after initial annotations were completed, it became clear to the two attorney-reviewers that a substantial number of paragraphs where they disagreed in their assessments were found in documents known as "Talking Points," or similar forms of "Qs and As," related to an imminent public appearance of an official announcing government policy. As it turns out, the inconsistent annotations reflect lack of settled

		Reviewer B		
		D0	**D1**	**Total**
Reviewer A	**D0**	212	69	281
	D1	6	160	166
	Total	218	229	447

Table 2. Annotator agreement, Cohen's Kappa, batch K2

precedent as to whether this type of document is "predecisional" or "deliberative."[21] Had paragraphs in "talking points" memos been consistently decided, measured annotator agreement would have been higher.

After computing annotator agreement, the two reviewers created an agreed final set of annotations for batch K2. We refer to these final annotations as having been created by "Reviewer AB." We note that the reviewers reached complete consensus after the following revised determinations were made: Reviewer A, D0 to D1 (58), D1 to D0 (2); Reviewer B, D0 to D1 (4), D1 to D0 (11).

3.4 Evaluation Measure

In the overall review task, performed by the human and machine working together, reviewers must be able to reach a high degree of confidence that all material that is exempt from release has been identified. The machine's task is different, however, since the goal of the machine is to suggest, not to decide. The machine's suggestions can be useful in two ways; they can help the human reviewer to avoid missing content exempt from release that they otherwise might miss, or they might help the human reviewer to decide more quickly on both exempt and non-exempt content. To be useful, the machine must therefore find a substantial portion of the exempt content, and it must avoid misclassifying much of the non-exempt content as exempt. With user studies we might be able to find the optimal balance between these two requirements, which correspond to recall and precision, respectively. Our focus in this paper, however, is on characterizing the performance of existing classifiers for this task. We therefore report recall and precision separately, and where a single objective is needed (as is the case when tuning parameters) we report the balanced harmonic mean of recall and precision (F_1). We compute 95% confidence intervals for F_1 using normal approximation. In result tables we bold the highest F_1 value and we underline values that are statistically significantly better than the "All 1s" baseline when the mean of each F_1 is outside the 95% confidence interval of the other.

3.5 Classifiers

We seek to support human review for FOIA Exemption 5 by highlighting paragraphs that the machine suggests may be exempt from release. This is a binary classification in which each paragraph is to be given a label that indicates whether the paragraph is within the scope of the privilege (1 or positive class) or not (0 or negative class). Four approaches to text classification were employed for purposes of finding material subject to the deliberative process privilege. These consisted of Logistic Regression (LR) [7], Support Vector Machine (SVM) [38], Begin-Inside-Outside (BIO) tagger using Conditional Random Fields[37], and keyword search [39].

[21] Cf. American Center for Law and Justice v. US Dep't of Justice, 325 F. Supp. 3d 162, 173 (D.D.C. 2018) ("A government employee drafting talking points . . . needs to know that her advice will remain privileged regardless of whether the [speaker] ultimately sticks to the script or decides to extemporize . . . Moreover, sticking to talking points often does not entail a verbatim recitation, leaving open the possibility that 'a simple comparison' of the talking points with the official's public remarks would reveal the agency's deliberations'), with Judicial Watch v. Dep't of State, 349 F.Supp.3d 1, 7-8 (D.D.C. 2018) ("This discussion proves too much . . . government officials give hundreds of speeches every day, all of which are important, though many elude recording or transcription. So stretching the deliberative process privilege [to include talking points] would put many important public statements outside FOIA's grasp, even well after the statements were made.").

In both LR and SVM, a raw paragraph is converted into a vector of word counts for the words in the paragraph. Then, this vector can be used as a training or test sample for classification. In BIO, each word in a document is annotated with either B, I, or O according to the following rules.

(1) If the current word is the beginning of privileged content, the word's label is B.
(2) If the current word is inside privileged content, the word's label is I.
(3) If the current word is not part of privileged content, the word's label is O.

The BIO classifier aims to predict the correct label of a word given its context, e.g. previous word and next word. Because our evaluation is based on paragraphs (as marked by Reviewer A), while the BIO classifier can place begin and end labels at any word, we need a way to map the BIO results to paragraph boundaries. The BIO classifier therefore predicts a paragraph as privileged if the number of words predicted as B or I exceeds a certain percentage or non-privileged otherwise. This percentage is a hyper-parameter to the BIO classifier.

