Semi-Automated Text Classification for Sensitivity Identification

Giacomo Berardi[◇], Andrea Esuli[◇], Craig Macdonald[♣],
Iadh Ounis[♣], Fabrizio Sebastiani[♡][*]
[◇]Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy
[♣]School of Computing Science, University of Glasgow, Glasgow, UK
[♡]Qatar Computing Research Institute, Hamad bin Khalifa University, Doha, Qatar

ABSTRACT

Sensitive documents are those that cannot be made public, e.g., for personal or organizational privacy reasons. For instance, documents requested through Freedom of Information mechanisms must be manually reviewed for the presence of sensitive information before their actual release. Hence, tools that can assist human reviewers in spotting sensitive information are of great value to government organizations subject to Freedom of Information laws. We look at sensitivity identification in terms of semi-automated text classification (SATC), the task of ranking automatically classified documents so as to optimize the cost-effectiveness of human post-checking work. We use a recently proposed utility-theoretic approach to SATC that explicitly optimizes the chosen effectiveness function when ranking the documents by sensitivity; this is especially useful in our case, since sensitivity identification is a recall-oriented task, thus requiring the use of a recall-oriented evaluation measure such as F_2. We show the validity of this approach by running experiments on a multi-label multi-class dataset of government documents manually annotated according to different types of sensitivity.

1. INTRODUCTION

Government documents may be deposited in archives for public viewing after a period of years, or released into the public domain through Freedom of Information mechanisms. However, documents containing *sensitive* information should not be released, as they may reveal personal information, thereby infringing on someone's privacy, or reveal information that may offend other countries.

Classically, for paper documents, the identification of sensitive documents has taken place using human reviewers. However, with limited government budgets, the adoption of text classification techniques that aid in the identification of sensitive documents is attractive, since it can increase the efficiency of human reviewers. The possibility of treating sensitivity review as an automated text classification task has recently been shown in [7], where text classifiers were used in order to automatically detect sensitive documents, and where "sensitive" can have different interpretations (e.g., defence-related issues, or issues related to law enforcement).

The task of sensitivity identification bears strong resemblances with "review for privilege" in e-discovery [8], where expert attorneys must check that "privileged" (i.e., sensitive) information is not accidentally disclosed to a requesting party in the context of a civil litigation process [3, 10]. Another task that bears resemblances with sensitivity identification is record anonymisation, as when e.g., medical records have to be anonymised before they are released for epidemiological studies; in this case, sensitive information such as patients' names and medical doctors' names have to be spotted in order to be redacted [9]. Sensitivity identification and privilege identification are text classification tasks, while record anonymisation is an information extraction task. Notwithstanding the differences, all these cases are characterized by the fact that the costs of accidental disclosure of sensitive information are high.

In this paper we follow in the steps of [7] and investigate automatic techniques for aiding sensitivity review. However, while [7] was concerned with automatically classifying documents by sensitivity, here we are concerned with aiding a human annotator who validates (i.e., inspects and corrects where appropriate) these automatically classified documents, with the goal of maximizing the cost-effectiveness of the annotator's work. In other words, while [7] was concerned with "Step 1" in the workflow, we tackle "Step 2".

We frame the task of aiding our human annotator as a *semi-automatic text classification* (SATC) task. SATC (see [2, 5, 6]) is defined as the task of ranking a set $\mathcal{D}$ of automatically classified textual documents in such a way that, if a human annotator validates the documents in a top-ranked portion of $\mathcal{D}$ with the goal of increasing the overall classification accuracy of $\mathcal{D}$, the expected increase in accuracy is maximized. Therefore, we envisage our annotators as validating documents by sensitivity, starting from the top of the ranked list we generate, and working downwards (until they are confident that the dataset has been cleared up, or until the budget for annotation work has been spent).

We approach SATC by adopting the utility-theoretic approach of [2] (hereafter: U-Theoretic). Essentially, U-Theoretic ranks the automatically labelled documents by taking two factors into consideration, i.e.,

[*] Fabrizio Sebastiani is currently on leave from Consiglio Nazionale delle Ricerche, Italy.

CIKM'15, October 19–23, 2015, Melbourne, Australia.
© 2015 ACM. ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806597.

1. the probability that the document has been misclassified by the classifier (documents with high probability of misclassification should be ranked higher, since no benefit will occur from validating a correctly classified document), and
2. the increase (or *gain*) in the overall accuracy of the automatically labelled set that occurs if the document is validated (documents that bring about a higher gain should be ranked higher).

