Journal of Digital Imaging (2022) 35:1694–1698
https://doi.org/10.1007/s10278-022-00673-0

Ensemble Approaches to Recognize Protected Health Information in Radiology Reports

Hannah Horng[1] · Jackson Steinkamp[2] · Charles E. Kahn Jr.[2,3] · Tessa S. Cook[2]

Received: 29 March 2022 / Revised: 2 June 2022 / Accepted: 7 June 2022 / Published online: 17 June 2022

Abstract

Natural language processing (NLP) techniques for electronic health records have shown great potential to improve the quality of medical care. The text of radiology reports frequently constitutes a large fraction of EHR data, and can provide valuable information about patients' diagnoses, medical history, and imaging findings. The lack of a major public repository for radiological reports severely limits the development, testing, and application of new NLP tools. De-identification of protected health information (PHI) presents a major challenge to building such repositories, as many automated tools for de-identification were trained or designed for clinical notes and do not perform sufficiently well to build a public database of radiology reports. We developed and evaluated six ensemble models based on three publically available de-identification tools: MIT de-id, NeuroNER, and Philter. A set of 1023 reports was set aside as the testing partition. Two individuals with medical training annotated the test set for PHI; differences were resolved by consensus. Ensemble methods included simple voting schemes (1-Vote, 2-Votes, and 3-Votes), a decision tree, a naïve Bayesian classifier, and Adaboost boosting. The 1-Vote ensemble achieved recall of 998 / 1043 (95.7%); the 3-Votes ensemble had precision of 1035 / 1043 (99.2%). F1 scores were: 93.4% for the decision tree, 71.2% for the naïve Bayesian classifier, and 87.5% for the boosting method. Basic voting algorithms and machine learning classifiers incorporating the predictions of multiple tools can outperform each tool acting alone in de-identifying radiology reports. Ensemble methods hold substantial potential to improve automated de-identification tools for radiology reports to make such reports more available for research use to improve patient care and outcomes.

Keywords Natural language processing · De-identification · Protected health information (PHI) · Reporting · Machine learning · Ensemble models

Introduction

Developments in machine learning have enabled rapid advancement in natural language processing (NLP) of text-based clinical notes in electronic health records (EHR) and radiology reports [1, 2]. Many NLP applications are trained on large, publicly available repositories of clinical notes such as i2b2/n2c2 and MIMIC; such repositories can be essential to the development of new NLP tools to improve patient care [3]. Although numerous repositories of radiological images are publicly available, there are few public repositories of radiology reports [4]. Such reports contain a wealth of information – such as imaging findings, diagnoses, conclusions, and recommendations – that could be leveraged by new NLP tools. Already, NLP tools trained using radiology reports can aid in diagnostic surveillance, cohort building, query-based case retrieval, and quality assessment [5].

Laws such as the U.S. Health Insurance Portability and Accountability Act (HIPAA) [6] and the European Union's General Data Protection Regulation (GDPR) [7] impose penalties for the release of protected health information. De-identification of radiology reports to make them publicly available poses a major obstacle to the construction of a public multi-site repository [8].

De-identification of clinical text – the identification and removal of protected health information (PHI) such as names, dates, patient numbers, and other identifiers – remains a difficult

Charles E. Kahn Jr. and Tessa S. Cook are Co-senior authors

✉ Charles E. Kahn Jr.
ckahn@upenn.edu

[1] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA

[2] Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA

[3] Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA

task in part due to the variations in structure, format, language, and distribution of PHI. Although some de-identification tools are publicly available, most have been trained on and developed for clinical notes [9–11]. Recent studies have shown that domain adaptation is required to adequately de-identify texts from other domains, and that these existing tools exhibit reduced performance when applied to radiology reports [12].

Ensemble methods combine the predictions of multiple decision models. Such ensembles use a wide range of methods of voting and weighting each vote to achieve greater performance in a variety of contexts [13, 14]. In this work, we developed and evaluated ensemble methods for de-identification of PHI from radiology reports that combined the predictions of three publicly available de-identification tools originally made for clinical notes. We tested the hypothesis that performance of an ensemble method could surpass that of individual de-identification tools.