As is obvious from the above, these factors can result in a huge number of combinations. As a result, we handle them in groups and present the results in the following subsections. It is worth noting that any training set is further split into training and validation sets. The validation set is used for hyper-parameters tuning. Table ... lists the hyper-parameters we tune and their corresponding ranges. We exhaustively consider all parameter combinations and extract the best parameters with the highest F_1 score on the validation set. Then we use the best parameters to estimate the class labels for test samples.

LR		SVM		BIO	
Parameter	Parameter Space	Parameter	Parameter Space	Parameter	Parameter Space
use_idf	{False, True}	use_idf	{False, True}	C1	{0.01, 0.1, 1, 5, 10}
stemmer	{None, Porter}	stemmer	{None, Porter}	C2	{0.01, 0.1, 1, 5, 10}
C	{0.01, 0.1, 1, 5, 10}	C	{0.01, 0.1, 1, 5, 10}	overlap %	{10, 20, ... 90, 100}
probability threshold	{0.1, 0.2, ... 0.9, 1}	kernel	{linear, rbf}		
		γ (for rbf)	{1, 0.1, 0.01, 0.001, 0.0001}		

Table 3. Parameter tuning by grid search.

Finally, keyword searches were performed using the following terms: *option OR recommendation OR proposal OR suggest OR suggestion OR discuss OR discussion OR upcoming OR alternative OR frank OR candid OR ongoing*. If a paragraph contains any of those keywords, it was predicted to be privileged. If none of the keywords were found, the paragraph was predicted to be non-privileged.

In addition to the above four methods, we compared each against assignment of "all 1's" (i.e., simply treating every paragraph as within the scope of the privilege). This simple approach achieves perfect recall, and is thus a useful baseline that other methods must exceed to be useful.

4 EXPERIMENT RESULTS

We might imagine several scenarios in which text classification might be employed. We have organized the presentation our our results around the four basic scenarios shown in Table 4. These explore the following dimensions of variation:

- Optimally the classifier would be trained and tested on annotations from the same reviewer, although that may not always be possible.
- Optimally the classifier would be trained and tested on documents that address a similar range of topics, although that may not always be possible. In our experiments we use the custodian as a proxy for the set of topics on which that custodian worked.

- We are most interested in the effectiveness of a classifier when it is presented with cases that call for the most difficult decisions, but it is also important that the classifier not make errors on the decisions that human reviewers find easy or trivial.

To explore these conditions, we conducted experiments for the conditions shown in Table 4. As notation, we describe a condition by the reviewer followed by the batches. for example, Train AB: $K_{2,3}$ would refer to training on Elena Kagan's batches 2 and 3, as annotated by the consensus of reviewers A and B. We begin with experiments in which the task is to detect exempt paragraphs in documents that can not be categorically excluded (i.e., without the "Easy" documents in batch E5).

Condition	Documents	Reviewers	Method
A	Same custodian	Same reviewer	Cross-validation
B	Different custodians	Same reviewer	Train-test split
C	Same custodian	Different reviewers	Train-test split
D	Different custodians	Different reviewers	Train-test split

Table 4. Primary experiment conditions.

Table 5 shows the results for Condition A, with the classifier trained on the same reviewer and the same custodian that it is tested on. All results are for 5-fold paragraph-scale cross-validation. Results, summarized in Table 5, show that the SVM performs well (by F_1). There is no evidence in the results table favoring further consideration of the keyword classifier, at least with the specific keywords that we chose.

	Train Cross-validate Test A: $K_{1,2,3,5}$			**Train Cross-validate Test A: R_4**			**Train Cross-validate Test B: K_2**			**Train Cross-validate Test AB: K_2**		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
LR	63.2	79.9	**70.3±1.9**	62.5	76.4	67.7±4.2	72.8	86.2	78.2±3.8	77.5	83.9	80.1±3.7
SVM	67.2	72.2	69.5±1.9	70.4	72.2	**70.7±4.1**	79.8	84.7	**82.1±3.6**	80.2	87.2	**83.5±3.4**
BIO	46.2	85.2	59.8±2.0	46.8	94.2	61.6±4.4	64.6	99.1	78.2±3.8	64.9	98.3	78.1±3.8
Keyword	53.0	32.1	39.9±2.0	56.9	38.8	45.0±4.5	86.2	29.3	43.1±4.6	86.2	29.8	43.5±4.6
All 1s	32.7	100	49.3±2.1	24.3	100	38.7±4.4	51.2	100	67.7±4.3	50.8	100	67.3±4.3

Table 5. Condition A: same reviewer, same custodian, cross-validation; D1 vs {D0, T0}.

