Secure Information Exchange of Patient's Health Records Using Anonymization Techniques

Saman Hina[†], Hafiz Muhammad Anas A. Wahab[††], Raheela Asif[†††], Sheikh Muhammad Uzair[††††], Waqar Mansoor[†††††], Muhammad Sufyian[†††††††] and Farrukh Ahmed[†††††††]

[†, ††, †††, ††††, †††††, ††††††, †††††††] Department of Computer Science and Software Engineering, NED University of Engineering and Technology, Karachi, 75270 Pakistan

Summary
In developing countries, data used for research is not considered as an option of risk of leaking important personal information. This act should be considered as highly unethical if conducted in the research domain. This paper focuses on securing patient's health records before using it for research purposes. The presented system is developed for the identification and classification of potential information that can identify any individual. After the identification of information and their classification with respective categories, this system developed using an open sources platform, is able to de-identify each single entity by replacing it with their respective categories. The presented system has been successfully validated by domain experts and tested on different test cases

Keywords:
De-identification; patient records; NLP

1. Introduction

The reality of data science and research is proliferating and sprawling over many domains whether it's medical, sports or educational field, it is creating great impact. Data is regarded as a kernel in every domain and plays a pivotal role as everything today is escalating towards automation and supposed to be handled by software. This expeditious growth needs tremendous amount of research in data field. The techniques and methods used in data research whether it is machine learning or NLP (Natural Language Processing) Algorithms usually requires plethora of research related data, this excessive need improves the efficacy of the software by increasing the accuracy of its results. This research data must be acceptable and provided by an authentic source because data is supposed to provide the connotation of actual problem and it is very crucial in order to achieve best results. Fake data can be helpful but eventually will fail to deliver the intended results.

It is also a fact that data cannot be provided impromptu to the researchers as it may contains personal information about the stakeholders and has a potential risk of insecurity of data which can be exploited in several ways. Business companies such as Stock Exchange Company or bank inherits datasets in form of statistical surveys, information of individual's bank balance, their shares and properties. These datasets can't be delivered freely for research as it incorporates sensitive information.

In medical domain, where doctor patient secrecy has a paramount importance, the data of patient should not be spread impulsively for research as it will be devastating both for doctor and patient.

Similarly, educational data of students will be detrimental if delivered without proper arrangements; it will not only expose the educational backgrounds of the students but can also bring negative impact on their seminaries.

From the above discourse it can be culminated that compromise on data security is not any option for using data in research domain. Before spreading data for research it must be ensured that the information pertinent to the stakeholders and has high risk of exposure must be ousted, it is the only way to shrug off this problem. But it will be a painstaking task to discern that information from research document, we can't simply omit that because it has a risk to remove the whole document; it is required to forge it with false information but this falsification must be done subtly and should not affect the integrity and actual image of the document. It is utterly impossible to process each document manually and also it is not a pragmatic approach. It is imperative to devise a method or system that has the potential to do this task adequately and more accurately.

Terms like Data de-Identification or anonymization are used in this mainstream; data de-identification is a process of removing information from datasets that can be perilous for identification of an individual, de-identification and anonymization often use interchangeably.

The Notion of data de-identification is not maiden, in 1996 United States of America (USA) enacted a legislation known as HIPAA (Health Insurance Portability and Accountability Act) [1], it entails all necessary procedures to safeguard medical related information among which data privacy and data security rules has sheer importance, these rules incorporates national standards for protection and security of patient related information, moreover it regulates the usage of PHI (Protected Health Information).The definition of PHI as cited by HIPAA:[2] "Public Health Information (PHI) entails information of an

Manuscript received May 5, 2017
Manuscript revised May 20, 2017

individual, manifests its present, past and future physical and mental health condition, it is collected and managed by healthcare provider, that information has the efficacy to either identify an individual directly or can cater plausible basis to believe that it can be utilized to trail its source."
HIPAA suggested 18 identifiers that can be a source to identify an individual and supposed to be falsified as shown in Table 1

Table 1: HIPAA Identifiers

S.no	Identifiers
1	All types of person names
2	Includes names of state, city, county, street addresses, precinct, zip code, geocodes etc
3	Dates regarding person like date of birth, date of death, date of admission and date of discharge, all ages above 89 or older by aggregation.
4	All types of contact numbers.
5	Fax numbers
6	Emails (Electronic Mails)
7	Social Security numbers
8	Medical record numbers
9	Health plan beneficiary numbers
10	Bank account numbers
11	Certificate/license numbers
12	Vehicle identifiers include serial numbers and license plate numbers
13	Electronic devices serial numbers
14	URI and URLs
15	IP Addresses
16	Biometrics e.g. thumb print.
17	Photographs, Images
18	Distinctive identifier, characteristic or cipher that has high risk of disclosure

Authors participated in i2b2 NGRID shared NLP challenge 2016 for research purpose of medical datasets. The datasets provided by i2b2 were fully secured and completely de-identified.

