FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

De-identification of Clinical Text Using Sentence Embeddings

Vasco Marinho Rodrigues Gomes Alves

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Henrique Lopes Cardoso

July 18, 2024

De-identification of Clinical Text Using Sentence Embeddings

Vasco Marinho Rodrigues Gomes Alves

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Daniel Augusto Gama de Castro Silva
Referee: Ricardo Daniel Santos Faro Marques Ribeiro
Supervisor: Henrique Lopes Cardoso

July 18, 2024

Abstract

Ensuring privacy when handling and sharing clinical data within Electronic Health Records (EHR) and clinical notes is crucial for compliance with data protection regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. Manual de-identification is labor-intensive and resource-demanding, prompting the exploration of Natural Language Processing (NLP) methods for automated solutions.

De-identification in the clinical domain involves removing Protected Health Information (PHI) from clinical text to prevent subject identification. PHI includes names, addresses, and dates, among other identifiers. This task is intrinsically challenging due to the vast volumes of clinical data and the diverse, often inconsistent terminology, which can include various formats, structures, and misspellings. Despite numerous de-identification systems developed over the years, their implementation remains limited in healthcare practice, as they are unable to guarantee the removal of all sensitive information.

In this work, we propose a novel technique that ensures the removal of all sensitive information by utilizing sentence embeddings. This method substitutes each sentence in a clinical document with semantically similar counterparts from an embedding space created from a de-identified dataset, ensuring no sensitive information remains in the final document. We evaluate the performance of different models using metrics that assess both anonymization sensitivity and the retention of clinical information.

Our results indicate that sentence replacement preserves relevant medical information more effectively than the previously proposed word replacement strategy, which, while better at anonymization sensitivity, often compromises the retention of clinical information. The best sentence embedding model obtains gains from 20 to 32% over the tested word embedding model on clinical information retention metrics. This research offers a promising direction for improving automated de-identification in clinical practice.

Keywords: Natural Language Processing, Protected Health Information, De-Identification, Anonymization, Sentence Embeddings

Resumo

Garantir a privacidade ao lidar e compartilhar dados clínicos em Registos Eletrónicos de Saúde e notas clínicas é crucial para a conformidade com regulamentos de proteção de dados, como a Health Insurance Portability and Accountability Act (HIPAA) nos Estados Unidos e o Regulamento Geral de Proteção de Dados (RGPD) na União Europeia. A desidentificação manual é uma tarefa intensiva e exige muitos recursos, o que motiva a exploração de métodos de Processamento de Linguagem Natural (PLN) para soluções automatizadas.

A desidentificação no domínio clínico envolve a remoção de Informação de Saúde Protegida (ISP) do texto clínico para evitar a identificação dos sujeitos. ISP inclui nomes, moradas e datas, entre outros identificadores. Esta tarefa é intrinsecamente desafiadora devido ao grande volume de dados clínicos e à terminologia diversificada e muitas vezes inconsistente, que pode incluir vários formatos, estruturas e erros ortográficos. Apesar dos inúmeros sistemas de desidentificação desenvolvidos ao longo dos anos, a sua utilização permanece limitada na área da saúde, pois são incapazes de garantir a remoção de toda a informação sensível.

Neste trabalho, propomos uma técnica inovadora que assegura a remoção de toda a informação sensível utilizando representações de frases. Este método substitui cada frase de um documento clínico por contrapartes semanticamente semelhantes obtidas de um espaço de representações criado a partir de um conjunto de dados desidentificado, garantindo que nenhuma informação sensível permaneça no documento final. Avaliamos o desempenho de diferentes modelos utilizando métricas que avaliam tanto a sensibilidade da anonimização quanto a retenção de informação clínica.

Os nossos resultados indicam que a substituição de frases preserva mais eficazmente a informação médica relevante em comparação com a estratégia de substituição de palavras proposta anteriormente, que, embora melhor em termos de sensibilidade de anonimização, muitas vezes compromete a retenção de informações clínicas. O melhor modelo de representações de frases obteve ganhos entre 20 a 32% em relação ao modelo de representações de palavras testado nas métricas de retenção de informação clínica. Esta pesquisa oferece uma direção promissora para melhorar a desidentificação automatizada na prática clínica.

Palavras-chave: Processamento de Linguagem Natural, Informação de Saúde Protegida, Desidentificação, Anonimização, Representações de Frases

Agradecimentos

To my supervisors, Prof. Henrique Lopes Cardoso and Vitor Rolla, for the support, feedback and knowledge sharing.
To Fraunhofer Portugal AICOS, for the support provided during this project. Specifically to Bruno, João and Duarte for their close monitoring and involvement.
To my sister, my parents, and the rest of my family, for everything.
To my friends, for all the shared moments.

Vasco Marinho Rodrigues Gomes Alves

"You should be glad that bridge fell down.
I was planning to build thirteen more to that same design"

Isambard Kingdom Brunel

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CRF	Conditional Random Fields
EHR	Electronic Health Records
ELMo	Embeddings from Language Models
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
HIPAA	Health Insurance Portability and Accountability Act
i2b2	Informatics for Integrating Biology and the Bedside
ICD-10	International Classification of Diseases, Tenth Revision
LLM	Large Language Model
LSTM	Long-Short Term Memory
NER	Named Entity Recognition
NLP	Natural Language Processing
PHI	Protected Health Information
PII	Personally Identifiable Information
POS	Part-of-speech
Regex	Regular Expression(s)
RNN	Recurrent Neural Network

Chapter 1

Introduction

De-identification refers to the removal of personally identifiable information (PII) from records or datasets so that individual persons cannot be identified. In the clinical domain, de-identification typically consists of removing protected health information (PHI) from clinical notes or health records. PHI is a type of PII that specifically pertains to health information. It is any information in a medical context that can identify an individual and is related to their health status, provision of healthcare, or payment for healthcare services. With the recent increase in the usage of digital tools by health institutions, clinical notes and health records have become largely available electronically and need to be handled with caution, as they often contain sensitive information about patients and healthcare practitioners [36]. Automated solutions based on Artificial Intelligence (AI) and Natural language Processing (NLP) methods have been explored to de-identify Electronic Health Records (EHR), which contain large amounts of unstructured text data and are therefore impossible to be manually de-identified. The purpose of de-identification is to allow the usage of healthcare data for research and development by different entities and agencies.

1.1 Context

This master's thesis is the result of a collaboration between Fraunhofer Portugal and Faculdade de Engenharia da Universidade do Porto. Fraunhofer Portugal is a non-profit private association funded by Fraunhofer-Gesellschaft, the largest organization for applied research in Europe. It has a presence in the Health, Information and Communications Technology (ICT) & Electronics, Manufacturing, and Public Administration sectors.

The INCOGNITUS [54] platform was developed with the healthcare industry in mind, as it is a privacy-demanding area due to the high volume of personal and sensitive information contained in health records. It allows the user to upload the text they want to de-identify and select one of three methods for such. Two of them are based on Named Entity Recognition (NER), in which a trained classification model identifies sensitive entities, and these are then replaced by categorized tags. The third method relies on a word embedding model that replaces every word in the text with one of the nearest ones in the embedding space.

1.2 Motivation

De-identifying clinical text is crucial to mitigate privacy concerns when dealing with sensitive clinical data in Electronic Health Records and clinical notes. The General Data Protection Regulation (GDPR) [15] and the Health Insurance Portability and Accountability Act (HIPAA) [64] are two pieces of legislation that regulate the handling and sharing of personal and health data and are highly relevant in the context of de-identification.

Manual de-identification is a laborious task that requires substantial human and time resources, so there is a need for automatic solutions. Solutions based on AI and NLP have been developed over the years and have achieved great performance in this task. However, there is a lack of adoption of these systems for real-world use cases, and this area of research remains mainly academic. It is not guaranteed that these systems can reliably detect the personal information contained in the datasets to which they are applied, and performance may vary across different types of clinical notes. There is a need for solutions that health institutions can effectively use. Still, it should also be considered if perfect de-identification, i.e., removing all the sensitive information while keeping the non-sensitive information intact, is a realistic goal [57].

Emphasizing this issue, Abdalla et al. [1] introduced an innovative solution involving proximity measures between word embeddings. Their method replaces each token in a clinical note with a semantically similar one, ensuring the removal of sensitive information. However, this approach raises concerns about potential information loss and readability issues. This method, known as K-Nearest Embeddings Obfuscation (KNEO), is one of the strategies implemented in the INCOGNITUS toolbox.

This work follows their approach but extends it to the usage of sentence embeddings, replacing full sentences instead of words. Additionally, new and adapted metrics for anonymization sensitivity and clinical information loss will be utilized.

1.3 Objectives

The principal objective of this work is to develop a method for clinical text de-identification that relies on sentence replacement to remove the sensitive information contained in them. As the INCOGNITUS platform is already developed and running, the idea is not to create a system from scratch but instead to integrate this new method into the platform. Our strategy is to replace the original sentences with sentences that do not contain sensitive information but are semantically similar, in order to maintain the relevant medical information, coherence, and readability of the clinical notes.

For this approach, an extensive vector database will be created from de-identified clinical and biomedical text, where each vector corresponds to a sentence from such text. Every sentence will be stored in its original and encoded format through the usage of an embedding model. An extensive, de-identified and publicly available clinical dataset will be used to create this vector database, the MIMIC-III [23] dataset.

To de-identify a clinical note, each sentence is encoded into the embedding space and replaced by one of the closest in the embedding space. After this process, we obtain a different clinical note, but hopefully, its clinical and medical information will be retained. Furthermore, as our vector database does not contain sentences with personal and sensitive information, the new version of the clinical note will also not contain it, as every sentence is being replaced.

With this strategy, we aim to answer the following research questions:

- Can sentence embeddings and embedding space-based replacements retain clinical and medical data while ensuring anonymity?
- Is sentence replacement an improvement over word replacement regarding clinical information retention?

This solution could potentially be useful in terms of data uniformization. By using a sentence database, a more standardized dataset could be generated, as the sentences are coming from the same source, which could benefit model training for future downstream tasks, such as clinical information extraction. Additionally, it aims to tackle the flaws of the existing methods, such as the failure to detect some of the sensitive information for the NER models and the information loss and lack of readability for the word replacement approach.

1.4 Contributions

A text anonymization pipeline based on sentence substitution was successfully implemented on top of the existing anonymization toolkit's code and structure. This toolkit could be made available for health and research institutions in the future, aiming to facilitate the sharing and secondary usage of clinical data. Additionally, the realization of this work in collaboration with Fraunhofer Portugal AICOS resulted in the contribution as a co-author to the writing of a scientific paper [51] and in the accepted submission of a short paper to the PrivateNLP@ACL 2024[1] workshop as the main author.

1.5 Document Structure

After this introductory chapter, the document consists of the following chapters:

- Chapter 2 provides an overview of the state-of-the-art in clinical de-identification.
- Chapter 3 gives an outline of what embeddings are, their usage in NLP tasks and some of the most used models.
- Chapter 4 describes the INCOGNITUS platform.