Materials and Methods

Dataset

The study was approved by the organization's Institutional Review Board. The study data were collected during routine clinical care. A random sample of 2,503 radiology reports from January 1, 2012 to January 8, 2019 was assembled retrospectively from a large, multi-hospital U.S. medical system, which included academic and community radiology practices with urban, suburban, and rural practice sites. Reports were created by more than 300 attending radiologists and radiology trainees. The reports included a variety of imaging modalities. We only evaluated final reports, and did not distinguish between reports with and without involvement of a trainee. Reports were partitioned into disjoint training (n = 1480) and testing (n = 1023) sets.

Annotation

PHI was defined to match the Safe Harbor criteria of the HIPAA Privacy rule [8], with the addition of three categories: names of healthcare workers, names of hospitals, and names of software/tools. All reports were labelled by two annotators to ensure inter-annotator reliability and produce labels of higher accuracy; differences of opinion were resolved by consensus. The frequency of PHI in the testing set is shown in Table 1.

De-Identification Software

We incorporated three publicly available de-identification tools, all developed originally for clinical notes. MIT deid and Philter incorporate a variety of dictionaries, rules, and expressions to identify PHI and do not incorporate machine learning [9, 11]. NeuroNER is a machine learning model that uses recurrent neural networks to identify various forms of PHI [10]; in this work, we used the NeuroNER model pre-trained on radiology reports, as described by Steinkamp et al. [14].

Ensemble Methods

We defined three simple voting approaches (1-Vote, 2-Votes, and 3-Votes) as positive if one or more, two or more, or all three of the primary models were positive, respectively. The value of each entry in the array was "1" if the existing tool had identified the corresponding token as PHI, and "0" otherwise (Fig. 1). The simple voting approaches were applied to all tokens in the 1023 reports in the test set.

The three features were then used as inputs into three traditional machine learning classifiers (Decision Tree, Bayesian, and Boosting) with a 60/40 train/test split. In a

Table 1 Distribution of PHI in the 1023 radiology reports used for evaluation. The Safe Harbor method from the HIPAA Privacy rule was used to define standardized PHI, with the addition of three categories: names of healthcare workers, names of hospitals, and names of software/tools

PHI Type	Number of documents	Percentage of all documents in Training set (n = 1023)	Number of tokens (n = 254,862)
Dates	603	58.9%	1127
Names of healthcare workers	99	9.7%	162
Names of hospitals	50	4.9%	66
Names of software/tools	22	2.2%	29
Any other numerical identifier	12	1.2%	18
Location	24	2.3%	31
Names of patients or family members	8	0.8%	12
Medical record numbers	7	0.7%	11
Phone numbers	4	0.4%	4
Other	1	0.1%	1
Any type of PHI	646	63.1%	1461