1.1 Literature

*A. **Similar Work***

Researchers used B-o-B(best-of-breed) clinical text de-identification system for de-identifying VHA (Veterans Health Administration) documents , this system takes benefit of rule based module and CRF(conditional random field) module after passing through NLP preprocessing techniques i.e. segmentation, tokenization etc. It sought out any PHI (protected health information) present in the document and makes it usable for research purpose[3].
Other researchers described the methodology the context of Heritage Health Prize (HHP). HHP is a data mining competition. The aim of the competition is to predict the number of patients that should be hospitalized in the subsequent year. Their objective was to de-identify to HHP dataset to satisfy the requirements of US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. They used competition dataset to diminish the risk of re-identification attacks by simulating re-identification attacks. They performed three specific attacks and calculated their re-identification probability[4].
Researchers have used Name Entity Recognition (NER) technique that is used to categorize information such as dates person's names, locations, and places in text data. However, there is no doubt the medical data contains more personal information than NER refers to.
NER usually works on linguistic grammar-based techniques and statistical models but these models requires an enormous amount of training data annotated manually. [5] [6]
After conducting a tremendous amount of effort in June 30, 2011 to find all available prevailing strategies of health record anonymization researchers pointed that many de-identification systems are facing problems in case of wrong spellings, typographical mistakes and in the names that create bewilderment with non-PHI entities, some systems only works with particular types of data therefore this realm needs improvement. They suggested statistical system over pattern based system having more advantages and requires less amount of development than the later system[7]. Table 2 shows tools and techniques used by other researchers for de- identification tasks[8].

Table 2: 12b2 tools

Tools	Rules & Features	Machine learning system and features	External resources
OpenNLP CRF++	Regular expressions for tokenization	CRF: lexical, syntactic, orthographic	-
MedEx	Regular expressions for categories such as PHONE,FAX,MEDICAL RECORD,EMAIL and IPADDR	CRFs: bag-of-words; part-of-speech (POS) tags; combinations of tokens and POS tags; sentence information; affixes, orthographic features; word shapes; section information; dictionary features	-
MIST Stanford NER	Regular expressions for categories such as PHONE,EMAIL,ZIP	MIST, Stanford NER: features not mentioned	Personal de-id corpus
Tree Tagger MEDINA toolkit	Rules to correct output of CRF	CRF: surface features, morpo-syntactic, semantic, distributional	-
Preprocessing: CTAKES and GATE	JAPE system: orthographic, pattern, contextual, entity	CFR: lexical, orthographic, semantic, positional	Dictionaries collected from Wikipedia, GATE, and de-id
Python packages Numpy and Scipy	-	Non-parametric Bayesian Hidden Markov Model: token, word token, number token	-
Pre-processing: CRF++	Yes, for categories such as FAX,EMAIL,DEVICE,BIOID	CRF: Word-token, context, orthographic, sentence-level, task-specific	Self-compiled dictionary

2. Method

The presented system is developed using GATE (General Architecture for Text Engineering) tool [9]. The dataset used in this system is collected from diverse sources but mainly from i2b2 NLP 2016 challenge. We have taken advantage of pattern based approach for text mining, it doesn't involve any statistical results besides it gives well annotated documents at the end. This approach have given benefit in minimizing conflicting errors in comparison with the baseline approach that uses gazetteer/dictionaries. This system aimed to process patient records present in the raw format and its performance may vary from machine to machine depends upon its processing power. The main objective is to target the PHI (Protected Health Information) shared by HIPAA (Health Insurance Portability and Accountability Act) that can be found in the patient's medical records. PHI roughly covers all the information that has a risk of disclosing patient's identity (For instance, name, location, age etc.). Stepwise implementation of this system is elaborated in the following sections:

B. Data Gathering

As mentioned in the previous section, data used in this system was mainly provided by i2b2 challenge 2016 in the form of xml files. At about 350 patient records were selected in the unstructured format consisting of patient's history and his prevailing condition and created from patient's psychiatric analysis thus replete by psychiatric terms and terminologies. Moreover, some data was collected by local doctors and hospitals. This data was in the both electronic and handwritten form , handwritten data was further scanned and passed by the OCR (Optical Character Recognition) software, some handwritten data typed manually that were not recognized by the OCR software.