[1] https://sites.google.com/view/privatenlp/

- Chapter 5 explains the methodology used for the de-identification of clinical text and outlines the evaluation metrics used.
- Chapter 6 presents the obtained results and discusses them.
- Chapter 7 concludes the work, discusses limitations, and proposes lines for future work.

Chapter 2

De-identification of Clinical Text

This chapter explores the application of NLP techniques to the problem of text de-identification in the clinical domain. Several methodologies will be analyzed, from simpler rule-based systems and traditional machine learning techniques to more complex deep learning architectures and transformer-based models. Different evaluation methods and metrics will be discussed to analyze the systems' performances. Additionally, multiple replacement strategies for the detected PHI terms will be evaluated, alongside their impact on subsequent tasks for which the clinical text may be used. Finally, the capability of generalizing these solutions to different datasets or languages in order to ensure real-world usage will also be covered.

2.1 Overview

With the recent increase in the use and adoption of EHR systems, significant amounts of patient clinical data have become available to be used by clinicians, researchers, and for operational purposes [36]. This has led to an advancement in the technologies being used and an improvement in medical investigations, diagnosis, and treatment processes, which results in a better healthcare service being provided to the patients [20].

However, the sharing of clinical data is severely limited due to the fact that it contains sensitive information about patients, doctors, institutions, and other entities. Consequently, sensitive information must be removed or obfuscated to allow the sharing of clinical data in order to preserve patient confidentiality [12].

Ensuring privacy when handling and sharing clinical data within EHR and clinical notes is crucial to complying with data protection regulations, such as the HIPAA [64] in the United States of America and the GDPR [15] in the European Union.

2.1.1 Health Insurance Portability and Accountability Act

The HIPAA Privacy Rule sets nationwide guidelines safeguarding people's medical records and personally identifiable health details, known as protected health information (PHI). It covers health

plans, healthcare clearinghouses, and providers engaging in specific electronic healthcare transactions [43]. The rule also permits the de-identification of PHI through two established methods [42]:

Expert Determination Involves a qualified expert applying generally accepted statistical and scientific principles and methods to ensure that the risk of re-identification of individuals is very small.

Safe Harbor Requires the removal of 18 specific identifiers of the individual or of relatives, employers, or household members of the individual, and ensuring that the information cannot be used alone or in combination with other information to identify the individual.

The following identifiers must be removed when utilizing the Safe Harbor method [42]:

- Names
- All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes
- All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age
- Telephone numbers, fax numbers, email addresses, Web Universal Resource Locators (URLs), Internet Protocol (IP) addresses
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers, serial numbers, Social Security numbers, medical record numbers, health plan beneficiary numbers, account numbers and certificate/license numbers
- Biometric identifiers, including finger and voice prints
- Full-face photographs and any comparable images
- Any other unique identifying number, characteristic, or code

The Privacy Rule allows the use and disclosure of de-identified health information for secondary purposes, as it is no longer considered protected health information.

2.1.2 General Data Protection Regulation

The GDPR classifies personal health data as a special category of data that needs robust data protection safeguards in order to be protected [59]. Its secondary usage is prohibited without individual consent. However, there is an exception to the ruling in Recital 26, as it states that "The principles of data protection should therefore not apply to anonymous information, namely

information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable." [45] [11].

Accordingly, once the data has been appropriately anonymized and individuals are no longer identifiable, it no longer falls within the scope of GDPR. It can, therefore, be used for secondary purposes.

2.1.3 Anonymization vs De-identification

In 2019, Chevrier et al. [10] conducted a review of the usage of the terms "anonymization" and "de-identification" in the literature. The authors suggest that the appropriate usage of the terms should be incentivized since several publications use the terms interchangeably and do not provide any definitions.

Lison et al. [28] define the two terms as follows:

Anonymization Complete and irreversible removal from a dataset of any information that, directly or indirectly, may lead to a subject's data being identified.

De-identification Process of removing specific, predefined direct identifiers from a dataset.

The GDPR does not provide an actual definition for "anonymization" or "de-identification" but specifies the requirements for data to be considered anonymized, as mentioned in Section 2.1.2. However, it defines "pseudonymization" as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person." [44].

Anonymized data and pseudonymized data differ in their classification under the GDPR. While anonymized data is not considered personal data, pseudonymized data still falls within the realm of personal data as per GDPR guidelines. Preserving the distinction between these two concepts is essential under this regulation [62].

On the other hand, the HIPAA Privacy Rule clearly states and describes the methods from which de-identification can be achieved and lists the 18 types of PHI that must be removed in order for the data to be considered de-identified when using the Safe Harbor method, as mentioned in Section 2.1.1.

2.2 Challenges and Data

To foster research in the area of clinical data de-identification, two significant events were created: the i2b2 (Informatics for Integrating Biology and the Bedside) de-identification tracks of 2006 [65] and 2014 [57]. These challenges focused on the de-identification of unstructured clinical data and resulted in the submission of systems with impressive results.

The data for the 2006 i2b2 de-identification challenge was composed of medical discharge summaries. Eight categories of PHI were present in the dataset: *Patients, Doctors, Hospitals, IDs, Dates, Locations, Phone Numbers,* and *Ages*. The goal of this challenge was to remove the PHI while maintaining the integrity of the data. It was composed of 889 records, from which 669 were used for training and the remaining 220 for testing. For the challenge, the records were tokenized, broken into sentences and converted into XML representation [65]. Table 2.1 provides an overview of the corpus, and Figure A.1 contains an example clinical note.

PHI Category	**Complete Corpus**	
	Instances	**Tokens**
Non-PHI	-	444 127
Patients	929	1 737
Doctors	3 751	7 697
Locations	263	518
Hospitals	2 400	5 204
Dates	7 098	7 651
IDs	4 809	5 110
Phone Numbers	232	271
Ages	16	16

Table 2.1: Distribution of instances and tokens in the 2006 i2b2 de-identification corpus (obtained from [65]).

Authors of the 2014 i2b2/UTHealth de-identification track used the HIPAA-PHI categories as a starting point and augmented them to obtain the following i2b2-PHI categories and types [57]:

- NAME (types: PATIENT, DOCTOR, USERNAME)
- PROFESSION
- LOCATION (types: ROOM, DEPARTMENT, HOSPITAL, ORGANIZATION, STREET, CITY, STATE, COUNTRY, ZIP, OTHER)
- AGE
- DATE
- CONTACT (types: PHONE, FAX, EMAIL, URL, IPADDRESS)
- IDs (types: SOCIAL SECURITY NUMBER, MEDICAL RECORD NUMBER, HEALTH PLAN NUMBER, ACCOUNT NUMBER, LICENSE NUMBER, VEHICLE ID, DEVICE ID, BIOMETRIC ID, ID NUMBER)

The corpus was composed of 1,304 individual records from 296 patients, with an average of 617.4 tokens per file [58]. Table 2.2 contains the distribution of the i2b2-PHI categories in the corpus, and Figure A.2 exemplifies the clinical text present in the challenge, with the appropriate XML tags (simplified version).

PHI Category	# in training data	# in test data	Total # in corpus
NAME: PATIENT	1,316	879	2,195
NAME: DOCTOR	2,885	1,912	4,797
NAME: USERNAME	264	92	356
PROFESSION	234	179	413
LOCATION: HOSPITAL	1,437	875	2,312
LOCATION: ORGANIZATION	124	82	206
LOCATION: STREET	216	136	352
LOCATION: CITY	394	260	654
LOCATION: STATE	314	190	504
LOCATION: COUNTRY	66	117	183
LOCATION: ZIP CODE	212	140	352
LOCATION: OTHER	4	13	17
AGE	1,233	764	1,997
DATE	7,507	4,980	12,487
CONTACT: PHONE	309	215	524
CONTACT: FAX	8	2	10
CONTACT: EMAIL	4	1	5
CONTACT: URL	2	0	2
CONTACT: IPADDRESS	0	0	0
ID: SSN	0	0	0
ID: MEDICAL RECORD	611	422	1,033
ID: HEALTH PLAN	1	0	1
ID: ACCOUNT	0	0	0
ID: LICENSE	0	0	0
ID: VEHICLE	0	0	0
ID: DEVICE	7	8	15
ID: BIO ID	1	0	1
ID: ID NUMBER	261	195	456
Total # of tags	**17,410**	**11,462**	**28,872**
Average PHI per file	**22.03**	**22.3**	**22.14**

Table 2.2: PHI distributions in the 2014 i2b2/UTHealth de-identification corpus (obtained from [58]).

A more recent de-identification challenge was the 2016 CEGS N-GRID shared tasks Track 1, which focused on the de-identification of a new corpus of 1,000 psychiatric intake records [56]. It was divided into two sub-tracks: one focused on how well existing systems generalize to new data by making nine teams run existing de-identification systems, without any modifications or training, on new records, and the other was the traditional de-identification task where the participating teams could train and test their systems. The PHI categories for this track were the same as the ones for the 2014 i2b2/UTHealth de-identification challenge. However, the 2016 corpus contained three times as many tokens per record compared to the 2014 corpus due to the extensive notes psychiatrists take about the patients [56].

Another dataset that is commonly used for health-related tasks and clinical studies is the

MIMIC-III[1] clinical database. It comprises information about patients admitted to critical care units at a large tertiary care hospital, and before being incorporated into the dataset, this information was de-identified in accordance with the HIPAA standards so that PHI was removed from the text. MIMIC-III is distributed as a collection of CSV files, and researchers must complete a course in protecting human research participants and sign a data user agreement to access it [23].

2.3 De-identification Systems

De-identification is an intrinsically difficult task due to the large volumes of unstructured data available from clinical notes. Manual de-identification is an extremely laborious task, so automated systems must be developed and adopted. Most authors and researchers treat de-identification as a NER problem, where the entities are the personal information that needs to be detected and masked, such as names, IDs, contact information, and dates, among others. [74, 28].

Early de-identification systems were based on pattern matching and dictionary look-ups. Machine learning algorithms then improved on those simple approaches, while lately, the focus has shifted towards the use of deep learning techniques.

2.3.1 Manual

Having experts manually perform the de-identification of data is not a feasible option, as it is an exhaustive task that would require substantial human and temporal resources. It typically requires many annotators, resulting in a performance that may be highly variable and prone to errors [40].

Dorr et al. [14] evaluated the time cost for manual de-identification and concluded it was a tedious and time-consuming task, as manually de-identifying a note took 87.3 ± 61 seconds on average.

Human annotators are, however, often used to annotate data for tasks or challenges. Uzuner et al. [65] use an automatic system in the first stage and three annotators in the second stage to mark PHI tags in the 2006 i2b2 de-identification challenge. Stubbz and Uzuner [58], for the 2014 i2b2/UTHealth corpus, had six annotators and randomly assigned each patient's records to two independent annotators for them to work in parallel.

2.3.2 Rule-Based Systems

Rule-based systems typically rely on hand-crafted patterns and dictionaries/lists, often with limited generalizability. Regular expressions are frequently used, mainly to detect personal information that has standardized formats, such as emails or ZIP codes. Dictionaries can be built and used to look up terms that are usually considered personally identifiable information, such as names and locations, or, following an opposite approach, use a biomedical thesaurus and classify the terms contained in them as non-sensitive [36].

