Assessing models for de-identification of Electronic Discharge Summary using Machine Learning tools

Tshilisanani Mudau (11640828)

Supervisor(s):
Professor Winston Garira
Dr Rendani Netshikweta

A Project report submitted in partial fulfillment of the requirements for the degree of Master of Science in the field of e-Science

in the

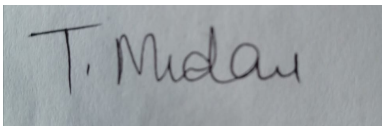
Department of Computer Science and Applied Mathematics
University of Venda

18 July 2024

Dedicated to my late mom.
How i wish you were here to witness God's grace upon your only daughter.

Declaration

I, Tshilisanani Mudau (11640828), declare that this report is my own, unaided work. It is being submitted for the degree of Master of Science in the field of e-Science at the University of Venda. It has not been submitted for any degree or examination at any other university.



Tshilisanani Mudau (11640828)

18 July 2024

Abstract

Background: De-identification is a technique that eliminates identifying information from Clinical Records in order to protect individual privacy. This procedure decreases the chance of personal information being collected, processed, distributed, and published from being used to identify the person. When Machine Learning techniques were included in the de-identification process, it substantially improved over the previous method.

Research Problem: The Electronic Discharge Summary(EDS) has evolved into a significantly improved technique of providing discharge summaries though this information contains Protected Health Information (PHI), which poses a risk to patients' privacy. This makes the process of de-identification to be mandatory. There have lately been several Machine Learning approaches to de-identify data. This study focuses on applying Machine Learning techniques to figure out which model can best de-identify a data set.

Methods: The open source data set from Harvard Medical School was used. This data set contains 899 Electronic Health Records (EHR), 669 for training and 220 for test purpose. The Conditional Random Fields (CRF), Long Short Term Memory (LSTM) and Random Forest models were used, and the performance of each model was assessed.

Findings: In order to assess each model's performance, evaluation metrics were used to compare F-measure, Recall and Precision at token level to determine which Machine Learning model performed best. The Long Short Term Memory was found to outperform both Conditional Random Fields and Random Forest with micro average F-measure, Recall and precision of 99%, and macro average F-measure of 77%, Recall of 73% and Precision of 90%.

Acknowledgements

Above all, I would like to thank the Almighty God for everything, his unconditional love, mercy and grace. If it wasn't of him, this would not have been possible.

I'd like to express my sincere gratitude to my supervisor, professor Winston Garira and my co-supervisor Dr Rendani Netshikweta for their assistance.

I would also like to thank everyone who assisted and supported me in any way.

My heartfelt gratitude goes out to my family for their unfailing support.

Funding was provided by the DSI-NICIS NEPTTP and is highly appreciated.

Contents

List of Figures

List of Tables

List of Abbreviations

PHI	**P**ersonal **H**ealth **I**nformation
EDS	**E**lectronic **D**ischarge **S**ummary
GP	**G**eneral **P**ractitioner
HIPPA	**H**ealth **I**nformation **P**ortability and **A**ccoutability **A**ct
NLP	**N**atural **L**anguage **P**rocessing
n2c2	**N**ational **N**atural Language Processing **C**linical **C**hallenges
AL	**A**rtificial **I**ntelligence
NLTK	**N**atural **L**anguage **T**ool **K**it
MIMIC	**M**edical **I**nformation **M**art and **I**ntensive **C**are
CRF	**C**onditional **R**andom **F**ields
i2b2	**I**nformatics for **I**ntegrating **B**iology and the **B**edside
SGD	**S**tochastic **G**radient **D**escent
LSTM	**L**ong **S**hort **T**erm **M**emory
NTCIR-10	**N**II **T**estbeds and **C**ommunity for **I**nformation access **R**esearch-10
MedNLP	**M**edical **N**atural **L**anguage **P**rocessing
SVM	**S**upport **V**ector **M**achine
MIRA	**M**edical **I**nstrument **R**esearch **A**ssociates
POS	**P**art **O**f **S**peech
BFGS	**B**royden **F**letcher **G**oldfarb **S**hanno
PRP	**P**ersonal **P**ronoun
CC	**C**oordinating **C**onjunction
NER	**N**amed **E**ntity **R**ecognition
PPV	**P**ositive **P**redictive **V**alue

Chapter 1

Introduction

De-identification is the process used to detect identifiers such as the names of Patients and date of birth that can be used to point an individual (or entity) in a direct or indirect way and then delete those identifiers from data [1]. The deleted identifiable fields can be replaced with the data describing the type of that identifiable field (e.g., “Tshilisanani Mudau" replaced by [person name]) or replaced with a fake identifiable field called pseudonyms [2] (e.g., “Tshilisanani Mudau" replaced by “Mudinda Munyai"). This process of removing the identifiable fields in the clinical data with non-identifiable fields eliminates the chance of re-identification of the patient during the secondary use of data. Protected Health Information is defined as health data containing personal identifiers linked to the data subject (PHI) [3]. Clinical Report generated by health care professionals summarizing information concerning a patient’s hospital stay and or a sequence of treatments received is known as an electronic discharge summary (EDS). Discharge medication mistakes caused by manual transcribing of medicine between discharge summaries and medication records can be reduced with electronic discharge summaries [4]. For General Practitioners (GP), the EDS increases the quality of discharge information [5].

Advice is provided by "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule" [6] on the procedures and techniques for achieving de-identification in conformity with the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), with two de-identification approaches briefly outlined. The purpose of this is to assist covered entities in

comprehending the procedures involved in de-identification, as well as how de-identified data is obtained and existing de-identification methods. For Health Records to be regarded as de-identified, eighteen PHI categories to be removed [7] are shown below.

TABLE 1.1: PHI Categories [6]

No.	PHI Category
1	Names
2	Every geographic subdivision which is smaller than a state
3	All elements of dates (except year)
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/license numbers
12	Vehicle identifiers and serial numbers, including license plate numbers
13	Device identifiers and serial numbers
14	Web Universal Resource Locators (URLs)
15	Internet Protocol (IP) addresses
16	Biometric identifiers, including finger and voice prints
17	Full-face photographs and any comparable image
18	Any other unique identifying number, characteristic, or code

When an EDS is considered de-identified, it must be free from HIPAA-mandated 18 PHIs [8]. After these PHI categories are removed from an EDS or data set, its use or exposure can no longer jeopardize an individual's privacy. However, it should be noted that while this procedure does not fully eliminate the possibility of re-identification of a data set, it does, however, provide data sets with a minimal chance of being re-identified [9].

1.1 Background

The process to de-identify Clinical Data has now shifted from being a manual task, which was prone to error, time consuming, a labour intensive task and an expensive method [10] to be a machine based process. Machine Learning-based algorithms are preferred for de-identification of clinical data, with Conditional Random Fields and Recurrent Neural Networks being most commonly used models. Patient records used in this research are in the form of unstructured data set. Though it is not easy to work with these narrative texts, they contain more information, unlike structured data set which is in the form of titled columns and rows, making it to be easily processed.

1.1.1 The purpose of De-identification

De-identification is essential because it allows agencies to access data sources and use information while maintaining an individual's privacy. Community expectations on how organizations manage personal data are also shaped by de-identification. De-identification is also useful for preventing the identity of an individual or a group of individuals from being exposed. The are several reasons behind the de-identification of personal information which include [11]:

1. Open Data
   When an organization does not have the authority to reveal personal information, de-identification may be implemented to permit data sharing. By proactively releasing data sets as well as to make this data publicly available for use and republishing, open data initiatives aim to promote government transparency and accountability. Given the growing amount and accessibility of information provided by these projects, it is critical that institutions release their data sets in a manner that protects individuals' privacy. Research, innovation, and the development of new applications and services are all objectives of open data initiatives. The greater the utility of open data sets, the more likely researchers, start-up companies, and entrepreneurs that want to utilize public data will succeed.

2. Access to information request
   De-identification may also be effective in responding to requests for data set access or information. Institutions can respond to requests in a privacy-protective manner while protecting the information's utility by applying de-identification. De-identification is a unique approach that could provide organizations with an opportunity to improve transparency.

3. Data sharing within and among institution
   While open data initiatives and access to information requests provide information to the public, there is a rising demand in government services for institutions to break down their "communication barriers" and exchange more information within and among their own organizations. While information sharing among institutions might help provide better, more efficient services, it may also have the unintended consequence of diminishing people's privacy by reducing their control over their personal data. Institutions should always consider de-identifying data sets before sharing them as a best practice.

4. Other reasons behind de-identification can include reducing the danger of a data breach and the harm it causes individuals, and to increase community trust in how agencies keep and handle data

1.1.2 Natural Language Processing (NLP)

Natural language processing (NLP) is specifically a field of computer science , a field of Artificial Intelligence (AI) concerning the ability of computers to interpret spoken words as well as text in a similar manner that humans can. It is used to enable programs on the computer to convert the text from one language to the other, respond to verbal commands and quickly sum up vast amounts of material [12]. This NLP is necessary since it assists in the resolution of linguistic ambiguity as well as providing data with a quantitative structure that is useful for a variety of downstream applications, like text analytics and speech recognition. Natural language processing encompasses a wide range of methods for analyzing human language, including statistical and machine learning techniques as well as rules-based and algorithmic approaches. This is due to the fact that text and voice-based data, as well as applications in practice require a diverse set of methods. Basic NLP tasks

include tokenization and parsing, lemmatization or stemming, part-of-speech tagging, language detection and semantic relation identification.

1.1.3 Privacy Requirement

There is a need for privacy. In this work, we looked at de-identification process, which is a way of recognizing (and eliminating) PHI instances from unstructured textual clinical records as defined by HIPAA [13]. These medical records contain sensitive information about patients that a physician, a nurse or a lab technician has recorded. Notably, because this is an unstructured text with no formal format, how the patient notes are written is entirely up to the medical practitioner. As a result, we followed the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule to recognize and sanitize sensitive information in predefined categories (like patient names, phone numbers, location and the dates).

1.1.4 Ethical Considerations

The Research data set used was open source data (n2c2 NLP Research Data Sets) from Harvard Medical School. The request to access Data Set has been approved upon signing the NLP Data Use Agreement. The Ethical Clearence certificate (ETHICAL CLEARENCE NO: FSEA/23/MCS/03/1707) was also obtained from University of Venda.