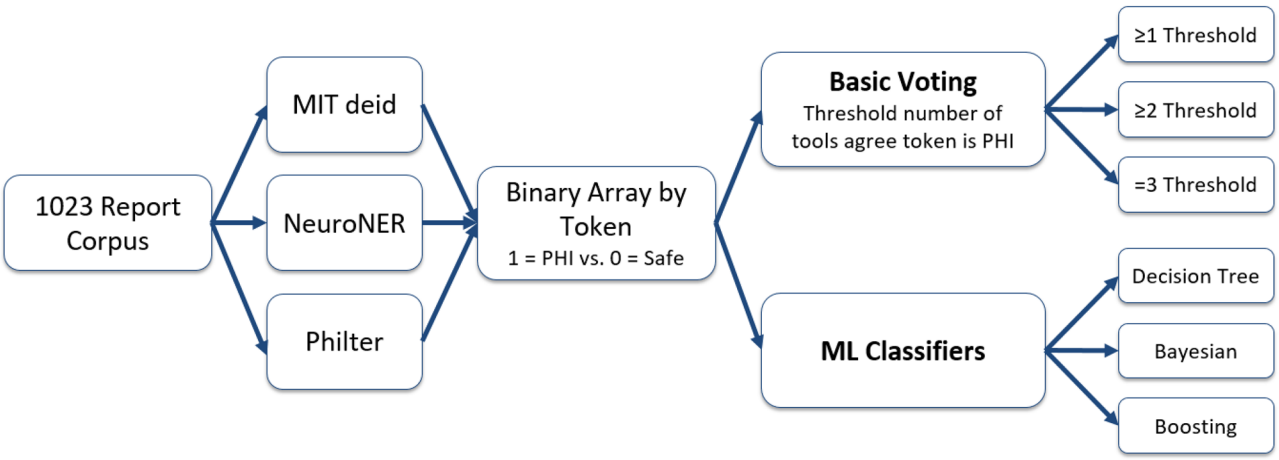

Fig. 1 Workflow for ensemble de-identification methods. PHI was detected at the token level, and the distribution of PHI in the dataset at both the report and token level can be found in Table 1

Decision Tree classifier, a set of features is used as input into a tree structure, where each internal node tests a specific attribute and is assigned to a class [14]. In a Naïve Bayes (Bayesian) classifier, the distribution of the outputs as well as the conditional distribution of the inputs on the classes are estimated, and Bayes rule is applied to obtain the posterior probability of a sample belonging to a class given its corresponding inputs [14]. Here we use a Gaussian Naïve Bayes classifier, where the likelihood of the inputs is assumed to be based on a Gaussian (normal) distribution. In an Adaboost (boosting) classifier, multiple weak classifiers are successively trained with increasing weight placed on misclassified observation, then combined into an ensemble (15). Precision, recall, and F1 score were computed for the voting algorithms and machine learning ensembles.

Results

The performance of MIT deid, NeuroNER, and Philter on their original datasets and on the testing set of radiology reports is shown in Table 2. Performance metrics included precision (positive predictive value, the fraction of relevant instances among the retrieved instances) and recall (sensitivity, the fraction of relevant instances that were retrieved), and F1 (the harmonic mean of precision and recall). All three tools demonstrated lower recall on radiology reports than their reported performance on clinical notes.

The performance metrics for the ensemble classifiers are shown in Table 3. The simple voting classifiers were evaluated on all 1,023 testing-partition reports; the other three ensemble classifiers were evaluated on the 40% test split of the testing partition reports. Not surprisingly, the 3-Votes ensemble achieved greater precision and the 1-Vote algorithm had greater recall than any of the underlying classifiers. The Decision Tree and Boosting classifiers also demonstrated greater precision than the three individual tools, while the Bayesian classifier showed greater recall.

Discussion

Summary

The three individual publicly available de-identification tools – MIT *deid*, NeuroNER, and Philter – all demonstrated lower recall on radiology reports than on clinical

Table 2 Performance of the three publicly available de-identification tools on their original datasets and on the test set of radiology reports [7–9]. Performance metrics for three publicly available de-identification tools on the testing set of 1,023 reports, with 95% confidence intervals [10]

	Performance on Original Dataset			Performance on Test Set of 1023 Radiology Reports		
Tool	Description	Precision	Recall	Precision	Recall	F1
MIT deid	2434 nursing notes	74.9%	96.7%	81.7% (79.9%–83.3%)	67.6% (65.7%–69.5%)	74.0%
NeuroNER	1635 free-text medical notes	98.8%	99.4%	94.5% (93.5%–95.4%)	92.6% (91.5%–93.6%)	93.6%
Philter	4500 clinical notes	79.4%	95.4%	31.2% (30.1%–32.4%)	83.0% (81.4%–84.4%)	45.4%

Table 3 Performance of the ensemble models with 95% confidence intervals

Ensemble Classifier	Testing Set Size	Precision	Recall	F1
1-Vote	1043	57.5% (56.0%–59.0%)	95.7% (94.8%–96.4%)	71.8%
2-Votes	1043	93.7% (92.6%–94.6%)	82.8% (81.3%–84.3%)	87.9%
3-Votes	1043	99.2% (98.6%–99.5%)	58.5% (56.5%–60.4%)	73.6%
Decision Tree	417	97.7% (96.4%–98.5%)	89.5% (87.5%–91.3%)	93.4%
Bayesian	417	56.8% (54.4%–59.2%)	95.3% (94.8%–96.4%)	71.2%
Boosting	417	98.2% (97.0%–98.9%)	78.9% (76.2%–78.9%)	87.5%

notes, likely because radiology reports can differ from clinical notes in both structure and language. The distribution of PHI in radiology reports can differ substantially from the general-purpose clinical notes used to train many text de-identification tools [14]. For example, in our radiology reports dataset, dates were the most common form of PHI and can vary widely in format, making them difficult to capture with solely rule-based de-identification methods (Table 1). The performance of existing de-identification tools made for clinical notes in radiology reports is inadequate for clinical and research use, indicating a need for improved de-identification methods specifically for radiology reports.