[1] https://physionet.org/content/mimiciii/1.4/

In 1996, Sweeney [60] proposed the Scrub System, which used templates (e.g., phone numbers and dates) and local knowledge sources, such as lists of first names or U.S. states. It uses parallel PHI-detection algorithms, one for each category. The algorithm with the highest precedence and certainty prevails, and its results can be shared with the other algorithms. The Scrub System was able to find 99-100% of the sensitive information in a database of 275 patient records and 3 198 letters to referring physicians.

Beckwith et al. [5] created a HIPAA-compliant de-identification system for free-text clinical notes. They implemented 50 regular expressions to identify patterns that commonly represent PHI, such as dates, addresses, and emails. Additionally, regular expressions to detect terms after the prefix "Dr" or "Doctor" are also implemented since it is very likely that a name follows them. The system was able to remove 98.3% of the unique identifiers but with a low precision (4 671 over-scrubs).

Similarly, the Medical De-identification System (MeDS) [18], developed by Friedlin et al., uses several regular expressions and lists. However, the authors implement additional processes to deal with ambiguous names (a name might be a non-name in a different context) and misspellings, which list implementations often have trouble dealing with. When evaluating 7 193 surgical pathology reports, 99.47% of the HIPAA identifiers were detected.

On the 2006 i2b2 de-identification challenge, the top-performing systems were based on machine learning but complemented with regular expressions templates [65]. The same was observed on the 2014 i2b2/UTHealth de-identification challenge [57].

Rule-based approaches require no labeled data (only for evaluation) and can easily be changed, and new rules can conveniently be added to detect new PHI tokens and improve performance [12]. It is also easy to know why the system classified a token as PHI or non-PHI, as it necessarily falls within at least one of the rules in order to be labeled as sensitive [36].

However, methods based on rules also have some significant disadvantages. For example, PHI that does not fall into the defined rules will always go undetected. It is also necessary to account for the multiple patterns that can occur for a PHI category, which requires different and complex algorithms in order to be detected [36]. These systems are also sensitive to misspellings and abbreviations and don't take context into account, resulting in bad performance when ambiguous terms are present [12]. Another problem is the generalizability since many of the rules are fine-tuned to one particular type of data and, therefore, may not apply to a different system [30].

For example, if a character is missing on a city name, it will not be found in the list of cities being used and, therefore, go undetected. Additionally, PHI and non-PHI can overlap and result in ambiguous terms. For instance, "Alzheimer" can be both the name of a disease or the name of a person. The former should not be considered PHI, but the latter should.

2.3.3 Machine Learning Systems

The 2006 i2b2 de-identification challenge [65] saw the first machine learning solutions being constructed for this problem. These are normally based on supervised machine learning methods,

where large amounts of labeled data with extractable features are available for a training phase [36, 9].

Conditional Random Fields (CRF) [25] are the most used method and took over as the best-performing systems in the 2006 and 2014 i2b2 de-identification challenges, but Support Vector Machines, Decision Trees, and Maximum Entropy are also commonly found. The algorithms are trained on a large corpus of annotated text and are accompanied by several feature engineering techniques in order to produce lexical, syntactic, and semantic features [36, 9]. These features capture different aspects of the text, such as the morphology, casing, symbols, and part-of-speech (POS) of the words [74].

These systems were also often complemented with regular expression and dictionary lookup modules, in addition to the main machine learning model [54]. It allowed the system to detect sensitive information matching the regular expressions or included in searchable lists that could have otherwise gone undetected by the machine learning algorithm.

One of the top-performing systems in the 2006 i2b2 de-identification challenge was developed by Wellner et al. [69] and was based on the named entity recognition toolkit Carafe[2]. Carafe is an implementation of Conditional Random Fields targeted at text processing tasks. Wellner et al. complemented it with regular expressions to capture PHI with more standardized formats, such as dates, and used features such as orthography, special characters and lexical context. The authors submitted two runs of this system but added lexical cues and dictionaries for people, locations and dates in the second run, which resulted in a performance improvement [65].

Aramaki et al. [4] also participated in the 2006 i2b2 de-identification challenge and used CRF with local, non-local, and extra-resource features. Local features include information about the target word and its surroundings, such as POS, casing, length, special characters, and regular expression matching. Non-local features relate to sentence attributes, such as length or position, and extra-resource features come from extra resources, such as person and location dictionaries.

Szarvas et al. [61] proposed a NER model that used a decision tree learning algorithm with local features and dictionaries. It contained orthographical features (capitalization, word length, information about word form, regular expression matching, etc.), frequency information, phrasal information (preceding words, suffixes), dictionaries (names, geographical locations, etc.), and contextual information (sentence position, quotation marks, etc.). They implement an iterative learning process that uses Boosting and C4.5 to train three similar classifiers and use a decision function to obtain a final prediction of whether a certain token is PHI or not.

Uzuner et al. [76] de-identify medical discharge summaries using SVMs and local context to classify words as one of seven PHI categories or non-PHI. Their system uses orthographic, syntactic, and semantic features of the target and its surrounding words in order to capture contextual clues. This representation of local context allows the system to de-identify PHI that contains out-of-vocabulary words or ambiguous terms.

Yang et al. [73] developed the winning system of the 2014 i2b2/UTHealth de-identification challenge using a hybrid model that combined machine learning techniques with keyword-based

[2] https://sourceforge.net/projects/carafe/

and rule-based approaches. They extract a wide variety of linguistic features from the text, such as token, contextual, orthographic, and discourse features. This feature set is then complemented with task-specific features, such as lists of names and regular expression template features. The authors use the CRF++ package to employ a CRF algorithm that deals with the PHI categories that are sufficiently present in the training data. Additionally, keyword lists and regular expression patterns are manually generated. Finally, a post-processing step is implemented to correct wrong PHI identifications or find potential PHI candidates that were not identified. This step involves the creation of trusted PHI term lists, which are unambiguous terms that the system should also consider as PHI. The full system pipeline is illustrated in Figure 2.1.

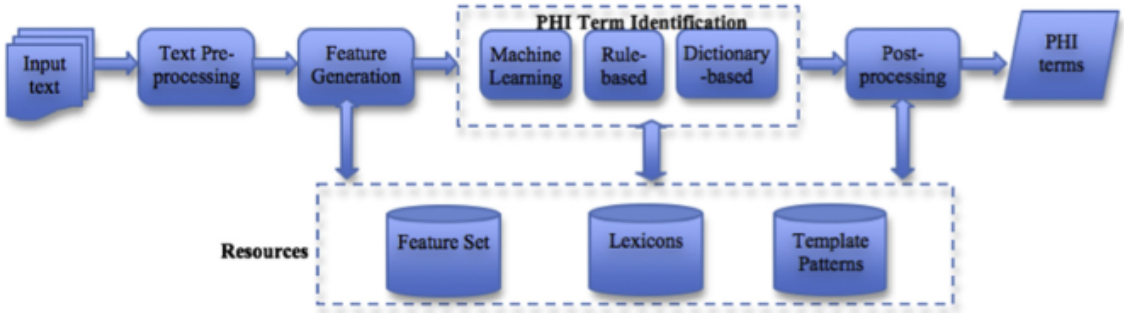


Figure 2.1: Pipeline of the 2014 i2b2/UTHealth de-identification challenge winning system (obtained from [73]).

Machine learning models are able to better generalize to new clinical text than most rule-based approaches and, therefore, can better identify ambiguous terms or PHI that is not present in the existing dictionaries and lists [74]. These systems are also able to recognize complex PHI patterns and keep the processing speed over time [36].

A downside of machine learning systems is that they need large amounts of labeled data in order to be properly trained and achieve good performance, as opposed to rule-based systems. It is also difficult to know why the system is classifying a term as PHI or non-PHI, and when an error is being made, adding more training data will not necessarily contribute to its correction. These systems also depend heavily on the quality of the features they are given during the training phase.

2.3.4 Deep Learning Systems

The appearance of deep learning methods revolutionized the NLP area and resulted in state-of-the-art performances for many tasks, including clinical NER [9, 74]. Neural networks and embeddings are two important elements of these deep learning systems.

A commonly used Recurrent Neural Network (RNN) for the de-identification problem is the Long Short-Term Memory (LSTM) strategy, which has a memory cell that can hold information. As such, LSTM networks are capable of learning long-term dependencies and the context of the text data. These are frequently implemented in a bidirectional way, which allows them to process the text in forward and backward directions, contributing to a better capture of dependencies.

Embeddings are vectorial representations of data that are useful for machine learning models to capture meaningful information about it. In NLP-related tasks, these are often representations of characters, words, or sentences and are used to capture semantic and syntactic similarities [35]. The smaller the distance between the vectors in the embedding space, the more related they are.

Liu et al. [30] proposed an ensemble method composed of a bidirectional LSTM without features, a bidirectional LSTM with features, a CRF classifier, and rules. The output of these four different subsystems is combined to obtain the PHI instances. This system outperformed the best model developed for the 2014 i2b2/UTHealth de-identification challenge and obtained first place in the 2016 CEGS N-GRID de-identification track.

In 2021, Catelli et al. [9] implement a Bi-LSTM + CRF architecture to de-identify Italian medical records. The Bi-LSTM obtains the overall representation of the context of the sentence at every word by concatenating the left and the right context representations obtained by the two unidirectional LSTM, which is then passed to the CRF layer for tagging in IOB format. The O-tag represents tokens that are not part of any PHI instance, the B-tag represents the beginning of a PHI instance and the I-tag is attributed to the tokens inside a PHI instance.

Yang et al. [74] use an LSTM-CRF model and compare five different word embeddings to evaluate de-identification results when training and testing on different datasets. Embeddings trained on general English text obtained better performance for de-identification than other embeddings trained on clinical and biomedical text, which is surprising but justified by the fact that the clinical and biomedical text had gone through a de-identification process, so many of the PHI from the input text was not found in those embeddings.

Dernoncourt et al. [12] present an approach consisting of an LSTM with three layers: a character-enhanced token embedding layer, a label prediction layer, and a label sequence optimization layer. These layers are responsible for mapping each token into a vector representation, obtaining the probability of each label for each token, and outputting the most likely predicted labels, respectively. This system outperforms a baseline CRF model and the best-performing system from the 2014 i2b2/UTHealth de-identification challenge. According to the authors, it presented more flexibility when dealing with language variations.

A novel and interesting approach was proposed by Abdalla et al. [1], which suggests the use of word embeddings in a way that achieves 100% recall on the removal of sensitive information. They argue that existing solutions based on NER techniques can never guarantee the removal of all sensitive information, as these methods are never perfect. To solve this issue, they suggest replacing every token with a similar token obtained from the word embedding space. This strategy assures that all the sensitive information is removed as every token is being replaced, and therefore, no original tokens are maintained. However, it comes at the cost of readability, as the precision is very low, and all the tokens that are not sensitive will still be replaced.

Over the years, embeddings and neural network architectures have significantly evolved, which resulted in the appearance of largely capable Language Models, such as ELMo (Embeddings from Language Models) [48], GPT (Generative Pre-trained Transformer) [52] and BERT (Bidirectional Encoder Representations from Transformers) [13].