1.2 Problem Statement

Patients privacy has been the primary concern whenever dealing with individually identifiable health information. The guarantee for privacy to patience whom the health information belongs to remains one of the main challenges to disseminating this information. The EDS has evolved into a significantly improved technique of providing discharge summaries [14]; nonetheless, this information contains Protected Health Information (PHI), which poses a risk to patients' privacy [15]. De-identification is required to access any record with PHI categories, according to Daumke P, etc [16]. However, given the sensitivity of the data and the lack of

software to automate the process, this has been a manual and labor-intensive procedure [7]. However, there have lately been several Machine Learning approaches that can conduct de-identification, and it is difficult to determine which one works better than the others. The purpose of this research is to apply Machine Learning approaches to figure out which model can best de-identify a data set.

1.3 Research Question

This research focuses on answering the following questions:

1. How does Conditional Random Fields, Long Short Term Memory and Random Forest perform de-identification as compared to each other given the same National Natural language processing (NLP) Clinical Challenges (n2c2) data set?

2. Among Conditional Random Fields, Long Short Term Memory and Random Forest, Which Machine Learning model best perform de-identification on EDS?

1.4 Research Aims and Objectives

In this section, the aim of research is discussed. The objectives of the research are also listed in accordance to how the steps to reach the aim or the purpose of this research will be taken.

1.4.1 Research Aims

The aim is to use Conditional Random Fields, Long Short Term Memory and Random Forest models to de-identify EDS and compare their results using evaluation metrics (F-measure, Precision, and Recall) to see how each model performs on the National Natural Language Processing (NLP) Clinical Challenges (n2c2) data set.

1.4.2 Objectives

1. To review the existing Literature on EDS to understand how Machine Learning techniques are used in de-identification.
2. Evaluate how effectively each model performs the de-identification of EDS.
3. Compare the results obtained from each model.
4. Using the Precision, Recall and F-measure ratings, conclude which model is the best.

1.5 Limitations

Machine Learning tools to perform de-identification are already in existence [9, 17].However, the development of these Machine Learning approaches is constrained because the system must have access to clinical data in order to be developed properly. This data is difficult to obtain for secondary purposes such as research, even if it is for de-identification without being de-identified first. But, the Harvard Medical School is one of the open source platforms where researchers can sign the NLP Data Use Agreement and gain access to data sets. This data set only have 669 health records for training which might not be enough and could cause poor generalisability.

1.6 Overview

There are five chapters in this writing. Following an explanation of what is meant by de-identification and how EDS are being de-identified comes chapter two. Chapter two discusses the related work as well as motivating the choice of models used. In chapter three, the research talks about how personal data is dealt with here in South Africa. The research design, information about the data set used, methods and analysis are provided in chapter four. Chapter five contains the results and discussion. The research conclusion and discussion of future work are contained in Chapter six.

Chapter 2

Literature Review

2.1 Related Work

The use of EDS is increasing together with the need for this data to be used for secondary purposes like research [18, 19]. In order for this data to be accessible, patient confidentiality must be protected, which cannot be overstated or neglected [20]. To support the patients privacy, there is a lot of work that has been done under the topic of de-identification. Existing methods to de-identify Clinical Data vary from rule-based systems where the patterns and dictionaries of phrases containing PHI are matched to Machine Learning models [21]. As a result of hard coding rules to a particular format of notes or the writing style of the health practitioner, rule-based systems (Expert Determination technique) have occasionally had poor performance in other settings [22]. As a result, Machine Learning approach is used as an alternative and the rules to remove PHI categories are "learned" through training the algorithm.

Clinical Records were de-identified using Recurrent Neural Networks, Conditional Random Fields, bidirectional Long-Short Term Memory [23]. A combination of these three models was used. They were assessed using 2014 i2b2 de-identification data set [24] and Medical Information Mart for Intensive Care (MIMIC) de-identification data set (customized version) [25]. On the basis of a CRF model, a de-identifier was constructed which produced state-of-the-art results with the i2b2 2014 data set, the reference data set for assessing de-identification systems. The Artificial Neural Network model was tested against this system as a challenging baseline. The training set sizes were varied. The i2b2 data set was analysed for six PHI

categories while the MIMIC data set was analysed for five PHI categories. During training handwritten features [26] were taken into consideration with the CRF model. Each patient note was tokenized with the Stanford CoreNLP tokenizer, and features were retrieved for each token in the CRF model. The CRF's parameters were tuned throughout the training phase to maximize the likelihood of the gold standard labels being found. The CRF predicted the labels throughout the testing phase. The quality of a CRF model's characteristics determines how well it performs. The features employed were lexical, morphological, temporal, semantic, gazetteer and regular expression. The regular expressions were primarily based on the i2b2 challenge's best-performing CRF-based competitors. The model which best performed was LSTM, outperforming the CRF model as well as the combined one. The LSTM obtained a F-measure score of 99.23% and a precision of 99.21% with a recall of 99.25% based on the customized MIMIC data set while the F-measure obtained with i2b2 data set was 97.88%. More experiments may be conducted to determine if over-fitting occurred.

Long Short-Term Memory and Conditional Random Fields were used to de-identify medical records [27]. The CRF system extracting features manually was used for training, while the LSTM system was using classifying tags for represented tokens. The threshold cut-off of the features -f was set to 4, the hyper-parameter -c to 10, and L2-norm was employed for regularization in the CRF-based system. A subset of training data was used to identify the hyper-parameter. The "validation set" was the name given to this subset. To avoid overfitting, a dropout with a probability of 0.5 was used. The stochastic gradient descent (SGD) approach was used to train the model and the learning rate was set to 0.005. The validation set was used to tune the dimension of pre-trained embedding, whereas other parameters were determined through experience. After fine-tuning, the system worked better. The results from this work showed that the performance of the systems was improved by pre-processing as well as showing that the model which was the best performer was LSTM, outperforming CRF model. The LSTM model obtained F-measure score of 91.45%, precision 0f 93.24% and a recall of 89.72%. This led to the conclusion that LSTM is the best model. However, based on the results of this research alone, this conclusion cannot be generalized unless there is evidence showing CRF model underperformed compared to LSTM in the challenge.

Three data sets were used to de-identify the free text from Japanese electronic health records [28]. NTCIR-10 MedNLP De-identification Task [29], a synthetic data set annotated by researchers and a combination of the two data sets containing eighty two records make up these data sets. The five PHI categories from the data set were removed using CRF, bidirectional LSTM and a rule-based method. The CRF model was utilized to process the output for the LSTM technique. The hybrid model produced an F-measure of 80.61% and it was chosen over the CRF and rule-based models. Though the rule-based method obtained the highest F-measure, the recall and Precision scores were not provided, weakening the standard of this work since the missing scores might provide additional insight.

Ensemble-based methods were used to enhance the electronic health record narratives de-identification [30]. These are the methods combining machine learning and rule-based methods to perform de-identification on the seven PHI categories. These methods involve decision template [31], stacked ensemble [32] and voting [33]. The Stanford CoreNLP tool [34] was used to carry out Pre-processing. Training of Machine Learning models was done individually on the 2014 i2b2 challenge data set including support vector machine (SVM) [35], LSTM, CRF, LSTM-CRF and margin enhanced relaxed algorithm (MIRA) [36]. The stacked ensemble had the highest F-measure score of 95.73% with a 97.73% and the recall score of 94.45%.

Evaluation of a recurrent neural network construction for de-identification of clinical data was done [37]. Seven PHI from the 2014 i2b2 data set were de-identified using RNN model versions. Jordan [38] and Elman [39] RNN variants are two types of RNN. A comparison was made between these two variations and a standard CRF model. Every state in an Elman-type network was made up of information about prior hidden layer states and output posterior probabilities were utilized to get inputs to recurrent convolutional networks. Hyper-parameter tuning and stochastic gradient descent were used to train RNN models. A typical set of hand-crafted characteristics were utilized to train the CRF model and both RNN variations outperformed the CRF model. With an F-measure of 93.84%, precision of 97.26% and recall of 90.67%, the Jordan-type was the best performer. However, because both of these variants had previously been proven to have similar performance [40], an

alternative RNN variation should have been utilized to replace one of the variants to increase the credibility of this work.

Deep Neural Networks, rule-based and feature-based methods to de-identify Dutch medical records were Compared [41]. Investigations on de-identifying the eight PHI categories using a CRF, a bidirectional LSTM and a rule-based approach were done upon the construction of data set of the records from nine Dutch healthcare institutions. Twelve experts annotated the data collection, which included 1260 records and 17500 PHI instances. Nursing notes corpus data set [10] with 2434 records, 1800 PHI instances with 2014 i2b2 data set were used in evaluating the performance between Dutch and English data sets. This resulted in the development of a rule-based strategy for the Dutch medical records [42]. The CRF approach uses a subset of features from [26] token-based approach. Because the method was integrated with CRF architecture, the bidirectional LSTM model was referred to as a hybrid. The hybrid model outperformed the other models in all three data sets, with an F-measure score of 91.20%, a precision of 95.90% and a recall of 86.90%. However, it was discovered that there is a PHI category called "other" in this work. Because the “other” PHI group was not clearly defined, a study comparability issue arose.

Deep learning algorithms outperformed traditional strategies in de-identification of German medical records [43]. The performance of a CRF, Rule-Based method and bidirectional LSTM model [44] with trained statistical capabilities was compared to that of a CRF, Rule-Based method and bidirectional LSTM model [44] with trained statistical skills as a baseline. Seventy-five percent of the recordings were used for training, with the remaining percentage being used for testing. The CRF algorithm was accessed using the sklearn-cfsuite wrapper 2. The LSTM technique was implemented using Tensor Flow 4 and Keras 3. The LSTM model, with an F-measure score of 96.0%, a recall of 95.5% and a precision of 97.0% outperformed the other models among the eight de-identified PHI categories. An extra set of gold-standard data, such as the 2014 i2b2, could have helped.

Though it is similar to POPIA in many ways, the related US law (HIPAA’s Privacy

Rule) only applies to a specific field and aims to protect identifiable health information [45]. In accordance with the terms of this Agreement, information will be considered de-identified if it is stripped of 18 specific identifiers or if a professional statistical analyst with the requisite expertise and knowledge determines that there is a very low likelihood that the information, either by itself or in combination with other information, could be used to identify a data subject [46].

It is suggested that a good general rule to achieve de-identification in South Africa would be to make sure the 18 identifiers of an individual listed in HIPAA are removed from the data set; the removal of the 18 identifiers is known as the safe harbor method, while the second route to achieving compliance is known as the expert determination method and depends on a statistician verifying that the risk of identification of a data subject is very low. Names, addresses, phone numbers, fax numbers, email addresses, identity numbers, medical record numbers, medical aid details, account numbers, certificate/license numbers, vehicle identifiers and serial numbers, device identifiers and serial numbers, URLs, IP addresses, biometric identifiers and photographs are among the identifiers to be removed (as modified for South Africa) [47]. Once these have been removed and cannot be re-identified, then one can assume that the data is de-identified.

