Mehran University Research Journal of Engineering and Technology
Vol. 39, No. 3, 612 - 624, July 2020
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.2003.16

Anonymization Framework for Securing Protected Health Information in a Complex Dataset of Medical Narratives

Saman Hina[1], Raheela Asif[2], Syed Abbas Ali[3]

RECEIVED ON 04.07.2017, ACCEPTED ON 26.07.2019

ABSTRACT

It is imperative in a medical domain that protection of information does not allow an individual to be overlooked. In medical domain, research community encourages use of real-time datasets for research purposes. These real-time datasets contain structured and unstructured (natural language free text) information that can be useful to researchers in various disciplines including computational linguistics. On the other hand, these real-time datasets cannot be distributed without anonymization of Protected Health Information (PHI). The information of PHI (such as Name, age, address, etc.) that can identify an individual is unethical.

Therefore, we present a rule-based Natural Language Processing (NLP) anonymization system using a challenging corpus containing medical narratives and ICD-10 codes (medical codes). This anonymization module can be used for pre-processing the corpus containing identifiable information. The corpus used in this research contains '2534' PHIs in '1984' medical records in total. 15% of the labelled corpus was used for improvement of guidelines in the identification and classification of PHI groups and 85% was held for the evaluation. Our anonymization system follows two step process: (1) Identification and cataloging PHIs with four PHI categories ('Patients Name', 'Doctors Name', 'Other Name [Names other than patients and doctors]', 'Place Name'), (2) Anonymization of PHIs by replacing identified PHIs with their respective PHI categories.

Our method uses basic language processing, dictionaries, rules and heuristics to identify, classify and anonymize PHIs with PHI categories. We use standard metrics for evaluation and our system outperforms against human annotated gold standard with 100% of F-measure by increasing 39% from baseline results, which proves the reliability of data usage for research.

Keywords: **Security, Anonymization, Medical Narratives, Classification, Protected Health Information, De-Identification.**

1. INTRODUCTION

In medical domain, researchers are keen to use real-time data instead of fictional data. At the same time, distribution of real-time patient's data (such as "progress notes", "discharge summaries") is unethical without anonymizing personal information of patient. A few research documents and datasets have been shared for NLP research [1-5]. The people responsible for research distribute datasets after the

[1] Department of Computer Science and Information Technology, NED University of Engineering and Technology, Karachi, Pakistan. Email: samhaque@neduet.edu.pk, (Corresponding Author)

[2] Department of Software Engineering, NED University of Engineering and Technology, Karachi, Pakistan. Email: rahmed@neduet.edu.pk

[3] Department of Computer and Information Systems Engineering, NED University of Engineering and Technology, Karachi, Pakistan. Email: saaj@neduet.edu.pk

approval of a particular user agreement based on ethical requirements. These datasets are developed for specific research tasks and may not be reused for other research problems. For other research tasks, researchers face unavailability of real-time datasets or have option of developing their own dataset which is time-consuming.

The reason behind absence of actual information for research is the privacy concerns of an individual (such as patient, doctor and patient's relative). This actual information cannot be dispersed without pre-processing the corpus by anonymization of PHI. PHI is the information that can identify an individual. According to [6], the terms unidentified and anonymization can be used for each other but the term de-identification states to remove or hide identifiers (PHIs) from data while in anonymization, information is completely anonymous. This means that in case of de-identification, it is feasible to connect information with identification while anonymization process does not run any connection to information.

In this research, we have followed the anonymization for our purposes. This anonymization module was developed as part of 'e-Health gateway to the Clouds' project'(http://www.jisc.ac.uk/whatwedo/programmes/di_research/researchtools/ehealth.aspx) so that it can be safely used by researchers following best practice in ethics and governance. In this project, the anonymization of PHIs was carried out as a two step process on a novel corpus containing mixture of medical narratives and medical codes: (1) Identification of PHI categories, (2) Anonymization of PHI categories [7], as shown in Fig. 1.

Documentation of this research is categorically prepared to flow through sections; Section 2 presents work linked to automated systems characterized by anonymization/de-identification; Section 3 includes the comment of best quality level corpus for the advancement and assessment of this anonymization module. Section 4 contains method for the anonymization of PHI categories in medical narratives. Finally, Section 5 concerns itself assessing the key module identifying the anonymization system.

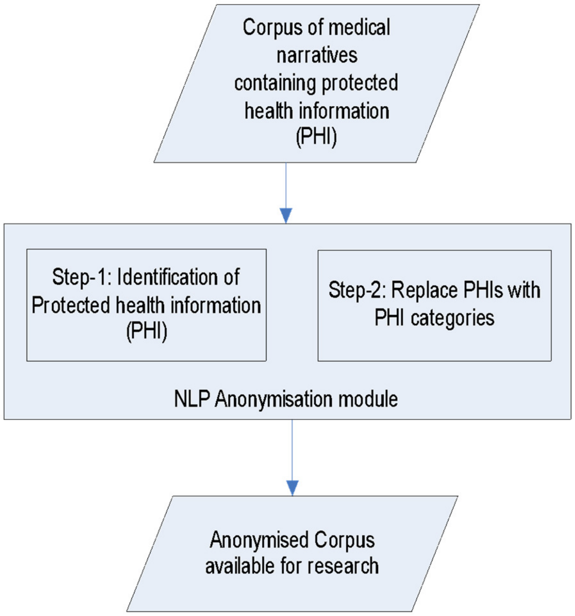

Corpus of medical narratives containing protected health information (PHI)
Step-1: Identification of Protected health information (PHI)
Step-2: Replace PHIs with PHI categories
NLP Anonymisation module
Anonymised Corpus available for research

FIG. 1. STEPS OF ANONYMIZATION

2. RELATED WORK

Security of data that distinguishes an individual ought to not be neglected in medical domain. In a published article [8], authors have addressed serious concerns of confidentiality of patient data even if the strong PHIs are already anonymized, other related information can reveal the identity of an individual. In USA, Health Insurance Portability and Accountability Act provides 18 HIPPA categories listed in Table1 for de-identification of medical information [9, 10]. This implies after the removal of these 18 PHI categories, the information can be considered as protected to utilize.

