Protected Health Information Removal from Text Documents: A Proposal for Raising F-measure, Precision, and Recall Scores

Joffrey Leevy, MSc
Florida Atlantic University, USA, jleevy2017@fau.edu

***Abstract*– Text de-identification, also known as scrubbing, is the removal or concealing of personal identifiers in narrative text documents. Under the HIPAA Privacy Rule, documents containing "identifiable" information are subject to certain legal protections that safeguard the privacy of patients. However, medical documents with de-identified information do not require such protections. De-identification systems based on machine-learning algorithms can automatically learn to recognize patterns that may involve protected health information data. The goal of this proposal is to improve upon currently published F-measure, precision, and recall scores of processes that perform automated text de-identification.**

***Keywords*—protected health information, text de-identification, machine learning, F-measure, precision, recall, HIPAA**

I. Introduction

The Health Insurance Portability and Accountability Act (HIPAA) protects the privacy of patient data. If this clinical data is de-identified as per the HIPAA Safe Harbor regulations, the data can freely be used by the public. These regulations require the removal of 18 data elements [1], collectively called Protected Health Information (PHI): PHI includes names, social security numbers, account numbers, e-mail addresses, geographic subdivisions smaller than a state, and several more elements.

Text de-identification, also known as scrubbing, is the removal or concealing of personal identifiers in narrative text documents. Under the HIPAA Privacy Rule, documents that have "identifiable" information are subject to certain legal protections that safeguard the privacy of patients. However, medical documents with de-identified information do not require such protections. If de-identification is done manually, it involves skilled and/or unskilled human workers who must meticulously analyze each document for removable PHI. This process is usually labor and cost-intensive, particularly when dealing with large documents that must conform to HIPAA privacy requirements. The preferable alternative, therefore, is automated text de-identification.

Automated text de-identification of clinical data take two main approaches: pattern matching and machine learning [1]. Typically, pattern matching utilizes regular expressions to check for a given sequence of tokens. Pattern matching also uses dictionaries constructed from publicly availably sources, including the Social Security Death Index, spell-checking lexicons, and biomedical thesauruses. As per HIPAA considerations, pattern matching requires databases of dictionaries and rules that can grow prohibitively large. Consequently, academic researchers and industry are now favoring the machine-learning approach. De-identification systems based on machine-learning algorithms can automatically learn to recognize patterns that may involve complex PHI data. For PHI, approaches based on machine learning are usually more generalizable than pattern-matching methods.

The following section briefly discusses relevant seminal research work on machine-learning techniques for the automated removal of PHI.

II. Seminal Research

Aramaki et al. [2] investigated a method that merged non-local features, such as label consistency and sentence features, with local features. The system used a machine-learning method called Conditional Random Fields (CRF) to learn the link between features and labels in a training set. Compared to other systems that were entered in a de-identification challenge (i2b2 2006 Challenge), the performance of this system is above average.

Guo et al. [3] adopted an approach that analyzed support vector machines. The researchers used the open source GATE and ANNIE systems for natural language processing and information extraction respectively. This system performed below average in the i2b2.

Hara [4] came up with a system that used support vector machines and text classifiers to identify PHI. The system also incorporated pattern matching and regular expressions. This approach showed an average performance in the i2b2.

Szarvas et al. [5] used an iterative technique based on decision trees, yet another machine-learning approach. This system also used regular expressions. It examined information in report headers in order to improve PHI recognition in the report body. The Sarvas system demonstrated above-average performance in the i2b2.

Wellner et al. [6] carried out their research as a sequence-labeling problem, where labels are assigned to individual words. The implementation of Conditional

Digital Object Identifier: (to be inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).

Random Fields (CRF) was a major factor that caused the Wellner system to earn top place in the i2b2. In addition, bias parameters and regular expressions were used to tune the balance of recall versus precision and also the confidence score. This approach was based on an adaptation of the Carafe toolkit developed by MITRE. It is worth noting that the precision, recall, and F-measure scores for this system were all greater than 96%.

III. PROJECT PROPOSAL

The goal of this project is to improve upon currently published F-measure, precision, and recall scores of processes that perform automated text de-identification. I intend to use hospital discharge summaries as the document source. Several advanced machine-learning techniques will be analyzed, such as the use of artificial neural networks (ANNs), conditional random fields (CRFs), ensemble and model-stacking methods, and the utilization of Bayesian optimization of hyperparameters. Phrase extraction evaluations will be performed at the phrase and token levels, and it is expected that selected PHI elements (names, phone numbers, IDs, zip codes, hospital names) will be efficiently removed.

One potential approach for my proposed research is based on the recent work of Dernoncourt et al. [8], which details a de-identification system based on ANNs. The primary components of this model are recurrent neural networks. The F-measure, precision, and recall scores were comparatively high when ranked against other competing systems. Another approach involves the training of CRFs with distinct categories of linguistic features [9]. Surface features are qualified by descriptors such as token length, punctuation and typographic case, whereas distributional analysis features use descriptors such as corpus and document section. A third possible approach implements a non-parametric Baynesian Hidden Markov Model [10]. This technique uses latent variables to group words with the same label more efficiently, thus enabling minor variations in the data to be better captured. For this model, a Dirichlet process selects the optimum number of latent variables contributing toward PHI removal. The three approaches mentioned in this paragraph are by no means an exhaustive list, but are expected to be techniques that will be investigated first during the experimental phase of my research.

REFERENCES

[1] Meystre, S., Friedlin, F., South, B., Shen, S., Samore, M.: Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology* 10(70) (2010) 1-16

[2] Aramaki, E., Imai, T., Miyo, K., Ohe, K.: Automatic de-identification by using sentence features and label consistency: *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC (2006)

[3] Guo, Y., Gaizauskas, R., Roberts, I., Demetriou, G., Hepple, M.: Identifying personal health information using support vector machines: *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC (2006)

[4] Hara, K: Applying a SVM based chunker and a text classifier to the Deid Challenge*: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC (2006)

[5] Szarvas, G., Farkas, R., Busa-Fekete, R.: State-of-the-art anonymization of medical records using an iterative machine learning framework: *J Am Med Inform Assoc* 14(5) (2007) 574-580

[6] Wellner, B et al.: Rapidly retargetable approaches to de-identification in medical records*: J Am Med Inform Assoc* 14(5) (2007) 564-573

[7]] Stubbs, A., Kotfila, C., Uzuner, O.: Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1: *Journal of Biomedical Informatics* 58 (2015) 511-519

[8] Dernoncourt, F., Lee, J., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks: *J Am Med Inform Assoc* 24(3) (2017) 596-606

[9] Grouin, C.: Clinical records de-identification using CRF and rule-based approaches*: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC (2014)

[10]Chen, T., Cullen, R., Godwin, M.: Hidden Markov model using Dirichlet process for de-identification: *Journal of Biomedical Informatics* 58S (2015) S60-S66