C. Preparing Gazetteer

In this step we emphasized on constructing gazetteer, gazetteers are essentially dictionaries in first format used in GATE tool, for our system we sought out all gazetteers required for the annotation of all HIPAA identifiers that include Name, Profession, Address, Countries, areas, regions and all the other similar attributes which should be good enough to cover almost every related information. The very first problem that we face is collision between Name and Location category, as many location names are based on person name. For example Jason Hospital, Taylor street. This can be resolved in later steps by writing JAPE rules. We classify our gazetteer for each words in two types' i.e. major type and minor type so that these gazetteer can be differentiate easily when writing JAPE rules. Required identifiers are divided into categories, for example Name is divided into first name, last name, middle name. Location is divided into area, city, region and country. This categorization is done for better accuracy purpose. The major purpose for designing gazetteer is that each token in the document is matched with these gazetteer following JAPE rule to identify the context and highlight its correct category.

D. JAPE Rules

JAPE is an acronym of Java Annotation Patterns Engine, it is a rule based approach used in GATE.JAPE is deemed as a predetermined state transducer that handles clarifications that are based on some regular expressions. It includes pattern-matching, semantic findings, and several other actions on syntactic trees similar to those that are produced by natural language parsers.

The main focus behind implementing a rule-based approach was the unapproachability of a large set of annotated document which would be required for testing, training and machine learning methods. The general syntax of JAPE rule is

Rule: Rule Name {Pattern} → Rule {Action}

The left hand side rule must be a pattern which is meant to perform the right hand side action subject to match the rule. A combination of these JAPE rules generates a phase and a number of these phases are combine to form a grammar. Below is the example of JAPE rule used in author's system.

```
Phase:  Profession
Input: Lookup Token
Options: control = applet
Rule: Profession1
(
{Lookup.majorType == title}
(
 {Lookup.majorType == title}
)?
)
:profession1
-->
```

:profession1.PROFESSION = {rule = "Profession1"}.

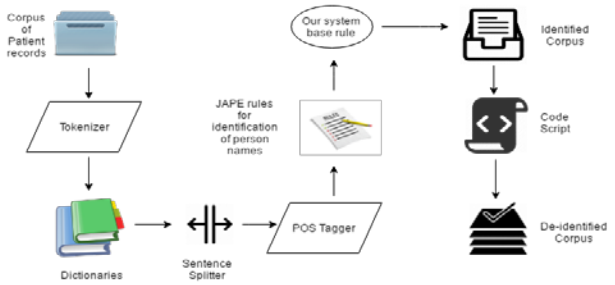


Figure 1: System Flow Diagram

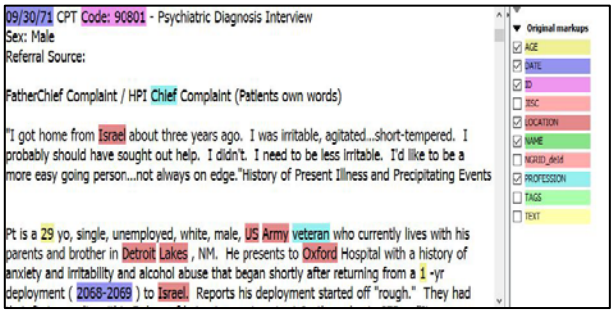
E. Executing GATE Application

For the Execution of De-identification application a conditional corpus pipeline needs to develop in GATE tool. The application pipeline of our De-identification application as shown in Figure 1 includes tokenization of the corpus, sentence splitting, dictionaries and gazetteers, corpus tagging with English grammar which is meant to be a part of speech tags and Java Annotation Pattern Engine – JAPE transducers for the development of annotation rules. GATE works in a series of steps as mention in fig b below that involves Document Reset, it resets the document by removing named annotation sets, English Tokenizer incorporates both normal tokenizer and JAPE transducer, Sentence Splitter breaks the text into sentences followed by period or any other symbol indicates the end of sentence like? etc, POS (Parts of Speech) Tagger works in tandem with English Tokenizer, Morphological Analyzer identifies lemma and affix, HIPAA(Health Insurance Portability and Accountability Act)[10] Gazetteer consist of all the required gazetteer of HIPAA identifiers to be use in the system. Flexible Gazetteer manages grammar of HIPAA gazetteer like singular and Plurals of gazetteer. Finally De-identification rules written in JAPE (Java Annotation Pattern Engine) format uses results generated in earlier steps and annotates the pertinent fields of HIPAA in the document.

3. RESULTS and discussion

After applying pipeline on patient's health records, potential personal health information was successfully identified as shown in Figure 2. This system provides specific details about each identified HIPPA entity by its exact position in the corpus as shown in Figure 3. Moreover, after the identification of potential HIPPA entities in the corpus, authors have written addition python script to replace these entities with their respective category (For instance, AGE, LOCATION, etc.). This retains the corpus in more understandable format and it does not lose the contextual information even after removal of identifiers. Significant validation from domain experts have been done and several test cases have been tested on this system.

Our system is not only able to de-identify these HIPPA entities but also is able to track original information by using start and end offset of each entity. This tracking is only provided to authentic domain users for better analysis of an individual's case.