In 2022, Meaney et al. [35] explored the performance of different transformer models in the 2014 i2b2/UTHealth de-identification challenge. The authors compare six different BERT variants: BERT-Base, BERT-Large, RoBERTa-Base, RoBERTa-Large, ALBERT-Base and ALBERT-XXLarge. These models were fine-tuned and tested on the 2014 i2b2/UTHealth corpus (using a random split), and it was observed that larger models performed better than their smaller counterparts within the same class. RoBERTa-Large was the best-performing model, and the authors highlight the importance of hyperparameter tuning. Table 2.3 illustrates the F_1-score obtained in the test set.

Model	**Test F_1-score**
RoBERTa-Large	0.9675
ALBERT-XXLarge	0.9644
BERT-Large	0.9543
RoBERTa-Base	0.9522
BERT-Base	0.9410
ALBERT-Base	0.9386

Table 2.3: F_1-score of different transformer models in the 2014 i2b2/UTHealth de-identification challenge (adapted from [35]).

Alsentzer et al. [3] demonstrate that training BERT with domain-specific data, such as clinical notes or biomedical literature, improves its performance across different clinical/medical tasks. The authors train two varieties of BERT on the MIMIC-III dataset: Clinical BERT, which uses all note types, and Discharge Summary BERT, which uses only discharge summaries. Additionally, they train two varieties of BioBERT [27] using the same strategy: Clinical BioBERT and Discharge Summary BioBERT. Table 2.4 shows the F_1-score each model obtained for the 2006 and 2014 i2b2 de-identification tasks. BioBERT was the best-performing model, and Clinical BERT performed worse than general BERT. Models initialized from BioBERT also showed better performance than when initialized from BERT.

Model	**i2b2 2006**	**i2b2 2014**
BERT	93.9%	92.8%
BioBERT	**94.8%**	**93.0%**
Clinical BERT	91.5%	92.6%
Discharge Summary BERT	91.9%	92.8%
Clinical BioBERT	94.7%	92.5%
Discharge Summary BioBERT	94.8%	92.7%

Table 2.4: F_1-score of different BERT variants in the 2006 and 2014 i2b2 de-identification challenges (adapted from [3]).

DeID-GPT [31] is a framework enabled by GPT-4 that automatically identifies and removes identifying information. The HIPAA guidelines and PHI identifiers are incorporated into the designed prompts, which are sent to the large language model (LLM) along with the original clinical reports to generate de-identified reports. This approach does not require any code or procedural

changes when being applied to different scenarios but instead relies on good prompt design. There is, however, a big limitation, as the GPT models can only be accessed through APIs, which results in the data being transmitted to and stored by an external party.

Deep learning architectures have demonstrated better performance in many NLP tasks in the clinical domain, such as de-identification [55, 72]. They also do not require time-consuming feature engineering, as the algorithms are capable of capturing various useful features automatically [74].

However, deep learning architectures are often complex and harder to interpret, and their training and fine-tuning are normally more computationally and time-demanding.

2.4 Evaluation

After the detection of the sensitive information to be de-identified, it is necessary to evaluate the performance of the system. This is commonly done by comparing the data classified as sensitive by the system against gold standard annotations produced by field experts and then measuring the overlap.

2.4.1 Metrics

Metrics such as precision, recall, and F-score are typically used since they provide an accurate and understandable overview of the system's performance.

Precision reflects the percentage of correctly identified sensitive data in relation to the total number of data identified as sensitive by the system:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2.1}$$

Recall is the percentage of correctly identified sensitive data in relation to the existing sensitive data:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2.2}$$

The recall of the systems needs to be balanced against its precision. Low recall results in a lot of sensitive information not being detected, and low precision results in the corruption of non-sensitive information [67]. As such, recall can be viewed as measuring the degree of privacy protection, while precision can be seen as reflecting the data utility [49].

The use of F-scores assesses this balance. F_1-score is the most used, as it is the harmonic mean between precision and recall:

$$F_1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.3}$$

Some authors and researchers ([6, 17, 16]) have suggested that recall is the most decisive metric, as privacy should be the priority, and propose to assign more weight to the recall of the system, which results in using a F_β-score different than the traditional F_1 [49].

These metrics can be measured at the macro and micro levels. Micro score calculations evaluate all the instances across classes as a single set, while macro score calculations evaluate each class independently and then average the scores across all classes in the corpus [58].

2.4.2 Strategies

When evaluating the correctness of the system, multiple strategies can be employed and combined. Since much of the sensitive information we are trying to detect can have different formats or lengths, we can evaluate it at an entity or token level:

Entity-based Also known as instance-based, it requires that the system detects the exact beginning and end locations of the sensitive information. The detected entity must be an exact match of the annotated label.

Token-based This approach requires that the system simply detects the tokens contained in a sensitive information instance.

For example, if the gold standard has "Vasco Alves" annotated as a name, an entity-based evaluation would require the system to detect "Vasco Alves" as a name, whereas a token-based evaluation would allow the detection of the individual tokens "Vasco" and "Alves" as names [57].

Furthermore, one can choose to assess the system's performance based on the simple detection of sensitive information or also on the correct identification of the category of such information:

Category Requires the system to correctly identify the category/class/type of the detected sensitive information.

Binary Simply requires the system to classify the information as sensitive vs non-sensitive.

When it is important to preserve the integrity of the data, the correct identification of categories should be accomplished. However, de-identification can still be successful even if the correct categories are not identified. For example, a system that incorrectly classifies a name as a location will still successfully de-identify the data [57].

On the 2006 i2b2 de-identification challenge [65], the authors used precision, recall, and F-measure to evaluate the systems submitted, both at the entity and token levels. They also analyze the ability to differentiate between PHI categories or to distinguish only from PHI and non-PHI. The authors of the 2014 i2b2/UTHealth de-identification track [57] use micro-averaged entity-based F_1-score over i2b2-PHI categories as the primary metric for the system's comparison.

Table 2.5 shows what system outputs would be considered correct or incorrect identifications according to entity or token-based evaluation, over the categories or simply binary. Entity-based evaluation is stricter than token-based, and binary evaluation accepts any detected PHI independently of the type.

Gold standard: <LOCATION>Brooks Infirmary</LOCATION>	**Entity-based**		**Instance-based**	
	PHI categories	**PHI/not PHI**	**PHI categories**	**PHI/not PHI**
<LOCATION>Brooks Infirmary</LOCATION>	✓	✓	✓	✓
<NAME>Brooks Infirmary</NAME>	✗	✓	✗	✓
<LOCATION>Brooks</LOCATION>	✗	✗	✓	✓
<NAME>Brooks</NAME>	✗	✗	✗	✓

Table 2.5: Correctness under the different evaluation strategies (adapted from [57]).

2.5 Sensitive Information Removal

The de-identification process consists of identifying the sensitive information and subsequently obscuring it in some fashion [6]. Several strategies can be employed, ranging from straightforward removal to surrogate substitution. Lothritz et al. [32] have proposed and evaluated six different strategies based on the usage of generic tokens or random names. Similarly, Berg et al. [6] explore four different concealment strategies in their study: *Pseudo*, *Class*, *Mask* and *Removal*.

Here are described some of these strategies:

Universal Tag This approach consists of replacing every detected entity with an identical tag.

Class Tag Entities are replaced with a tag according to their respective classes.

Pseudonymization This method involves replacing every detected entity with a realistic surrogate according to its type.

Term Removal This strategy consists of simply removing the identified PHI terms from the sentences containing them.

Sentence Removal Completely remove the PHI terms identified and the sentences in which they are contained.

Table 2.6 outlines the application of the described removal strategies to the original sentence: "The patient Vasco is 23 years old.".

Original	**"The patient Vasco is 23 years old."**
Universal Tag	"The patient <ENTITY> is <ENTITY> years old."
Class Tag	"The patient <NAME> is <AGE> years old."
Pseudonymization	"The patient Vítor is 32 years old."
Term Removal	"The patient is years old."
Sentence Removal	" "

Table 2.6: Removal Strategies.

2.6 Beyond Traditional Metrics

The process of de-identification is an important component in managing EHR that contain sensitive and personal information. However, it is crucial that the usability of the data is maintained

for future tasks and applications. Furthermore, the generalization of the developed systems should also be considered to ensure their adoption across a wider range of health institutions.

2.6.1 Downstream Evaluation

Training an NLP model for a specific downstream task on de-identified data may lower the performance of the resulting model when compared to a model trained on the original data [32]. This is due to the fact that automated techniques are imperfect, which results in the introduction of noise into the data [66]. Assessing this aspect is crucial for potential secondary use of the de-identified data.

Information loss can be evaluated by analyzing the performance of the model on certain tasks when trained on the original data and comparing it to when trained on the de-identified data. Obeid et al. [41] evaluate the impact of de-identification on an ICD-10 code recognition task, using both traditional machine learning methods and deep learning models. They conclude that the impact of the de-identification results in a negligible difference in performance across both types of machine learning.

Vakili & Dalianis [66] follow a similar approach and examine the impact of the noise introduced in the de-identification process by training and testing BERT models. They train the models on pseudonymized and unaltered data and then evaluate their performance on two classification tasks and one NER task. The performance of the models was indistinguishable.

The degree of information loss is also dependent on the removal strategy (Section 2.5) applied to deal with the detected PHI. For example, Lothritz et al. [32] argue that de-identification does have a negative impact on the performance of NLP models, but it is relatively low. However, they conclude that pseudonymization techniques involving random names lead to better performance across most tasks.

Berg et al. [6] find that the choice of the concealment strategy has a large impact on downstream clinical NER tasks. Pseudonymization has the least impact, while removing the sentences containing PHI terms has a much higher negative impact.

The INCOGNITUS [54] platform incorporates a novel metric for information loss assessment. A pre-trained BioBERT [27] model fine-tuned on MIMIC-III [23] data is used to identify ICD-10 code categories both on the original and the de-identified versions of the documents. Information loss is then calculated by comparing the number of classes simultaneously present in both versions of the same document.

Vakili et al. [67] evaluate the performance on downstream tasks of the KB-BERT [33] model by training it using either the original or the de-identified data. Three versions were compared: training the model with the actual dataset, training the model with the pseudonymized dataset, and training the model using the dataset but with its sentences containing sensitive information removed. Six classification and NER tasks were used to compare the models, and it was concluded that de-identification does not lead to a discernable drop in performance. In fact, the pseudonymized version even outperforms the actual version in some tasks.

A more penalizing performance was covered by Abdalla et al. [1], which observed a decrease of up to 5% in F_1-score for different classification tasks when using the obfuscated text. However, it is important to notice that a word embedding approach was used to replace every token in the documents, not just the ones containing PHI.

Usually, the lower the precision, the bigger the information loss, as it means that more data is incorrectly classified as sensitive (false positives) and, as a result, obfuscated.