Table 6 shows results for condition B, in which we trained classifiers using batches from one custodian annotated by one attorney, and then tested it using batches from a different custodian but annotated by the same attorney. As can be seen, the BIO classifier does well in this condition relative to other classifiers, but note that the F_1 values for the trained classifiers are well below the corresponding cross-validation values in Table 5.

	Train A: $K_{1,2,3,5}$ Test A: R_4			Train **Train A: R_4 Test A: $K_{1,2,3,5}$**		
	P	R	F_1	P	R	F_1
LR	47.0	61.9	53.4±4.5	37.9	84.4	52.3±2.1
SVM	43.7	83.2	**57.3±4.5**	42.1	63.5	50.7±2.1
BIO	44.1	78.8	56.5±4.5	39.6	83.7	**53.8±2.1**
Keyword	56.8	37.2	44.9±4.5	52.9	32.0	39.9±2.0
All 1s	24.2	100	39.0±4.4	32.7	100	49.3±2.1

Table 6. Condition B: same reviewer, different custodian; D1 vs {D0, T0}.

Table 7 shows results for condition C, in which we have trained our classifiers using batches from one custodian annotated by one reviewer, and then we test using batches from the same custodian, but annotated by a different reviewer. As was the case for condition A, the SVM classifier performs well (by F_1).

	Train A: $K_{1,3,5}$ Test B: K_2			**Train A: $K_{1,3,5}$ Test AB: K_2**			**Train B: K_2 Test A: $K_{1,3,5}$**			**Train AB: K_2 Test A: $K_{1,3,5}$**		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
LR	78.3	31.4	44.9±4.6	76.1	30.8	43.9±4.6	44.8	63.8	**52.6±2.3**	42.7	63.4	51.0±2.3
SVM	66.8	93.9	**78.0±3.8**	66.5	94.3	**78.0±3.8**	36.5	91.0	52.1±2.3	40.8	76.3	**53.1±2.3**
BIO	66.8	86.9	75.5±4.0	67.1	88.1	76.2±3.9	39.1	80.2	**52.6±2.3**	39.4	76.6	52.0±2.3
Keyword	86.1	29.7	44.2±4.6	86.1	30.0	44.4±4.6	48.5	31.2	38.0±2.2	48.5	31.2	38.0±2.2
All 1s	51.2	100	67.8±4.3	50.8	100	67.4±4.3	31.7	100	48.1±2.3	31.7	100	48.1±2.3

Table 7. Condition C: different reviewer, same custodian; D1 vs {D0, T0}.

Table 8 shows results for category D, in which we trained using batches from one custodian annotated by some reviewer, and then tested using batches from a different custodian that were annotated by a different reviewer. As the results show, the Logistic Regression classifier consistently does well in this condition (as measured by F_1), with the BIO classifier not far behind.

	Train A: R_4 Test B: K_2			**Train A: R_4 Test AB: K_2**			**Train B: K_2 Test A: R_4**			**Train AB: K_2 Test A: R_4**		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
LR	63.5	87.3	**73.5±4.1**	63.5	88.1	**73.8±4.1**	46.3	77.0	**57.8±4.5**	46.2	79.6	**58.4±4.5**
SVM	65.7	62.0	63.8±4.5	67.6	64.3	65.9±4.4	36.2	98.2	52.9±4.5	42.1	90.3	57.5±4.5
BIO	63.9	82.1	71.9±4.2	63.9	82.8	72.2±4.2	41.6	87.6	56.4±4.5	42.4	84.1	56.4±4.5
Keyword	86.1	29.7	44.2±4.6	86.1	30.0	44.4±4.6	56.8	37.2	44.9±4.5	56.8	37.2	44.9±4.5
All 1s	51.2	100	67.8±4.3	50.8	100	67.4±4.3	24.2	100.0	39.0±4.4	24.2	100.0	39.0±4.4

Table 8. Condition D: different reviewer, different custodian, D1 vs {D0, T0}.

4.1 Including Paragraphs from Easily Recognized Non-Exempt Documents

Table 9 shows the effect of adding batch E5, which consists entirely of the documents in Elena Kagan's Batch 5 that a human reviewer would easily recognize as categorically non-exempt. This has the effect of adding 286 additional paragraphs, all with E0 annotations, to the training set (in condition B), to the test set (in condition C), or to both (in condition A). Adding batch E5 to the test set (as happens for Conditions A and C) reduces F_1 values, suggesting that documents that human reviewers can easily recognize as categorically non-exempt are not actually easy for classifiers that see only isolated paragraphs. This suggests that it may be useful to automatically recognize categorically non-exempt documents as a preprocessing stage, something we did not try in our experiments. Interestingly, adding batch E5 to training (as happens for condition B) seems to have only small effects on F_1.