Uzuner, *et. al.* [5] reported a study of anonymization undertakings completed as a piece of global NLP challenge organized by i2b2 project organizers. In that paper, authors reported the de-identification challenge, procedure of explaining best quality level, evaluation metrics, summary of systems participated in the challenge and analysis of future directions in this area of research.

The challenge organizers organized documents by explaining PHIs and anonymized all PHIs by replacing them with reasonable surrogates. The highest quality level information was aggregated for the de-identification of following eight categories; "Patient, Doctors, Hospitals, IDs, Dates, Locations, Phone numbers and Ages". This highest quality level was first clarified by a programmed framework and after that approval was physically done by three annotators. After approval annotation PHI were replaced by real surrogates. Inter-annotator was not testified by authors.

TABLE 1: HIPPA SAFE HARBOUR CATEGORIES

1.	Names [immediate issue of names of drugs administered]
2.	Geographic locations smaller than a state/country, including zip codes or post codes
3.	All elements of dates except years relating to individuals e.g. admission/discharge dates. Also all elements of dates including year indicative of age > 89 [aggregation and age banding permitted as substitutions]
4.	Telephone numbers
5.	Fax numbers
6.	Email addresses
7.	Social security numbers
8.	Medical record numbers
9.	Health plan beneficiary numbers
10.	Account numbers
11.	Certificate and license numbers
12.	Vehicle identifiers e.g. Serial numbers and license plate numbers
13.	Device identifiers and serial numbers (not restricted to medical devices)
14.	University resources locators (URLs)
15.	Internet Protocol address (IP addresses)
16.	Biometric identifiers (including finger and voice prints)
17.	Full face photographs and comparable images
18.	Any other unique identifying number, characteristic or code.

A few analysts proposed techniques for utilizing dictionaries, NLP utilizing qualities and heuristics for the finish of anonymization in medical records [11]. In this research, they identified personal information but mainly focused in patient record on the anonymization of general practitioner records. The Norwegian corpus utilized in the research was tested in light of the fact that semantic features shift from English dialect and existing methodologies could not be utilized. In first step, they developed dictionaries utilizing their very own corpus and some external word references from different sources, for example, word references of medical names from International Collegiate Programming Contest (ICPC), natural names from National Map and postal administrations and Norwegian individual names and so forth. In second step, all word references were arranged in a solitary dictionary to perform correct matching of names. Also, suffix tree was utilized to enhance performance and the organized names were labeled with their particular kinds. Nonliterary sorts, for example, dates, telephone numbers, security numbers, and so forth were recognized utilizing suffix tree. At that point every labeled word having various sorts were explored and untagged words were physically audited by a nearby clinician for labeling. At long last, all labeled words were replaced by pseudonyms. Scientists have not talked about any part of approval or assessment of their work. Researchers also reported a rule-based de-identification program to anonymize patient's personal information [12]. The technique depended on the recognizable proof of patients by their surname, forenames and dates of birth. This methodology may fit to the organized patient's records in which each archive contains surname, forename and date of birth, however it will not be fit for unstructured reports that contain random pieces of information about personal data. The authors have revealed number of records in assessment and talked about the issues experienced amid the anonymization of information however this technique did not promise its general relevance on unstructured content. One particular and helpful usage in this exploration was the creation of key code for every patient with the goal that the patient's record can be reused in future.

In contrast with above mentioned systems, [13] exhibited a de-identification proof strategy which was introduced in first i2b2 worldwide NLP challenge on clinical information. They announced a novel iterative machine learning approach for named element acknowledgment Named Entity Recognition (NER) utilizing semi-organized archives. This technique first labeled all substances which were available is organized piece of report then this data was additionally used to discover different PHIs in unstructured piece of the content. To discover PHIs, the researchers utilized orthographical highlights, frequencies, PHI phrases, queries (dictionaries of areas names, infections, non-PHI tokens, and so on) for word-level arrangement. Using this feature set, combination of two classifiers (Boosting, C4.5) was trained in three phases and successfully achieved 99.7534% of f-measure on evaluation set.

In other research, researchers reported an anonymization system to de-identify name, address, phone number and date of birth. They noticed the problem of ambiguity in PHI identifiers [14]. For example, in some cases 'River' may be a person name but can be mistakenly identified as common noun instead of proper noun. Such issues do not generally contain signs yet ought to be expelled/replaced if present as PHI in the corpus. The corpus used in the research was a mix of German and English language and was split into two sets;

(1) 20% of the corpus was used to set up system.
(2) 80% of the corpus for evaluation.