2.2 Conclusion

Our model Conditional Random Fields utilizes prior labels' contextual information to increase the amount of data the model has to make a good prediction. Long Short Term Memory has the ability to overcome the vanishing gradient problem of standard Recurrent Neural Networks. Random Forest model is capable of both regression and classification and also generates accurate predictions that are simple to follow. In comparison to the decision tree method, the random forest algorithm is more accurate at predicting outcomes. Most of the previous work using i2b2 data set were having only six Protected Health Information (PHI) which are patients, IDs, locations, phone numbers, dates and ages but for this work, two additional categories, hospital and doctor, were included, bringing the total to eight.

Chapter 3

Protection of Health Information in South Africa

3.1 How Personal Health Information is collected

In the process of consultations, examinations, and treatments, healthcare professionals gather patient health information. Health facilities including hospitals, clinics, and other health facilities collect patient information as part of their registration and record-keeping responsibilities. As part of the enrollment process and to manage claims, medical plans also acquire private health information. Participants in health surveys are asked for personal health information in order to learn more about the general health of the population. Participants' private health information is gathered for research reasons in health studies. The confidentiality and privacy of personal health information are protected by laws and regulations in South Africa.

3.2 How Personal Health Information is stored and protected

South Africa has its own rules and regulations to protect health information. Guidelines for the gathering, using, and sharing of personal health information are provided by both the **Protection of Personal Information Act (POPIA)** and **the National Health Act (NHA)** [48].

According to the NHA [49], all patient health information must be transmitted and maintained securely. This comprises paper documents, electronic health records,

and other patient data. Health workers are required to uphold rigorous ethical guidelines and patient confidentiality. Health information may be processed lawfully under the guidelines established by POPIA. Before collecting, using, or disclosing patients' information, medical establishments are required by POPIA to get their express consent. Moreover, health data must be true, appropriate, and kept on file for no longer than is absolutely necessary [48]. Healthcare providers are advised to put security measures in place, such as encryption, regular data backups, access controls, and staff training, to ensure compliance with these regulations. In addition, serious fines and penalties may be imposed for POPIA infractions.

3.3 Logistics behind getting South African dataset

Identifying the health information dataset you require is important if you want to obtain one for study in South Africa. The first step is figuring out what kind of health information dataset you require. After you are aware of the type of health information you need, you must locate the data source. The National Department of Health, the Medical Research Council, and the Health Systems Trust are just a few of the entities that gather and maintain medical data in South Africa. You must verify that your research topic is ethical and complies with all applicable regulations before you can access the data. Any research involving human subjects needs ethical approval, and data access may need the owners' written authorization. Data request should be sent for data access. The data owners will assess the application after submission of data access request, and if it satisfies all the criteria, they will provide access to the dataset. Once given permission to use the dataset, one can get it and begin research.

From my side, given the time frame to do the research, it was not possible to obtain South African dataset since the application for Ethical clearance would have taken about six months. I was informed that i should wait for the meeting of the commit which check ethical clearence applications. After, i had to send a letter and request approval from the hospital CEO, which no one knows how long it will take to get the signature. After I have to spend at least three months going to the hospital shadowing, to see how they capture the data. After this, when Ethical clearance

certificate has been granted, it is then that the institution would have released the dataset to me. I was advised to opt for open source dataset since the process of obtaining it is less than a week.

Also, the South African National Health Research Ethics Council has issued a thorough handbook on ethics in health research, in accordance with section 72 of the NHA, in the context of patient data used in health research [50]. The Health Protection Act (HPA) creates a Health Professionals Council in addition to the NHA, which is in charge of directing the medical industry [51]. Although the NHA contains the enabling requirements that necessitate confidentiality, the HPA is required to offer assistance to health care practitioners on a number of significant topics pertaining to medical practice.

Prior to POPIA, the National Health Act (NHA), the Health Professions Act (HPA), and the South African Medical Research Council Act (SAMRC Act) offered - and continue to provide - a framework for processing patient data discreetly and responsibly [48].

3.4 De-identification and its Rationale in USA

The United States is rapidly adopting health information technology, which accelerates their potential to support advantageous investigations combining sizable, intricate data sets from many sources. De-identification, the process of removing identifiers from health information, reduces privacy risks for individuals and facilitates the secondary use of data for projects like policy evaluation, life sciences research, comparative effectiveness studies, and other initiatives.

The purpose of the Privacy Rule is to safeguard personally identifiable health information by allowing only specific uses and disclosures of PHI that are either permitted by the Rule or approved by the specific person whose information they contain. Nonetheless, acknowledging the possible use of health data even in the absence of personal identification. By adhering to the de-identification standard and implementation requirements, a covered entity or its business associate may produce information that is not personally identifiable under the Privacy Rule. These clauses

permit the organization to use and reveal data that neither uniquely identifies a person nor offers a solid foundation for doing so.

3.5 Sufficient degree of identification risk for a professional assessment

The "very small" level specified by the approach is not universally satisfied by any clear numerical level of identification risk. When determining the risk associated with a data set, an expert must consider a variety of elements, including the recipient's capacity to identify the individual (i.e., subject of the information). This is due to the possibility that the risk of identification established for a certain data set within a given environment may not apply to another data set within the same environment or to a different data set within a different environment. Consequently, a specialist will determine what constitutes a reasonable "very small" risk by considering the likelihood that an individual might be identified.

3.6 Validity period determined by experts for a given dataset

The Privacy Rule does not state clearly that the determination of whether a data set or the process that produced it is de-identified information must have an expiration date; nonetheless, experts have acknowledged that social conditions, technology, and the availability of information are dynamic. As a result, some de-identification professionals employ the strategy of expiring credentials. In this regard, the specialist will evaluate the anticipated shift in computational capacity and access to diverse data sources, and then ascertain a suitable duration that will let the health information to be deemed reasonably shielded from personal identification.

Once the certification limit has been reached, information that was previously de-identified may still be sufficiently de-identified. It is not implied that data that has

already been distributed is no longer adequately protected in compliance with the de-identification standard when the certification deadline expires. To meet the very low risk criteria, covered organizations will need to have an expert assess whether additional or different de-identification techniques that are consistent with present conditions should be applied to data releases to the same recipient in the future (e.g., monthly reporting).

Chapter 4

Research Methodology

The key components of the procedure that were used in this study are described in this section. The research methods that were used in carrying out the research are discussed. A brief description of the data set used is provided.

4.1 Research design

This research focused on de-identification of n2c2 NLP Research Data Set using Conditional Random Fields, Long-Short Term Memory and Random Forest. These models' performance was assessed, and their results were compared. F-measure, Precision and Recall ratings at token level were computed to assess each model's performance. These metrics were used to determine which model outperformed the other.

4.2 Models

4.2.1 Conditional Random Fields (CRF)

CFR is regarded as an un-directed graph-based model that considers words that occur before the entity as well as after it. Graphical models have become a better way to show how output variables are connected [52]. In this area, a lot of work focuses on models such as the hidden Markov model (HMM), characterizing a collective probability distribution over observed input features x and corresponding annotated output y.

Parameter Learning: The maximum likelihood learning for $P(Y_i|X_i\theta)$ will be used to learn the parameters θ. The optimization is convex if all the nodes are noted as well as having exponential family distributions during training.

4.2.2 Random Forest (RF)

A supervised learning algorithm is random forest. It creates a "forest" out of an ensemble of decision trees, which are commonly trained using the "bagging" method. The bagging method's basic premise is that combining different learning models improves the overall output. Random forest increases the model's randomness while generating the trees. When splitting a node, it looks for the best feature from a random subset of features rather than the most essential feature. As a result, there is a lot of variety, which leads to a better model.

The **n_jobs** hyper-parameter specifies how many processors the engine is permitted to employ. It can only use one processor if the value is one. There is no limit if the value is "-1". The random state hyper-parameter allows the output of the model to be replicated. When random state is set to a fixed value and the model is given the same hyper-parameters and training data, the model will always deliver the same results. Finally, there's the random forest cross-validation method **oob_score** (also known as oob sampling). About one-third of the data in this sample is not utilized to train the model but can be used to assess its performance. Out-of-bag samples are what they're called. It's comparable to the leave-one-out cross-validation method, except it comes with nearly no additional computational overhead.

4.2.3 Long Short-Term Memory (LSTM)

LSTM is selected because of its ability to overcome the vanishing gradient problem of standard RNNs [53] . We can apply the LSTM on the sequence both forwards and backwards since we have access to the whole text. Thus, a bidirectional LSTM is used. A token-level and a character-level input is used and embedding layers transforms them to word vectors. Thus, "The patient has. . . " will look like ["the", "patient", "has"] for the token-level input and [["t", "h", "e"], ["p", "a",

“t”, “i”, “e”, “n”,“t”], [“h”,“a”, “s”]] for the character-level input. Token-level input is a sequence of tokens x1,...,xn, whereas character-level input is a sequence of sequences of characters, [[x1,1,..., x1,l(1)],...,[xn,1,..., xn,l(n)]], where the character sequence forms a token. n is the length of the token sequence, and l is the length of the character sequence of the i-th token (i).
Token-Based Embedding: From the token sequence, each token will be passed through a token embedding function, expecting the results to be a sequence of token-based embedding. This embedding function will be pre-trained, probably using GloVe or word2vec.

Character-Based Embedding: Each sequence of characters will be passed through a character embedding layer. The purpose being the production of second embedding for the token based on its sequence of characters. An embedding layer is used, thus, a mapping from a character to a vector. The sequence of vector embedding for the sequence of characters will then be passed through a bidirectional LSTM, producing one output vector. The LSTM activation is thus combined to one output for the whole sequence instead of one per character. This way, the method produces embedding per token which is based on each of its characters combined.

Bidirectional Long Short-Term Memory:
A bidirectional LSTM is used to process the sequence of combined embeddings and token-based embeddings, producing an output per token. Each of these outputs is the result of concatenating the forward and backward LSTM outputs for each token. A softmax function will convert the output into probabilities for tokens belonging to a specific class. The entire network’s output will be a sequence of vectors, with elements representing the token’s chances of belonging to one of the output classes.

Variations: The character-level input will be enhanced by concatenating white spaces, symbols or newlines. This is to help the network understand the sentence structure better, also in cases where dates are written in the format “DD/MM/YY”, which is quite common. Adding this information to the input gives the network the chance to learn features which use the presence of the characters. During pre-processing, all tokens will be transformed to a lower case. The network will also be given additional input indicating the presence of salutation words such as “Mr”, “Mrs”, or

"Ms" [53] . The motive behind this is to help the network recognise names more easily since some of the trained networks do not recognise patient names despite the presence of these strong indicator words.