4.2 Excluding Trivially Non-Exempt Paragraphs

Table 10 shows one representative result for each condition when trivially non-exempt paragraphs such as headers and signature blocks (i.e., those with T0 annotations) are excluded from both training and testing. For condition A, in which cross-validation yields training and test conditions that are remarkable similar, the three trained classifiers yield results very similar to those of the corresponding condition in Table 5. However, the best

	Train Cross-validate Test A: $K_{1,2,3,5}$+E_5			**Train A: $K_{1,2,3,5}$+E_5 Test A: R_4**			**Train B: K_2 Test A: $K_{1,3,5}$+E_5**		
	P	R	F_1	P	R	F_1	P	R	F_1
LR	61.7	74.7	**67.4±1.8**	52.9	63.7	57.8±4.5	40.0	63.8	**49.2±2.1**
SVM	62.8	71.8	66.9±1.8	45.6	77.9	57.5±4.5	31.7	91.0	47.0±2.1
BIO	45.8	71.4	55.6±1.9	47.5	76.1	**58.5±4.5**	34.0	80.2	47.8±2.1
Keyword	51.0	32.1	39.3±1.9	56.8	37.2	44.9±4.5	46.6	31.2	37.4±2.1
All 1s	29.1	100	45.0±1.9	24.2	100	39.0±4.4	27.3	100	42.9±2.1

Table 9. Effects of adding Batch E5; Left to right: one example each for conditions A, B, C; D1 vs {D0, T0, E0}

trained classifier results for conditions B, C and D, do not exceed "all 1s" baseline. This occurs in part because the "all 1s" baseline naturally improves when any zero annotations (including T0) are excluded, and we expect in part because the resulting classification task is harder.

	Train cross-validate Test A: $K_{1,2,3,5}$			**Train A: R_4 Test A: $K_{1,2,3,5}$**			**Train A: $K_{1,3,5}$ Test AB: K_2**			**Train A: R_4 Test AB: K_2**		
	P	R	F_1±CI	P	R	F_1±CI	P	R	F_1±CI	P	R	F_1±CI
LR	65.3	75.9	70.1±2.0	38.9	48.9	43.3±2.2	79.9	57.0	66.5±4.9	67.2	72.2	69.6±4.8
SVM	72.0	68.9	**70.4±2.0**	40.0	51.3	44.9±2.2	79.4	43.5	56.2±5.2	66.6	95.2	78.4±4.3
BIO	53.0	77.3	62.2±2.1	41.0	73.0	52.5±2.2	75.9	70.0	72.9±4.7	67.6	55.2	60.8±5.1
Keyword	55.0	32.1	40.5±2.2	55.2	32.2	40.7±2.2	93.2	29.6	44.9±5.2	93.2	29.6	44.9±5.2
All 1s	37.8	100	54.8±2.2	37.8	100	**54.9±2.2**	65.7	100	**79.3±4.2**	65.7	100	**79.3±4.2**

Table 10. Effects of exclusion of T0 (trivially non-exempt) paragraphs; left to right: conditions A, B, C, D; D1 vs D0.

4.3 Topic Effects

Table 11 shows F_1 values for classifiers that were tested using Reviewer A's judgments for all paragraphs that had been labeled with the topic shown. Each classifier was trained using Reviewer A's judgments for all other paragraphs (i.e., paragraphs labeled with any other topic). As in Tables 5 through 8, trivial zeros were included and easy zeros (i.e., batch E5) were excluded. For 13 of the 16 topics, covering 95% of all paragraphs (2,590 of 2,735), some trained classifier statistically significantly outperformed the "all 1s" baseline. The keyword classifier exhibited the greatest variation, with F_1 values between 0.059 (for Education) and 0.600 (for Tax Proposals and for Family).

4.4 Efficacy of Keyword Searching vs. Machine Learning

Tables 5 through 11 consistently indicate that classifiers based on matching any one of a manually selected set of keywords were consistently outperformed by classifiers that learned from training examples. This result is consistent with numerous other studies [2, 9], suggesting either that human designers lack the sufficient insight into vocabulary used in the items that are sought in the collection, or perhaps that such "flat" disjunctive classification rules are not sufficiently expressive. It may, however, ultimately be possible to augment human performance at this task by suggesting terms for human designers to consider. As an example, Table 12 shows the keywords with the greatest positive and negative weights for the exempt class for one fold of the first logistic regression classifier shown in Table 5. Strikingly, there is only one word (option) in common between the keywords chosen by Reviewer A before annotating anything and the words on which a classifier most heavily relies after training on Reviewer A's annotations.

