Classifying Cyber-Risky Clinical Notes by Employing Natural Language Processing

Suzanna Schmeelk[1], Martins Samuel Dogo[2], Yifan Peng[3], and Braja Gopal Patra[3]
[1]St. John's University, Queens, New York
[2]Queen's University Belfast, United Kingdom
[3]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA
schmeels@stjohns.edu, mdogo01@qub.ac.uk, yip4002@med.cornell.edu, bgp4001@med.cornell.edu

Abstract

Clinical notes, which can be embedded into electronic medical records, document patient care delivery and summarize interactions between healthcare providers and patients. These clinical notes directly inform patient care and can also indirectly inform research and quality/safety metrics, among other indirect metrics. Recently, some states within the United States of America require patients to have open access to their clinical notes to improve the exchange of patient information for patient care. Thus, developing methods to assess the cyber risks of clinical notes before sharing and exchanging data is critical. While existing natural language processing techniques are geared to de-identify clinical notes, to the best of our knowledge, few have focused on classifying sensitive-information risk, which is a fundamental step toward developing effective, widespread protection of patient health information. To bridge this gap, this research investigates methods for identifying security/privacy risks within clinical notes. The classification either can be used upstream to identify areas within notes that likely contain sensitive information or downstream to improve the identification of clinical notes that have not been entirely de-identified. We develop several models using unigram and word2vec features with different classifiers to categorize sentence risk. Experiments on i2b2 de-identification dataset show that the SVM classifier using word2vec features obtained a maximum F1-score of 0.792. Future research involves articulation and differentiation of risk in terms of different global regulatory requirements.

1. Introduction

Medical data can be misused in many ways, from business profiteering and mandatory governmental controls to an individual's identity theft. Due to the enormous impacts of identity theft and other unauthorized usages of medical data, governments are gradually developing privacy and security regulations. Examples of such regulations include the United States' Health Insurance Portability and Accountability Act (HIPAA) of 1996 [1], The Health Information Technology for Economic and Clinical Health Act (HITECH) of 2009 [2], the European Union's General Data Protection Regulation (GDPR) [3], and the Chinese Cybersecurity Law [4]. In addition to regulations protecting the privacy and security of clinical data, states are conforming to the United State's 21st Century Cures Act [5], requiring healthcare providers to provide clinical data information. As healthcare data becomes more freely available, data providers need to identify security and privacy risks within shared Electronic Health Records (EHR) and to lower privacy and security risks to patients and healthcare entities.

We classify clinical note sentences in terms of risk of containing patient health information (PHI). We solve a novel problem as prior to our contribution, researchers reported solely on methodologies for clinical note de-identification. We discuss de-identification in the literature review as we could not find any clinical note PHI risk classification literature. Our classification methodology is closely related to de-identification but unique in that classification can be employed for other purposes. For example, classification could transpire upstream or downstream from de-identification to classify risk type pre- or post- de-identification. If we implement risk classification pre- or upstream from de-identification, then we could perhaps classify different types of risks to be passed to different de-identification methodologies. If we implement risk classification post- or downstream from de-identification, then we could perhaps catch sentences that were not properly de-identified. In addition, we predict that classification may simplify the computing resources needed to keep pace with the large quantity of data being produced daily by healthcare facilities as sharing clinical notes directly with patients and for research, while maintaining privacy and security,

URI: https://hdl.handle.net/10125/79841
978-0-9981331-5-7

is an international challenge. Recently, for example, *OpenNotes* became an "international movement to create partnerships toward better health and health care by giving everyone on the medical team, including the patient, access to the same information [6]."

We employ natural language processing (NLP) to identify the risk of containing protected PHI in clinical notes by classifying sentences in terms of risk based on the presence of sensitive information such as PHI. We use the Harvard Clinical NLP dataset for our analysis. The datasets were initially created at a former National Institute of Health (NIH)-funded National Center for Biomedical Computing, known as Informatics for Integrating Biology and the Bedside (i2b2) [7, 8]. We present current literature, data, and our novel risk-classification based methodology. Finally, we present the results where the top-performing system obtained an F1-score of 0.792 for correctly identifying the risk using word2vec and SVM classifier.

The rest of the paper is organized as follows. We first present related work in Section 2. Then, we describe the data and methods in Section 3, followed by our experimental setup, results, and discussion in Section 4. We conclude with future work in the last section.

