Comparing Intrinsic and Extrinsic Evaluation of Sensitivity Classification

Mahmoud F. Sayed(✉), Nishanth Mallekav, and Douglas W. Oard

University of Maryland, College Park, MD 20742, USA
{mfayoub,oard}@umd.edu, nishum@terpmail.umd.edu

Abstract. With accelerating generation of digital content, it is often impractical at the point of creation to manually segregate sensitive information from information which can be shared. As a result, a great deal of useful content becomes inaccessible simply because it is intermixed with sensitive content. This paper compares traditional and neural techniques for detection of sensitive content, finding that using the two techniques together can yield improved results. Experiments with two test collections, one in which sensitivity is modeled as a topic and a second in which sensitivity is annotated directly, yield consistent improvements with an intrinsic (classification effectiveness) measure. Extrinsic evaluation is conducted by using a recently proposed learning to rank framework for sensitivity-aware ranked retrieval and a measure that rewards finding relevant documents but penalizes revealing sensitive documents.

Keywords: Evaluation · Sensitivity · Classification

1 Introduction

The goal of information retrieval is to find things that a searcher wants to see. Present systems are fairly good, so content providers need to be careful to exclude things that should not be found from the content being searched. As content volumes increase, segregation of sensitive content becomes more expensive. One approach is to ask content producers to mark sensitive content, but that suffers from at least two problems. First, the producer's interests may differ from those of future searchers, so producers may not be incentivized to label sensitivity in ways that would facilitate future access to content that is not actually sensitive. As an example, some lawyers note at the bottom of every email message that the message may contain privileged content. Doing so serves the lawyer's general interest in protecting privileged content, but there is no incentive for the lawyer to decide in each case whether such a note should be added. Second, sensitivity can change over time, so something marked as sensitive today may not be sensitive a decade from now. For both reasons, post-hoc sensitivity classification is often required. This paper explores measurement of the utility of a post-hoc classifier, comparing intrinsic evaluation (asking whether the classifier decided correctly in each case) with extrinsic evaluation (measuring the

M. Hagen et al. (Eds.): ECIR 2022, LNCS 13186, pp. 215–222, 2022.
https://doi.org/10.1007/978-3-030-99739-7_25

effect of a sensitivity classifier on a search engine that seeks to protect sensitive content [1]).

2 Related Work

This problem of deciding what information can be shown in response to a request arises in many settings [2], including protection of attorney-client privilege [3], protection of government interests [4], and protection of personal privacy [5]. Three broad approaches have been tried. The first is detecting sensitive content at the point of creation, a type of pre-filtering. For example, social media posts can be checked before posting to detect inappropriate content [6]. One problem with pre-filtering is that the effort required to detect errors is spread equally over all content, including content nobody is ever likely to search for. A second approach is to review search results for sensitive content before their release. This post-filtering approach is used when searching for digital evidence in lawsuits or regulatory investigations [7–9] and for government transparency requests [10–13]. Post-filtering and pre-filtering yield similar results, but with different operational considerations. Some limitations of post-filtering are that the initial search must be performed by some intermediary on behalf of the person requesting the content, and that review of retrieved results for sensitivity may be undesirably slow. There has been some work on a third way, integrating sensitivity review more closely with the search process [1]. The basic idea in this approach is to train a search system to balance the imperatives to find relevant documents and to protect sensitive documents. In this paper we compare post-filtering with this approach of jointly modeling relevance and sensitivity.

Determining whether a document is sensitive is a special case of text classification [14]. Many such techniques are available; among them we use the sklearn implementation of logistic regression in this paper [15]. More recently, excellent results have been obtained using neural deep learning techniques, in particular using variants of Bidirectonal Encoder Representation from Transformer (BERT) models [16]. In this paper, we use the DistilBERT implementation [17]. In text classification, the most basic feature set is the text itself: the words in each document, and sometimes also word order. Additional features can also be useful in specific applications. For example, in email search, senders and recipients might be useful cues [18–21]. Similarly, in news the source of the story (e.g., the New York Times or the National Enquirer) and its date might be useful. For this paper we limit our attention to word presence and, for BERT, word order.