Consequently, to handle these issues, a de-identification framework was created on in excess of 40 rules. This framework depends on;

(1) MEDTAG dictionary for lexical resources.
(2) Rule-based MS (Morphosyntactic) and WS (Word Sense) tagger for disambiguation task. 99% of success rate was reported.

In spite of the fact that, greater part of the work was done on organized data but [15] also created reasonable system named HIDE Health Information De-Identification Proof (HIDE) for de-identification proof of PHI data in both organized and unstructured information. They utilized Bayesian classifier, testing based systems and restrictive irregular fields based strategies for the extraction and identifiable proof of sensitive data. Their technique gave the advantage of data linkage by utilizing an identifier for an individual record and furthermore gave three adaptable alternatives to the de-identification proof: (1) full de-identification, incomplete de-identification and factual de-identification proof. Preliminary outcomes demonstrated generally speaking exactness of 75-98% for the de-identification proof of name (begin, intermediate), age, account number, restorative record number and date [16]. For identification of PHI in their corpora, these researchers exhibited a de-identifier named Stat De-id, based on Support Vector Machine (SVM) and local context. The methodology was fruitful in demonstrating that Stat De-id utilizing SVM and local context outperformed more than four frameworks: (1) SNoW, (2) IdentiFinder, (3) Dictionaries + Heuristics and (4) Conditional Irregular Fields (CRF). Detail De-identify following seven PHI classes in discharge outlines; Patients, Doctors, Hospitals, IDs, Dates, Locations, Phone numbers. This methodology used huge number of features (syntactic, syntactic bigrams and semantic highlights). Limitation of Stat De-id framework was accounted for [16]as far as nonattendance of local context in sentence. In the referenced i2b2 de-recognizable proof test, [17] learned nearby, worldwide and external highlights by utilizing conditional random fields-CRF. They utilized Beginning Inside Outside (BIO) labeling to distinguish pieces in tokens. External highlights utilized in this work included dictionary of individuals, location and dates; global highlights included sentential highlights to stamp sentences and tokens from the previous sentence.

An ongoing audit was finished by [6] on programmed de-distinguishing proof of frameworks created after 1995. This survey helped us in the completion of writing an audit on programmed de-identification frameworks/devices. As indicated by this survey, dominant part of work was done on organized data and not many specialists have concentrated on narratives. This review concludes that de-identification frameworks primarily contains regular PHI class of names yet in addition have other diverse PHI classifications. Having distinctive PHI classifications is one reason why one context cannot be contrasted with other de-identification frameworks. They examined 18 frameworks and we previously talked about some of them prior in this area [14, 17-19] and other 12 will be discussed in the following text. All of the 18 frameworks broke down in this survey de-identify general classifications of names, ages, dates, contact subtleties, emergency clinics and medical services providers, areas and identification. Regarding techniques, principally design coordinating calculations and machine learning was embraced, yet a few frameworks utilized the two methodologies. An open source resources was developed for the de-identification utilizing pathology reports [20]. The framework was named Healthcare Management System (HMS) scrubber and pursued three stages to finish the procedure of de-identification. In initial step, all pathology reports were assessed into Extensible Markup Language (XML) arrangement to isolate headers and content. This progression gives an organized look by isolating critical PHI into headers, for example, date of birth, medical record number, government managed savings number, promotion number and pathology division. At that point in second step, design coordinating was performed utilizing standard articulations to discover examples of date, phone number *etc*. In last step, string matching was done to identify and removed person names and location names. HMS scrubber accomplished 98% of review on 1800 reports. In another investigation, scientists have additionally utilized standards, query tables and customary articulations to de-identify PHIs in medicinal reports [21, 22].

Similar to the system developed by [20, 21], another system named MeDS was reported which utilized ordinary articulations, headers and word references (people, areas) [23]. They utilized around 50 normal articulations to recognize and dismiss incorrectly spelled names in the corpus. MeDS was assessed on two distinct informational data.

(1) 2400 reports (laboratory reports, narrative reports, mixed source reports).
(2) 1193 surgical reports.

On first data set, MeDS was able to de-identify 99.06% of Health Insurance Portability and Accountability Act (HIPPA) identifiers and 98.26% of non-HIPPA identifiers. On second data set, MeDS identified 99.47% of HIPPA identifiers and 96.23% of One more framework called idea coordinate scrubber was created by [24]. At first, archives were pre-handled by parsing content into words, sentences and stop words at that point utilized open source terminology Unified Medical Language System (UMLS) to coordinate and replace standard terms with their separate terms and codes. Then, all non-matching terms were replaced by a blocking tag. This framework may help researchers dealing with actual investigation however probably will not be useful for relevant check of documents. Authors of this work did not report any real standard assessment of precision and review, accepted to accomplish a high evaluation since archives just held identifier containing stop words.

The survey that additionally incorporated a standard based de-identification framework named Scrub was produced which utilized a few parallel detection algorithms and local word references to de-identify proper names (first names, last names, full names), addresses, states, countries and cities [25]. The two sets of investigation. In first investigation, a human was used to recognize specific identification data in letters composed by doctors. Second test was based on a computer methodology that used detection algorithm and knowledge sources. There was separate algorithm for every element (PHI) and the calculation announcing most astounding estimation of probability that was considered if there should be an occurrence of ambiguity. The Scrub framework effectively de-identified specifically recognizing data up to 99%-100% in correlation with straight pursuit (accomplished 32-37%) and straight inquiry with signals (32-84%).