Concerning (2), while the same gain occurs from validating either a true positive (TP) or a true negative (TN) – this gain is 0, since the human annotator will not change their labels – the gain that occurs from validating a false positive (FP) or a false negative (FN) may be, as shown in [2], different. When the two gains are different, the utility-theoretic approach tested in this paper:

- is indeed different from an approach based on (1) only, which we describe as the *purely probabilistic approach* (indeed, the two approaches instead coincide when the two gains are the same), and
- is intuitively superior to the latter, since the goal of SATC is increasing the overall labelling accuracy of $\mathcal{D}$, and this means that we must bring to bear the increase in this overall labelling accuracy that validating a given document brings about.

Note that the gains from validating TPs and TNs are different (as shown in Equations (3) and (4) below) when accuracy is measured via functions such as F_1 (which pays equal importance to precision and to recall). However, this is even more true when using metrics that give more importance to recall than to precision (or vice versa), since the imbalance between the gains deriving from validating a FP or a FN is even higher. This is indeed our case, since sensitivity identification is a recall-oriented task (there is a higher cost involved in missing a sensitive document than in erroneously catching a non-sensitive document), which means that a measure (such as e.g., F_2) that emphasizes recall over precision needs to be adopted.

Hence, in this paper, our application of the utility-theoretic approach for semi-automated text classification to the recall-oriented task of sensitivity identification, brings two contributions: (i) it verifies the utility-theoretic approach on the more difficult task of sensitivity identification, and (ii) it moves the state-of-the-art in sensitivity identification from automatic text classification to an assistive approach that can benefit the efficiency of an annotator examining documents for sensitivities. The rest of the paper is structured as follows. In Section 2 we describe our approach to top-ranking sensitive information via semi-automated text classification. Section 3 describes our experiments carried out on a dataset of documents annotated according to different types of sensitivity. Section 4 provides concluding remarks.

2. SATC SENSITIVITY IDENTIFICATION

The U-Theoretic approach described in [2] tackles SATC in a "multi-label multi-class" context, i.e., there is a set of classes $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ with $|\mathcal{C}| > 1$ (this makes it a "multi-class" problem) and each document d_i may belong to zero, one, or several among the classes in $\mathcal{C}$ (this makes it a "multi-label" problem). The U-Theoretic approach assumes that binary classifiers h_j, one for each $c_j \in \mathcal{C}$, have classified the set of unlabelled documents (which, for the purpose of our experiments, will here be equated with the test set Te), also returning a confidence score for each classification decision; the binary decisions returned by classifier h_j will be collectively denoted as $h_j(Te)$.

U-Theoretic ranks documents in terms of *total utility*, i.e.,

$$U(d_i) = \sum_{c_j \in \mathcal{C}} U_j(d_i) \tag{1}$$

where $U_j(d_i)$ (*class-specific utility*) is defined as:

$$U_j(d_i) = \sum_{\omega_j^i \in \{tp_j^i, fp_j^i, fn_j^i, tn_j^i\}} P(\omega_j^i) G(\omega_j^i) \tag{2}$$

Here, $P(\omega_j^i)$ is the system's estimate of the probability that event ω_j^i occurs, tp_j^i is the event "d_i is a true positive for c_j", and fp_j^i, fn_j^i, tn_j^i are defined similarly. If the classifier is a probabilistic one, the $P(\omega_j^i)$'s are the posterior probabilities directly generated by the classifier (e.g., if d_i is a positive example of c_j, then $P(tp_j^i)$ is $P(c_j|d_i)$, $P(fn_j^i)$ is $(1 - P(c_j|d_i))$, and $P(fp_j^i) = P(tn_j^i) = 0$). If the classifier is not a probabilistic one, these probabilities can be obtained from the confidence scores output by the classifier via the application of a generalized logistic function (see [2, Section 3.3] for details).