2. Related Work

NLP for patient information de-identification and risk mitigation has been developed extensively over the past decade. However, direct research for clinical note classification as introduced by our research does not exist. Therefore, to provide a state-of-the art literature review we built it on entirely different but closely related privacy and security challenges. These challenges revolve around three other pillars of literature related to de-identifying medical notes. First, there has been a need to develop accessible medical data benchmarks to evaluate de-identification methods. Second, techniques have been produced for de-identifying English language medical records using NLP. Third, the industry has expanded de-identification English-based NLP techniques to de-identify non-English medical records. Finally, to reiterate, these three pillars of de-identification research literature solve entirely different privacy and security challenges than our particular research contribution.

Gold Standard PHI Benchmarks: Building gold standard PHI de-identification benchmarks for NLP tasks has been an ongoing effort over the past two decades. In the early 2000s, most corpora developed to measure the performance of de-identification systems were either not publicly available for privacy reasons, or were incomplete as they were either synthetically generated or composed of only a few document types (e.g., discharge summaries, pathology reports, nursing progress notes, outpatient follow-up notes, and medical message boards). Mayer et al. [9] introduced a human/manual inductive creation of an annotation schema and subsequent reference standard for de-identification of United States Veterans Affairs (VA) electronic medical records. The benchmark was created for private use only within the national VA network. The researchers manually marked PHI as risky based on the type of identifier. Deleger et al. [10] introduced a public gold standard benchmark based on annotating different types of clinical narrative texts from the Cincinnati Children's Hospital Medical Center. Kumar et al. [8] introduced a new longitudinal corpus of clinical narratives based on shared task corpus from the 2014 Informatics for Integrating Biology and the Bedside (i2b2) foundation, and University of Texas Health Science Center (UTHealth) at Houston. This corpus consisted of discharge summaries and medical correspondences of 1.3K medical records for 296 patients. Further, Kumar et al. [8] developed a new annotated de-identification benchmark from the i2b2/UTHealth corpus. The developed de-identified datasets are limited, which affected the system performances [8, 9, 10].

English De-identification Methods: Next, we describe research on NLP systems developed on English PHI healthcare records for de-identification. In 2007, Uzuner et al. [11] facilitated an analysis of submissions into the i2b2 project which was organized to research automating PHI removal from medical discharge records. The tool evaluation consisted of precision, recall, and F-measures on the ground truth. The systems with the best performance scored above 98% in F-measure for all PHI categories. A few years later, Hanauer et al. [12] examined the MITRE Identification Scrubber Toolkit (MIST) for record de-identification. After manual training rounds on a clinical social work and history/physical medical corpus of 360 documents, the tool improved performance based on interactive rounds. After eight hours of annotation time (round 21), MIST achieved an F1-score of 0.95. Meystre et al. [13] extended the MIST analysis by evaluating five de-identification state-of-the-art systems: MIT, MIST (Mitre), HIDE, HMS, and MeDS on two corpora: the 2010 i2b2 NLP challenge corpus and a corpus of VA clinical notes. Overall, the authors found that the different tools had certain strengths/limitations. Liu et al. [14, 15] performed de-identification from unstructured clinical texts via recurrent neural networks (RNNs). Liu et al. [14] built and analyzed a long-short term memory (LSTM) and RNN model on 2010, 2012,

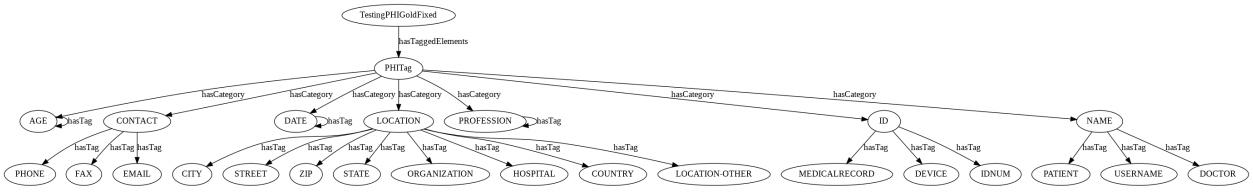

Figure 1. i2b2 PHI tag knowledge-graph

and 2014 i2b2 challenges. Liu et al. [15] extended the systems and analyzed the new system based on a corpus from the 2016 Centers of Excellence in Genomic Science (CEGS) Neuropsychiatric Genome-scale and RDOC Individualized Domains (N-GRID) clinical NLP challenge comprised of 1,000 (600 train and 400 test) annotated mental health records. The authors report competitive performance metrics, with limitations handling abbreviations, and specific PHI categories.