Implementation This method will be implemented in Python using version 3.5.3. We will use the "high-level neural network API" Keras version 2.0.2, with TensorFlow version 1.0, as its deep learning back-end. The Graphics Processing Unit (GPU) mode will be used to run both the training and testing of the model. For the implementation of the model we will use the Keras layers Dropout, Input, Embedding, Time Distributed, Dense LSTM, Bidirectional. The model will be trained using the function fit and used using the function predict of the Model. When training a model using Keras, target values have to be represented using a 1-hot-encoding.

4.3 Data

The Data Set used was the **National Natural language processing Clinical Challenges (n2c2) NLP Research Data Set**, which is an open source data set from Harvard Medical School. These are called Medical Discharge Summaries and are unstructured notes from Partners Health Care's Research Patient Data Registry. The data was created as a portion of the (Informatics for Integrating Biology and the Bedside) i2b2 project (link here). This data was created by substituting plausible surrogates for any authentic PHI and annotating the results.

The Annotation Format is as follows:
Data was tokenized, then broken down into phrases and then converted to XML. Every record was contained within RECORD> tags and assigned a random ID number. Every record's text was contained within TEXT> tags. PHI instances were contained within PHI> tags. Each PHI category was represented by the TYPE attribute of the beginning PHI> element.

From the eighteen PHI categories specified by HIPAA, only six were found to appear in the data but two extra categories which are hospital and doctor were added,

making them eight. These eight PHI categories are given below:

1. Patients: Patients' names, family members' names and health proxy names
2. IDs: Any mixture of numbers, special characters and letters
3. Locations: Geographic locations such as building names, street names, states, cities and zip code.
4. Phone numbers: The pager, the telephone and fax numbers
5. Hospitals: The names of nursing homes and medical organizations offering treatment to patients.
6. Doctors: The medical practitioners and doctors included in the records.
7. Dates: All elements of a date excluding the year. identifying patients, medical records or doctors.
8. Ages: Years above 90.

The following is a sample of the n2c2 data set before pre-processing.

```
<?xml version="1.0"?>
<ROOT>
 - <RECORD ID="502">
    - <TEXT>
        <PHI TYPE="ID">113416550</PHI>
        <PHI TYPE="HOSPITAL">PRGH</PHI>
        <PHI TYPE="ID">13523357</PHI>
        <PHI TYPE="ID">630190</PHI>
        <PHI TYPE="DATE">6/7</PHI>
        /1999 12:00:00 AM Discharge Summary Signed DIS Admission Date :
        <PHI TYPE="DATE">06/07</PHI>
        /1999 Report Status : Signed Discharge Date :
        <PHI TYPE="DATE">06/13</PHI>
        /1999 HISTORY OF PRESENT ILLNESS : Essentially , Mr.
        <PHI TYPE="PATIENT">Cornea</PHI>
        is a 60 year old male who noted the onset of dark urine during early
        <PHI TYPE="DATE">January</PHI>
        . He underwent CT and ERCP at the
        <PHI TYPE="HOSPITAL">Lisonatemi Faylandsburgnic, Community Hospital</PHI>
        with a stent placement and resolution of jaundice . He underwent an ECHO and endoscopy at
        <PHI TYPE="HOSPITAL">Ingree and Ot of Weamanshy Medical Center</PHI>
        on
        <PHI TYPE="DATE">April 28</PHI>
        . He was found to have a large , bulging , extrinsic mass in the lesser curvature of his stomach . Fine needle aspiration showed atypical cells , positively reactive mesothelial cells . Abdominal CT on
        <PHI TYPE="DATE">April 14</PHI>
        , showed a 12 x 8 x 8 cm mass in the region of the left liver , and appeared to be from the lesser curvature of the stomach or left liver . He denied any nausea , vomiting , anorexia , or weight loss . He states that his color in urine or in stool is now normal . PAST MEDICAL HISTORY : He has hypertension and nephrolithiasis . PAST SURGICAL HISTORY : Status post left kidney stones x2 , and he has had a parathyroid surgery . ALLERGIES : He has no known drug allergies . MEDICATIONS PRIOR TO ADMISSION : Hydrochlorothiazide 25 mg q.d. , Clonidine 0.1 mg p.o. q.d. , baclofen 5 mg p.o. t.i.d. HOSPITAL COURSE : Basically , patient underwent a subtotal gastrectomy on the
        <PHI TYPE="DATE">7th of June</PHI>
        by Dr.
        <PHI TYPE="DOCTOR">Kotefooksshuff</PHI>
        . He had an uncomplicated postoperative course and he was transferred . Advanced his diet on postop day # 4 to a transitional diet . His PCA was discontinued on postop day # 4 , and essentially he was started on his pre-op medications on the postop day # 5 . PHYSICAL EXAMINATION : CHEST : Clear . HEART : Regular . ABDOMINAL INCISION : Clean , dry and intact . No drainage . VITAL SIGNS : He is afebrile and otherwise vital signs are stable . He is having good p.o. intake on present diet and he is moving his bowels . DISPOSITION : He is going to a rehabilitation facility until he is able to live independently . DISCHARGE MEDICATIONS : Same as pre-op , with the addition of Roxicet elixir . Dictated By :
        <PHI TYPE="DOCTOR">THAMETO DOYLE</PHI>
        , M.D.
        <PHI TYPE="ID">OS43</PHI>
        Attending :
        <PHI TYPE="DOCTOR">PRO R. KOTEFOOKSSHUFF</PHI>
        , M.D.
        <PHI TYPE="ID">RE6 7Y525/1920</PHI>
```

FIGURE 4.1: **Sample of the data**

There are 889 annotated records where 669 are for training while 220 is for testing.The distribution of PHI categories throughout the entire corpus, training and test sets is shown below.

TABLE 4.1: The Instances and Tokens for Complete Challenge Corpus, Training and Test Data

PHI Category	Complete Corpus		Training Data		Test Data	
	Instances	Tokens	Instances	Tokens	Instances	Tokens
Non-PHI	—	444127	—	370989	—	73183
Patients	929	1737	684	1340	245	397
IDs	4809	5110	3666	3063	1143	2047
Locations	263	518	244	491	19	27
Phone Numbers	232	271	174	160	58	111
Doctors	3751	7697	2681	6409	1070	1288
Hospitals	2400	5204	1724	4429	676	775
Dates	7098	7651	5167	5358	1931	2293
Ages	16	16	13	13	3	3

4.3.1 Pre-processing

As shown on the Sample of data in Figure 3.1 above, the Data set used was unstructured texts and required pre-processing. During pre-processing of this data set, we performed the following:

1. Lower Casing
   All the uppercase characters where converted to lowercase characters and where there were no uppercase characters in the data set, the original strings were returned. This was all done using built-in Python method .lower() primarily used for string handling.

2. Sentence Tokenization
   The technique of splitting text into individual sentences is known as sentence tokenization. This was done to split all the texts in the data set into sentences. A library called the Natural Language Tool Kit (NLTK) was utilized to accomplish this.

3. Word Tokenization
   The technique of splitting sentences into words is known as word tokenization. This was done to split all the sentences into words since it is required in natural language processing activities where each word must be collected and subjected to further analysis such as classification and counting for a specific sentiment.

4. Punctuation Removal
   In the English language, there are 14 different punctuation marks. Punctuation signs include the period, exclamation point, dash, question mark, semicolon, colon, hyphen, comma, braces, brackets, apostrophe, parentheses, ellipsis, and quote mark. All these were removed from the data set.

5. Stop words Removal
   Stop words (such as articles, prepositions, pronouns, conjunctions, and so on) are the most common words in any language and do not add much information to the text. Stop words in English include "the", "a", "an", "so", and "what". We removed the low-level information from our text by omitting these terms, allowing us to focus more on the crucial information. Removing stop words reduces the data set size and hence reduces training time.

Upon completion of pre-processing, we converted our pre-processed data into a DataFrane.

	record_id	word_id	word	pos_tag	labels	num_chars
0	0	0	<phi type="id">123547445</phi>	NN	id	30
1	0	1	<phi type="hospital">fih</phi>	NN	hospital	30
2	0	2	<phi type="id">7111426</phi>	NN	id	28
3	0	3	<phi type="id">47933/f911</phi>	NN	id	31
4	0	4	<phi type="id">557344</phi>	NN	id	27
5	0	5	<phi type="date">11/19</phi>	NN	date	28
6	0	6	am	VBP	Non-PHI	2
7	0	7	discharge	NN	Non-PHI	9
8	0	8	summary	NN	Non-PHI	7
9	0	9	unsigned	JJ	Non-PHI	8
10	0	10	dis	NN	Non-PHI	3
11	0	11	report	NN	Non-PHI	6
12	0	12	status	NN	Non-PHI	6
13	0	13	unsigned	JJ	Non-PHI	8
14	0	14	admission	NN	Non-PHI	9
15	0	15	date	NN	Non-PHI	4
16	0	16	<phi type="date">11/19</phi>	NN	date	28

FIGURE 4.2: **Sample of the pre-processed data**

4.4 Methods

Software and Data

Named Entity Recognition (NER) is a series of tasks for Information Extraction where Named entities are phrases that include the names of people, organizations, places, times, and quantities. We used the Computational Natural Language Learning CoNLL2002 corpus to build a NER system and is available in Natural Language Took KIT (NLTK). For this data presentation, each token (in this case, a single word) is placed on its own line. In the first column, the entire tokenized document was presented, the second column was the pos tag (The technique of marking up a word in a text as relating to a specific part of speech, also known as grammatical tagging, is known as part-of-speech tagging) and the third column was the named entity tag, simply named labels. The tokens were labeled in the sequence as the class "Non-PHI" or a specific PHI-category (id, date, hospital and so on). Columns were separated by a comma.

Features Extraction

The feature extraction stage involves extracting and producing feature representations that are appropriate for the type of NLP task one is tackling and the model that will be used. For this process, tokenization was required. After tokenization, the following features were extracted:

- Bag-of-words
  In NLP, the bag-of-words model is a simplified representation. A text is represented as a bag of its words in this approach, which ignores syntax and even word order while maintaining multiplicity.

- Part-of-speech (POS) tags
  The practice of assigning a word in a corpus to a corresponding component of a speech tag based on its context and definition is known as POS tagging. This is a difficult endeavor because a word may have a varied part of speech depending on the context in which it is used. Here, averaged perceptron tagger was used for POS tagging. We began by downloading a handful of NLTK language processing tools. The averaged perceptron tagger was used

to tag words with their parts of speech, whereas punkt was utilized to tokenize phrases (POS).