09/30/71 CPT Code: 90801 - Psychiatric Diagnosis Interview
Sex: Male
Referral Source:

FatherChief Complaint / HPI Chief Complaint (Patients own words)

"I got home from Israel about three years ago. I was irritable, agitated...short-tempered. I probably should have sought out help. I didn't. I need to be less irritable. I'd like to be a more easy going person...not always on edge."History of Present Illness and Precipitating Events

Pt is a 29 yo, single, unemployed, white, male, US Army veteran who currently lives with his parents and brother in Detroit Lakes , NM. He presents to Oxford Hospital with a history of anxiety and irritability and alcohol abuse that began shortly after returning from a 1 -yr deployment (2068-2069) to Israel. Reports his deployment started off "rough." They had

Original markups
AGE
DATE
ID
LOCATION
NAME
PROFESSION
TAGS
TEXT

Figure 2: Output of Document no 0003

This is an important factor that an individual's personal information should be highly secure when it comes to sharing of data for research. The act of sharing data without anonymization/de-identification is considered as highly unethical and should not be ignored. The awareness of securing personal health information is lacking in developing countries and it is advised to establish "Ethical Review Committee" in each public/private sector conducting research so that data sharing can be done in ethical way.

Type	Set	Start	End	Id	Fixture
PROFESSION	Original markups	19	26	6039	(rule=Profession1)
DATE	Original markups	90	98	6053	(rule=FullDateNumbers)
ID	Original markups	103	114	6032	(CATEGORY=ID)
PROFESSION	Original markups	206	211	6040	(rule=Profession1)
LOCATION	Original markups	262	268	6072	(Rule=Place)
AGE	Original markups	544	546	6033	(CATEGORY=AGE)
LOCATION	Original markups	584	586	6073	(Rule=Place)
LOCATION	Original markups	587	591	6074	(Rule=Place)
PROFESSION	Original markups	592	599	6041	(rule=Profession1)

Figure 3: Identified HIPAA entities in Document no 0003

4. Comparision with gold standards

In comparison with human-annotated gold standard, the presented system achieved overall accuracy of 88.96%. Accuracy of individual annotation type is also recorded as shown in Table 3

Table 3: Accuracy of anonymization system against human annotated gold standard

Annotation	Total number of Annotations (Gold Standard)	Annotation Matches (Anonymization system)	Accuracy
AGE	74	68	91.89 %
CONTACT	18	12	66.66 %
DATE	475	420	88.42 %
LOCATION	132	124	93.93 %
NAME	224	202	90.17 %
PROFESSION	18	17	94.44 %
ID	47	36	76.59 %
Total	988	879	88.96 %

Acknowledgment

Authors would like to give credit to the following two grants which made the organization of the challenge possible:
NIH P50 MH106933 (PI: Isaac Kohane)
NIH 4R13LM011411 (PI: Ozlem Uzuner)

References

[1] Atchinson, Brian K.; Fox, Daniel M. (May–June 1997). "The Politics Of The Health Insurance Portability And Accountability Act"

[2] III, E. J., Hartley, C., & Roberts, W. (n.d.). HIPAA 'Protected Health Information': What Does PHI Include. Retrieved from HIPAA: https://www.hipaa.com/hipaa-protected-health-information-what-does-phi-include/

[3] Ferrández, O., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H., & Meystre, S. M. (2012, September 4). BoB, a best-of-breed automated text de-identification system for VHA clinical documents. Retrieved from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555325/

[4] Emam, K. E., Koru, G., Eze, B., Gaudette, L., Neri, E., Rose, S., . . . Gluck, J. (2012, February 27). De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset. Retrieved from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3374547/

[5] Lin, Dekang; Wu, Xiaoyun (2009). Phrase clustering for discriminative learning (PDF). Annual Meeting of the ACL and IJCNLP. pp. 1030–1038.

[6] Nothman, Joel; et al. (2013). "Learning multilingual named entity recognition from Wikipedia". Artificial Intelligence. 194: 151–175. doi:10.1016/j.artint.2012.03.006.

[7] Kushida, C. A., Nichols, D. A., Jadrnicek, R., Miller, R., Walsh, J. K., & Griffin, K. (2012, july). Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies. Retrieved from Medical Care: http://journals.lww.com/lww-medicalcare/Fulltext/2012/07001/Strategies_for_De_identification_and_Anonymization.17.aspx.

[8] Stubbsa, A., Kotfilab, C., & Uzunerb, Ö. (2015, March 10). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. Retrieved from Science Direct: http://www.sciencedirect.com/science/article/pii/S1532046415001173

[9] GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications", by Cunningham H., Maynard D., Bontcheva K. and Tablan V. (In proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002) https://gate.ac.uk/sale/acl02/acl-main.pdf

[10] Atchinson, Brian K.; Fox, Daniel M. (May–June 1997). "The Politics Of The Health Insurance Portability And Accountability Act"