Research on jointly modeling relevance and sensitivity has been facilitated by test collections that model both factors. We are aware of four such collections. Two simulate sensitivity using topic annotations in large collections of public documents (news [8] or medical articles [1]). Although using topicality to simulate sensitivity may be a useful first-order approximation, higher fidelity models are also needed. Two email collections have been annotated for relevance and sensitivity (the Avocado collection [5,22] and the Enron collection [23,24]). However, the use of content that is actually sensitive requires policy protections.

Some sensitivities decline over time, so a third approach is to annotate content that was initially sensitive but is no longer so. We are aware of two collections annotated for former sensitivities (national security classification [25] and deliberative process privilege [12]), but neither case includes relevance annotations.

3 Test Collections

The test collections used to train and evaluate the models are the Avocado Research Email Collection, and the OHSUMED text classification test collection. The OHSUMED test collection is a set of 348,566 references from MEDLINE, an on-line medical information database, consisting of titles and abstracts from 270 medical journals for the period 1987 through 1991. Each document is categorized based on predefined Medical Subject Heading (MeSH) labels, from which Sayed and Oard [1] selected two categories to represent sensitivity: C12 (Male Urogenital Diseases) and C13 (Female Urogenital Diseases and Pregnancy Complications). The Avocado Research Email Collection consists of emails and attachments taken from 279 accounts of a defunct information technology company referred to as "Avocado". The collection includes messages, attachments, contacts, and tasks, of which we use only the messages and the attachments (concatenating the text in each message and all of its attachments). There are in total of 938,035 messages and 325,506 attachments. The collection is distributed by the Linguistic Data Consortium on a restricted research license that includes content nondisclosure provisions [26].

Sayed et al. [22] created a test collection based on the Avocado email collection. Each email that is judged for relevance to any topic is also judged for sensitivity according to one of two predefined personas [5]. The persona represents the sender if the email was sent from an Avocado employee, or the recipient if the email was sent from outside the company network. The sensitivity of an email was annotated based on the persona's expected decision whether to allow the email to appear in search results. The John Snibert persona was motivated to donate his email to an archive because it documents his career; he was careful in his use of email, but worried that he may have overlooked some kinds of information about which he was sensitive (e.g., romantic partners, peer reviews, and proprietary information). 3,045 messages are annotated for a total of 35 topics, 1,485 of which are sensitive. The Holly Palmer persona, by contrast, had originally been reluctant to donate her email because she knows how much sensitive information they contain (e.g., family matters, receipts that contain credit card numbers, and conversations that might be taken out of context). 2,869 messages were annotated for a total of 35 topics, 493 of which are sensitive.

4 Intrinsic Evaluation

In this section, we measure the effectiveness of three models for classifying sensitivity: logistic regression, DistilBERT, and a combination of the two. For the OHSUMED collection, all documents have sensitivity labels, but only a subset

have relevance labels. So we use the subset that has both sensitivity and relevance labels as our test set, we use 85% of the documents that lack relevance labels for training sensitivity classifiers, and we use the remaining 15% of those documents that lack relevance labels as a validation set for sensitivity classifier parameter selection. Avocado is smaller, so in that case we evaluated classifiers using cross-validation. For each persona (John Snibert or Holly Palmer) in the Avocado collection, we first randomly split the annotated query-document pairs into 5 nearly equal partitions. We then iteratively chose one partition for evaluation and randomly selected 85% from the remaining four partitions as the corresponding training set, reserving the remaining 15% as a validation set.

Our logistic regression classifiers use sklearn's Logistic Regression library to estimate sensitivity probabilities [15]. The logistic regression model was trained on the union of the title and abstract for each document in the OHSUMED dataset, and on the union of the subject, body, and attachments for the Avocado collection. Our neural classifier estimates sensitivity probabilities using huggingface's DistilBERT, a pre-trained classification model trained on a large collection of English data in a self-supervised fashion [17]. DistilBERT is a distilled version of BERT large that runs 60% faster than BERT large while still retaining over 95% of its effectiveness. For the OHSUMED collection, fine-tuning DistilBERT for this classification task was performed using the training set. Many email messages have more text in the union of their subject, body and attachments than DistilBERT's 512-token limit, so for Avocado we divided the text of each item into 500-token passages with a 220-token stride. For fine-tuning each of the 5 Avocado classifiers for this task on the 5 training folds, we considered a passage sensitive if the document from which it had been extracted was marked as sensitive; for testing, we considered a document sensitive if any passage in that document was sensitive, and the probability of sensitivity for a document to be the maximum sensitivity probability for any passage in that document.