In general, the current research reflects on limitations. First, models may be built with only a small number of representative document types limiting analysis precision on document types not represented during training. Second, system evaluations trained from a certain institution document structure may not perform well on other document structures not-represented during training. Third, some parts of clinical notes, such as the medication section, may have higher error-rates than other sections of the same clinical notes. Fourth, reported models indicate limitations to how entity recognition was performed. Lastly, benchmark datasets used for evaluation can be further improved by evaluating multi-rater interrater reliability interclass correlation coefficients (ICCs) [16] for more than one annotator to improve tag accuracy.

Non-English De-identification Methods: Non-English NLP healthcare records PHI de-identification has emerged within published research during the last few years. Jian et al. [17] published research on developing a private intra-organizational Chinese de-identification benchmark, comprising 3K+ heterogeneous clinical documents, for unstructured narrative data. They reported the challenges and limitations of such English language-based techniques: (1) PHI is sparse in Chinese medical records, and (2) word segmentation and word morphological features are more difficult in Chinese than in other languages. For example, word capitalization is nonexistent in Chinese. Foufi et al. [18] introduced a de-identification of French unstructured clinical narrative data. Their techniques applied a Named Entity Recognition (NER) model to 11K+ French discharge summaries. They reported a limitation in the analyzed discharged summaries as they are often written in a hurry and contain as a consequence, spelling, orthographic and typographic errors, which can affect the de-identification process.

3. Data and Methods

NLP-centric risk identification appears to remain a needed, but unmet, endeavor. In this section, we explain the risk identification problem, the dataset used in our experiments, and describe the methodology that we used to identify risk in clinical notes.

3.1. Problem

The goal of this research is to classify clinical note sentences into two risk categories, high or low, based on whether or not each sentence is at risk of potentially leaking sensitive information about patients. The input is a sentence in a medical note. The output is one of the two classes of risks depending on the presence of sensitive words within the sentence.

3.2. i2b2 De-identification Data

In this project, we used the i2b2 de-identification dataset to evaluate the proposed models [8]. We used BeautifulSoup [19] to parse each clinical note into sections for analysis of the actual text and analysis of the gold-standard PHI tags.

Each gold-standard dataset file consists of an XML file with a section for the actual clinical note wrapped as a <TEXT><![CDATA[...]]></TEXT> and a separate section <TAGS></TAGS> listing the gold standard sensitive information locations within the note. For example, tags include gold standard identifiers and locations within the clinical notes for sensitive PHI information such as specifics related to HOSPITAL, DATE, DOCTOR, USERNAME, NAME, MEDICALRECORD, AGE, and IDNUM. Figure 1 shows our created knowledge-graph representation of the PHI categories and tags within the analyzed i2b2 dataset. This dataset has labeled PHI into eight main categories: Age, Contact, Date, Location, Profession, ID and Name. Figure 2 shows the layout of a gold-standard PHI labeled clinical note. All sensitive information has been removed in the figure. As shown

```
<?xml version="1.0" encoding="UTF-8" ?>
<deIdi2b2>
<TEXT><![CDATA[
...
CLINICAL NOTE
...
]]></TEXT>
<TAGS>
<DATE id="P0" start="16" end="26" text="xxx" TYPE="DATE" comment="" />
<ID id="P1" start="36" end="44" text="xxx" TYPE="MEDICALRECORD" comment="" />
<NAME id="P2" start="45" end="57" text="xxx" TYPE="PATIENT" comment="" />
<DATE id="P3" start="58" end="66" text="xxx" TYPE="DATE" comment="" />
<NAME id="P4" start="67" end="82" text="xxx" TYPE="DOCTOR" comment="" />
<AGE id="P5" start="379" end="381" text="xx" TYPE="AGE" comment="" />
<DATE id="P6" start="1092" end="1096" text="xxx" TYPE="DATE" comment="" />
<NAME id="P7" start="2346" end="2361" text="xxx" TYPE="DOCTOR" comment="" />
<DATE id="P8" start="2371" end="2379" text="xxx" TYPE="DATE" comment="" />
<DATE id="P9" start="2384" end="2392" text="xxx" TYPE="DATE" comment="" />
<NAME id="P10" start="2407" end="2422" text="xxx" TYPE="DOCTOR" comment="" />
<ID id="P11" start="2443" end="2460" text="xxx" TYPE="IDNUM" comment="" />
</TAGS>
</deIdi2b2>
```

Figure 2. A sample clinical note structure in the i2b2 de-identification dataset

Category	# of sent.
Low	11,750
High	10,791
Total	**22,541**

Table 1. Characteristics for the dataset

in the XML markups, each tag provides a location for the sensitive information within the clinical note.