2.6.2 Generalization

Although systems and methods with excellent performance have been developed over the years, there is a lack of adoption in real-world scenarios. This can be attributed in part to the uncertainty that these systems will perform equally as well when presented with different types or formats of clinical notes since existing studies often utilize training and testing data from the same institution [74]. Many of these systems also need to be fine-tuned in accordance with the scenario they are being applied to, which limits their wider use [31]. Additionally, most of the top-performing systems are developed for English, a high-resource language. Therefore, the development of these systems is severely limited by the lack of resources, such as datasets, in other languages, which are consequently defined as low-resource languages [8].

When analyzing the systems submitted for the 2006 i2b2 de-identification challenge, Uzuner et al. [65] claim having strong reasons to believe that extrapolating the systems would be difficult, since many of them took advantage of the specific characteristics of the discharge summaries and the institution from which these were drawn. This led to changes in the 2014 i2b2/UTHealth de-identification task, where the data contained a wider variety of clinical records and PHI categories, which made it more challenging and the systems more robust [57].

Yang et al. [74] developed deep learning de-identification models using the 2014 i2b2/UTHealth corpus for training but evaluated them against a test corpus built using 500 clinical notes from the University of Florida instead. The performance dropped from entity-based and token-based F_1-scores of 0.9547 and 0.9646 when evaluated on the 2014 i2b2/UTHealth validation set, to 0.8568 and 8958, respectively.

With the objective of investigating the ability of methods to transfer knowledge between different languages for the de-identification task, Catelli et al. [8] created an Italian de-identification dataset from COVID-19 clinical records. They then explored four different training approaches (EN, IT, MIX, EN-IT), using the English 2014 i2b2/UTHealth de-identification corpus and the Italian COVID-19 de-identification corpus, with testing being performed always on the Italian dataset. For both a bidirectional LSTM + CRF and a Multilingual BERT (M-BERT[3]) architecture, training in English and testing in Italian did not obtain good results. For the BiLSTM + CRF model, the best approach was to train first with the high-resource language (English) and then with the low-resource language (Italian). In contrast, for the M-BERT model, training in Italian provided the best results.

[3] https://github.com/google-research/bert/blob/master/multilingual.md

Recently, some efforts have been made to create clinical corpora for languages other than English. Miranda-Escalada et al. [39] created the CodiEsp[4] corpus for ICD-10 code assignment, and Marimon et al. [34] prepared the MEDDOCAN[5] corpus, to be used in a de-identification track. Both corpora are provided in Spanish with gold standard annotations.

Using Multilingual BERT, Pires et al. [50] perform experiments to study the generalization of linguistic representations across languages by fine-tuning the model using task-specific training data from one language and evaluating the same task in a different language. The obtained results show that high lexical overlap between languages and similar typologies (subject/object/verb order and adjective/noun order) improves cross-lingual generalization.

2.7 Summary

In this chapter, we provide an overview of the state-of-the-art in the de-identification of clinical text, exploring the many factors and concerns that are taken into account. Initially, we highlight the importance of the problem and analyze its framing in accordance with two major regulations. We describe major events that boosted interest in this topic and datasets that are available for researchers to address this problem.

We identify and describe the multiple strategies that are commonly employed when developing de-identification systems, as well as their advantages and disadvantages. We examine their evolution throughout the years according to innovations in the AI and NLP fields.

Furthermore, we discuss the different evaluation techniques that can be applied to this problem, including different metrics and strategies. We also discuss the best approach to removing sensitive information, as we want to maintain the usability of the documents.

Finally, we take a look at other aspects that should be considered when developing the de-identification systems, as they might influence and limit their adoption by health institutions.

Table 2.7 summarizes some of the systems described in the previous sections. It is important to mention that these systems often differ in implementation, training, and testing data, and sometimes even in terms of the identifiers they aim to detect, so direct comparison should be performed with caution.

[4] https://zenodo.org/records/3837305
[5] https://zenodo.org/records/4279323

Name/Authors	**Strategy**	**Train**	**Evaluation**	**Results**
Scrub System [60]	Regex + Lists	-	3 473 Medical Documents	99-100% Recall
MeDS [18]	Regex + Lists	-	7 193 Surgical Reports	99.47% Recall
Wellner et al. [69]	CRF + Regex + Dictionaries	i2b2 2006 Train Set	i2b2 2006 Test Set	97.36% F_1-score
Szarvas et al. [61]	DT + Dictionaries	i2b2 2006 Train Set	i2b2 2006 Test Set	99.75% F_1-score
Yang et al. [73]	CRF + Regex + Lists	i2b2 2014 Train Set	i2b2 2014 Test Set	93.6% F_1-score
Liu et al. [30]	Bi-LSTM + CRF + Regex	i2b2 2014 Train Set	i2b2 2014 Test Set	95.11% F_1-score
Liu et al. [30]	Bi-LSTM + CRF + Regex	2016 N-GRID Train Set	2016 N-Grid Test Set	91.43% F_1-score
Dernoncourt et al. [12]	Embeddings + Bi-LSTM	i2b2 2014 Train Set	i2b2 2014 Test Set	97.85% F_1-score
Abdalla et al. [1]	Replacement using Embedding Similarity	-	-	100% Recall
Alsentzer et al. [3]	BioBERT	i2b2 2006 Train Set	i2b2 2006 Test Set	94.8% F_1-score
Alsentzer et al. [3]	BioBERT	i2b2 2014 Train Set	i2b2 2014 Test Set	93.0% F_1-score

Table 2.7: Summary of different de-identification systems and their respective results.

Chapter 3

Embeddings

In this chapter, we explore the concept of embeddings and their critical role in NLP tasks. We begin by introducing word embeddings, which are vectorized representations of words that allow us to capture relationships between words. Then, we delve into sentence embeddings, which extend the concept to entire sentences, enabling more complex and nuanced text representations. We present different algorithms for generating word and sentence embeddings, such as Word2Vec, Doc2Vec, and Sentence Transformers, among others. These will be the ones used and compared throughout this work.

3.1 Overview

Converting text into representations that machine learning algorithms can use is a challenging but necessary step in most NLP tasks [24]. Embeddings are dense, distributed and fixed-length vectors of real numbers that represent pieces of text, such as words or sentences. The value of each dimension corresponds to a text feature that allows these representations to capture useful syntactic and semantic properties [63]. As a result, the vectors for semantically or syntactically related text pieces will be close to each other, and distant vectors represent differing meanings [24]. Additionally, such vectorial representations also allow the text pieces to be the subject of mathematical operations that wouldn't otherwise be possible [2], aiding in finding similarities between text pieces.

Embeddings are learned directly from running text in an unsupervised fashion, as they do not require any manually crafted features, thus saving effort and time in the domain-specific feature engineering and extraction typically performed in traditional NLP [24].

3.2 Word Embeddings

One of the most common forms of text representation in machine learning, and the most deployed in medical NLP, is word embeddings [46]. Word embeddings are based on the distributional hypothesis [19], which states that words that occur in the same contexts tend to have similar

meanings. As a result, it is expected that synonyms appear close to each other in the vector space, and non-related words are distant. Representing words as vectors allows us to perform arithmetic operations over them. One classic example is that if we were to subtract the vector for the word "man" from the vector for the word "king" and add the vector "woman", we would obtain a vector that is close to the vector for the word "queen": Vector("king") - Vector("man") + Vector("woman") ≈ Vector("queen") [38].

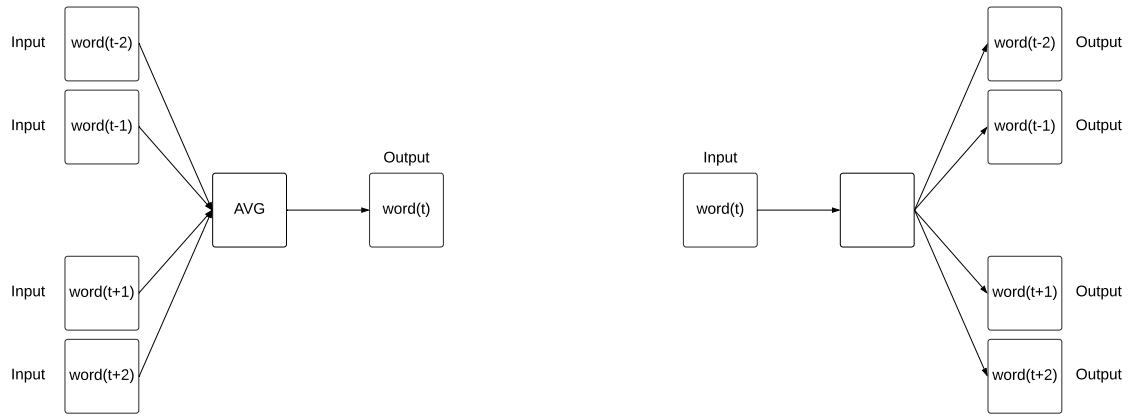
Although all word embedding techniques use context during the training stage, they can be categorized into two major subgroups: contextual and non-contextual. Once trained, non-contextual embedding approaches obtain a single fixed representation for each word, which does not change according to its actual surrounding context [29, 75]. For example, the word "mouse" will always be represented by the same vector, even if it is being used with two different meanings — computer object and animal. In contrast, contextual embeddings are capable of capturing the multiple meanings of the same word by using a representation based on an entire sequence, thus changing a word's representation according to its surrounding context. Contextual word embeddings typically outperform non-contextual ones and have obtained state-of-the-art results in many NLP problems [29].

Word2Vec [37] is a non-contextual word embedding algorithm based on neural networks that produce continuous vector representations of words by learning relationships between them using large amounts of plain text. The authors introduce two innovative model architectures, Continuous Bag-of-Words (CBOW) (Figure 3.1a) and Continuous Skip-gram (Figure 3.1b), which significantly improve the efficiency and accuracy of learning word representations from large datasets. The CBOW model predicts the current word based on the context of surrounding words by averaging their vectors, simplifying the traditional neural network structure by removing the non-linear hidden layer. This design reduces computational complexity and accelerates training. The Skip-gram model, conversely, predicts surrounding words given a target word, capturing more complex relationships by considering a broader context. Both models exhibited substantial improvements in word similarity tasks, achieving state-of-the-art performance while drastically cutting down computational costs. These advancements made it feasible to train on extensive datasets, enhancing various NLP applications like machine translation, speech recognition, and information retrieval by providing high-quality word vectors that reflect deep linguistic relationships.

GloVe (Global Vectors for Word Representation) [47] is a model developed by the Stanford NLP group that produces non-contextual word embeddings. It leverages statistical information from a large corpus by constructing a word-word co-occurrence matrix, where each element represents the frequency with which two words appear together. This approach allows GloVe to generate word vectors that capture meaningful semantic and syntactic relationships, resulting in a robust and interpretable vector substructure.

ELMo [48] and BERT [13] are examples of language models that build context-sensitive word embeddings, which aid in dealing with ambiguities. These models process the text in a bidirectional manner (forward and backward), resulting in better word representations.

ELMo is implemented using a bidirectional language model that consists of two layers of



(a) CBOW architecture.

(b) Skip-gram architecture.

Figure 3.1: Word2Vec architectures (adapted from [37].