To assign a binary (yes/no) value for sensitivity to each test document, we learned one probability threshold for each classifier. We learned this threshold using the single validation partition on OHSUMED. For Avocado, we learned 5 thresholds, one for each validation fold. In each case, we used a grid search in the range [0, 1] with step size 0.01 to find the threshold that optimized the F_1 measure.

Our third model, a disjunctive combination of our logistic regression and DistilBERT models, used the **or** function between the decisions of the DistilBERT and logistic regression models. For example, if Logistic Regression identified an Avocado email message as sensitive but DistilBERT classified it as not sensitive, the combined model would declare it as sensitive.

Table 1 reports four intrinsic measures of classification effectiveness: precision, recall, F_1, and F_2. Our experiment showed the DistilBERT classifier having the best F_1 score on the OHSUMED dataset, logistic regression having the best F_1 for John Snibert, and the combined model having the best F_1 for Holly Palmer. The reason for DistilBERT excelling on OHSUMED for F_1 is likely related to the number of training samples ($> 250k$). Neural methods

Table 1. Intrinsic sensitivity classification results (percent, ↑ indicates higher is better). Superscripts indicate statistically significant improvement in accuracy over that system by McNemar's test [27] at $p < 0.05$.

Classifier	OHSUMED				
	Precision↑	Recall↑	F_1↑	F_2↑	Accuracy↑
(a) LR	76.72	73.29	74.96	73.95	94.01
(b) DistilBERT	**82.75**	80.08	**81.39**	80.60	**95.52**^{a,c}
(c) Combined	74.61	**83.81**	78.94	**81.8**	94.53^{a}
Classifier	Avocado: Holly Palmer				
	Precision↑	Recall↑	F_1↑	F_2↑	Accuracy↑
(a) LR	**72.29**	69.98	71.12	70.43	**90.34**^{b,c}
(b) DistilBERT	66.20	67.85	67.02	67.52	88.65
(c) Combined	64.15	**80.11**	**71.25**	**76.31**	89.02
Classifier	Avocado: John Snibert				
	Precision↑	Recall↑	F_1↑	F_2↑	Accuracy↑
(a) LR	**80.53**	84.85	**82.63**	83.95	**83.06**^{b,c}
(b) DistilBERT	72.87	87.00	79.31	83.75	78.44
(c) Combined	70.86	**93.73**	80.71	**88.05**	78.72

tend to perform better with more data, and the Avocado collection only contains around 2,000 training samples. The combined model performed the best by F_2 for all three datasets. F_2 emphasizes recall, and as expected, the combined model yielded the best recall in every case.

5 Extrinsic Evaluation

In this section, we study the effect of sensitivity classification on search among sensitive content. The post-filter approach works by applying the sensitivity classifier on the ranking model's output as a filter, so that any result that is predicted to be sensitive is removed from the result list. We build ranking models using the Coordinate Ascent ranking algorithm [28], optimizing towards normalized Discounted Cumulative Gain (nDCG). The joint approach works by having the ranking model optimized towards a measure that balances between relevance and sensitivity. This can be achieved by leveraging listwise learning to rank (LtR) techniques. In our experiments, we used the Coordinate Ascent listwise LtR algorithm, which outperforms other alternatives on these collections [1]. We use the nCS-DCG@10 measure for both training and evaluating models in this approach [1].

For the Combined joint model, we calculate the sensitivity probability using an independence assumption as $P_{Combined} = 1 - (1 - P_{LR})(1 - P_{DistilBERT})$, where P_x is the sensitivity probability of classifier x. Logistic regression produces

Table 2. Extrinsic nCS-DCG@10 ($C_s = 12$) (percent, higher is better), 5-fold cross validation. Superscript: significant improvement over that system by 2-tailed paired t-test ($p < 0.05$) [30].