In Equation (2), $G(\omega_j^i)$ is the *gain*, defined as the average increase in the value of the evaluation function that derives when an event of type ω_j^i occurs, where the average is computed across all documents of the same type (e.g., if $\omega_j^i \equiv fp_j^i$, the average is computed across all of the false positives for c_j). For instance, we take $G(fp_j^i)$ to be the average increase in the accuracy of Te that derives by correcting a false positive for c_j – i.e. removing from Te a false positive and adding to Te a true negative – calculated as:

$$\begin{aligned} G(fp_j^i) &= \frac{1}{FP_j}(F_1^{FP}(h_j(Te)) - F_1(h_j(Te))) \\ &= \frac{1}{FP_j}(\frac{2TP_j}{2TP_j + FN_j} - \frac{2TP_j}{2TP_j + FP_j + FN_j}) \end{aligned} \tag{3}$$

where by TP_j we indicate the number of true positives for class c_j (and analogously for FP_j, FN_j, TN_j).We assume F_1 to be the chosen evaluation function, where $F_1(h_j(Te))$ denotes the value of F_1 computed on $h_j(Te)$, and where by $F_1^{FP}(h_j(Te))$ we indicate the value of F_1 that would derive by correcting all false positives in $h_j(Te)$ (i.e., turning all of them into true negatives). Similarly, we take $G(fn_j^i)$ to be:

$$\begin{aligned} G(fn_j^i) &= \frac{1}{FN_j}(F_1^{FN}(h_j(Te)) - F_1(h_j(Te))) \\ &= \frac{1}{FN_j}(\frac{2(TP_j + FN_j)}{2(TP_j + FN_j) + FP_j} - \\ &\qquad \frac{2TP_j}{2TP_j + FP_j + FN_j}) \end{aligned} \tag{4}$$

Here, the values of TP_j, FP_j, FN_j, TN_j are unknown (since in a real application the true labels of the examples we have automatically classified are unknown), but may be estimated via k-fold cross-validation (see [2, Sect. 3.4] for details).

Note that the method generates a single ranking (according to the document scores computed via Equation (1)), and not $|\mathcal{C}|$ different ones; this allows the human annotator to scan the document list only once, validating a document for all the classes in $\mathcal{C}$ before moving on to the next document.

3. EXPERIMENTS

In this section we report on experiments that we have conducted in order to test how well U-Theoretic performs on the task of sensitivity identification.

3.1 Experimental setting

For our experiments we have used the same dataset as used in [7]. This dataset contains 1,111 government records sampled from a larger corpus of documents addressing international relations. Unlike the rest of the corpus, which is still unlabelled, these 1,111 records have been manually labelled by government experts according to two types of sensitivity identified in the UK's Freedom of Information Act 2000[1], namely "Section 40", which deals with the occurrence of personal information and "Section 27", which describes material damaging to international relations. Of the 1,111 judged records, 104 were judged to be sensitive for Section 27 and 86 for Section 40 (see [7] for more details on this dataset).

For each of the 1,111 records we use the same features employed in [7] for text classification purposes. In particular, we use the contents of the document represented as a vector of term frequencies, along with four other features, namely the presence of dictionary last names, presence of internationally famous persons as listed in DBpedia (e.g., Royal Family, government leaders, etc.), number of subjective sentences as identified by the OpinionFinder toolkit [11], and a risk score for the record depending on the countries mentioned within the record. Again, see [7] for more details on the feature representations used.

As a measure of classification accuracy we use F_2, an instance of the well-known F_β function defined as:

$$\begin{aligned} F_\beta &= \frac{(\beta^2+1)TP}{(\beta^2+2)TP+(\beta^2+1)FN+FP} \\ F_2 &= \frac{5TP}{6TP+5FN+FP} \end{aligned} \tag{5}$$

We adopt F_2 since (as already observed in the introduction) ours is a recall-oriented task, and F_2 is a standard choice for this kind of tasks. Note that, as a consequence, in Equations (3) and (4) we use F_2 (and its definition from Equation (5)) in place of F_1; in other words, for us a gain $G(\omega_j^i)$ is defined in terms of increases in F_2 (and not F_1) obtained from the annotator's validation activity.

As the learning algorithm we use SVMs, c.f. Joachims' SVM-light implementation [4]; in all our experiments we use a linear kernel. Due to the imbalance of the training data, we oversample each training set by generating duplicates of sensitive records until we obtain the same numbers of sensitive and non-sensitive records in the training set. This solution (see e.g., [1]) turns out to be effective in improving the SVM performance despite its simplicity, especially when the number of the minority class examples is small. (Initial experiments we have performed without oversampling have yielded radically worse accuracy results.)

We perform our experiments using repeated random subsampling validation, i.e., (a) we generate multiple (in our case: 10) random training / test splits of the original dataset (in our case: always using 80% of the data for training and 20% for testing), (b) for each such split we train our classifiers on the training set and (c) classify + rank the test data using the trained classifier, and finally (d) we compute the final effectiveness results as the average effectiveness across the different splits.