- General NER information:
  The NER tag of the word generated by labeling a word as 'Non-PHI' or assigning it to a specific PHI category

Other things that were considered during feature extraction were, if the word is a digit, the length of the word - number of characters (shorter words were expected to be more likely to belong to a particular part of speech, like prepositions or pronouns), Stemmed version of the word, which deleted all vowels along with g, y and n from the end of the word, but left at least a 2 character long stem, Features mentioned above for the previous word, the following word and the words two places before and after. The extracted features were then converted to sklearn-crfsuite format.

Training

Python-crfsuite and sklearn-crfsuite were installed. We imported some necessary libraries like sklearn-crfsuite, pycrfsuite, RandomForestClassifier, scipy.stats, classification report and RandomizedSearchCV as well as visualization libraries such as eli5, matplotlib.pyplot and seaborn. The Keras API with TensorFlow as its backend to build and train a bidirectional LSTM neural network model to recognize named entities in text data were also used.Then our training data was loaded. The models training was as follows: Algorithm was L-BFGS (default) which is a quasi-Newton optimization procedure that uses a limited amount of computer memory to simulate the Broyden Fletcher Goldfarb Shanno algorithm (BFGS) [54]. All possible states were none while all possible transitions were true, averaging was none. c was none while c1 was 0.1 and c2 was also 0.1. Calibration candidates were none as well as delta and epsilon. Maximum iterations were 100 with variance being none and verbose being false. c1 and c2 were the parameters for L1 and L2 regularization respectively.

But as for LSTM model, for model building, Embedding, Dense,TimeDistributed, SpatialDropout1D and Bidirectional were imported in order to check the trainable

parameters, and all the parameters were found to be trainable. For training, ModelCheckpoint and EarlyStoppingwere imported. The model was trained as follows: for ModelCheckpoint monitor was val loss, verbose was 1, savebest only was True, save weights only was True and mode was min. For early stopping, monitor was val accuracy, min delta was 0, patience was 1, verbose was 0, mode was max, baseline was None, restore best weights was False. Then batch size was 32, epochs was 3 and verbose was 1.

We examined the models by looking at the algorithm and visualizing the odds of transitioning from one label to another. We also looked at which characteristics were crucial in predicting a specific label. The investigation was carried out with the help of the eli5 library. It appeared like the model just remembered a large number of words. For example, the algorithm remembered '111', '104' and '108' for the label 'age'. Also the algorithm remembered 'oaksgekesser memorial hospital', 'retelk county medical center' and 'ph university of medical center' for the category 'hospital'. To overcome this issue, we opted for hyperparameter optimization. And as we expected, the model stopped to rely on words and used the context more. The model generalized better.

Hyperparameter Optimization
To increase quality, We selected regularization parameters using randomized search and used 3-folds cross-validation to train and validate our models. We were fitting a model 50 * 3 = 150 times. Though this consumed a significant amount of CPU time and RAM, the performance was greatly improved.

A simple tree-based model with a simple feature map was employed for Random Forest. Simple tree-based models have shown to be effective in the development of Named Entity Recognition and Classification systems. Random Forest, one of the most often used tree-based models, can learn the basic principles for labeling concepts. Features were given into a Random Forest classifier. The performance was computed using evaluation matrix based on token level and classification report was obtained for both models.

4.5 Performance metrics

After training the models, their performance were evaluated. The standard evaluation metrics which are F-measure, Precision and Recall were taken into consideration when measuring the performance of the system according to the instance and token level of ground truth.

Precision
This is referred to as Positive Predictive Value (PPV). It's the proportion of tokens or entities in a category that are correctly identified compared to the total number of tokens or things in that category.

Recall
This is referred to as sensitivity. It is the proportion of correctly identified tokens or entities in a category to the total number of tokens or entities in that category.

The confusion matrix was used to represent the output classifier, with true positive (TP) indicating when a positive instance is correctly identified as positive, true negative (TN) indicating when a negative instance is correctly identified as negative, false positive (FP) indicating when a positive instance is incorrectly identified as positive, and false negative (TN) indicating when a negative instance is incorrectly identified as negative. False Negative (FN), also known as the Type II mistake, occurs when a negative occurrence is incorrectly classified as positive, and False Positive (FP) occurs when a positive instance is incorrectly detected as negative. Below is an illustration of how the confusion matrix will appear.

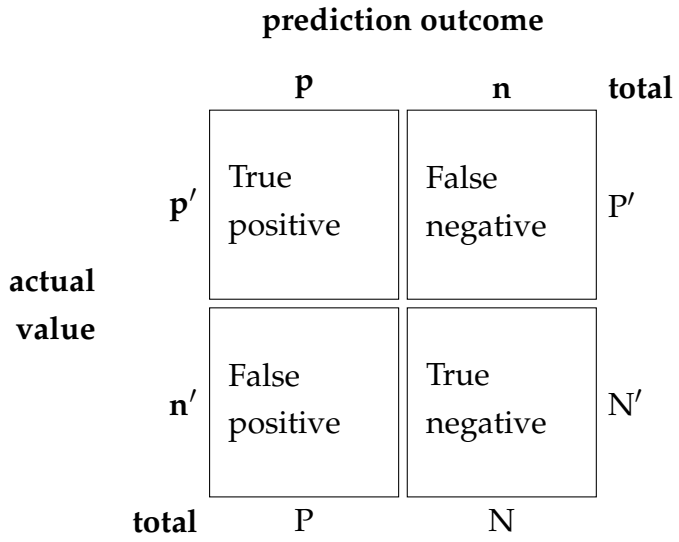


FIGURE 4.3: **Confusion Matrix**

From the Confusion Matrix, we computed:

$$Precision : P = \frac{True\ Positive}{True\ Positive + False\ Positive} = \frac{TP}{TP + FP} \tag{4.1}$$

$$Recall : R = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{TP}{TP + FN} \tag{4.2}$$

Having the above , we then computed the F-measure as follows.

$$F - measure : F_\beta = \frac{(1 + \beta)\ PXR}{\beta P + R} \tag{4.3}$$

4.6 Analysis

Regarding Descriptive Statistics, in order to understand our data better or compare the variables, we have the distribution of all the labels, PHIs and Part of speech tags. These results give an idea of how the variables and words themselves look like in the Data. As for experimental results, the performance was evaluated using

the standard evaluation metrics which are F-measure, Precision and Recall. This was done based on token level for each PHI category. Also, the micro, macro and weighted averages were computed.

Chapter 5

Results and Discussion

This section outlines the outcomes of the experiments that were performed on the National Natural language processing Clinical Challenges (n2c2) NLP Research Data Set from Harvard Medical School. The findings of this study are presented and discussed in relation to the study's aim, thus, to use Conditional Random Fields, Long-Short Term Memory and Random Forest models for de-identification of EDS and compare their results using evaluation metrics (F-measure, Precision, and Recall) to see how each model performs on the National Natural Language Processing Clinical Challenges (n2c2) data set.

5.1 Descriptive statistics

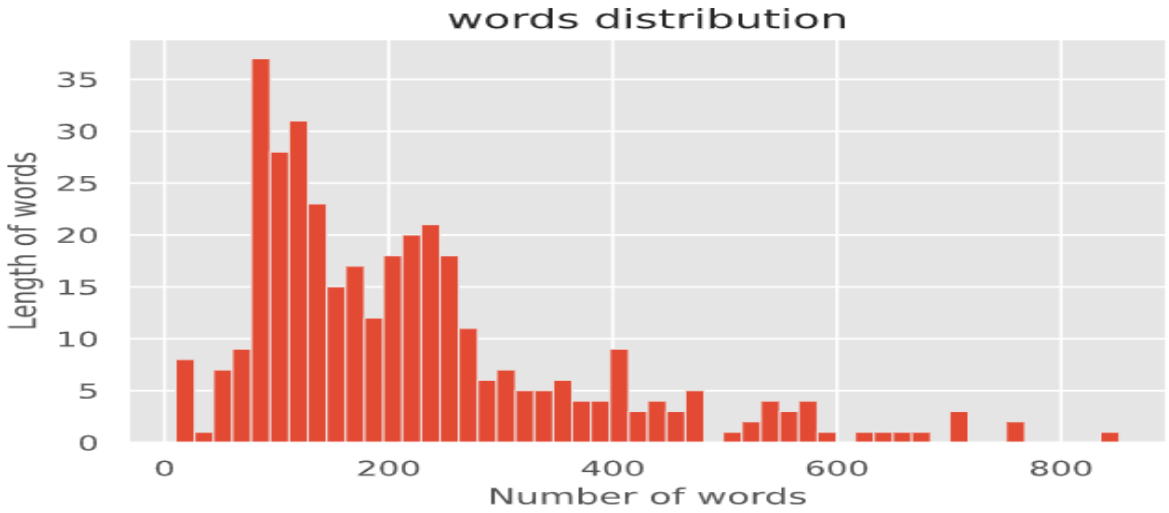


FIGURE 5.1: **The length of words**

The above Figure shows how words were distributed according to their length. The distribution is positively skewed to show that most words had shorter length. The word with the greatest length were PHIs (e.g, Hospital names, dates, IDs and location). Those which had a shorter length were Non-PHI and this can be seen on a sample of pre-processed data in Figure 3.2.