In correlation with surely understood principle based and design coordinating frameworks, an alternate methodology was received by [26] who utilized general NLP framework, MedLEE to recognize and remove restorative ideas in reports. Because of this extraction, the corpus just contained therapeutic ideas without PHIs. The yield MedLEE was checked on by a doctor and just 3.2% of PHIs were identified in the corpus.

A method dependent on vocabulary of names and UMLS Metathesaurus was introduced by [27] which utilized increased inquiry and supplant calculation to distinguish legitimate names in the corpus. Their technique additionally included utilization of ordinary articulations to recognize prefixes and postfixes of names related –proper names. This framework was assessed against physically created best quality level of 1001 pathology reports and distinguished 98.7% of names in story area and 92.7% of names in entire corpus.

Other than utilization of dictionaries and pattern-matching utilizing normal expression, few analysts incorporated machine learning and factual methodologies for the undertaking of de-identification in medical records [16, 19, 27-31]. Among these referenced places [30] utilized factual demonstrating to distinguish names in patient reports while [31] utilized two toolboxs (Lingpipe and Carafe) for identification of named identity in the corpus.

Another group of researchers used "Support Vector Machine- SVM" for recognition of PHIs [28]. For pre-processing of their corpus, ANNIE-"A Nearly New Information Extraction" system was employed [32]. ANNIE was utilized to pre-comment on the preparation set with individual name, date, and so forth. At that point different features were utilized to prepare SVM learning classifier, for example, dates, specialist name and so on. This framework took an interest in first i2b2 NLP challenge and accomplished precision, review and f-measure more prominent than 86%. A system resembled to this, [29] additionally utilized SVM to build up a de-identification framework. SVM was utilized to perform NER in medical reports. This framework likewise took part in i2b2 test of de-identification and accomplished 92% (roughly) of f-measure. Their strategy originally utilized pattern matching to recognize subdivision headers, at that point normal expressions were utilized to distinguish dates and telephone numbers. A sentence classifier was likewise used to recognize PHIs in sentence. Lastly a SVM based content chunker was utilized to distinguish location, patient, age, and so on.

In an examination with related work done by different analysts, the curiosity of our anonymization module lies in handling the unique and identification issues on novel corpus containing blend of regular language and ICD-10 codes. The other factor is that 18 HIPPA classes were not straightforwardly relevant on this corpus; hence, altered PHI classifications were utilized in the advancement of anonymization module.

3. MATERIALS AND METHOD

The corpus used in this research was provided by 'Leeds Institute of Health Sciences'. This corpus was created as a result of lab session for medical students. A structure containing patient's health data subtleties was given to every person understudy and a recorded consultation video was shown to medical students. Medical scholars were approached to record this interview in an Electronic Restorative Record (EMR) framework as clinical codes (ICD-10 or READ codes). This data contained system generated fictional names of patients and was created as a practice exercise for coding clinical texts. A large portion of students utilized normal language rather than clinical codes to record their perceptions. Some of them utilized both clinical codes and normal language. This corpus was challenging for anonymization project because it contained free text and alphanumeric ICD-10 codes for medical concepts. The ICD-10 codes were written within free text consultation record.

After collection of this corpus, two annotators were employed to analyze corpus manually and produce labeled corpus for the development of anonymization module. The corpus contained following four labels;

(1) Patient Names
(2) Doctor Names
(3) Place Names
(4) Other Names (other than patient names and doctor names)

There were few names in the corpus which do not fit under classes of 'Patient Name' and 'Specialist Name'. Subsequently, all such names were included under the classification of 'Other Names'. All the names recognized by annotator were then manually mentioned on to deliver best quality level corpus utilizing an open source explanation instrument 'GATE-General Architecture for Text Engineering' [33]. 15% of the named corpus was utilized for the improvement of guidelines for the recognizable proof of PHI classifications and 85% was held for the assessment. Details of corpus measurements for development set and test set is given in Table 2. The next section explains the development of this anonymization module using medical narratives.

TABLE 2: CORPUS MEASUREMENTS

	Types of Corpus Measurements	Development Corpus	Test Corpus	Whole Corpus
Size of Corpus	Patient records	301	1683	1984
	Tokens	23031	167065	190098
	Sentences	1298	9889	11187
Size of Gold Standard Annotations	Patient Names	376	2117	2493
	Doctor Names	1	6	7
	Other Names	2	5	7
	Place Names	2	25	27

3.1 Anonymization of Protected Health Information in Patient Records

This anonymization module is produced for the anonymization of four protected health data (PHI) classifications in corpus containing medical narrative and ICD-10 codes. In the main phase of anonymization module, identification and arrangement of PHIs were required. Consequently, we originally explored existing named entity recognizer, ‘A Nearly New Information Extraction (ANNIE) gave in open source GATE tool [34] as baseline system; at that point based on confinements saw in baseline framework, we executed a standard based framework for the distinguishing proof and arrangement of four PHIs.

The baseline system reported 76% accuracy on all categories. After the implementation of baseline system, we tried to improve the accuracy by developing more rules to resolve following highlighted language issues observed in development set that were missed by baseline system;

(1) There were some **Asian names** and **Nick Names** in the corpus which were absent in the dictionary.
(2) A few names in the corpus were not written in proper **format**.

For instance, 'Davina TRN Smith' is a patient name which is not in the word reference yet 'Davina' and 'Smith' were in dictionary. For this situation, 'TRN' can be accepted as an **initial** yet was not reflecting starting of this name. Another model was of patient name 'Mrs. Parsons' which was missed by word reference application since it was written in the **small letter case**.