Collection: topics classifier	OHSUMED: 106		Holly Palmer: 35		John Snibert: 35	
	Post-filter	Joint	Post-filter	Joint	Post-filter	Joint
(a) LR	83.11	83.81	79.92	87.38	76.32	80.87
(b) DistilBERT	84.57[a]	**85.95[a,c]**	82.41	86.30	75.48	80.74
(c) Combined	**84.97[a]**	84.44	**84.40[a]**	**90.67[a]**	**79.65**	**83.46[a]**
Oracle	89.44	88.70	92.19	89.64	95.40	91.91

well calibrated probabilities, but DistilBERT probabilities can benefit from calibration. For this, we binned the DistilBERT sensitivity estimates on the validation set into 10 uniform partitions (0-10%, 10%-20%, ..., 90%-100%). The fraction of truly sensitive documents in each partition was then computed using validation set annotations. We then found an affine function to transform system estimates to ground truth values, minizing Mean Square Error (MSE) over the 10 points. At test time, this function was used to map DistilBERT sensitivity probability estimates to better estimates of the true sensitivity probability. This is similar to Platt scaling [29], but with a linear rather than sigmoid model.

Table 2 compares the effect of the three classifiers, and an oracle classifier that gives 100% probability to the ground truth annotation, on the two sensitivity-protecting search approaches. As expected, the oracle classifier with post-filtering consistently yields the best results because it never makes a mistake. However, with real classifiers, jointly modeling relevance and sensitivity consistently yields better results than post-filtering. We also see that using DistilBERT for sensitivity classification yields strong extrinsic evaluation results for our largest training data condition (OHSUMED). However, for both of our smaller training data conditions (Holly Palmer and John Snibert) the combined model outperforms DistilBERT numerically (although not statistically significantly). Looking back to Table 1, we see that it was F_2 that preferred the combined classifier on those two smaller test collections, suggesting that when training data is limited, F_2 might be a useful intrinsic measure with which to initially compare sensitivity classifiers when optimizing for measures such as nCS-DCG that penalize failures to detect sensitive content which is our ultimate goal.

6 Conclusions and Future Work

It is tempting to believe that better sensitivity classification will yield better results for search among sensitive content, but we have shown that the truth of that statement depends on how one measures "better." Of course, there is more to be done. Our current classifiers use only words and word sequences; additional features such as relationship graphs and temporal patterns might help to further improve classification accuracy [31]. For our email experiments,

we have trained on sensitivity labels that are available only for documents that have been judged for relevance, but active learning might be used to extend the set of labeled documents in ways that could further improve classification accuracy. Finally, although we have tried a neural classification technique, we have combined this with a traditional approach to learning to rank for integrating search and protection. In future work, we plan to experiment with neural ranking as well.

Acknowledgments. This work has been supported in part by NSF grant 1618695.