5.1.1 Distribution of all the labels

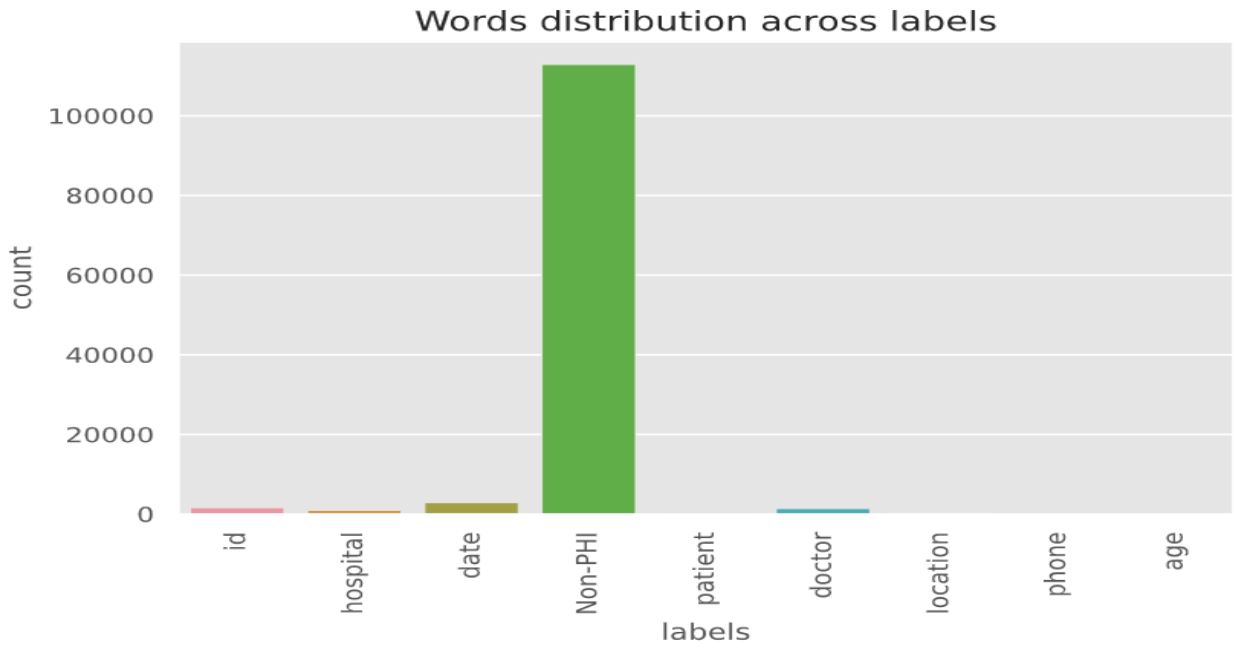


FIGURE 5.2: **Distribution of all the labels**

From the above Figure, Non-PHIs appeared to be far more than all the labels. This is due to the fact that in each Medical Report, there were paragraphs describing the patient's state, medical history, procedures taken as well as medication administered at each stage. Only a few words in that record were the personal information (PHI). When these paragraphs were split into words (tokens), they resulted in a large number of words being Non-PHIs. This was expected that the Non-PHI labels will be far more than all the PHI labels. Due to a few number of some labels, on this Figure, it appeared as if they were not there whereas they just had a few number of tokens like age and phone.

5.1.2 Distribution of all the PHIs

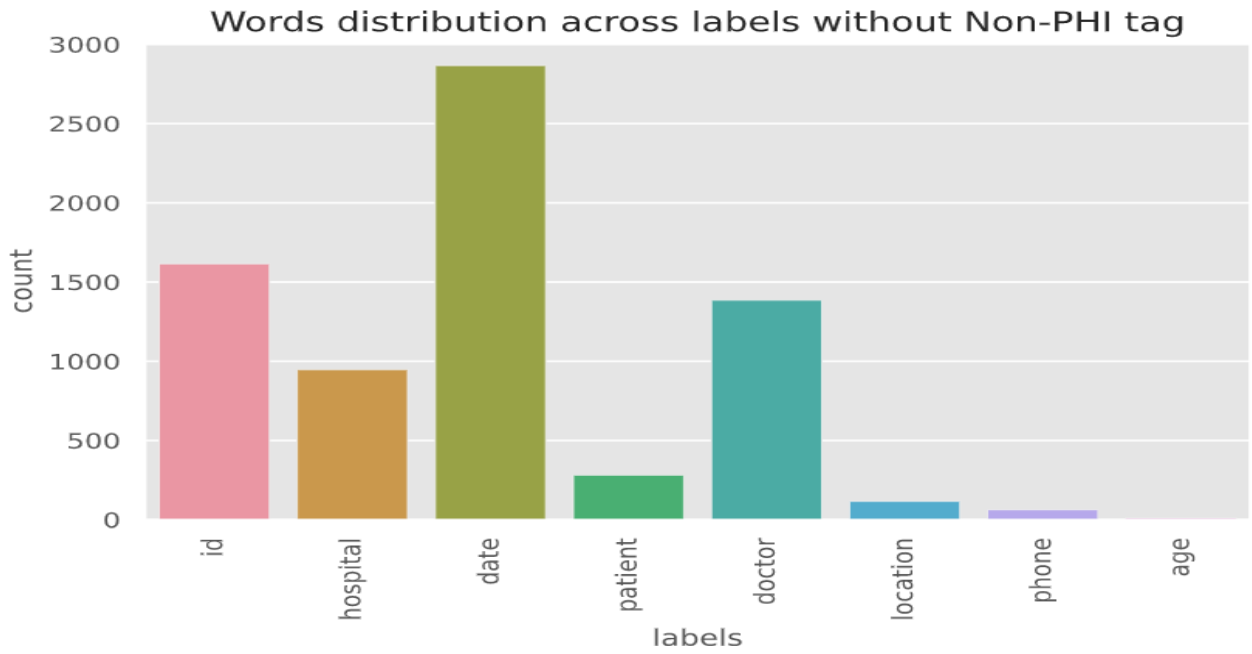


FIGURE 5.3: **Distribution of all the Personal Health Information**

The distribution of PHI labels is very skewed, as shown in the Figure above. Dates, IDs and doctor names are extremely common PHI kinds, whereas age, phone numbers and location are rather rare. On the Medical Reports, dates were mentioned all the time. This includes admission, discharge, the date when the medical problem was noticed, all the dates were medical steps (eg, scans or any medical examination) were taken, the date on which transfer to the other hospital took place if the patient was transferred as well as dates were changes in patient's condition were noticed. All these dates could fall under one patient's record, resulting in more dates than PHIs like names of patients, location and phone numbers. A very few number of tokens considered under age PHI category was found. This is due to the fact that only ages above 89 were considered to be PHI. This shows that most patients were of age 0-89 and only a few on the entire data set were above 89 years. One can conclude that a higher number of doctor names is as a result of different doctors treating one patient at the same time or at different times according to their area of expertise. Also, several times, one patient was transferred from one hospital

to another, resulting in a higher number of hospital names appearing as compared to the patient names.

5.1.3 Distribution of all the Part of speech tags

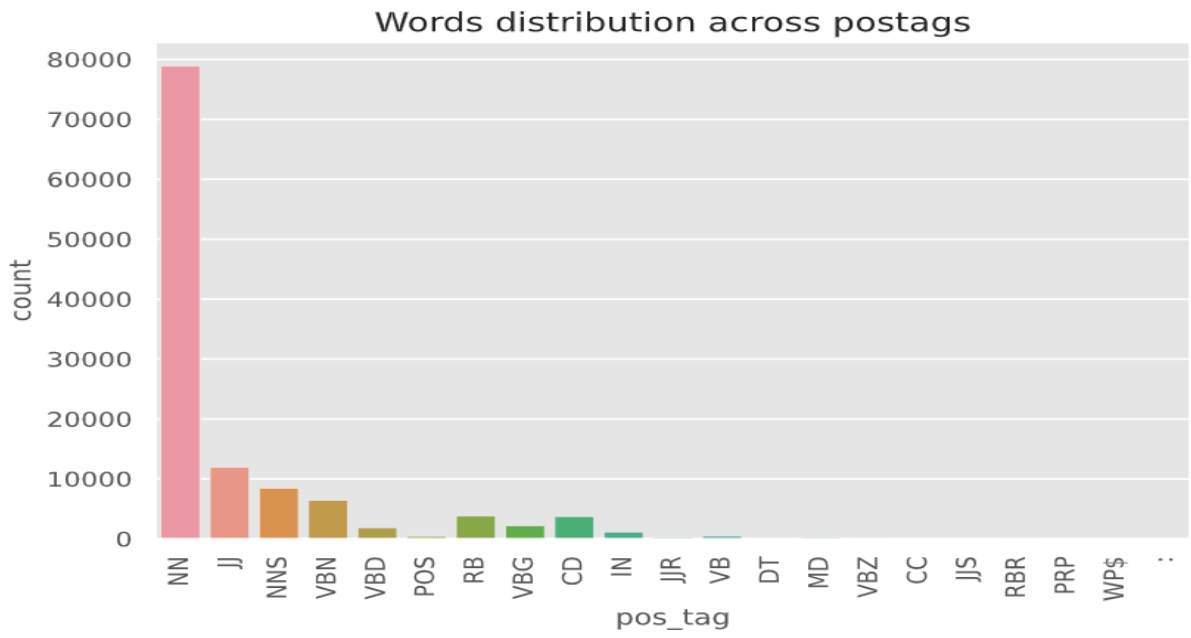


FIGURE 5.4: **Distribution of all the Part of speech tags**

As shown in the Figure above, the distribution of Part of speech tags is very skewed. Words which were tagged as nouns appeared to be extremely more than any other word. The fundamental issue with Part Of Speech tagging is that it is ambiguous. Many common words in English have various meanings and thus multiple Part Of Speech. Apart from words with multiple Part Of Speech, these words were tagged when they were just individual tokens, not as sentences. In a sentence, it was easy to pick up the meaning of a word. But as an individual word, names were correctly tagged as nouns but so many other words that have lost their meaning due to the fact that they are no longer in a sentence like most verbs were mistakenly tagged as nouns. Some parts of speech appeared in a very small number like coordinating conjunction (CC), personal pronoun (PRP) (hers, herself, him, himself) and so on.

5.1.4 Visualization to Inspect weights

From \ To	Non-PHI	age	date	doctor	hospital	id	location	patient	phone
Non-PHI	1.092	0.106	-0.033	0.464	0.048	-0.859	0.146	0.516	0.422
age	0.393	0.0	-0.101	-0.044	-0.015	-0.093	0.0	0.0	0.0
date	0.866	0.091	-0.088	-0.273	-0.976	-2.647	-0.758	-0.043	-1.288
doctor	0.089	-0.054	-0.958	-1.059	1.042	-0.25	0.703	-0.167	1.243
hospital	0.053	-0.027	-0.344	-0.185	-0.177	0.949	2.113	-0.965	0.348
id	-1.499	-0.145	1.373	0.214	0.257	0.733	-1.13	-1.52	-1.192
location	0.399	0.0	0.0	0.424	-0.041	-2.293	-0.082	0.324	-0.0
patient	0.283	1.689	-1.325	0.012	-0.562	0.865	-0.065	-0.028	0.0
phone	0.44	0.0	-0.131	0.576	0.453	-0.701	-0.005	0.0	-0.0

y=Non-PHI top features

Weight?	Feature
+4.196	postag=NNS
+3.007	word[-2:]=al
+2.726	postag[:2]=VB
+2.617	word[-2:]=nt
+2.559	word[-2:]=on
+2.416	word[-3:]=ion
+2.307	word[-2:]=ed
+2.159	word[-3:]=ent
+2.148	postag[:2]=JJ
	... 1007 more positive ...
	... 186 more negative ...
-2.336	+1:word.lower=m.d.