(3) **Coded data** in the corpus was picked as short form for location names and patient names. For instance, in Read code 'Xa0NZ', 'NZ' was chosen as short form for New Zealand which was in word reference.

(4) Clinicians can utilize distinctive conceivable organization of names in therapeutic accounts which cannot be put away in word references as appeared in figure. Additionally, proper names, for example, 'May', 'little', 'Short', 'Long', and so on can be utilized as **adjectives** in medical stories.

(5) The **medical terms** can be resolved as proper names in medical records for example, 'Ray' was recognized as appropriate name in 'X Ray'. Another model was 'TIA' which is a condensing of medical term ' transient ischemic attack' was resolved as legitimate name in the corpus.

Along with the above-mentioned problems, two examples of misspelled names were present in the corpus therefore; we did not develop any rule for spell-checking.

Similar to a standard framework, word references of locations, person_first (male) and person_first (female) names were utilized from ANNIE application in GATE. As referenced before, the corpus was a tab delimited record, along these lines all epithets and asian names were extracted by sending out corpus into spread sheet. All extracted names were then added to the word reference of names.

Dictionary of location names was utilized to identify 'Place Names' and the staying two word references of ‘names’ were incorporated in a solitary word reference of names. This single word reference was gathered in light of the fact that the classification of 'Patients Name' and 'Doctors Name' do not require difference of male and female names. Dictionary application was tested over a small sample of 300 patient records (15% of the corpus). It was observed that dictionaries predictably lost required information but also identified wrong information. Example is shown in Fig. 2.

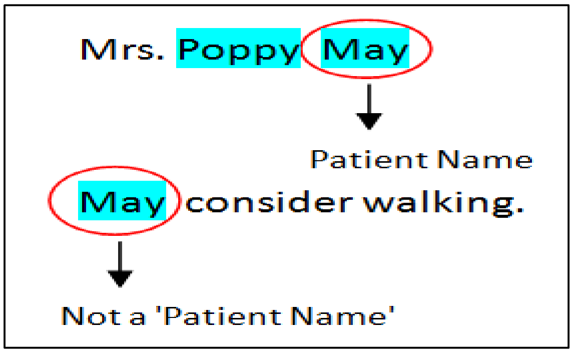


FIG. 2: OUTPUT OF DICTIONARY APPLICATION

Due to referenced issues, existing general named element acknowledgment frameworks are improper for medical stories. Along these lines, notwithstanding names in word references, a standard based anonymization module was created by investigating in the development corpus (clarified in next segment). The pre-preparing of this anonymization module incorporates essential language handling steps (tokenisation, split sentences, grammatical form labeling). In the wake of applying fundamental language preparing word references were included for the application pipeline look into coordinating names in the corpus. Finally, grammar rules were added to identify categories of 'Patient Name', 'Doctor Name', 'Other Names' and 'Place Names' using 'JAPE-Java Annotation Pattern Engine" [35].

The word reference named 'others' was added to differentiate names other than patient names and doctor names. This dictionary included jobs and occupations which do not speak to patient name or doctor name. For example, 'Nurse', 'Nurse practitioner', 'brother', 'partner', and so forth can speak to individual names in the corpus. The purpose behind isolating dictionary of 'others' was to recognize the standards for ID of names other than patient names and doctor names. Another word reference of 'Noplace' was accumulated containing terms which were wrongly recognized by dictionary application and for which general guidelines were not pertinent. Rules were developed to restrict false positives for names and places.

To develop useful patterns of proper names, first, a dictionary application was tested on the corpus which used dictionary to match names in the corpus. The perceptions demonstrated that numerous bogus positives were set apart by word reference application and the 'Full scale' rule settled the bogus positives. First all single word proper names of patients were separated by 'Full scale' rule by token coordinating in the dictionary. At that point false positives were limited utilizing designs that checked classification highlights of tokens. Examples of false positives are shown in Table 3 with the patterns (identified by POS tagger based on GATE's built-in regular expressions) developed for restrictions. Notwithstanding word reference coordinating, the right distinguishing proof of individual names required various examples. Consequently, a general 'Macro' rule was created to distinguish singular names independent of their significance with PHI classification, shown in Table 4.

This 'Macro' guideline will recognize single name (First name/Center name/Last name) in the corpus on the grounds that these single word names show up in regular language free content. This general 'Macro' guideline would then be able to be utilized being developed of examples for names under PHI classifications of 'Patient Name', 'Doctor Name' and 'Other names'.

In the development test of medical stories, it was seen that doctors utilized diverse arrangements to compose appropriate names. For example, some medics utilized first name or full name while others utilized initials of center name in full names. Thus, scope of organizations has been seen in the improvement corpus to create helpful tenets for the identification of patient names.

In the examination of corpus, a piece of information that was noticed is that the names of medical specialists were written with their occupations (Dr, GP, Doctor, etc.). As a result, a separate dictionary was created to store titles. The name of occupations were

TABLE 3: EXAMPLES OF FALSE POSITIVES IDENTIFIED BY DICTIONARY APPLICATION

Patterns	False Positives
Token.category!=VBP	Not drinking in 'am'.
Token.category!=VBN	She has 'read' over the info regarding smoking.
Token.category!=VB	...cut 'short' the consumption of alcohol.
Token.category!=RB	She feels a 'little' guilty for drinking.
Token.category!=MD	...that diabetes 'may' be related to cough.
Token.category!=JJ	'green' spit
Token.category!=lowercase	'little' or no exercise.

