Journal of Biomedical Informatics 63 (2016) 174–183

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A unified framework for evaluating the risk of re-identification of text de-identification tools

Martin Scaiano [a,b], Grant Middleton [b], Luk Arbuckle [d], Varada Kolhatkar [a,b], Liam Peyton [a], Moira Dowling [e], Debbie S. Gipson [f], Khaled El Emam [a,b,c,d,*]

[a] School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada
[b] Privacy Analytics Inc., Ottawa, Canada
[c] Department of Pediatrics, University of Ottawa, Ottawa, Canada
[d] Children's Hospital of Eastern Ontario Research Institute, Ottawa, Canada
[e] Michigan Institute for Data Science (MIDAS), University of Michigan Medical School, Office of Research, Ann Arbor, United States
[f] Department of Pediatrics, University of Michigan, Ann Arbor, United States

ARTICLE INFO

Article history:
Received 10 December 2015
Revised 14 June 2016
Accepted 13 July 2016
Available online 15 July 2016

Keywords:
De-identification
Re-identification risk
Medical text
Evaluation framework
Natural language processing
Data sharing

ABSTRACT

Objectives: It has become regular practice to de-identify unstructured medical text for use in research using automatic methods, the goal of which is to remove patient identifying information to minimize re-identification risk. The metrics commonly used to determine if these systems are performing well do not accurately reflect the risk of a patient being re-identified. We therefore developed a framework for measuring the risk of re-identification associated with textual data releases.

Methods: We apply the proposed evaluation framework to a data set from the University of Michigan Medical School. Our risk assessment results are then compared with those that would be obtained using a typical contemporary micro-average evaluation of recall in order to illustrate the difference between the proposed evaluation framework and the current baseline method.

Results: We demonstrate how this framework compares against common measures of the re-identification risk associated with an automated text de-identification process. For the probability of re-identification using our evaluation framework we obtained a mean value for direct identifiers of 0.0074 and a mean value for quasi-identifiers of 0.0022. The 95% confidence interval for these estimates were below the relevant thresholds. The threshold for direct identifier risk was based on previously used approaches in the literature. The threshold for quasi-identifiers was determined based on the context of the data release following commonly used de-identification criteria for structured data.

Discussion: Our framework attempts to correct for poorly distributed evaluation corpora, accounts for the data release context, and avoids the often optimistic assumptions that are made using the more traditional evaluation approach. It therefore provides a more realistic estimate of the true probability of re-identification.

Conclusions: This framework should be used as a basis for computing re-identification risk in order to more realistically evaluate future text de-identification tools.

1. Introduction

There has been significant research on developing tools for the de-identification of free-form medical text [1,2]. The evaluation methods currently used to determine whether these tools are performing well enough are borrowed from the areas of entity extraction and information retrieval [3]. There has been some recognition that these evaluation approaches are not always the most appropriate for measuring the probability of re-identification nor are the benchmarks typically used to decide what is "good enough" directly relevant to the de-identification task [4]. Such concerns triggered the current work.

In this paper we critically examine the methods that are currently used to evaluate medical text de-identification tools [1,2], identify their weaknesses, and propose improvements. We then propose a unified framework for evaluation in terms of the probability of re-identification when medical text is de-identified using automated tools. Our framework builds on existing work, and its

* Corresponding author at: Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1H 8L1, Canada.
E-mail address: kelemam@uottawa.ca (K. El Emam).

http://dx.doi.org/10.1016/j.jbi.2016.07.015
1532-0464/

main contribution is that it brings multiple concepts together from the disclosure control literature, the information retrieval literature, and the risk modeling literature to provide a more detailed evaluation scheme for measuring re-identification risk.

The issues we identify in current evaluation methods can in some instances inflate the performance of de-identification tools by making them look better than they really are, and in other instances may also penalize them by making them seem much worse than they really are. This means that our proposed evaluation framework will not consistently give higher risk values or lower risk values than currently used methods, although we argue that it represents a more accurate modeling of the probability of re-identification because it better accounts for the distribution of identifiers in documents. We illustrate the differences between our framework and conventional evaluation approaches using theoretical and empirical examples. We then illustrate the application of this framework on a clinical data set, and compare the findings to what would be obtained using current evaluation methods.