For each of the 10 splits, as a substep of step (b) we perform parameter optimization on the training data. We do this by first optimizing (via 10-fold cross-validation, and independently for each class) the C parameter of SVMs, which determines the tradeoff between the training error and the margin. Then, using the value of C deemed optimal we optimize the parameter of the generalized logistic function (used for calibrating the $P(\omega_j^i)$ needed in Equation (2) – see Section 2) via a second 10-fold cross-validation phase (in this phase we also obtain the estimates of TP_j, FP_j, FN_j, TN_j which are needed for computing $G(fp_j^i)$ and $G(fn_j^i)$ in Equations (3) and (4)). The test sets are then classified and ranked by employing the parameter values deemed optimal.

In order to evaluate the effectiveness of our approach, we use the $ENER_\rho^M(\xi)$ ("expected normalized error reduction") measure proposed in [2]. $ENER_\rho^M(\xi)$ measures the reduction in classification error (of the automatically labelled document set) obtained when a human annotator validates (i.e., inspects and corrects where appropriate) the top-ranked documents in the ranking generated by ranking method ρ; here, the M superscript stands for "macroaveraging", to indicate that the error is computed for each class separately and the results are then averaged. $ENER_\rho^M(\xi)$ is based on a probabilistic user model, and ξ represents the expected percentage of the ranking that the human annotator inspects; for instance, $ENER_\rho^M(0.10)$ examines a scenario in which the expected number of documents that the user validates is one tenth of the entire automatically labelled set.

In $ENER_\rho^M(\xi)$, "classification error" may be measured according to any desired measure of error; differently from [2], which used $(1-F_1)$, we obviously use $(1-F_2)$. The values on which $ENER_\rho^M(\xi)$ ranges are a strict subset of [0,1], with better rankings corresponding to higher values of $ENER_\rho^M(\xi)$. We refer the reader to [2, Section 4] for more details on $ENER_\rho^M(\xi)$.

We follow [2] in using, as the baseline for our experiments, a probabilistic (as opposed to utility-theoretic) system that ranks documents according to the probability of misclassification only, i.e., the documents most likely to be misclassified are top-ranked. This is obtained by using U-Theoretic with both $G(fp_j^i)$ and $G(fn_j^i)$ set to 1. This is a "lower bound" baseline, which we expect the U-Theoretic approach to outperform. Along with [2], we also report two "upper bound", idealised baselines (named Oracle1 and Oracle2), that we expect our system to underperform; both baselines are versions of the U-Theoretic approach endowed with foreknowledge (Oracle1 has foreknowledge of the true values of TP_j, FP_j, FN_j, TN_j and can thus compute $G(fp_j^i)$ and $G(fn_j^i)$ precisely, while Oracle2 even has foreknowledge of the true labels in the test set, and can thus use binary values in place of probabilities in Equation (2)). These two "upper bound" baselines thus indicate how far U-Theoretic is from the ideal performance.

3.2 Results

The results of our experiments are reported in Table 1, for three different values of ξ. As an example, the result $ENER_\rho^M(0.05) = 0.215$ obtained by U-Theoretic can approximately be interpreted as saying that, when using this ranking method, by only validating 5% of the automatically classified documents an annotator is expected to obtain a

[1] http://www.legislation.gov.uk/ukpga/2000/36/contents

Table 1: Results of different ranking methods in terms of $ENER_{\rho}^{M}(\xi)$, for $\xi \in \{0.05, 0.10, 0.20\}$, with F_2 as the measure of classification error. Improvements are relative to the baseline.

	$\xi = 0.05$	$\xi = 0.10$	$\xi = 0.20$
Baseline	0.189	0.263	0.306
U-Theoretic	0.215 (+14%)	0.283 (+8%)	0.316 (+3%)
Oracle1	0.223 (+18%)	0.289 (+10%)	0.321 (+5%)
Oracle2	0.517 (+174%)	0.629 (+139%)	0.647 (+111%)

reduction in classification error (measured in terms of (1-F_2)) strictly higher than 0.215[2].

The first interesting fact we may observe is that the utility-theoretic method substantially outperforms the baseline, with achieved improvements with respect to it ranging from +3% to +14%. A second interesting fact is that the results of the utility-theoretic method are very close to the results of Oracle1, an idealized method that has foreknowledge of the true values of TP_j, FP_j, FN_j, TN_j; this means that U-Theoretic does a good job in estimating these quantities.