TABLE 4: MACRO RULE FOR THE IDENTIFICATION OF PROPER NAMES

Macro: Name
({Lookup.majorType == Names, Token.category! = VBP, Token.category! = VBN, Token.category! = VB, Token.category! = RB, Token.category! = MD, Token.category! = JJ, Token.orth! = lowercase})

used as a hint to recognize the names of doctor, practitioners in the corpus. Similarly, a 'Macro' rule for names was used to create guidelines for the identification of doctors.

The quantity of medical history used in the investigation also contains a few names other than patients and doctors. These names were identified with different roles and responsibilities, for example, nurse, partner, husband, and so forth. These other names can not be classified regarding their roles and responsibilities in light of the fact that the corpus don not contain enough models identified with explicit roles and responsibilities. All these examples needed to be identified and categorized as 'Other Names'. Consequently, rules were produced by logical examination for the distinguishing proof and anonymization of 'Other Names' in the corpus.

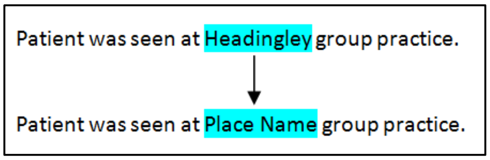
For example,

Seen by Nurse practitioner *Lara Jones* . // Other Names
Split up with partner, Mark.

In spite of the fact that, there were not very many instances of 'Other Names' found in the corpus however word reference of 'others' alongside standards will probably give solid insights for the recognizable proof of 'Other Names'. So also, there were few names of locations found in the corpus however leads were created based on these precedents.

As referenced before, the corpus contained few yet intriguing precedents place names showed up with general terms (general practice, emergency clinic, amass practice, and so forth). For example, ' Headingley gather practice' was a spot name in the corpus in which 'Headingley' is the identification of location. This implies some other city name related with general terms can decide somewhere else name, for example, 'Meanwood group practice', 'Sherburn bunch practice', 'Yaxley group practice', and so on. These general terms do not recognize personal health information data in restorative stories and ought to be kept in the records. The anonymization of location name can be considered as finished by the recognition and anonymization of 'Headingley' with its PHI classification. Along these lines, general terms will keep up meaningfulness of the content for investigation, as shown in Table 5.

TABLE 5: ANONYMIZATION OF PLACE NAME



Patient was seen at Headingley group practice.

Patient was seen at Place Name group practice.

In this way, the current word reference of location names was refreshed with all spots barring any broad terms (medical clinic, college, transport station, and so on). This refreshed lexicon additionally helped in the distinguishing proof of single word place names by applying string coordinating from the dictionary. Then again, it was seen that some of spot names were wrongly distinguished by word reference application, shown in Table 6 and therefore patterns were developed to restrict wrong place names.

TABLE 6: ISSUES IDENTIFIED IN THE IDENTIFICATION OF PLACE NAMES USING DICTIONARY APPLICATION

1	She split up with husband
2.	I saw a nice Xa6nO (Blonder hair)
3.	Let us know //Highlighted place name
4.	Saw Mrs Abby Hayward in clinic today
5.	Read code: XE1hO

For precedents 1, 2, 3 and similar cases, a standard included example was added to confine orthographic component of 'lowercase' and for cases, for example, models 4 and 5, an example was added to limit orthographic element of 'mixedcaps' appeared as follows.

```
Rule: PlaceName
({ lookup. majorType==place,        //Dictionary
containing place names.
   Token.orth!=lowercase,
   Token.orth!=mixedCaps})
```

4. EVALUATION

After the fruitful recognizable proof and arrangement of four PHI classes, the subsequent stage was the de-identification of names related with these named entities. After identification of required information, PHI classes along with respective information can either be deleted to finish the procedure of anonymization or can be replaced by non-identifiers ('ABC', 'XXX', and so on). In the present investigation, the yield of the identity in PHI classifications was first sent out in XML positions. At that point, a python program was composed to replace the identity PHIs with their separate PHI class to finish anonymization. The replacing of PHIs with their separate classification will assist specialists with understanding the setting of the corpus. For the assessment of identity and characterization of PHIs, standard data extraction measurements of accuracy, review and f-measure were utilized [36].

The assessment was done against human explained highest quality level (85% of entire corpus) and accomplished generally speaking f-proportion of 99%. Performance measurements for each individual PHI category on development and evaluation set are shown in Table 7-8.

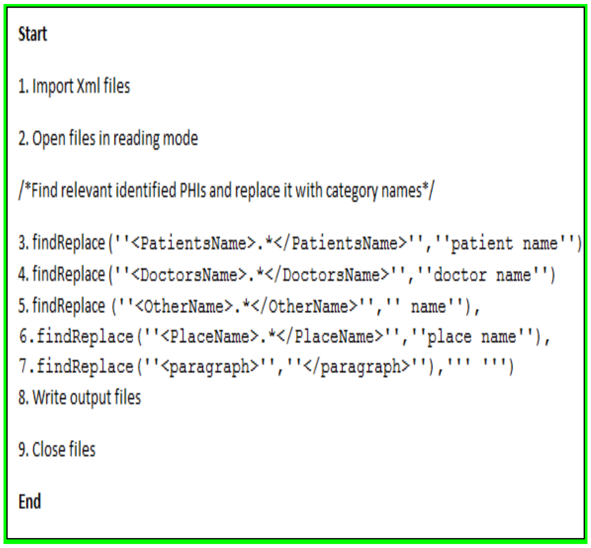
After identification of all entities under PHI categories, the output was exported in XML format and anonymization step was performed by using python script to replace identified entities with their respective PHI categories. The pseudocode of python script is provided in Fig. 3.