During the data cleaning, we split the notes into sentences, using the Natural Language Toolkit (NLTK) [20], and rearranged tags to the sentences. Each sentence was then given a risk ranking based on the gold-standard PHI labels. In fact, sentences may contain more than one tag and could further be classified on a larger risk scale in future research, but in this project, we only focus on binary classes: low (0) or high (1). In total, 22,541 sentences were given risk scores from the tagged 514 clinical note dataset (Table 1). We used five folds to group the dataset into training and testing. The characteristics of the dataset in terms of the number of sentences with tags (high risk) and the number of sentences without tags (low risk) is seen in Table 1.

3.3. Features

Bag-of-Words: We used the NLTK [20] to transform clinical note sentences into unigrams. After the transformation, the number of times a unigram appears in each text is counted to form the bag-of-words (BOW) feature [21]. The number of features in this model is equal to the vocabulary size found by analyzing the data. We employed scikit-learn `CountVectorizer` [22] with the `binary` parameter set to true so that all non-zero counts are set to one. Setting the binary parameter as such is useful for discrete probability models, which model binary occurrences as opposed to integer counts.

Word-Embeddings: We also employed word2vec to transform words into vectors. We used the word2vec model pre-trained on the Google News dataset which gives a 300-dimensional vector. The word vectors were averaged to form the final sentence vector using the spaCy word2vec library [23].

Experimental Settings: We employed classifiers with different parameters. The experimental settings and results obtained are discussed in Section 4.

The risk classification was performed on the Harvard I2B2 N2C2 project dataset. Specifically, the 2014 De-identification & Heart Disease dataset. Within the 2014 i2b2/UTHealth corpus we examined the tagged dataset testing-PHI-Gold-fixed which is part of the deIdi2b2 subgroup. Table 2 shows the PHI labeled tag counts within the dataset that we calculated matching the tag counts reported by He et al. [24].

We employed five classifiers from scikit-learn [22]. In the first experiment, for each feature set, we employed the Bernoulli Naïve Bayes (NB) Classifier, which

Tag Name	Count
EMAIL	1
FAX	2
DEVICE	8
LOCATION-OTHER	13
ORGANIZATION	82
USERNAME	92
COUNTRY	117
STREET	136
ZIP	140
PROFESSION	179
STATE	190
IDNUM	195
PHONE	215
CITY	260
MEDICALRECORD	422
AGE	764
HOSPITAL	875
PATIENT	879
DOCTOR	1,912
DATE	4,980
Total	11,462

Table 2. Characteristics for PHI tags in the dataset

accepts binary values, and is used for discrete data with a Bernoulli distribution. Next, we employed the Gaussian NB Classifier, which is useful when working with continuous values with probabilities that can be modeled using a Gaussian distribution. Third, we employed the AdaBoost Classifier, which is a meta-estimator that adjusts outlying cases during classification such that subsequent classifications focus on difficult cases. Fourth, we employed the Random Forest Classifier, which is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting, with the default value of 100 trees. Fifth, we employed the C-Support Vector Machine (SVM) classifier, a statistical learning framework with an implementation based on *libsvm*. Lastly, we employed LinearSVM which is computationally more efficient than standard SVM as it uses a linear kernel.

For all the classifiers, the default scikit-learn parameters were employed. For the Gaussian NB Classifiers, the prior probabilities of the classes were not specified and a portion of the largest variance of all features was used at $1e{-}9$. For the Bernoulli analysis, we employed the default parameters to enable Laplace/Lidstone smoothing. We also set the threshold for binarizing to the default of none as it was presumed to already consist of binary vectors. Similarly, the AdaBoost four parameters were set to the defaults as well as Random Forest, LinearSVM, and SVM.

3.4. Evaluation Metrics

We evaluated the mode using precision, recall, and F1-score. Precision is defined to be the proportion of cases classified as positive that were true positives. In our case, the precision identifies the number of high risk sentences that were in fact classified as having high risk, in that they did contain sensitive patient information. Recall, or sensitivity, is defined to be the proportion of positive cases in the gold standard that are correctly classified as positive. In our case, high risk sentences that were identified as having high risk.

4. Results and Discussion

We examine performances of classifiers. We were able to achieve different results with different classifiers using Bag-of-words (BOW) features as summarized in Table 3, which show the mean of 5-fold cross-validation for the five classifiers. The LinearSVM classifiers obtained the best cross-validated F1-score of 0.767.