Topic	Paragraphs	LR	SVM	BIO	Keyword	All 1s
Drugs	699	25.4±3.2	44.4±3.7	**55.2±3.7**	22.4±3.1	51.6±3.7
Health	297	**55.2±5.7**	47.7±5.7	41.1±5.6	41.8±5.6	37.7±5.5
Tax Proposals	253	54.3±6.1	**61.8±6.0**	57.7±6.1	60.0±6.0	54.6±6.1
Welfare	245	**51.3±6.3**	39.3±6.1	38.5±6.1	31.5±5.8	38.3±6.1
Child Support	216	**59.3±6.6**	44.0±6.6	39.3±6.5	54.0±6.6	29.2±6.1
Service	198	**62.0±6.8**	56.6±6.9	54.0±6.9	40.0±6.8	53.9±6.9
Miscellaneous Emails	188	**62.9±6.9**	50.7±7.1	43.8±7.1	44.8±7.1	33.6±6.8
Disability	105	**61.0±9.3**	51.8±9.6	51.7±9.6	50.0±9.6	35.9±9.2
Education	103	37.9±9.4	**53.2±9.6**	48.5±9.7	5.9±4.5	41.5±9.5
Budget	100	57.1±9.7	**80.7±7.7**	70.5±8.9	40.0±9.6	63.0±9.5
Kids	92	69.2±9.4	**91.9±5.6**	77.5±8.5	56.3±10.1	82.8±7.7
Environment	73	62.0±11.1	**66.7±10.8**	62.5±11.1	54.5±11.4	58.3±11.3
Social Security	72	64.2±11.1	**69.3±10.7**	65.0±11.0	57.8±11.4	54.5±11.5
Fathers	45	**38.5±14.2**	37.8±14.2	37.8±14.2	13.3±9.9	26.9±13.0
Family	30	**72.7±15.9**	55.6±17.8	48.0±17.9	60.0±17.5	33.3±16.9
Superfund	19	**84.6±16.2**	73.3±19.9	73.3±19.9	45.5±22.4	73.3±19.9

Table 11. F_1 for for testing on one topic when training on all other paragraphs; D1 vs. {D0, T0}.

Top words, positive weights	Top words, negative weights
option	today
counter	committee
options	state
doj	subject
authority	clinton
program	education
increase	human
splitting	family
idea	let
largest	police
accreditation	30
language	soon
initiatives	eop
coordinator	radiation
vouchers	00
necessary	experiments
targeted	prisoner
significant	approved
think	18
action	jose

Table 12. Top 20 words with the greatest positive and negative weights for LR in Table 5, leftmost case, first fold.

5 DISCUSSION

These experiments indicate that using supervised machine learning to help human reviewers identify content exempt from release under the FOIA Exemption 5 deliberative process privilege is feasible. As it stands, human

reviewers manually review records without the records being "weighted" or ranked in any fashion. Applying a classifier would introduce at least three efficiencies into the existing process. First, it could be used to push to the front of a large-scale review process those records least likely to contain exempt material, allowing reviewers to focus first on the portion of the collection that could perhaps be reviewed most rapidly. Second, for any portions of records that the classifier determines contain potentially withholdable material, highlighting could help to draw the reviewer's eye to passages that would benefit from the most careful review. Third, once the review has been completed, a classifier could be trained from all of the resulting decisions and then used to detect inconsistencies in those decisions [13]. Such a machine-assisted final review process could help to further improve confidence in the review process, possibly catching both inadvertent errors of commission (incorrect release) and errors of omission (incorrect withholding). Sampling and independent review can then be used as a final step to still further increase confidence in the process that was used in the review.

The results in Tables 5 through 8 clearly indicate that classifiers can be trained to reliably highlight the vast majority of content that is exempt from release, and that they can do so with sufficient precision to be directly useful. As with many applications of supervised machine learning, closely matching the training and test conditions, as in condition A, yields the best results. This suggests that iterative retraining of the classifier as more annotations are made would be a promising option to explore.