2. Background

2.1. Evaluation approaches used in text de-identification

Most of the current text de-identification systems treat Personal Health Information (PHI) identification as a named entity recognition problem. Consequently, they evaluate the identification performance with metrics used in the named entity recognition and information retrieval literature [3]. In particular, they typically annotate different types of entities (or categories), such as *date*, *patient name*, and *ID*, and report performance primarily using three metrics: *precision*, *recall*, and *f-measure*. Let *tp* be the number of true positive annotations, *fp* be the number of false positive annotations, and *fn* be the number of false negative annotations. Then, recall r is given by

$$r = tp/(tp + fn), \tag{1}$$

and precision p is given by

$$p = tp/(tp + fp). \tag{2}$$

Recall and precision answer two questions about a de-identification tool, respectively: "Did we find all that we were looking for?" and "Did we only label what we were looking for?" The metric *f-measure* combines precision and recall, typically by taking the harmonic mean of the two. To get a sense of the overall performance of a system, the most commonly used metrics are *micro-average* and *macro-average* precision, recall, and f-measure. To compute micro-average, one creates a confusion matrix for all categories and then computes precision and recall from this table, giving equal weight to each PHI instance irrespective of its category. To compute macro-average, one computes precision and recall for each category separately and then averages them over all categories, giving equal weight to each category, to get an overall measure of performance.

In Appendix A we summarize evaluation metrics currently used in the text de-identification literature. This review indicates that micro-average recall is a primary metric for evaluating such tools. We also conclude that the number of clinical notes (i.e., number of patients) used in different studies range from 100 to 7193, and that the number of test documents used in different studies range from 220 to 514.

In the context of text de-identification, current evaluation approaches are limited in three ways. First, they report performance on all instances of an entity across all documents. However, none of them consider the number of PHI elements missed within a document, which is an important aspect in de-identification, as a document typically corresponds to a patient and any leaks within a document mean potentially revealing the identity of that patient. In other words, current evaluation approaches do not truly reflect the risk of a patient being re-identified. Second, they evaluate all types of entities with the same evaluation metric, giving equal weight to each entity type even though directly identifying entities, such as name and address, have a higher risk of re-identification compared to indirectly identifying entities, such as age and race. Finally, they do not account for the distribution of PHI across documents. For example, an entity type that is rare and appears in very few documents will have a higher sensitivity to the performance of an information extraction tool than a more prevalent entity type. We examine each of these issues below.

2.2. Basic concepts

The key assumptions that we make in developing our evaluation framework are detailed below. Some of these assumptions are already made in the literature implicitly, but it is important in our context to make them explicit.

2.2.1. One document = one patient

We assume that every document that is being analyzed pertains to an individual patient (i.e., there is a one-to-one mapping between documents and patients). This means that if a document pertains to multiple patients then that information is split into multiple documents. This assumption simplifies the presentation of our framework and its rationale.

In the case where a simple split is not possible, as in the case of clinical study reports from clinical trials, then we assume that all of the information pertaining to an individual trial participant can be extracted as a unit and treated as a separate virtual document for the purposes of evaluation.

This assumption also means that each patient only has one document in the corpus. For example, if the evaluation corpus consists of hospital discharge records, then each patient has a single discharge record.

2.2.2. Information leak = re-identification

Furthermore, we assume that if an annotation is not detected (i.e., "leaked") then it can be used to re-identify a patient. So the probability of re-identifying a patient is conditional on a leak occurring. We have:

$$Pr(\text{reid}, \text{leak}) = Pr(\text{reid}|\text{leak}) \times Pr(\text{leak}) \tag{3}$$

The probability of a leak in a set of documents is directly related to recall, r, given by:

$$Pr(\text{leak}) = 1 - r \tag{4}$$

Based on our assumptions we can then say:

$$Pr(\text{reid}|\text{leak}) = 1 \tag{5}$$

We will examine further below *how much* information needs to be leaked to re-identify a patient. This simplifying assumption is conservative in that it will inflate the risk of re-identification.

2.2.3. Re-identification from correct information extraction

A corollary to the assumption above is that if an annotation is detected, or "caught", then it is either redacted or re-synthesized, such that the probability of re-identifying a patient from that information is zero.

We can formulate this probability as:

$$Pr(\text{reid}, \text{catch}) = Pr(\text{reid}|\text{catch}) \times Pr(\text{catch}) \tag{6}$$

where $Pr(\text{catch}) = 1 - \text{Pr}(\text{leak})$, which is recall. Clearly the annotations that were leaked versus those that were caught are mutually