MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231199

High Accuracy Open-Source Clinical Data De-Identification: The CliniDeID Solution

Stéphane MEYSTRE[a,b,1] and Paul HEIDER[c]
[a] *OnePlanet Research Center, Nijmegen & Wageningen, The Netherlands*
[b] *Clinacuity, Inc., Charleston, SC, USA*
[c] *Medical University of South Carolina, Charleston, SC, USA*
ORCiD ID: Stephane Meystre https://orcid.org/0000-0002-7632-9625, Paul Heider https://orcid.org/0000-0002-1589-4567

Abstract. Clinical data de-identification offers patient data privacy protection and eases reuse of clinical data. As an open-source solution to de-identify unstructured clinical text with high accuracy, CliniDeID applies an ensemble method combining deep and shallow machine learning with rule-based algorithms. It reached high recall and precision when recently evaluated with a selection of clinical text corpora.

Keywords. De-identification, privacy protection, natural language processing, AI

1. Introduction

The increased use and adoption of Electronic Health Records (EHR), and parallel growth in patient data available for secondary use by clinicians, researchers, and operational purposes, all cause patient data privacy protection to become an increasingly important requirement and expectation. The laws protecting patient privacy and confidentiality typically require the informed consent of the patient to use data for research purposes, a requirement that can be waived if the data are de-identified. Most clinical data in the EHR is stored in unstructured text format. Several methods to automatically remove identifying information from this text have been tested experimentally over the last 10 years, mostly guided by the U.S. HIPAA “Safe Harbor” methodology. As part of these efforts, *CliniDeID* automatically de-identifies clinical notes and structured data. It is described in more details below.

2. Methods

CliniDeID applies an ensemble method to combine a selection of deep and shallow machine learning with rule-based algorithms to obtain as high sensitivity as possible while conserving good positive predictive value when detecting personally identifiable information (PII) in unstructured clinical text notes. It is built on the Apache UIMA framework for enterprise-grade robustness and implemented in two versions: a standalone software application for on-premises uses and a service-based version for secure cloud computing. Input and output can be both text files or a relational database

[1] Corresponding Author: Stéphane Meystre, email: stephane.meystre@imec.nl.

(PostgreSQL with OMOP CDM for structured data to de-identify). Both use the same NLP and machine learning pipeline, starting with text pre-processing and a regular expressions annotator (e.g., phone numbers, emails, SSNs). A dictionary lookup component then follows (e.g., first and last names, cities, countries). Feature generation then produces the input for several machine learning models run in parallel: recurrent neural networks (Bi-LSTM), conditional random fields, support vector machines and a margin infused relaxed algorithm (MIRA). All are combined in an ensemble method, along with the dictionary lookup and regular expressions output. The final PII annotations can be replaced with tags or resynthesized PII and exported to text files or a relational database. A user-friendly user interface controls the application and its parameters. Various levels of de-identification can be selected (including a customized selection of PII) combining up to 21 categories of PII (e.g., patient name, provider name, relative name, street address, country, ZIP code, date, day of week, time of day). A random selection of notes for either training or testing was created from a combined corpus of 3,943 manually annotated clinical notes (2006, 2014 and 2016 i2b2 and n2c2 challenges and MUSC corpus of 750 notes). Evaluation metrics were computed using the ETUDE open-source tool (scripts at https://github.com/musc-tbic/article-addenda).

3. Results

The version of CliniDeID trained with the training sets from the combined corpus reached high recall and precision (Table 1). The on-premises standalone version has been released as free and open-source software (https://github.com/Clinacuity/CliniDeID).

Table 1. CliniDeID accuracy (test split only) and speed (train and test split) results.

Test corpus	Recall	Precision	F_1-measure	Speed (s/note)
i2b2 2006	0.908	0.815	0.859	0.7562
i2b2 2014	0.987	0.993	0.990	0.8099
CEGS N-GRID 2016	0.981	0.982	0.981	2.1355
MUSC	0.822	0.913	0.865	0.8702

4. Conclusions

CliniDeID enables high accuracy clinical data de-identification and even the highest recall and precision among several text de-identification tools recently compared.[1]

Acknowledgments

Funding provided by the U.S. NIGMS (R41GM116479 and R42GM116479). We thank Youngjun Kim and all software developers and engineers at Clinacuity and the Medical University of South Carolina who made the application development and release possible.

References

[1] Heider PM, Obeid JS, Meystre SM. A Comparative Analysis of Speed and Accuracy for Three Off-the-Shelf De-Identification Tools. AMIA Jt Summits Transl Sci Proc. 2020 May;2020:241-50.