As the results in Tables 9 and in 10, further improvements might be obtained by incorporating specialized classifiers for entire documents that are categorically non-exempt (what we have called "easy zeros") or for document elements that human reviewers would not need to spend time reviewing (what we have called "trivial zeros"). These additional classifiers could make use of features from, for example, layout analysis, genre detection, or authorship attribution. Future work should also explore whether those additional features and others (e.g., from opinion detection) might help to improve our base classifiers' abilities to make the most challenging decisions (between D1 and D0).

Our findings are subject to a number of caveats. First, this research was conducted on a small scale involving the records primarily of one senior official in government, coupled with a limited number of records from a second official. Before a classifier could be considered reliable, it would need to be shown that based on a given training set of records, the classifier could perform well across a range of records drawn from a greater set of key custodians in federal agencies thought to be holding responsive records to a given FOIA request.

Second, our findings are based on imposing a measure of artificiality in that paragraphs were manually identified for purposes of facilitating comparisons between human and machine annotations. Our experiments with Begin-Inside-Outside classification point the way toward more flexible approaches that could work at the scale of sentences, lines, or even words.

Third, although we have adopted the opinions of experienced reviewers as the gold standard for our experiments, there will always be room for human judgment in the process and necessarily a measure of uncertainty in the results. The key question to ask, therefore, is not how good is our classifier, as if it were operating on its own, but rather how useful is our classifier when used to support a real FOIA review process, by real FOIA reviewers who must grapple with the discretion and degree of ambiguity embodied in FOIA law. The prospect of automation should always be considered to be part of what is ultimately going to remain a human-centered review process. A natural next step would therefore be to study the use of systems that include classifiers like those that we have described here on a larger universe of records. McDonald et al. studied the effect of the accuracy of sensitivity classification on human reviewers performance [20]. Results show that quality of classification results affects reviewer's performance in terms of correctness and time to finish. We would like to perform a similar study and work closely with FOIA reviewers to understand what the best way is to couple their efforts with sensitivity classification.

Fourth, we have to date used the words found in passages as the features on which classification is based. This choice fit our goals well in this study since the materials we selected were from a small range of custodians

on a constrained set of topics, but when extended to a broader range of materials in full-scale applications a broader range of features might be explored. For example, Underwood and Isbel have experimented with semantic annotation of PRA materials [33], and access to such features might further improve classification accuracy.

Finally, we note that our experiments here have focused on the deliberative process privilege, which is one part of one exemption in the federal FOIA law. Our promising results suggest that further work on other exemptions at the federal and state levels, and on other aspects of FOIA (such as judging the relevance of a document to a request) justifiably deserve attention.

6 CONCLUSION AND FUTURE WORK

The approach adopted here represents an attempt to "reformulat[e] a seemingly difficult problem" – involving using human judgment to filter sensitive content in a large universe of data – through the use of computational tools [42]. This study has shown that classifiers trained using supervised machine learning can potentially be of benefit in highlighting portions of records that are within the scope of the deliberative process privilege under FOIA Exemption 5. Because considerable time and resources are currently devoted to manually reviewing records in responding to FOIA requests, including with respect to the statutory duty to reasonably segregate exempt and non-exempt material, automated ways with which to identify potentially exempt portions of text would be welcome. The use of such methods would not obviate the need for human review; rather, efficient and accurate "flagging" of material deemed by a classifier to be potentially within the scope of the privilege would assist reviewers in determining which records to review on the "front end" of any overall review effort.

Although our focus has been on training classifiers to detect deliberative process privilege material, this research exercise provides a path forward for applying a similar set of classifiers, or possibly an ensemble of them, to other FOIA exemptions. For example, FOIA Exemption 4 allows for withholding records containing trade secrets and confidential or proprietary information. FOIA Exemption 5 also allows for the withholding of attorney-client privileged information, as well as attorney work product. FOIA Exemption 6 allows records containing certain types of personal information to be withheld. FOIA Exemption 7 allows for withholding of various categories of law enforcement records the release of which would constitute an unwarranted invasion of privacy. All of these exemptions (and possibly others) could potentially benefit from automated classifiers being used to identify records or portions of records that might be deemed withholdable.

The need for automated classification for purposes of satisfying FOIA obligations is growing, especially given the digital turn that the records of our government are currently undergoing. After 2022, NARA will require agencies to transfer permanent records to the archives only in electronic form. Additionally, after 2022 agencies will be expected to manage all of their records (temporary and permanent) in electronic form [25]. Since 2016, e-mail records have been required to be similarly managed. With the introduction of what is known as the "Capstone" policy for e-mail retention [21], tens or hundreds of millions of e-mail records, including their many attachments, will soon be residing in any number of the larger agencies' repositories, all potentially subject to FOIA requests. Absent automated ways to reliably filter exempt from non-exempt materials to comply with FOIA obligations, access to a growing percentage of the government's records will be diminished. Any method that promises to facilitate the needed document review more efficiently through computational methods should be worthy of serious consideration.