y=age top features

Weight?	Feature
+3.577	-1:word.lower=age
+3.130	word[-3:]=hi>
+3.130	word[-2:]=i>
+2.983	+1:word.lower=man
+2.733	+1:word.lower=registration
+2.484	-1:word.lower=patient
+1.990	word.lower=<phi type="age">111y</phi>
+1.990	word.lower=<phi type="age">114y</phi>
+1.754	word.lower=<phi type="age">108-year-old</phi>
+1.753	-1:word.lower=<phi type="patient">kotereef</phi>
	... 26 more positive ...

y=date top features

Weight?	Feature
+7.681	word[-3:]=hi>
+7.681	word[-2:]=i>
+5.472	-1:word.lower=td
+5.086	-1:word.lower=date
+4.158	+1:word.lower=_____
+3.482	+1:word.lower=-94
+3.403	+1:word.lower=pm
+3.226	+1:word.lower=-92
+3.156	+1:word.lower=-91
+3.113	word.lower=<phi type="date">07/20</phi>
	... 2138 more positive ...
	... 141 more negative ...

y=doctor top features

Weight?	Feature
+8.751	+1:word.lower=m.d.
+8.301	-1:word.lower=dr.
+5.856	word[-2:]=i>
+5.856	word[-3:]=hi>
+4.618	+1:word.lower=provider
+4.375	-1:word.lower=dictated
+4.329	+1:word.lower=preliminary
+4.182	-1:word.lower=dicatated
+3.482	+1:word.lower=<phi type="date">05/26</phi>
+3.478	-1:word.lower=dr
	... 290 more positive ...
	... 52 more negative ...

y=hospital top features

Weight?	Feature
+5.741	word[-3:]=hi>
+5.741	word[-2:]=i>
+5.195	word.lower=<phi type="hospital">omh</phi>
+5.136	word.lower=<phi type="hospital">ph university of medical center</phi>
+4.917	word.lower=<phi type="hospital">ctmc</phi>
+4.907	word.lower=<phi type="hospital">retelk county medical center</phi>
+4.897	word.lower=<phi type="hospital">bh</phi>
+4.828	word.lower=<phi type="hospital">tgcho</phi>
+4.778	word.lower=<phi type="hospital">puomc</phi>
+4.758	word.lower=<phi type="hospital">oaksgekesser/ memorial hospital</phi>
	... 1102 more positive ...
	... 30 more negative ...

y=id top features

Weight?	Feature
+8.018	word[-2:]=i>
+8.018	word[-3:]=hi>
+5.183	-1:word.lower=batch
+4.637	+1:word.lower=room
+4.564	-1:word.lower=ref
+4.511	-1:word.lower=number
+4.042	-1:word.lower=report
+3.991	BOS
+3.954	-1:word.lower=no.
+3.762	+1:word.lower=cc
	... 1637 more positive ...
	... 70 more negative ...

y=location top features

Weight?	Feature
+5.277	word[-3:]=hi>
+5.277	word[-2:]=i>
+4.150	word.lower=<phi type="location">mississippi</phi>
+2.765	+1:word.lower=tobacco
+2.407	-1:word.lower=lives
+2.259	+1:word.lower=lived
+2.204	-1:word.lower=done
+2.133	word.lower=<phi type="location">texas</phi>
+2.096	-1:word.lower=neurologist
+2.088	-1:word.lower=hospital
	... 317 more positive ...
	... 5 more negative ...

y=patient top features

Weight?	Feature
+6.177	-1:word.lower=mr.
+5.585	+1:word.lower=unit
+4.518	word[-3:]=hi>
+4.518	word[-2:]=i>
+4.509	-1:word.lower=ms.
+4.409	+1:word.lower=arrived
+4.390	+1:word.lower=mrn
+3.912	-1:word.lower=******
+2.941	-1:word.lower=mrs.
+2.861	-1:word.lower=illness
	... 209 more positive ...
	... 10 more negative ...

y=phone top features

Weight?	Feature
+4.234	-1:word.lower=call
+4.090	word[-3:]=hi>
+4.090	word[-2:]=i>
+2.809	-1:word.lower=telephone
+2.801	-1:word.lower=number
+2.703	word.lower=<phi type="phone">381-389-5852</phi>
+2.526	+1:word.lower=pcp
+2.432	+1:word.lower=dictated
+2.228	word.lower=<phi type="phone">610-7597</phi>
+2.221	-1:word.lower=office
	... 114 more positive ...
	... 1 more negative ...

FIGURE 5.5: **The weights**

Eli5 is a python package and was used to check the weights. It can be seen from the figure above that it was most likely from the data set to find hospital name right after id (hospital -> id has a large positive weight). Also, it was unlikely to find doctor name right after date, date right after hospital name and hospital name right after patient name (doctor -> date, date -> hospital, hospital -> patient has a large negative weight).

5.2 Experimental Results

The standard evaluation metrics were taken into consideration when measuring the performance of the models. The F-measure, precision and recall scores were used to assess de-identification models (standard evaluation approach for NER systems). This was done at token level for each model. We defined these standard evaluation metrics as follows:

- **Precision**
  Precision refers to a classifier's ability to avoid labeling a negative instance as positive. It is calculated as the ratio of true positives to the sum of true positives and false positives for each class. This is referred to as Positive Predictive Value (PPV). It is the proportion of tokens or entities in a category that are correctly identified compared to the total number of tokens or entities in that category

- **Recall**
  The capacity of a classifier to discover all positive cases is known as recall. It is calculated as the ratio of true positives to the sum of true positives and false negatives for each class. Sensitivity is the term for this. It is the ratio of correctly identified tokens or entities to the total number of tokens or entities in a category.

- **F-measure (F1-score)**
  The F-measure is a weighted harmonic mean of precision and recall, with 1.0 being the highest and 0.0 being the lowest. Because precision and recall are factored into F1 scores, they are lower than accuracy measurements. When comparing classifier models, the weighted average of F1 was utilized rather

than global accuracy as a rule of thumb. This is a metric for how accurate a model is on a given data set. It's used to assess binary classification algorithms that categorize examples as either "positive" or "negative."

5.2.1 Classification Reports

We would desire high precision and recall values in a good classifier. The values of these metrics were obtained in a text format per label using classification reports. To ensure that we had a good model, we analyzed the model using these metrics per label. Support column was provided to show the number of tokens in each label. The following classification reports show the results obtained on the test set from each model (Conditional Random Fields, Random Forest and LSTM).

	precision	recall	f1-score	support
age	1.00	0.33	0.50	3
date	0.90	0.97	0.93	2293
doctor	1.00	0.87	0.93	1288
hospital	0.89	0.77	0.83	775
id	0.92	0.99	0.96	2047
location	0.43	0.11	0.18	27
patient	0.99	0.99	0.99	397
phone	0.95	0.49	0.64	111
micro avg	0.93	0.93	0.93	6941
macro avg	0.88	0.69	0.74	6941
weighted avg	0.93	0.93	0.92	6941
samples avg	0.08	0.08	0.08	6941

FIGURE 5.6: **Classification Report For Conditional Random Fields model**

The above Figure is a classification report obtained using Conditional Random Fields model. We can see that the weighted average is high, being 93% for precision and recall and 92% for f1-score. This means that our model was 93% accurate (from weighed average f1-score). Apart from high accuracy and a good average score, the model did perform well for each label. Also, for other labels with more

tokens under its category like date, doctor, ID and patient, the model performed very well, giving good results for precision, recall and F-measure. The model has a good precision for the labels which had a few tokens under its category like age, location and phone though the performance on recall and f1-measure is poor.

	precision	recall	f1-score	support
Non-PHI	1.00	1.00	1.00	73183
age	0.00	0.00	0.00	3
date	0.68	0.89	0.77	2293
doctor	0.57	0.76	0.65	1288
hospital	0.41	0.28	0.33	775
id	0.58	0.48	0.52	2047
location	0.00	0.00	0.00	27
patient	0.18	0.01	0.01	397
phone	0.00	0.00	0.00	111
accuracy			0.97	80124
macro avg	0.38	0.38	0.37	80124
weighted avg	0.96	0.97	0.96	80124

FIGURE 5.7: **Classification Report For Random Forest model**

The above figure is a classification report obtained using Random Forest model. Despite having high accuracy and a good average score, the model performed poorly. Age, location and phone labels had precision and recall levels of 0. It appeared that the features required for the model to make sound decisions were missing. The model essentially memorized words and tags, which was insufficient. This happened to the labels which had a few tokens under its category (e.g age, which only had 3 tokens). For Non-PHI label, with the highest number of tokens, the model greatly performed with precision, recall and F-measure giving 100% each. Also for labels with more tokens under its category like date, doctor and ID, the model performed much better, giving better results for precision, recall and F-measure.

	precision	recall	f1-score	support
Non-PHI	1.00	1.00	1.00	73183
age	1.00	0.33	0.50	3
date	0.90	0.97	0.93	2293
doctor	1.00	0.87	0.93	1288
hospital	0.89	0.77	0.83	775
id	0.92	0.99	0.96	2047
location	0.43	0.11	0.18	27
patient	0.99	0.99	0.99	397
phone	0.95	0.49	0.64	111
micro avg	0.99	0.99	0.99	80124
macro avg	0.90	0.73	0.77	80124
weighted avg	0.99	0.99	0.99	80124
samples avg	0.99	0.99	0.99	80124

FIGURE 5.8: **Classification Report For Long-Short Term Memory (LSTM) model**

The above Figure is a classification report obtained using LSTM model. We can see that the weighed average is higher, being 99% for precision, recall and f1-score. This means that our model was 99% accurate (from weighed average f1-score). Apart from high accuracy and a good average score, the model performed very well for each label. As for Non-PHI label, the classifier correctly labelled all the tokens, giving 100% for precision, recall and F1-measure. Also, for other labels with more tokens under its category like date, doctor, ID and patient, the model performed very well, giving good results for precision, recall and F-measure. The model also performed better even to the labels which had a few tokens under its category like age, location and phone since the performance was not 0%.

5.3 Practical implications as a result of limitation

Since the dataset only had 669 records for training, one can easily check the results from each model and notice that there were variables like age and location which had a very low count. For example, Age only appeared three times in the data. We can easily jump into conclusion that, maybe if we had more records, age count

could have been improved. On the graph, it seems not to appear at all just because of this low count. But apart from this, the rest of the variables appeared the way one could have assumed them to be.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Manual de-identification of Clinical Records proves to be error prone, labor intensive and takes a long time. In this research, we focused on de-identification of EDS using Machine Learning models (Conditional Random Fields, Random Forest and Long Short Term Memory). Since insufficient, erroneous and or missing data, can have an impact on data quality which might complicate the model training process as well as reducing the efficiency of the trained model, the exploration of these three models, thus Conditional Random Fields, Random Forest and Long Short Term Memory was the excellent way to detect data quality in EHR of free text. We compared and assessed the effectiveness and efficiency of two NLP approaches to de-identification in this study. We discovered that when training data set is available, even as a small set, machine learning approaches are still preferable. After evaluation of these three models, their performance was compared based on the classification reports provided above. It was concluded that the Long Short Term Memory model outperformed both Conditional Random Fields and Random Forest model. This is as a result of Long Short Term Memory model having high percentage of precision, recall and F-measure when the label had a high number of tokens as well as a low number of tokens. On the other hand, Conditional Random Fields performed well as compared to Random Forest model, which performed well only for those labels with a high number of tokens but very poor on labels with a low number of tokens. Also the accuracy score of Long Short Term Memory was 99%, Conditional Random Fields was 93% and for Random Forest was 97%. Based on these observations, we have concluded that Long Short Term Memory is the best.