TABLE 7: IDENTIFICATION OF PHI CATEGORIES ON DEVELOPMENT SET

PHI Categories	Correct Matches	Partial Matches	Recall (%)	Precision (%)	F-Measure (%)
Patients Name	374	1	100	99	99
Doctors Name	1	0	100	100	100
Other Name	2	0	100	100	100
Place Name	2	0	100	100	100
Micro Average	379	1	100	99	99

TABLE 8: IDENTIFICATION OF PHI CATEGORIES ON EVALUATION SET

PHI Categories	Correct Matches	Partial Matches	Recall (%)	Precision (%)	F-Measure (%)
Patients Name	2109	7	100	99	100
Doctors Name	6	0	100	100	100
Other Name	4	0	80	80	80
Place Name	24	0	96	92	94
Micro Average	2143	7	100	99	100



```
Start

1. Import Xml files

2. Open files in reading mode

/*Find relevant identified PHIs and replace it with category names*/

3. findReplace(''<PatientsName>.*</PatientsName>'',''patient name'')
4. findReplace(''<DoctorsName>.*</DoctorsName>'',''doctor name'')
5. findReplace (''<OtherName>.*</OtherName>'','' name''),
6.findReplace(''<PlaceName>.*</PlaceName>'',''place name''),
7.findReplace(''<paragraph>'',''</paragraph>''),''' ''')
8. Write output files

9. Close files

End
```

FIG. 3. PSEUDOCODE FOR ANONYMISATION AFTER IDENTIFICATION OF PHI CATEGORIES

5. CONCLUSIONS

The anonymization module developed in this research performed 99% on challenging corpus comprises of natural language text amalgamated with ICD-10 codes. Although the developed module achieved impressive results, still false positives and negatives analyzed in the corpus needs to be addressed for improvements. Naturally, the capacity of this study was focused in a direction where concern to false positives laid outside the perimeter. Misspelled names was the outcome of a lack of employment of a spelling checker dictionary such as Ispell (http://www.gnu.org/software/ispell/ispell.html), to state one example. Moreover, the system missed 'Worsley building', being a local name of a building because dictionary of 'place names' do not contain information regarding local buildings. For this research, dictionary of 'place name' is limited to identify city and country names as it helped in the development of rules/patterns that could identify some government locations based on city names (such as 'Leeds General infirmary', 'Bradford General Infirmary', etc.).

Other limitations arose from discrepancy in following procedures: students fell short of maintaining the integrity of the documentation format. Such discrepancies resulted in place names being identified incorrectly. For instance, capitalizing 'Nice' in the sentence expressing quality 'Seemed Nice' resulted in the outcome 'Nice' being confused with a place name, likely in France. For such examples we were unable to find general rules which are the limitation of our system. We also left categories of date intentionally because their examples in the corpus did not provide any clue about an individual (such as patient, doctor, etc.). These dates were found with incomplete format such as 'Seen in 2001', 'Scan in Jan 2001' which did not identify any individual and left in the text. Moreover, we also think that using any clinical vocabulary such as SNOMED CT, UMLs can be used to restrict medical terms which were wrongly identified as names. After publication, this anonymization module will be contributed in GATE open source tool.

ACKNOWLEDGEMENT

Authors would like to thank NED University of Engineering & Technology, Karachi, Pakistan, for providing moral support for this research.

REFERENCES

[1] "International Challenge: Classifiying Clinical Free Text using Natural Language Processing," 30-06-2012, 2012; http://www.computationalmedicine.org/challenge/index.php.

[2] "NLP Research Data Sets," 12-04-2010, 2010; https://www.i2b2.org/NLP/DataSets/Main.php.

[3] Pestian J. P., Brew C., Matykiewicz P., Hovermale D. J., Johnson N., Cohen K. B., and Duch W., "A shared task involving multi-label classification of clinical free text," Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Prague, Czech Republic, 2007, pp. 97–104.

[4] Uzuner O., Goldstein I., Luo Y., and Kohane I., "Identifying patient smoking status from medical discharge records," Journal of the American Medical Informatics Association :

JAMIA, Vol. 15, No. 1, pp. 14-24, Jan-Feb, 2008.

[5] Uzuner O., Luo Y., and Szolovits P., "Evaluating the state-of-the-art in automatic de-identification," Journal of the American Medical Informatics Association, Vol. 14, No. 5, pp. 550-63, Sep-Oct, 2007.

[6] Meystre S. M., Friedlin F. J., South B. R., Shen S., and Samore M. H., "Automatic de-identification of textual documents in the electronic health record: a review of recent research," BMC Medical Research Methodology, Vol. 10, pp. 70, Aug 2, 2010.

[7] Hina S., Atwell E., Johnson O., and Brierley C., "Identification, Classification and Anonymisation of 'Protected Health Information' in real-time medical data for research purposes," in The 23rd Meeting of Computational Linguistics in the Netherlands (CLIN 2013), Netherlands, 2013.