LSTM networks. These LSTMs are trained on a large corpus with a coupled forward and backward language model objective, which captures the context of a word by considering both its preceding and succeeding words in a sentence. Each word token is assigned a representation that is a function of the entire input sentence, incorporating both the complex characteristics of word usage and how these uses vary across different contexts. This deep, context-sensitive approach allows ELMo to effectively model polysemy and improve performance across a wide range of NLP tasks by providing rich semantic and syntactic representations

BERT is implemented using a deep bidirectional Transformer [68] encoder architecture, which allows the model to consider both left and right context simultaneously during training. This is achieved through two main pre-training tasks: the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In MLM, random tokens in the input sequence are masked, and the model is trained to predict these masked tokens based on their context, enabling it to learn deep bidirectional representations. NSP, on the other hand, involves training the model to understand the relationship between two sentences by predicting whether a given sentence follows another in the original text. This dual-task pre-training approach allows BERT to capture both the nuanced meaning of words and their inter-sentence dynamics.

3.3 Sentence embeddings

Sentence embeddings appeared due to the increasing interest in tasks that require representations of larger pieces of text and complete sentences [7]. Similarly to word embeddings, the idea is to encode sentences into vectors so that similar sentences are placed close in the vector space [53].

Doc2Vec [26] extends the concept of Word2Vec to complete sentences or documents. It enables, through unsupervised learning, the generation of fixed-length numerical representations, or vectors, for variable-length pieces of text, such as sentences, paragraphs, or documents. This is

achieved through two primary architectures: the Distributed Memory Model of Paragraph Vectors (PV-DM) and the Distributed Bag of Words version of Paragraph Vectors (PV-DBOW). In PV-DM, the paragraph vector is concatenated with several word vectors from the paragraph to predict a word, effectively capturing the context within the text. Conversely, PV-DBOW ignores the context words and instead predicts words randomly sampled from the paragraph using the paragraph vector alone, simplifying the model by reducing the amount of stored data. Both methods involve training through stochastic gradient descent and backpropagation, ensuring that the paragraph vectors can encapsulate semantic meanings and structural information of the text.

Sentence transformers are a cutting-edge approach in NLP that leverages pre-trained transformer models to encode sentences into dense vector representations. It originates from the work of Sentence-BERT [53], a modification of the pre-trained BERT network using siamese and triplet network structures in order to obtain semantically meaningful sentence embeddings that can be compared using cosine similarity. This approach obtained state-of-the-art results on common Semantic Textual Similarity (STS) tasks, outperforming other sentence embedding methods. Sentence transformers are trained on a labeled or structured dataset that informs the model if two sentences are similar or different. Afterward, these models can be used to obtain vectorial representations for a variety of sentences, making them highly versatile for numerous NLP tasks such as semantic search, paraphrase mining, and clustering.

Chapter 4

INCOGNITUS

This chapter describes the current state of the INCOGNITUS [54] platform for the automated anonymization of clinical notes. It offers different techniques and provides multiple performance assessment metrics.

4.1 Overview

The INCOGNITUS pipeline is illustrated in Figure 4.1. It allows the user to upload the text they want to anonymize and select one of three methods for such. Two of them are based on NER, in which the sensitive entities are recognized and replaced by categorized tags. The third method relies on a word embedding model that replaces every token of the text with one of the nearest ones in the embedding space. After the anonymization step, the anonymized version of the text is presented along with different evaluation metrics: recall, precision, F_1 score and information loss.

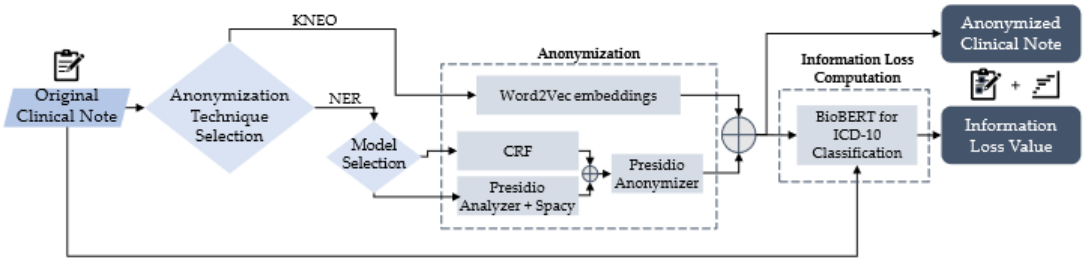


Figure 4.1: INCOGNITUS pipeline (obtained from [54]).

The user interface can be seen in Figure 4.2, where the technique, performance metrics, and both the original and anonymized versions of the note are shown.

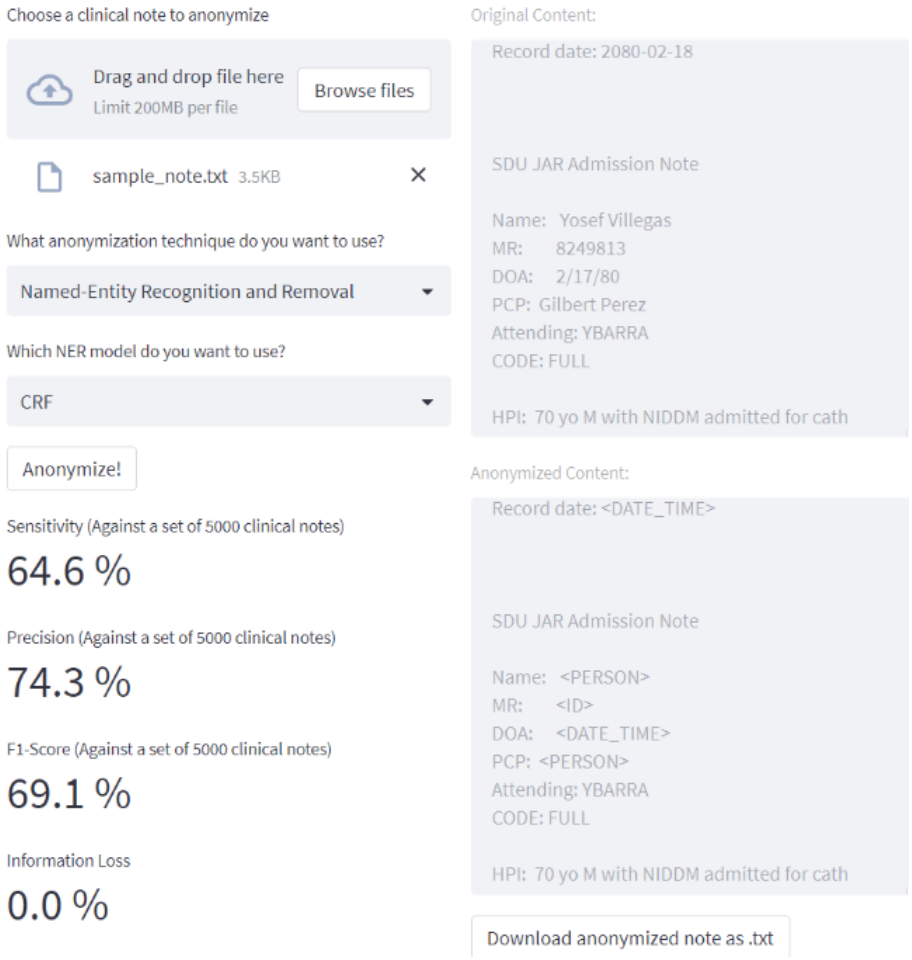


Figure 4.2: INCOGNITUS interface (obtained from [54]).

4.2 Architecture

In this section, we explain in further detail the two main components of the INCOGNITUS platform: the anonymization techniques and the evaluation metrics. We will describe the implementation of the two NER models and the word replacement strategy, as well as the novel evaluation metric introduced for the assessment of information loss.

4.2.1 Anonymization Techniques

The two NER models for the identification and classification of sensitive information are a CRF [25] classifier and a pre-trained spaCy[1] model. The word replacement technique uses a Word2Vec [37] embeddings model.

The CRF classifier was trained on the training sets of both i2b2 de-identification challenges, using the following features regarding each token and its two instant neighbors: POS tag, the last 2 or 3 characters, whether it starts with a capital letter, whether it is a title, and whether it is a digit.

[1] https://spacy.io/

For the second technique, a Microsoft Presidio[2] Analyzer receives a pre-trained spaCy language model as input and detects the PII.

After the recognition of sensitive entities through either of those two techniques, a Microsoft Presidion Anonymizer replaces them with categorized tags, as can be seen in the "Anonymized Content" text box of Figure 4.2.

The word replacement approach is called K-Nearest Embeddings Obfuscation (KNEO), and it uses word embeddings to replace every token of the text with one of the K most semantically similar in the embeddings space. As every token is being replaced, this method achieves a recall of 100% (all sensitive entities are removed), but at the cost of readability and data usefulness. This was implemented using a word embeddings model trained on 54,652 discharge notes from the de-identified MIMIC-III [23] database using a Word2Vec strategy. For this task, the Faker[3] library for Python was used to create fake entities according to the category tags present in the MIMIC-III notes, in order to obtain a more realistic text.

4.2.2 Evaluation Metrics

Besides providing the evaluation results on traditional metrics such as recall, precision and F_1-score, the INCOGNITUS framework also provides a new metric designed to assess the loss of information during the anonymization process. A pre-trained BioBERT [27] model with a set of 157 ICD-10 code categories is used. This model receives the clinical note and outputs the confidence for each of the categories being present in the text. So, it is used to identify the top 10 categories with the highest scores in the original and anonymized notes. Finally, information loss is calculated by analyzing the number of codes simultaneously present in both versions, as shown by Equation 4.1, where y_{anon} and y_{orig} represent the set of the 10 categories present in the anonymized note and the original note, respectively. For example, if 8 code categories are present in both versions of the note, we have an information loss of 20%.

$$IL = (1 - \frac{\sum_{i=1}^{10}(y_{anon_i} \in y_{orig})}{10}) \times 100 \tag{4.1}$$

4.2.3 Results

The performance of each strategy was tested against the test sets of both the i2b2 de-identification challenges and 5,000 discharge summaries from the MIMIC-III dataset. The F_1-score was calculated using a binary evaluation strategy, i.e., the system simply needs to classify information as sensitive or non-sensitive, ignoring the category. Table 4.1 presents the results.

The results show that a simple CRF model is capable of obtaining high performance in terms of F_1-score when trained and tested on the same dataset. However, there is a notable drop in performance when this model is evaluated on a different test set, which suggests a non-ideal adaption to the training data.

[2] https://microsoft.github.io/presidio/
[3] https://faker.readthedocs.io/en/master/

		F_1-score	IL
i2b2 2006	CRF	94.8	15.8 ± 11.4
	Presidio	73.0	21.6 ± 13.0
	KNEO	-	59.9 ± 21.3
i2b2 2014	CRF	87.8	15.7 ± 12.4
	Presidio	64.6	21.3 ± 14.0
	KNEO	-	58.4 ± 21.1
MIMIC	CRF	69.1	21.3 ± 13.8
	Presidio	66.6	24.9 ± 14.6
	KNEO	-	63.4 ± 18.4

Table 4.1: F_1-score values (percentage) obtained by each anonymization method on the test sets (obtained from [54]).