Next, classifiers using word2vec feature out performed previous models using BOW. The results in Table 3 show the mean of 5-fold cross-validation for the five classifiers. The SVM classifier obtained the best cross-validated F1-score of 0.792. We will further examine the measurements in Table 3 as interpreted in our case of classifying/identifying sentences within clinical notes that contain sensitive patient information.

Table 3 shows that different classifiers give different precision and recall results. In our case of identifying high-risk sentences, the recall measurement has stronger security ramifications than precision ramifications since recall measures the proper detection/classification of high risk sentences within the clinical notes. Recall takes into account the false negatives into the calculation. The F1-score compares precision measures with recall measures and in general higher F1-scores are considered stronger performance.

Figure 3 (a) BOW and 4 (b) word2vec show the confusion matrix for the top two performing classifiers as they had the overall best performance. The top-right corner is the measure of most importance for privacy and security as it indicates the number of sentences within the clinical notes which contained sensitive patient information that were labeled improperly as low risk. Figure 3 shows that the SVM classifier improperly labeled 551 sentences as low risk; and Figure 4 shows that the SVM classifier improperly labeled 572 sentences as low risk when in fact it contained sensitive patient information.

Model		Precision	Recall	F1-score
BOW	Bernoulli Naïve Bayes	**0.906**	0.498	0.650
	Gaussian Naïve Bayes	0.804	0.574	0.659
	AdaBoost	0.799	0.576	0.652
	RandomForest	0.867	0.678	0.757
	LinearSVM	0.827	**0.739**	**0.767**
	SVM	0.816	0.708	0.758
word2vec	Bernoulli Naïve Bayes	0.747	0.676	0.717
	Gaussian Naïve Bayes	0.695	**0.768**	0.742
	AdaBoost	0.762	0.724	0.752
	RandomForest	0.824	0.723	0.767
	LinearSVM	0.794	0.740	0.765
	SVM	**0.825**	0.758	**0.792**

Table 3. Model performances

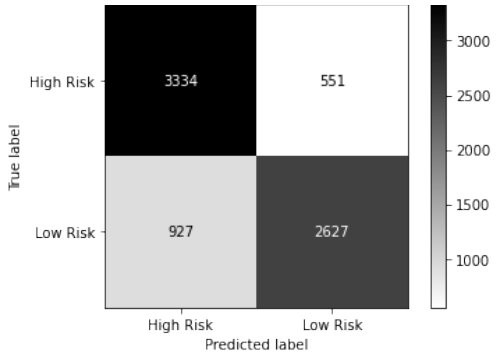

Figure 3. Confusion matrix for system using LinearSVM with BOW features

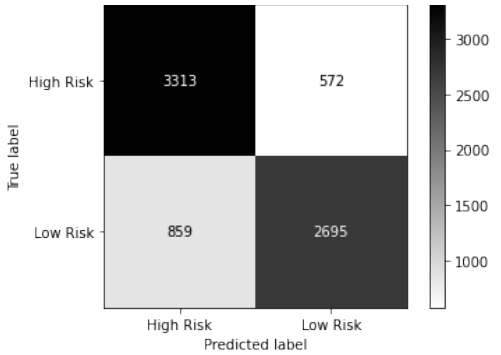

Figure 4. Confusion matrix for system using SVM with word2vec features

Error analysis: Further, we conducted error analysis to examine why our model may make errors. There are three steps in our model which may have errors. First, the data parsing of the clinical notes using NLTK saw the best performance over other parsing libraries, but this could still be improved for more properly parsing medical texts. Second, the feature extraction could also be improved. Perhaps, our feature extraction methods may not be adequately identifying short/long sentences within the clinical notes. Third, more advanced methods could be employed such as grid search to find optimal parameters for the classifiers used in this work. These directions of future work may improve our current state-of-the-art model.

5. Conclusion and Future Work

The identification of risk and the de-identification of medical records is now essential to uphold privacy and security regulations of patient health information employed for medical research and information sharing. Early literature in this domain consisted of developing industry benchmarks for the evaluations of de-identification techniques and methodologies starting with private datasets moving to public gold standard datasets. With the development of benchmarks, the healthcare industry was enabled to further openly analyze the performance of de-identification techniques employing NLP. As the research showed, de-identification success with English-based medical records. With the emergence of translational services such as Google Translate, the latest industry research has been in exploring NLP de-identification techniques in other languages such as Chinese, French, Serbian, German, Korean, Portuguese and Spanish. Privacy and security concerns will only mature with time.