[8] Affleck P., and Carrigan C., "Sharing patient data: understanding anonymisation," Bio Medical Journal, Vol. 362, pp. k2700, 2018.

[9] "Health Information Privacy," 20-04-11, 2011; https://www.hhs.gov/hipaa/index.html.

[10] Standards for privacy of individually identifiable health information: final rule, 67 Federal Register H. Office of the Civil Rights Standard 02-20554, 2002.

[11] Tveit A., Edsberg O., Røst T., Faxvaag A., Nytrø Ø., Nordgård T., Ranang M., and Grimsmo A., "Anonymization of General Practioner Medical Records," 01/01, 2004.

[12] Marciniak M., Mykowiecka A., and Rychlik P., "Medical Text Data Anonymization," Journal of Medical Informatics & Technologies, Vol. 16, pp. 83-88, 2010.

[13] Szarvas G., Farkas R., and Kocsor A., A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms, Berlin: Springer, Berlin, Heidelberg, 2006.

[14] Ruch P., Baud R. H., Rassinoux A. M., Bouillon P., and Robert G., "Medical document anonymization with a semantic lexicon," Proceedings AMIA Symposium, pp. 729-733, 2000.

[15] Gardner J., and Xiong L., "An integrated framework for de-identifying unstructured medical data," *Data &* Knowledge Engineering, Vol. 68, No. 12, pp. 1441-1451, 2009.

[16] Uzuner Ö., Sibanda T. C., Luo Y., and Szolovits P., "A de-identifier for medical discharge summaries," Artificial Intelligence in Medicine, Vol. 42, No. 1, pp. 13–35, 2008.

[17] Aramaki E., Imai T., Miyo K., and Ohe K., "Automatic Deidentification by using Sentence Features and Label Consistency," Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington DC, 2006.

[18] Gardner J., and Xiong L., "HIDE: An Integrated System for Health Information DE-identification," Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems, pp. 254-259, 2008.

[19] Szarvas G., Farkas R., and Busa-Fekete R., "State-of-the-art anonymization of medical records using an iterative machine learning framework," Journal of the American Medical Informatics Association : JAMIA, Vol. 14, No. 5, pp. 574-580, Sep-Oct, 2007.

[20] Beckwith B. A., Mahaadevan R., Balis U. J., and Kuo F., "Development and evaluation of an open source software tool for de-identification of pathology reports," BMC Medical Informatics and Decision Making, Vol. 6, pp. 12, Mar 6, 2006.

[21] Gupta D., Saul M., and Gilbertson J., "Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research," American Journal of Clinical Pathology, Vol. 121, No. 2, pp. 176-86, Feb, 2004.

[22] Neamatullah I., Douglass M. M., Lehman L.-w. H., Reisner A., Villarroel M., Long W. J., Szolovits P., Moody G. B., Mark R. G., and Clifford G. D., "Automated de-identification of free-text medical records," BMC Medical Informatics and Decision Making, Vol. 8, No. 1, pp. 32, 2008.

[23] Friedlin F. J., and McDonald C. J., "A Software Tool for Removing Patient Identifying Information from Clinical Documents," Journal of the American Medical Informatics Association, Vol. 15, No. 5, pp. 601-610, 2008.

[24] Berman J. J., "Concept-match medical data scrubbing. How pathology text can be used in research," Archives of Pathology & Laboratory Medicine, Vol. 127, No. 6, pp. 680-6, Jun, 2003.

[25] Sweeney L., "Replacing personally-identifying information in medical records, the Scrub system," Proceedings of the AMIA Fall Symposium, pp. 333-337, 1996.

[26] Morrison F. P., Li L., Lai A. M., and Hripcsak G., "Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-identify Clinical Notes?," Journal of the American Medical Informatics Association, Vol. 16, No. 1, pp. 37-39, 2009.

1. [27] Thomas S. M., Mamlin B., Schadow G., and McDonald C., "A successful technique for removing names in pathology reports using an augmented search and replace method," Proceedings AMIA Symposium, pp. 777-781, 2002.

[28] Guo Y., Gaizauskas R., Roberts I., and Demetriou G., "Identifying personal health information using support vector machines," Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.

[29] Hara K., "Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge," in i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2007.

[30] Taira R. K., Bui A. A., and Kangarloo H., "Identification of patient name references within medical documents using semantic selectional restrictions," Proceedings of the AMIA Symposium, pp. 757-61, 2002.

[31] Wellner B., Huyck M., Mardis S., Aberdeen J., Morgan A., Peshkin L., Yeh A., Hitzeman J., and Hirschman L., "Rapidly retargetable approaches to de-identification in medical records," Journal of the American Medical Informatics Association : JAMIA, Vol. 14, No. 5, pp. 564-573, Sep-Oct, 2007.

[32] Cunningham H., Maynard D., Bontcheva K., and Tablan V., "GATE: an architecture for development of robust HLT applications," Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002, pp. 168–175.

[33] Cunningham H., Maynard D., and Bontcheva K., Text Processing with GATE: Gateway Press CA, 2011.

[34] Cunningham H., "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Phildelphia, *PN*, 2002.

[35] Cunningham H., Maynard D., and Tablan V., "JAPE: a Java Annotation Patterns Engine," 12/01, 2000.

[36] Sokolova M., and Lapalme G., "A systematic analysis of performance measures for classification tasks," Information Processing & Management, Vol. 45, No. 4, pp. 427-437, 2009.