Subject to the above caveats, the prospect of having reliable machine learning methods in place for assisting human reviewers in making FOIA determinations means that the documentary heritage embodied in government records will be made more accessible to the American people. This will continue to be of great importance in an ever-expanding universe of public records, both during and after the COVID-19 era.

ACKNOWLEDGMENTS

We wish to thank Patricia Weth, a senior government attorney with substantial expertise in FOIA law, for her assistance in annotating documents. We also wish to thank Dana Simmons, a supervisory archivist at the Clinton Library, for her assistance in helping identify a suitable test collection within the Library's holdings. This work was supported in part by NSF grant IIS-1618695.

REFERENCES

[1] Daniel Abril, Guillermo Navarro-Arribas, and Vicenç Torra. 2011. On the Declassification of Confidential Documents. In *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 235–246.

[2] Jason R Baron, David D Lewis, and Douglas W Oard. 2006. TREC 2006 Legal Track Overview.. In *Proceedings of the Fifteenth Text Retrieval Conference*. NIST Special Publication 500-272.

[3] Bennett B Borden and Jason R Baron. 2016. Opening up Dark Digital Archives Through the Use of Analytics to Identify Sensitive Content. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 3224–3229.

[4] Gordon V Cormack and Maura R Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. *arXiv preprint arXiv:1504.06868* (2015).

[5] Gordon V Cormack and Maura R Grossman. 2017. Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 5–14.

[6] Gordon V Cormack, Maura R Grossman, Bruce Hedin, and Douglas W Oard. 2010. Overview of the TREC 2010 Legal Track.. In *Proceedings of the Nineteenth Text Retrieval Conference*. NIST Special Publication 500-294.

[7] Jan Salomon Cramer. 2002. *The Origins of Logistic Regression*. Technical Report. Tinbergen Institute. https://papers.tinbergen.nl/02119.pdf

[8] Manfred Gabriel, Chris Paskach, and David Sharpe. 2013. The Challenge and Promise of Predictive Coding for Privilege. In *ICAIL 2013 DESI V Workshop*. http://users.umiacs.umd.edu/~oard/desi5/research/Gabriel-final2.pdf

[9] Maura R Grossman and Gordon V Cormack. 2010. Technology-Assisted Review in E-discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review. *Richmond Journal of Law and Technology* 17 (2010), 1.

[10] Maura R Grossman, Gordon V Cormack, and Adam Roegiest. 2016. TREC 2016 Total Recall Track Overview. In *Proceedings of the Twenty-Fifth Text Retrieval Conference*. NIST Special Publication 500-321.

[11] Mark H Grunewald. 1998. E-FOIA and the Mother of All Complaints: Information Delivery and Delay Reduction. *Administrative Law Review* 50 (1998), 345.

[12] Brian Harris, Elizabeth Whitaker, and Robert Simpson. 2005. *Access Restriction Checker*. Technical Report. Georgia Tech Research Institute. https://www.archives.gov/files/applied-research/papers/access-restrictiom-checker.pdf

[13] Emi Ishita, Satoshi Fukuda, Yoichi Tomiura, and Douglas W Oard. 2020. Using Text Classification to Improve Annotation Quality by Improving Annotator Consistency. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020).

[14] Joanne Kaczmarek and Brent West. 2018. Email Preservation at Scale: Preliminary Findings Supporting the Use of Predictive Coding.. In *International Conference on Digital Preservation (iPRES)*.

[15] Sandra Laib and William Underwood. 2006. *FOIA Processing in the Presidential Electronic Records Pilot System*. Technical Report. Georgia Tech Research Institute. https://www.archives.gov/files/applied-research/papers/presidential-electronic-records-pilot.pdf

[16] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* (1977), 159–174.

[17] Graham McDonald. 2019. *A Framework for Technology-Assisted Sensitivity Review: Using Sensitivity Classification to Prioritise Documents for Review*. Ph.D. Dissertation. University of Glasgow. http://theses.gla.ac.uk/41076/

[18] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2015. Using Part-of-Speech N-grams for Sensitive-Text Classification. In *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*. 381–384.

[19] Graham McDonald, Craig Macdonald, and Iadh Ounis. 2017. Enhancing Sensitivity Classification With Semantic Features Using Word Embeddings. In *European Conference on Information Retrieval*. Springer, 450–463.