No ensemble method was able to outperform NeuroNER in both precision and recall, but both the Bayesian classifier and 1-Vote basic voting algorithm outperformed NeuroNER by exceeding 95% in recall. This performance represents a substantial improvement: de-identification tasks tend to prioritize recall over precision because a false negative requires significant manual review to identify. By using an ensemble method, one effectively pools the predictions of multiple tools such that if a PHI token goes undetected by one tool, it could still be detected as PHI by another tool to enable superior recall.

Limitations

Although these methods are promising, they were evaluated on a limited dataset from a single health system with a skewed distribution of PHI.

Future Work

Future work will include assessing performance of ensemble methods from a larger multicenter dataset to incorporate more variations in format. Data augmentation will be applied to better quantify tool performance on types of PHI that are under-represented in this dataset, such as the names of patients, family members, and healthcare workers. We also will examine whether the addition of more publicly available de-identification tools to the ensemble can improve recall performance.

Conclusions

In this work, we have developed ensemble methods for de-identification of PHI in radiology reports that incorporate publicly available de-identification tools developed for clinical notes. We have shown that the Bayesian classifier and ≥ 1 threshold basic voting algorithm were able to outperform the best individual de-identification tool (NeuroNER) in recall, indicating that these ensemble methods show substantial promise for larger-scale implementation in building a publicly available corpus of de-identified radiology reports. Future work includes evaluation on a larger multicenter dataset augmented with under-represented forms of PHI, as well as incorporation of additional de-identification tools into the ensembles.

Availability of Data and Material The radiology reports evaluated in this study contain protected health information (PHI), and thus cannot be made available publicly.

Code Availability The de-identification tools evaluated here are publicly available: MIT *deid* (https://physionet.org/content/deid/1.1/), NeuroNER (https://github.com/Franck-Dernoncourt/NeuroNER), and Philter (https://github.com/BCHSI/philter-ucsf).

Declarations

Ethics Approval This study was approved by an institutional review board.

Consent to Participate/Publication Informed consent was waived.

Conflicts of Interest The authors declare no conflicts of interest.

References

1. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR Med Inform 2019; 7(2):e12239.
2. Luo JW, Chong JJR. Review of natural language processing in radiology. Neuroimaging clinics of North America 2020; 30(4):447-458.

3. Wu S, Roberts K, Datta S, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 2020; 27(3):457-470.
4. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013; 26(6):1045-1057.
5. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016; 279(2):329-343.
6. United States Congress. Public Law 104–191. Health insurance portability and accountability act. 1996. Available online: https://www.govinfo.gov/app/details/PLAW-104publ191. Accessed 10 December 2021.
7. European Union. Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA. 2016. Available online: https://eur-lex.europa.eu/eli/dir/2016/680/oj. Accessed 10 December 2021.
8. U.S. Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability Act (HIPAA) Privacy Rule. 2012. Available online: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf. Accessed December 2021.
9. Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008; 8:32.
10. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017; 24(3):596-606.
11. Norgeot B, Muenzen K, Peterson TA, et al. Protected Health Information filter (Philter): accurately and securely de-identifying free-text clinical notes. NPJ digital medicine 2020; 3:57.
12. Lee HJ, Zhang Y, Roberts K, Xu H. Leveraging existing corpora for de-identification of psychiatric notes using domain adaptation. AMIA Annu Symp Proc 2017:1070–1079.
13. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013; 20(1):77-83.
14. Steinkamp JM, Pomeranz T, Adleberg J, Kahn CE, Jr., Cook TS. Evaluation of automated public de-identification tools on a corpus of radiology reports. Radiol Artif Intell 2020; 2(6):e190137.