As expected, the quantity of lost clinical information increases significantly when the KNEO strategy is applied. It goes from values in the area of 15-25% when applying the NER models to values around 60% when replacing words. This amount of lost information could harm the future use of this data for other downstream tasks, but it is the price to pay for guaranteeing the removal of all sensitive entities.

4.3 Summary

In this chapter, we summarize the first version of the INGOCNITUS platform, as described by Ribeiro et al. [54]. At Fraunhofer Portugal AICOS, this anonymization toolkit is a work in progress, of which this research is a part, so it is expected that new anonymization methods and evaluation metrics will be added throughout the realization of this work.

The sentence embeddings approach, which is the subject of this work, follows up on the word embeddings approach as a way to minimize the loss of relevant medical information while still maintaining the removal of all sensitive entities.

Chapter 5

Methodology

This chapter details the research methodology employed in this work. It outlines the data, processes, and techniques used to implement and analyze the proposed solution. The chapter is structured to provide a clear and systematic account of the methods and procedures undertaken, ensuring the research's reproducibility and reliability. The methodology describes the different components that are necessary for the de-identification of clinical text using sentence embeddings, such as data processing and analysis, text pre-processing steps, embedding space generation and analysis, and the different anonymization strategies and models. Two different anonymization strategies, word and sentence replacement, were implemented using one and four embedding models, respectively. Each section will provide an in-depth explanation of the technologies and resources used, as well as relevant implementation details.

5.1 Overview

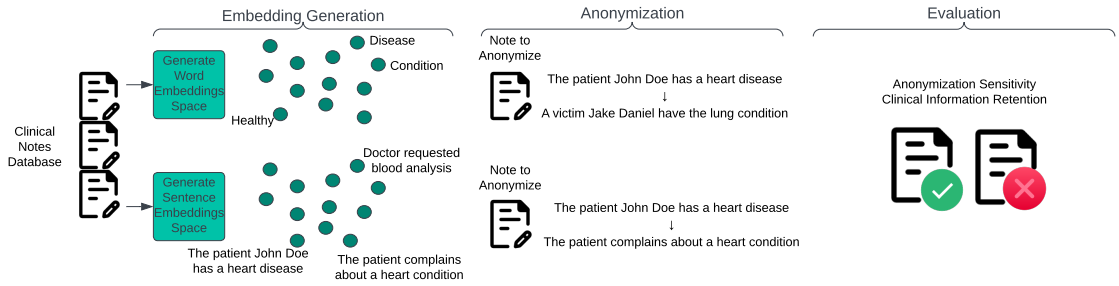


Figure 5.1: Pipeline for the anonymization of clinical notes using word (top) or sentence (bottom) substitution.

The bigger picture of the anonymization pipeline is illustrated in Figure 5.1. It is divided into two main modules: the embedding generation module and the anonymization module. Firstly, we create the vector space from a de-identified dataset of clinical notes, which stores the words/sentences and their respective embeddings. Then, to anonymize a new clinical note, we replace every word or sentence, depending on the strategy, with a similar one obtained from the embedding space.

These strategies will be described with further detail for each of the strategies in Sections 5.3, 5.4 and 5.5. Finally, we evaluate the performance of our strategies on a test set using evaluation metrics detailed in Section 5.6, aimed at anonymization sensitivity and clinical information retention.

5.2 Data

The MIMIC-III clinical database [23] is a large, de-identified and freely available dataset comprised of health-related data. During the de-identification process, its sensitive information was replaced by category tags. To obtain a more realistic version of the notes, the Faker[1] library for Python was used to create fake entities according to each category. A subset of 33,321 discharge summary notes were used to generate the embedding space, and another of 19,989 notes was used to evaluate the different approaches. MIMIC-III also contains different note types with varying proportions, and it was assured that both subsets have the same distribution.

5.2.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in the development of machine learning solutions. It involves summarizing and analyzing the main characteristics of a dataset with the objective of obtaining valuable insights that can help address the problem at hand. We conducted an EDA on the complete dataset that was used (33,321 notes for embedding generation and 19,989 for evaluation).

The rows of our dataset are constituted by three columns: note type, identified text, and annotations. The first column indicates the type of the note, the second column is the text of the clinical note, but with fake entities replacing the categorized tags, and the final column contains the annotations of the sensitive information. Figure 5.2 illustrates an example.

note_type	identified_text	annotations
Case Management	Insurance information Primary insurance: Hawarden Regional Healthcare HEALTH PLAN Secondary insurance: Insurance reviewer:: Free Care application: N/A Status: Medicaid application: N/A Pre-Hospitalization services: None prior to admission DME / Home O[2]: None prior to admission Functional Status / Home / Family Assessment: Pt. lives with his mother in Lawson. He is independent with his ADL's Primary Contact(s): Samantha Guzman Kim (Mother) +1-632-870-7204	TEXT='Hawarden Regional Healthcare'; START='41'; END='69'; SUBCATEGORY='HOSPITAL'; CATEGORY='INSTITUTION'; TEXT='Lawson'; START='355'; END='361'; SUBCATEGORY='NAME'; CATEGORY='NAME'; TEXT='Samantha Guzman'; START='416'; END='431'; SUBCATEGORY='NAME'; CATEGORY='NAME'; TEXT='Kim'; START='432'; END='435'; SUBCATEGORY='NAME'; CATEGORY='NAME'; TEXT='+1-632-870-7204'; START='445'; END='460'; SUBCATEGORY='PHONE_NUMBER'; CATEGORY='CONTACT_NUMBER';

Figure 5.2: Example of a row from our dataset.

Table 5.1 shows the fifteen different types of clinical notes and their distribution. This distribution was maintained on both subsets, except for the Consult and Pharmacy types, for which there is only one note.

We can observe the average text length of the identified text column, in characters, by looking at Figure 5.3. It is interesting to notice that discharge summaries and physician notes are much longer than the rest.

[1] https://faker.readthedocs.io/en/master/

Note Type	Proportion
Nursing/other	0.394935
Radiology	0.250778
Nursing	0.107334
ECG	0.100356
Physician	0.067998
Discharge summary	0.028625
Echo	0.021966
Respiratory	0.015213
Nutrition	0.004502
General	0.003958
Rehab services	0.002589
Social Work	0.001257
Case Management	0.000450
Consult	0.000019
Pharmacy	0.000019

Table 5.1: Note type distribution.

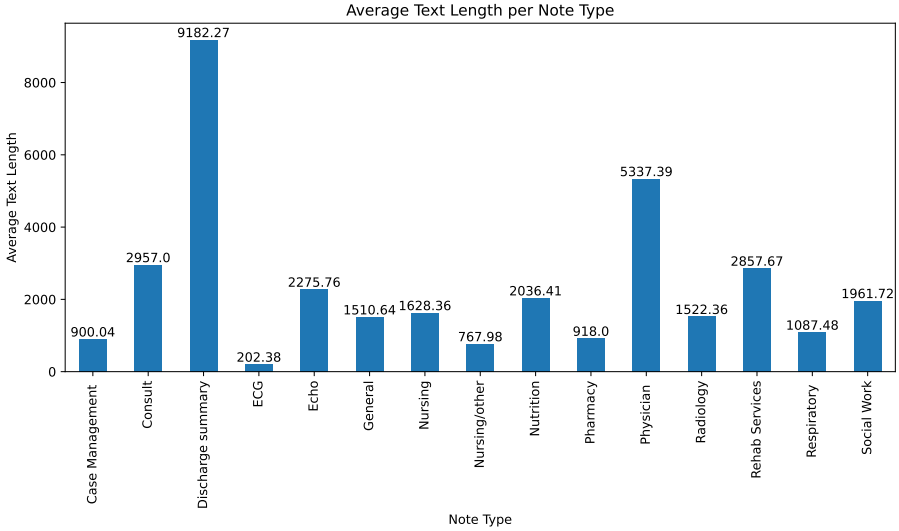


Figure 5.3: Average text length (characters) per type of note.

The dataset contains multiple categories of sensitive information present in the clinical text, which are registered in the annotations column. Table 5.2 displays the categories and the number of occurrences of its subcategories. Figure 5.4 shows the average length of each category's text. As expected, text categories such as addresses (both physical and virtual) and institutions typically have a higher length than number categories, such as dates, ages and IDs. It is also no surprise the dates and names are the two most common categories.

Category	Subcategory	Count
AGE_ABOVE_89	AGE_ABOVE_89	1186
CONTACT_NUMBER	PAGER	58
	PHONE_NUMBER	5403
DATE	DATE	182855
	DATE_RANGE	633
	DAY/MONTH	883
	DAY/MONTH/YEAR	348
	MONTH	2213
	MONTH/YEAR	725
	YEAR	917
EMAIL	EMAIL	20
HOLIDAY	HOLIDAY	44
ID	ID	3839
	JOB_NUMBER	504
	MED_NUMBER	16151
INSTITUTION	COMPANY	270
	HOSPITAL	31499
	UNIVERSITY/COLLEGE	72
LOCATION	COUNTRY	265
	LOCATION	3975
	STATE	316
	STREET_ADDRESS	292
NAME	ATTENDING/DICTATOR_NAME	5
	FIRST_NAME	15565
	LAST_NAME	35713
	NAME	17855
	NAME_INITIALS	2777
URL	URL	1

Table 5.2: Number of sensitive entities by category.

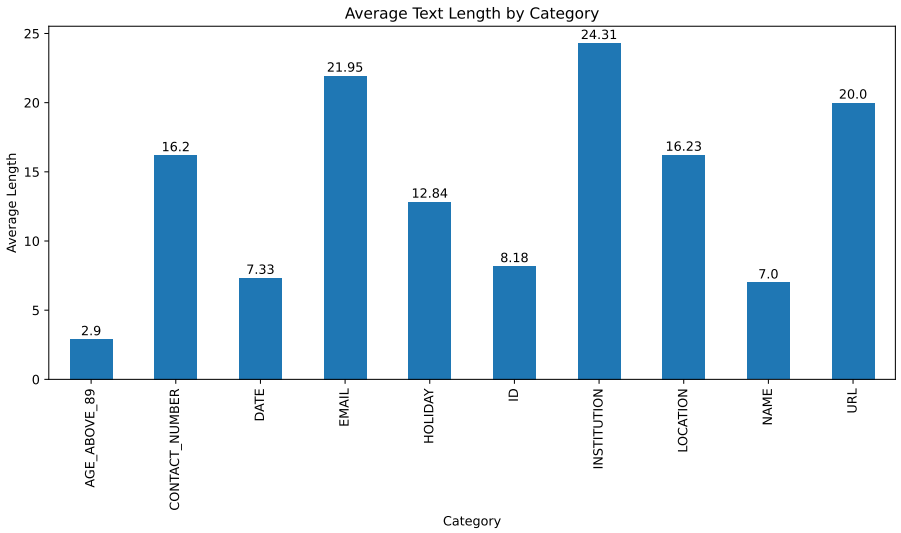


Figure 5.4: Average text length (characters) per sensitive information category.