[20] 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.

[21] National Archives and Records Administration. 2015. *White Paper on The Capstone Approach and Capstone GRS*. https://www.archives.gov/files/records-mgmt/email-management/final-capstone-white-paper.pdf

[22] Douglas W Oard, Fabrizio Sebastiani, and Jyothi K Vinjumur. 2018. Jointly Minimizing the Expected Costs of Review for Responsiveness and Privilege in E-discovery. *ACM Transactions on Information Systems (TOIS)* 37, 1 (2018), 1–35.

[23] Office of Government Information Services. 2020. *Webinar: FOIA Requests for CDC COVID-19 Records*. https://www.archives.gov/ogis/outreach-events/2020-05-12 Accessed: 2021-08-20.

[24] Office of Information Policy. 2019. *Summary of Annual FOIA Reports for Fiscal Year 2019*. https://www.justice.gov/oip/page/file/1282001/download Department of Justice, Accessed: 2021-08-20.

[25] Office of Management, Budget & National Archives, and Records Administration. 2019. *Memorandum M-19-21: Transition to Electronic Records.* https://www.archives.gov/files/records-mgmt/policy/m-19-21-transition-to-federal-records.pdf Accessed: 2021-08-20.
[26] David Sánchez, Montserrat Batet, and Alexandre Viejo. 2012. Detecting sensitive information from textual documents: an information-theoretic approach. In *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 173–184.
[27] 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*. 1605–1608.
[28] Mahmoud F Sayed and Douglas W Oard. 2019. Jointly Modeling Relevance and Sensitivity for Search Among Sensitive Content. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 615–624.
[29] Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the Scrub system.. In *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 333.
[30] Jacqueline Thomsen. 2020. *Watchdog Group Alleges FOIA Violations in Health Agencies' Handling of COVID-19 Records Requests.* https://www.law.com/nationallawjournal/2020/05/15/watchdog-group-alleges-foia-violations-in-health-agencies-handling-of-covid-19-records-requests/ Accessed: 2021-08-20.
[31] 2018-2020 FOIA Advisory Committee to the Archivist. 2020. *Final Report and Recommendations.* https://www.archives.gov/files/ogis/assets/foiaac-final-report-and-recs-2020-07-09.pdf Accessed: 2021-08-20.
[32] Amund Tveit, Ole Edsberg, TB Rost, Arild Faxvaag, O Nytro, T Nordgard, Martin Thorsen Ranang, and Anders Grimsmo. 2004. Anonymization of General Practioner Medical Records. In *Second HelsIT Conference.* https://www.researchgate.net/publication/228956524_Anonymization_of_General_Practioner_Medical_Records
[33] William Underwood and Sheila Isbell. 2008. *Semantic Annotation of Presidential E-Records.* Technical Report. Georgia Tech Research Institute. https://www.archives.gov/files/applied-research/papers/semantic-annotation.pdf
[34] William Underwood and Richard Marciano. 2019. Computational Thinking in Archival Science Research and Education. In *IEEE International Conference on Big Data*. 3146–3152.
[35] UNESCO. 2020. *Turning the Threat of COVID-19 into an Opportunity for Greater Support to Documentary Heritage.* https://en.unesco.org/news/turning-threat-covid-19-opportunity-greater-support-documentary-heritage Accessed: 2021-08-20.
[36] Jyothi K Vinjumur and Douglas W Oard. 2015. Finding the Privileged Few: Supporting Privilege Review for E-discovery. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.
[37] Hanna M Wallach. 2004. *Conditional random fields: An introduction.* Technical Report. University of Massachusetts. http://dirichlet.net/pdf/wallach04conditional.pdf
[38] Lipo Wang. 2005. *Support Vector Machines: Theory and Applications*. Vol. 177. Springer Science & Business Media.
[39] William Webber. 2010. Evaluating the Effectiveness of Keyword Search. *IEEE Data Engineering Bulletin* 33, 1 (2010), 54–59.
[40] David Weintrop, Elham Beheshti, Michael Horn, Kai Orton, Kemi Jona, Laura Trouille, and Uri Wilensky. 2016. Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology* 25, 1 (2016), 127–147.
[41] Rob Weychert. 2016. Delayed, Denied, Dismissed: Failures on the FOIA Front. *Propublica* (21 July 2016).
[42] Jeannette M Wing. 2006. Computational thinking. *Commun. ACM* 49, 3 (2006), 33–35.