Note that the improvements obtained by U-Theoretic over Baseline are lower than the ones (even exceeding +100%) reported in [2] for training sets of comparable size. The likely reason is that classification by sensitivity is much harder than classification by topic (which is the task [2] tackles), as shown by the fact that the SVM classifiers of [2] obtain (for training sets of comparable size) $F_1 = 0.527$ while our SVM classifiers here obtain (as an average across the 10 test sets) $F_2 = 0.213$. We conjecture that a classifier (such as ours) that obtains lower classification accuracy also generates less reliable confidence scores, which are the key input to U-Theoretic. We can thus expect U-Theoretic to achieve even bigger margins on the probabilistic baseline once training sets are larger and (as a consequence) both the classification accuracy and the quality of the confidence scores are higher.

While the improvements we have obtained are substantial, there are reasons to believe that in realistic applications, characterized by much larger *unlabelled* sets than the one we had access to, the improvements would be even larger; in fact, as [2] shows, the larger the set to be ranked, the more U-Theoretic shines with respect to the baseline.

4. CONCLUSION

The sensitivity review of “digital-born” textual records is an important process for Freedom of Information initiatives at the heart of the open-government agenda. Framing sensitivity identification as an automatic text classification task is challenging, as noted in previous work [7]. For this reason, assistive techniques that can aid a sensitivity reviewer – such as semi-automatic text classification (SATC) – play an important role. In this work, we investigate the applicability of the utility-theoretic approach for SATC, previously proposed in [2]. Our experiments indicate that, when ranking documents in order to maximize the cost-effectiveness of human post-inspection work for sensitivity identification, there are clear benefits to gain from using a utility-theoretic approach instead of the purely probabilistic approach.

[2] We say “strictly higher than 0.215” because $ENER_{\rho}^{M}(\xi)$ does not measure the “absolute” reduction in error, but a normalized version of it where we factor out the reduction in error (equal to 5%, i.e., 0.050) we would obtain by using a random ranker. The improvement obtained by U-Theoretic is thus 0.215+0.050=0.265.

5. REFERENCES

[1] G. E. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations*, 6(1):20–29, 2004.

[2] G. Berardi, A. Esuli, and F. Sebastiani. A utility-theoretic ranking method for semi-automated text classification. In *Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012)*, pages 961–970, Portland OR, US, 2012.

[3] M. Gabriel, C. Paskach, and D. Sharpe. The challenge and promise of predictive coding for privilege. In *Proceedings of the ICAIL 2013 Workshop on Standards for Using Predictive Coding (DESI V)*, Roma, IT, 2013.

[4] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. J. Burges, and A. J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, chapter 11, pages 169–184. The MIT Press, Cambridge, US, 1999.

[5] M. Martinez-Alvarez, A. Bellogin, and T. Roelleke. Document difficulty framework for semi-automatic text classification. In *Proceedings of the 15th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2013)*, Prague, CZ, 2013.

[6] M. Martinez-Alvarez, S. Yahyaei, and T. Roelleke. Semi-automatic document classification: Exploiting document difficulty. In *Proceedings of the 34th European Conference on Information Retrieval (ECIR 2012)*, Barcelona, ES, 2012.

[7] G. McDonald, C. Macdonald, I. Ounis, and T. Gollins. Towards a classifier for digital sensitivity review. In *Proceedings of the 36th European Conference on Information Retrieval (ECIR 2014)*, pages 500–506, Amsterdam, NL, 2014.

[8] D. W. Oard and W. Webber. Information retrieval for e-discovery. *Foundations and Trends in Information Retrieval*, 7(2/3):99–237, 2013.

[9] G. Szarvas, R. Farkas, and R. Busa-Fekete. State-of-the-art anonymisation of medical data with an iterative machine learning model/framework. *Journal of the American Medical Informatics Association*, 14(5):574–580, 2007.

[10] J. K. Vinjumur, D. W. Oard, , and J. H. Paik. Assessing the reliability and reusability of an e-discovery privilege test collection. In *Proceedings of the 37th ACM Conference on Research and Development in Information Retrieval (SIGIR 2014)*, pages 1047—1050, Gold Coast, AU, 2014.

[11] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of the HLT/EMNLP 2005 Interactive Demonstrations*, pages 34–35, Vancouver, CA, 2005.