6.2 Future Work

Other intriguing elements of de-identification are generalizability and robustness across many healthcare domains. We can expect more results and strategies on generalizability in the future because it is an on-going and popular area in NLP research. Finally, none of the de-identification approaches can guarantee perfect de-identification quality since they cannot fully eliminate the possibility of re-identification but produce data sets with a low probability of being re-identified. De-identification methods are extremely difficult to implement in practice because of this. To limit the privacy concerns of revealing PHI, further (non-technical) steps are required in future to enhance this process. Also, concerning data streaming, big data and data quality, there is a shortage of studies along these lines. So future work should also focus on these gaps.

Bibliography

[1] Mehmet Kayaalp. "Modes of De-identification". In: *AMIA Annual Symposium Proceedings*. Vol. 2017. American Medical Informatics Association. 2017, p. 1044.

[2] El Emam Khaled and Anita Fineberg. "An overview of techniques for de-identifying personal health information". In: *Access to Information and Privacy Division of Health Canada* (2009).

[3] Schweikart Scott J. "Should immigration status information be considered protected health information?" In: *AMA journal of ethics* 21.1 (2019), pp. 32–37.

[4] Callen Joanne, Jean McIntosh, and Julie Li. "Accuracy of medication documentation in hospital discharge summaries: a retrospective analysis of medication transcription errors in manual and electronic discharge summaries". In: *International journal of medical informatics* 79.1 (2010), pp. 58–64.

[5] SF Murphy et al. "Electronic discharge summary and prescription: improving communication between hospital and primary care". In: *Irish Journal of Medical Science (1971-)* 186.2 (2017), pp. 455–459.

[6] Insurance Portability and Accountability Act. "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule". In: (2012).

[7] Kaung Khin, Philipp Burckhardt, and Rema Padman. "A deep learning architecture for de-identification of patient notes: Implementation and evaluation". In: *arXiv preprint arXiv:1810.01570* (2018).

[8] Geetha Mahadevaiah et al. "De-Identification of Protected Health Information PHI from Free Text in Medical Records." In: *International Journal of Security, Privacy and Trust Management* 8.1/2 (2019), pp. 1–11.

[9] Latanya Sweeney. "Replacing personally-identifying information in medical records, the Scrub system." In: *Proceedings of the AMIA annual fall symposium*. American Medical Informatics Association. 1996, p. 333.

[10] Ishna Neamatullah et al. "Automated de-identification of free-text medical records". In: *BMC medical informatics and decision making* 8.1 (2008), pp. 1–17.

[11] Salma Berrada et al. "De-Identification". PhD thesis. University of British Columbia, 2017.

[12] Elizabeth D Liddy. "Natural language processing". In: (2001).

[13] SL Garfinkel. *De-identification of personal information. National institute of standards and technology. 2015*. 2015.

[14] Claire Pocklington and Loay Al-Dhahir. "A comparison of methods of producing a discharge summary: handwritten vs. electronic documentation". In: *BJMP* 4.3 (2011), a432.

[15] Leibo Liu et al. "De-identifying Hospital Discharge Summaries: An End-to-End Framework using Ensemble of De-Identifiers". In: *arXiv preprint arXiv:2101.00146* (2021).

[16] Katrin Tomanek et al. "An interactive de-identification-system". In: *Proceedings of SMBM* (2012), pp. 82–86.

[17] Margaret Douglass et al. "Computer-assisted de-identification of free text in the MIMIC II database". In: *Computers in Cardiology, 2004*. IEEE. 2004, pp. 341–344.

[18] Isaac S Kohane, Susanne E Churchill, and Shawn N Murphy. "A translational engine at the national scale: informatics for integrating biology and the bedside". In: *Journal of the American Medical Informatics Association* 19.2 (2012), pp. 181–185.

[19] Andrew J McMurry et al. "SHRINE: enabling nationally scalable multi-site disease studies". In: *PloS one* 8.3 (2013), e55811.

[20] Özlem Uzuner, Yuan Luo, and Peter Szolovits. "Evaluating the state-of-the-art in automatic de-identification". In: *Journal of the American Medical Informatics Association* 14.5 (2007), pp. 550–563.

[21] Bruce A Beckwith et al. "Development and evaluation of an open source software tool for deidentification of pathology reports". In: *BMC medical informatics and decision making* 6.1 (2006), pp. 1–9.

[22] Andrew J McMurry et al. "Improved de-identification of physician notes through integrative modeling of both public and private medical text". In: *BMC medical informatics and decision making* 13.1 (2013), pp. 1–11.

[23] Franck Dernoncourt et al. "De-identification of patient notes with recurrent neural networks". In: *Journal of the American Medical Informatics Association* 24.3 (2017), pp. 596–606.

[24] Amber Stubbs and Özlem Uzuner. "Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus". In: *Journal of biomedical informatics* 58 (2015), S20–S29.

[25] Mohammed Saeed et al. "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database". In: *Critical care medicine* 39.5 (2011), p. 952.

[26] Zengjian Liu et al. "Automatic de-identification of electronic medical records using token-level and character-level conditional random fields". In: *Journal of biomedical informatics* 58 (2015), S47–S52.

[27] Zhipeng Jiang et al. "De-identification of medical records using conditional random fields and long short-term memory networks". In: *Journal of biomedical informatics* 75 (2017), S43–S53.

[28] Kohei Kajiyama et al. "De-identifying free text of Japanese dummy electronic health records". In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. 2018, pp. 65–70.

[29] Mizuki Morita et al. "Overview of the NTCIR-10 MedNLP Task." In: *NTCIR*. Citeseer. 2013.

[30] Youngjun Kim, Paul Heider, and S Meystre. "Ensemble-based methods to improve de-identification of electronic health record narratives". In: *AMIA annual symposium proceedings*. Vol. 2018. American Medical Informatics Association. 2018, p. 663.

[31] Ludmila I Kuncheva, James C Bezdek, and Robert PW Duin. "Decision templates for multiple classifier fusion: an experimental comparison". In: *Pattern recognition* 34.2 (2001), pp. 299–314.

[32] David H Wolpert. "Stacked generalization". In: *Neural networks* 5.2 (1992), pp. 241–259.

[33] Jennifer D'Souza and Vincent Ng. "Ensemble-based medical relation classification". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 1682–1693.

[34] Christopher D Manning et al. "The Stanford CoreNLP natural language processing toolkit". In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014, pp. 55–60.

[35] Seyed Hamed Hashemi Mehne and Seyedali Mirjalili. "Support vector machine: Applications and improvements using evolutionary algorithms". In: *Evolutionary Machine Learning Techniques*. Springer, 2020, pp. 35–50.

[36] Koby Crammer and Yoram Singer. "Ultraconservative online algorithms for multiclass problems". In: *Journal of Machine Learning Research* 5.3 (2003), pp. 951–991.

[37] Ankit Srivastava et al. "A recurrent neural network architecture for de-identifying clinical records". In: *Proceedings of the 13th international conference on natural language processing*. 2016, pp. 188–197.

[38] Michael I Jordan. "Serial order: A parallel distributed processing approach". In: *Advances in psychology*. Vol. 121. Elsevier, 1997, pp. 471–495.

[39] JL Elman. "Finding structure in time. cognitive science, vol. 14". In: (1990).

[40] Joseph Chee Chang and Chu-Cheng Lin. "Recurrent-neural-network for language detection on twitter code-switching corpus". In: *arXiv preprint arXiv:1412.4314* (2014).

[41] Jan Trienes et al. "Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records". In: *arXiv preprint arXiv:2001.05714* (2020).

[42] Vincent Menger et al. "DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text". In: *Telematics and Informatics* 35.4 (2018), pp. 727–736.

[43] Phillip Richter-Pechanski et al. "Deep Learning Approaches Outperform Conventional Strategies in De-Identification of German Medical Reports." In: *GMDS*. 2019, pp. 101–109.

[44] Phillip Richter-Pechanski, Stefan Riezler, and Christoph Dieterich. "De-Identification of German Medical Admission Notes." In: *GMDS*. 2018, pp. 165–169.

[45] Lee Swales. "The Protection of Personal Information Act and data de-identification". In: *South African Journal of Science* 117.7-8 (2021), pp. 1–3.

[46] Ilene N Moore et al. "Confidentiality and Privacy in Health Care from the Patient's Perspective: Does HIPPA Help". In: *Health Matrix* 17 (2007), p. 215.

[47] Fida K Dankar, Andrey Ptitsyn, and Samar K Dankar. "The development of large-scale de-identified biomedical databases in the age of genomics—principles and challenges". In: *Human genomics* 12 (2018), pp. 1–15.

[48] Lee Swales. "The Protection of Personal Information Act 4 of 2013 in the context of health research: enabler of privacy rights or roadblock?" In: *Potchefstroom Electronic Law Journal/Potchefstroomse Elektroniese Regsblad* 25.1 (2022).

[49] Helen L Walls et al. "Understanding healthcare and population mobility in southern Africa: the case of South Africa". In: *South African Medical Journal* 106.1 (2016), pp. 14–15.

[50] Ames Dhai and Boyce Mkhize. "The Health Professions Council of South Africa and the medical practitioner". In: *CME: Your SA Journal of CPD* 24.1 (2006), pp. 8–11.

[51] Dirk T Hagemeister. "Employer-generated complaints to the statutory registration authority: The regulatory framework for the supervision of employed health professionals in the South African public sector". In: *South African Journal of Bioethics and Law* 11.1 (2018), pp. 11–14.

[52] Hanna M Wallach. "Conditional random fields: An introduction". In: *Technical Reports (CIS)* (2004), p. 22.

[53] Eva-Lisa Meldau. *Deep neural networks for inverse de-identification of medical case narratives in reports of suspected adverse drug reactions*. 2018.

[54] Robert Malouf. "A comparison of algorithms for maximum entropy parameter estimation". In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. 2002.