Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210917

Clinical Notes De-Identification: Scoping Recent Benchmarks for n2c2 Datasets

Taridzo CHOMUTARE[a,1]
[a]*Norwegian Centre for E-health Research, Tromsø, Norway*

Abstract. Publicly shared repositories play an important role in advancing performance benchmarks for some of the most important tasks in natural language processing (NLP) and healthcare in general. This study reviews most recent benchmarks based on the 2014 n2c2 de-identification dataset. Pre-processing challenges were uncovered, and attention brought to the discrepancies in reported number of Protected Health Information (PHI) entities among the studies. Improved reporting is required for greater transparency and reproducibility.

Keywords. Natural language processing, NLP, i2b2, de-identification

1. Introduction

Removing identifying information from data sources is defined as de-identification, where the goal is to make re-identification of individuals impossible. For healthcare, data privacy and security are the primary concerns. In this regard, a number of legislative guidelines exist in many territories around the world. For instance in USA, where Health Insurance Portability and Accountability Act (HIPAA) mandates that certain direct identifiers and quasi-identifiers, aptly named Protected Health Information (PHI), be removed from any health data before such data is shared. In Europe, the General Data Protection Regulation (GDPR) provides similar guidelines to protect the data of citizens.

For most datasets, a typical de-identification pipeline includes removing direct identifiers such as IDs or email. Even after these are removed, the data may still contain quasi-identifiers such as the date of birth. Therefore, an anonymization process is required to transform the data in order to reduce the risk of disclosure. For structured data, this process can be straightforward, for example by using statistical methods to reduce the risk of re-identification, e.g, by generalization or suppression.

In contrast, unstructured data, such as clinical text, requires more complex methods to reduce disclosure risk. There is an abundance of unstructured clinical text that could help shed light on some of the most key healthcare challenges today. In view of this important problem, a challenge to promote and disseminate natural language processing (NLP) methods for de-identifying clinical notes was launched in 2006, by Informatics for Integrating Biology and the Bedside (i2b2) [10], now called n2c2. Since then, the task has generated wide interest, and different research groups continue to work on improving performance. Even though much has been published on this topic, we still

[1] Corresponding Author, Taridzo Chomutare, Norwegian Centre for E-Health Research, 9019 Tromsø, Norway; E-mail: Taridzo.chomutare@ehealthresearch.no.

lack an overview of the progress. This study reviews some of the recent work on the topic.

2. Methods

A search was conducted for studies that had used the 2014 i2b2 de-identification datasets for benchmarking their new algorithms. The search string used variations of the core string "i2b2 AND de-identification". This excludes studies that participated in the challenge itself, and only includes studies conducted in the past five years (2017 - 2021). Search on Google Scholar, PubMed, IEEE and ACM were conducted. Studies that did not fully describe the properties of the data were excluded, for instance, if they did not provide the PHI entity numbers after pre-processing.

The following data items were collected: (i) methods used, (ii) data properties (number of PHI entities), and (iii) performance data. In addition to this data, it is also noted if each respective study publishes its code in a public repository.

3. Results

Based on the search hits, duplicates were screened out and 9 studies that met the selection criteria were included in this study. Results of the data properties are shown in Table 1, where we can observe only two studies that agree on the number of PHI entities for some of the HIPAA categories (bold print). The rest of the studies ended up with varying number of entities based on the same i2b2 dataset.

Table 1. Counts of entities as a total (*or just the test set) datasets as reported in the studies.

PHI	[1]	[2]*	[3]*	[4]	[5]	[6]	[7]	[8]	[9]
DATE	12482	4951	4980	-	-	12468	12381	12473	12532
NAME	7348	4131	2883	-	-	-	7258	7361	4839
AGE	**1997**	-	764	-	-	1991	2028	**1997**	790
CONTACT	**541**	171	218	-	-	-	610	**541**	419
ID	**1506**	576	625	-	-	1039	1549	**1506**	1126
LOCATION	4580	1177	1813	-	-	-	3986	4578	3001
PROFESSION	**413**	-	179	-	-	-	420	**413**	340
All entities	28867	10861	11462	28872	26787	28862	28205	28869	23047

Disregarding these discrepancies in number of entities, the reported performances are shown in Table 2. From the table, performance measured by F1, improved from a 2017 high of 0.983 [1], to a high of 0.985 [2] in 2021. However, because of the different evaluation methods, this improvement cannot be taken at face value. Studies used different evaluation methods, from token-based binary [1,6,7,8] classification to entity classification based on HIPAA PHI [5,9].

In terms of methods, all the studies were deep learning-based, and invariably used bidirectional long-short term memory (Bi-LSTM), Conditional Random Fields (CRF) and Gated recurrent units (GRU). With these base methods, studies developed multiple innovative ensemble methods through voting mechanisms [3], stacking [3,7] and novel attention mechanisms based on transformer models [7, 8], and use of rule-based methods and dictionaries [1, 2, 9].

Table 2. Performance as reported in the studies, and the respective methods used.