In this research, we identified NLP as a useful methodology to identify sentence risk of containing sensitive information within clinical notes. Our developed systems were able to achieve the state-of-the-art of such a novel need. Future work involves improving the data cleaning, sentence

vectorization, classification, and the elaboration of sentence risk levels. For risk levels, we could classify sensitive information on a risk scale where identifiers such as hospital name, doctor name, and perhaps age carry less patient identifying risk than a patient name, a patient medical record number, and other direct patient identifiers. As clinical notes need to be de-identified for data sharing and research, this work provides fundamental contributions for employing NLP to enable more robust healthcare.

Acknowledgment

This work is supported by the National Library of Medicine under Award No. 4R00LM013001.

References

[1] U. States, "The Health Insurance Portability and Accountability Act (HIPAA)," 2004.

[2] "Health information technology for economic and clinical health (hitech) act, title xiii of division a and title iv of division b of the american recovery and reinvestment act of 2009 (arra), pub. l. no. 111-5, 123 stat. 226," 2009.

[3] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, p. 3152676, 2017.

[4] H. Zhang, Z. Tang, and K. Jayakar, "A socio-technical analysis of china's cybersecurity policy: Towards delivering trusted e-government services," *Telecommunications Policy*, vol. 42, no. 5, pp. 409–420, 2018.

[5] "Health and human services (2020) hhs extends compliance dates for information blocking health it certification requirements in 21st century cures act."

[6] "Launching opennotes in new york state." https://nyshealthfoundation.org. Accessed: 2021-08-23.

[7] A. Stubbs and Ö. Uzuner, "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus," *Journal of biomedical informatics*, vol. 58, pp. S20–S29, 2015.

[8] V. Kumar, A. Stubbs, S. Shaw, and Ö. Uzuner, "Creation of a new longitudinal corpus of clinical narratives," *Journal of biomedical informatics*, vol. 58, pp. S6–S10, 2015.

[9] J. Mayer, S. Shen, B. R. South, S. Meystre, F. J. Friedlin, W. R. Ray, and M. Samore, "Inductive creation of an annotation schema and a reference standard for de-identification of va electronic clinical notes," in *AMIA Annual Symposium Proceedings*, vol. 2009, p. 416, American Medical Informatics Association, 2009.

[10] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, *et al.*, "Building gold standard corpora for medical natural language processing tasks," in *AMIA Annual Symposium Proceedings*, vol. 2012, p. 144, American Medical Informatics Association, 2012.

[11] Ö. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association*, vol. 14, no. 5, pp. 550–563, 2007.

[12] D. Hanauer, J. Aberdeen, S. Bayer, B. Wellner, C. Clark, K. Zheng, and L. Hirschman, "Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs," *International journal of medical informatics*, vol. 82, no. 9, pp. 821–831, 2013.

[13] S. M. Meystre, Ó. Ferrández, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Text de-identification for privacy protection: a study of its impact on clinical text information content," *Journal of biomedical informatics*, vol. 50, pp. 142–150, 2014.

[14] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, and H. Xu, "Entity recognition from clinical texts via recurrent neural network," *BMC medical informatics and decision making*, vol. 17, no. 2, pp. 53–61, 2017.

[15] Z. Liu, B. Tang, X. Wang, and Q. Chen, "De-identification of clinical notes via recurrent neural network and conditional random field," *Journal of biomedical informatics*, vol. 75, pp. S34–S42, 2017.

[16] P. Shrout and J. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, pp. 420–428, Mar. 1979.

[17] Z. Jian, X. Guo, S. Liu, H. Ma, S. Zhang, R. Zhang, and J. Lei, "A cascaded approach for chinese clinical text de-identification with less annotation effort," *Journal of biomedical informatics*, vol. 73, pp. 76–83, 2017.

[18] V. Foufi, C. Gaudet-Blavignac, R. Chevrier, and C. Lovis, "De-identification of medical narrative data," *Stud Health Technol Inform*, vol. 244, pp. 23–27, 2017.

[19] L. Richardson, "Beautiful soup documentation," 2007.

[20] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[21] D. Jurafsky and J. Martin, *Speech and Language Processing, 2nd Edition*. Prentice Hall, 2009.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[23] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol. 7, no. 1, pp. 411–420, 2017.

[24] B. He, Y. Guan, J. Cheng, K. Cen, and W. Hua, "CRFs based de-identification of medical records," *J Biomed Inform*, vol. 58 Suppl, pp. S39–S46, Dec 2015.