References

1. Sayed, M.F., Oard, D.W.: 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, pp. 615–624. ACM (2019)
2. Thompson, E.D., Kaarst-Brown, M.L.: Sensitive information: a review and research agenda. J. Am. Soc. Inf. Sci. Technol. **56**(3), 245–257 (2005)
3. Gabriel, M., Paskach, C., Sharpe, D.: The challenge and promise of predictive coding for privilege. In: ICAIL 2013 DESI V Workshop (2013)
4. Mcdonald, G., Macdonald, C., Ounis, I.: How the accuracy and confidence of sensitivity classification affects digital sensitivity review. ACM Trans. Inf. Syst. (TOIS) **39**(1), 1–34 (2020)
5. Iqbal, M., Shilton, K., Sayed, M.F., Oard, D., Rivera, J.L., Cox, W.: Search with discretion: value sensitive design of training data for information retrieval. Proc. ACM Human Comput. Interact. **5**, 1–20 (2021)
6. Biega, J.A., Gummadi, K.P., Mele, I., Milchevski, D., Tryfonopoulos, C., Weikum, G.: R-susceptibility: an IR-centric approach to assessing privacy risks for users in online communities. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 365–374 (2016)
7. Oard, D.W., Webber, W.: Information retrieval for e-discovery. Found. Trends Inf. Retrieval **7**(2–3), 99–237 (2013)
8. Oard, D.W., Sebastiani, F., Vinjumur, J.K.: Jointly minimizing the expected costs of review for responsiveness and privilege in e-discovery. ACM Trans. Inf. Syst. (TOIS) **37**(1), 11 (2018)
9. Vinjumur, J.K.: Predictive Coding Techniques with Manual Review to Identify Privileged Documents in E-Discovery. PhD thesis, University of Maryland (2018)
10. McDonald, G., Macdonald, C., Ounis, I.: Enhancing sensitivity classification with semantic features using word embeddings. In: Jose, J.M., Hauff, C., Altıngovde, I.S., Song, D., Albakour, D., Watt, S., Tait, J. (eds.) ECIR 2017. LNCS, vol. 10193, pp. 450–463. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56608-5_35
11. Abril, D., Navarro-Arribas, G., Torra, V.: On the declassification of confidential documents. In: Torra, V., Narakawa, Y., Yin, J., Long, J. (eds.) MDAI 2011. LNCS (LNAI), vol. 6820, pp. 235–246. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22589-5_22
12. Baron, J.R., Sayed, M.F., Oard, D.W.: Providing more efficient access to government records: a use case involving application of machine learning to improve FOIA review for the deliberative process privilege. arXiv preprint arXiv:2011.07203, 2020

13. McDonald, G., Macdonald, C., Ounis, I., Gollins, T.: Towards a classifier for digital sensitivity review. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C.X., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 500–506. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06028-6_48
14. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002)
15. Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., Hutter, F.: Auto-sklearn 2.0: the next generation. arXiv preprint arXiv:2007.04074 (2020)
16. Adhikari, A., Ram, A., Tang, R., Lin, J.: DocBERT: BERT for document classification. arXiv preprint arXiv:1904.08398 (2019)
17. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
18. Alkhereyf, S., Rambow, O.: Work hard, play hard: email classification on the Avocado and Enron corpora. In: Proceedings of TextGraphs-11: The Workshop on Graph-based Methods for Natural Language Processing, pp. 57–65 (2017)
19. Crawford, E., Kay, J., McCreath, E.: Automatic induction of rules for e-mail classification. In: Australian Document Computing Symposium (2001)
20. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. In: Learning for Text Categorization: Papers from the 1998 workshop, Madison, Wisconsin, vol. 62, pp. 98–105 (1998)
21. Wang, M., He, Y., Jiang, M.: Text categorization of Enron email corpus based on information bottleneck and maximal entropy. In: IEEE 10th International Conference on Signal Processing, pp. 2472–2475. IEEE (2010)
22. Sayed, M.F., et al.: A test collection for relevance and sensitivity. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1605–1608 (2020)
23. Cormack, G.V., Grossman, M.R., Hedin, B., Oard, D.W.: Overview of the TREC 2010 legal track. In: TREC (2010)
24. Vinjumur, J.K., Oard, D.W., Paik, J.H.: Assessing the reliability and reusability of an e-discovery privilege test collection. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1047–1050 (2014)
25. Brennan, W.: The declassification engine: reading between the black bars. The New Yorker (2013). https://www.newyorker.com/tech/annals-of-technology/the-declassification-engine-reading-between-the-black-bars
26. Oard, D., Webber, W., Kirsch, D., Golitsynskiy, S.: Avocado research email collection. Linguistic Data Consortium, Philadelphia (2015)
27. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
28. Metzler, D., Croft, W.B.: Linear feature-based models for information retrieval. Inf. Retrieval **10**(3), 257–274 (2007)
29. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classifiers **10**(3), 61–74 (1999)
30. De Winter, J.C.F.: Using the Student's t-test with extremely small sample sizes. Pract. Assess. Res. Eval. **18**(1), 10 (2013)
31. Sayed, M.F.: Search Among Sensitive Content. PhD thesis, University of Maryland, College Park (2021)