Study	Precision	Recall	F1	Evaluation	Code	Methods used
[1]2017	0.993	0.973	0.983	Binary token	no	Bi-LSTM, CRF, rule-based
[9] 2017	0.983	0.973	0.979	HIPAA PHI	Yes	Bi-LSTM, CRF, dictionaries
[6] 2018	0.989	0.972	0.981	Binary token	No	Bi-GRU
[8] 2019	0.990	0.983	0.987	Binary token	No	Bi-LSTM, CRF, transformers
[3] 2020	-	-	0.959	Strict entity	No	Bi-LSTM, CRF, voting, stacking
[7] 2020	0.980	0.984	0.982	Binary token	Yes	Bi-GRU, GRU-LSTM, stacking, self-attention
[2] 2021	0.979	0.992	0.985	-	No	DL, iterative fine-tuning, dictionaries
[4] 2021	0.947	0.918	0.933	Strict entity	No	Bi-LSTM, CRF, n-gram moving window
[5] 2021	0.839	0.818	0.828	HIPAA PHI	No	Bi-LSTM, CRF

4. Discussion

Perhaps the most unexpected finding was the large discrepancies in the reported number of PHI entities. It appears the problem stems from the need to re-format the original datasets, to satisfy the input format requirements for specific algorithms. This re-formatting or pre-processing during a typical de-identification pipeline, appear to yield significantly different results for each study. Therefore, it is difficult to measure the overall performance improvements if the data are not consistent. This is a key point since most reported improvements will only be small fractions of a percent. Even small discrepancies in test datasets will skew results.

This problem stems from the very nature of the dataset, which is generally considered sensitive and requires individuals to sign non-disclosure agreements. Therefore, the data or respective transformations into new data formats cannot be uploaded to the Internet. One solution could be to ask the data proprietor to update the repository with new updates on data formats. This could be useful since multiple studies reported minor errors in some annotations.

In terms of performance, different studies use different evaluation methods, and this has a large effect on the overall results, and makes it difficult to compare results across studies. While there is debate regarding the best evaluation method, it could be beneficial if multiple evaluation methods were used and reported. So far, however, the general reporting appears insufficient, since some studies are not specific about the evaluation methods used. Further, most of the studies provided neither the algorithm code nor the evaluation script. It is therefore nearly impossible to reproduce their work. This is especially important because there are many implementations of an algorithm, and small variations of an algorithm can have a significant impact. In addition, there are multiple combinations of hyper-parameters, and optimization processes were never reported. Advancement of scientific knowledge depends on full disclosure of such information, but only two studies provided a code repository [7, 9] of their work.

Turning to the methods used, a possible explanation for the common use of deep-learning is the scientific progress in the field, especially with the development of contextual embeddings and large language models like BERT [12]. These developments have changed the game for NLP, and much of emerging new innovation centers around their use. However, it is interesting to note the use of dictionaries, where some studies

use rule-based systems combined with deep learning algorithms as part of a whole system [1], or as part of a post-processing step.

5. Conclusions

While interest in this de-identification task appear to continue to increase, there are still challenges that distract the scientific community from fully realizing the ideals of shared datasets. Perhaps prioritizing better reporting and full code-sharing could be a starting point. This is an important step for reproduction of work and to make further scientific progress by building on current knowledge.

Acknowledgments

This work was partially supported by the Northern Norway Regional Health Authority, (Helse Nord); research grant HNF1395-18, NorKlinTekst project. The funding body does not have any role in the study.

References

[1] Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. Journal of biomedical informatics. 2017 Nov 1;75:S34-42.
[2] Murugadoss K, Rajasekharan A, Malin B, Agarwal V, Bade S, Anderson JR, Ross JL, Faubion Jr WA, Halamka JD, Soundararajan V, Ardhanari S. Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. Patterns. 2021 Jun 11;2(6):100255.
[3] Kim Y, Heider PM, Meystre SM. Comparative Study of Various Approaches for Ensemble-based De-identification of Electronic Health Record Narratives. InAMIA Annual Symposium Proceedings 2020;2020:648.
[4] Lee K, Kayaalp M, Henry S, Uzuner Ö. A Context-Enhanced De-identification System. arXiv preprint arXiv:2102.08513. 2021 Feb 17.
[5] Ahmed A, Abbasi A, Eickhoff C. Benchmarking Modern Named Entity Recognition Techniques for Free-text Health Record De-identification. arXiv preprint arXiv:2103.13546. 2021 Mar 25.
[6] Zhao YS, Zhang KL, Ma HC. et al. Leveraging text skeleton for de-identification of electronic medical records. BMC Med Inform Decis Mak. 2018;18.
[7] Ahmed T, Al Aziz MM, Mohammed N. De-identification of electronic health record using neural network. Scientific reports. 2020 Oct 29;10(1):1-1.
[8] Tang B, Jiang D, Chen Q, Wang X, Yan J, Shen Y. De-identification of Clinical Text via Bi-LSTM-CRF with Neural Language Models. AMIA Annu Symp Proc. 2019.
[9] 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).
[10] Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform. 2015;58.
[11] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.