Figure 5.5 shows the ten words that are the most frequent in the identified text column, excluding stopwords and words composed of one or two characters. We can observe that none of the words represent potentially sensitive information and are well suited to the type of text we are dealing with.

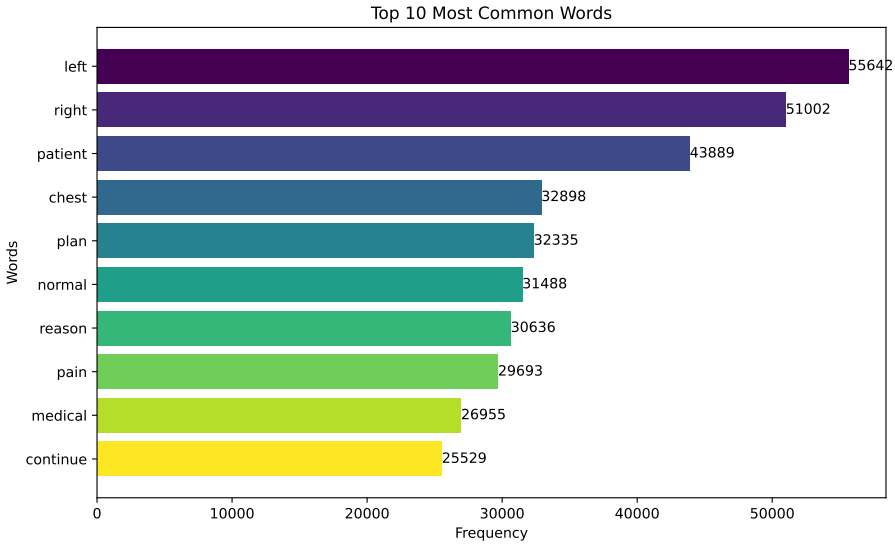


Figure 5.5: 10 most common words present on the identified text column.

5.2.2 Pre-Processing

Common and data-specific text pre-processing steps were implemented:

Lowercasing Convert all tokens to lowercase.

Non-Alphanumeric Characters Removal Remove characters that are not letters or digits and replace them with a white space.

Consecutive White Spaces Removal Replace multiple white spaces with a single one.

Lowercasing and removing non-alphanumeric characters are standard text pre-processing steps in NLP tasks. Additionally, consecutive white spaces frequently appear in the text, so their removal was also applied. These pre-processing steps are performed on the text before the respective embeddings are calculated.

5.3 Word2Vec Anonymization

In this study, a word embedding model was trained on a corpus of 33,321 clinical notes using Gensim's implementation of Word2Vec[2]. This process created a dense vector representation, known

[2] https://radimrehurek.com/gensim/models/word2vec.html

as the embedding space, for each word in the clinical notes. The Word2Vec model was trained over 100 epochs using the continuous bag-of-words architecture, with specific parameters set as follows: a vector size of 256, a context window of 15 words, a minimum word count threshold of 1, and the model utilized a single worker thread for training. Although this work focuses on the de-identification of clinical text using sentence embeddings, the word embeddings strategy serves as a comparison point to answer our research question concerning the possible improvement in the capability of retaining clinical information offered by the sentence substitution strategy.

To generate the embedding space using Word2Vec, the following algorithm was implemented:

Algorithm 1 Word Embedding Space Generation

```
sentences ← {}
for note in train notes do
    note_sentences ← sentence_tokenize(note)        ▷ Each note as a list of sentences
    for sentence in note_sentences do
        preprocess(sentence)                                              ▷ optional
        word_tokenize(sentence)                     ▷ Each sentence as a list of tokens
        sentences ← sentences + sentence
model ← Word2Vec(sentences)
train(model)
```

Algorithm 1 can be translated to the following steps:

1. **Sentence Segmentation**: we iterate through the 33,321 clinical notes and employ NLTK's sentence tokenizer[3] to divide each note into individual sentences.

2. **Pre-processing**: optionally, the pre-processing steps described in Section 5.2.2 are applied to the sentences.

3. **Tokenization**: each sentence is tokenized into a list of words (tokens).

4. **Model Training**: These tokenized sentences are used to train the Word2Vec model. During training, the model learns to generate a vector representation for each word in the vocabulary, capturing semantic relationships between words based on their context in the corpus.

Once the embedding space is established, it is utilized for the anonymization of the 19,989 clinical notes reserved for testing. Algorithm 2 describes the anonymization process using this word replacement strategy:

[3] https://www.nltk.org/api/nltk.tokenize.sent_tokenize.html

Algorithm 2 Word Embedding Anonymization

```
for note in test notes do
    anonymized_note ← ""
    note_sentences ← sentence_tokenize(note)          ▷ Each note as a list of sentences
    for sentence in note_sentences do
        preprocess(sentence)                                              ▷ Optional
        sentence ← word_tokenize(sentence)            ▷ Each sentence as a list of tokens
        for word in sentence do
            new_word ← choice(model.most_similar(word, 5))   ▷ 1 from the 5 most similar
            anonymized_note ← anonymized_note + new_word
```

We iterate through the notes in the test set and replace every token in a note with a similar one obtained from the embeddings space:

1. **Sentence Segmentation**: divide each note into sentences using NLTK's sentence tokenizer.
2. **Pre-processing**: apply to the sentences the same pre-processing steps that were used when creating the embedding space.
3. **Tokenization**: obtain the tokens of each sentence.
4. **Similarity-based Replacement**: each token is then replaced with a different token. This replacement token is randomly selected from the top 5 most similar tokens, based on cosine similarity, within the embedding space. This method ensures that the replacement token maintains a semantic similarity to the original (as far as possible), thereby preserving the contextual integrity of the note while obfuscating sensitive information.

After this process, we obtain a new version of each clinical note where every token has been replaced with a different one, resulting in the removal of the sensitive information.

5.4 Doc2Vec Anonymization

In a similar approach to Word2Vec anonymization, a document embedding model was trained on the same corpus of 33,321 clinical notes using Gensim's implementation of Doc2Vec[4]. This method created an embedding space for entire sentences, capturing semantic relationships at the sentence level. The process of training the Doc2Vec model involves the following steps to ensure the creation of meaningful embeddings:

1. **Sentence Segmentation**: initially, each clinical note is divided into sentences using NLTK's sentence tokenizer.

[4] https://radimrehurek.com/gensim/models/doc2vec.html

2. **Pre-processing**: optionally, the pre-processing steps detailed in Section 5.2.2 are applied to the sentences.

3. **Model Training**: The Doc2Vec model is trained on these pre-processed and tokenized sentences for 100 epochs. The training parameters for the baseline model were set as follows: a vector size of 256, a distributed bag-of-words for the training algorithm, a context window of 15 words, a minimum word count of 1, and a single worker thread.

As can be seen in Algorithm 3, the process is very similar to the generation of embeddings using Word2Vec:

Algorithm 3 Doc2Vec Embedding Space Generation

```
sentences ← { }
for note in train notes do
    note_sentences ← sentence_tokenize(note)        ▷ Each note as a list of sentences
    for sentence in note_sentences do
        preprocess(sentence)                                              ▷ Optional
        sentences ← sentences + sentence
model ← Doc2Vec(sentences)
train(model)
```

In Doc2Vec, each document/sentence has another vector in addition to the word vectors, the sentence vector. This sentence vector is used for training predictions and is updated just like the word vectors. As we are using distributed bag-of-words, these vector representations for each sentence are obtained by training a neural network on the task of predicting a target word just using the sentence vector and not the other word vectors. The sentence vectors were stored alongside their respective sentences, resulting in an embedding space with 644,052 sentences that are then used for the anonymization process that relies on sentence substitution. To anonymize a clinical note, we follow these steps:

1. **Sentence Segmentation**: divide the note into individual sentences.

2. **Pre-processing**: apply the same pre-processing steps that were used during the embedding space creation to the sentences.

3. **Sentence Embedding**: Each sentence is embedded using the trained Doc2Vec model to obtain its vector representation.

4. **Similarity-based Replacement**: Each sentence is then replaced with a different one. This replacement sentence is randomly chosen from the top 5 most similar sentences, based on cosine similarity, within the embedding space. This method ensures that the new sentence maintains a semantic relationship with the original, thus preserving the contextual integrity of the clinical note while anonymizing sensitive information.

Algorithm 4 illustrates this process in pseudo-code:

Algorithm 4 Doc2Vec Embedding Anonymization

```
for note in test notes do
    anonymized_note ← ""
    note_sentences ← sentence_tokenize(note)          ▷ Each note as a list of sentences
    for sentence in note_sentences do
        preprocess(sentence)                                                  ▷ Optional
        sentence_embedding ← model.infer_vector(sentence)
        new_sentence ← choice(model.most_similar(sentence_embedding, 5)) ▷ 1 from the 5 most similar
        anonymized_note ← anonymized_note + new_sentence
```

When replacing every sentence with a similar one, some parts of the sentence might be equal. Ideally, this would happen on the parts that contain relevant medical information, but it could also happen that the parts that remain the same are the ones containing personal information. This is, however, tackled by the fact that our sentence embedding space is obtained from already de-identified clinical notes containing fake entities. As we use the same dataset for training and testing, it is possible that there is an overlap in terms of these fake entities, as some of them can appear in both subsets, so some of them might still be present after the replacement process. In a real-world scenario, it is very unlikely that this overlap would happen.

5.5 Sentence Transformers Anonymization

We experiment with different pre-trained sentence-transformer models available in the Sentence-Transformers Python framework[5]. The following three models were used:

all-MiniLM-L6-v2[6] Baseline model that maps sentences into a 384-dimensional dense vector space.

avsolatorio/GIST-large-Embedding-v0[7] Model that has a good performance on the BIOSSES (biomedical sentence similarity estimation) benchmark. Generates embeddings with 1024 dimensions.

pritamdeka/S-PubMedBert-MS-MARCO[8] Model trained on biomedical text from PubMed that maps sentences to a 768-dimensional dense vector space.

NLTK's sentence tokenizer was employed to extract individual sentences from the 33,321 clinical notes. Unlike previous methods, no model training was performed in this approach due to the lack of a labeled dataset of sentence pairs, which is essential for training a sentence transformer model. Instead, pre-trained sentence transformer models were utilized to encode the sentences into dense vector representations, as can be seen in Algorithm 5.

The process of anonymizing a clinical note involves the following steps:

[5] https://sbert.net/

Algorithm 5 Sentence Transformer Embedding Space Generation

```
sentences ← {}
model ← SentenceTransformer()
for note in train notes do
    note_sentences ← sentence_tokenize(note)          ▷ Each note as a list of sentences
    for sentence in note_sentences do
        preprocess(sentence)                                                  ▷ Optional
        sentences ← sentences + sentence
embeddings ← model.encode(sentences)
return embeddings
```

1. **Sentence Segmentation**: divide the note into sentences using NLTK's sentence tokenizer.
2. **Pre-processing**: the same pre-processing steps used during the embedding space generation are applied to the sentences.
3. **Sentence Embedding**: the pre-trained sentence transformer model is then used to encode each pre-processed sentence into a vector representation.
4. **Similarity-based Replacement**: Each sentence embedding is compared within the vector space, and the sentence is replaced with a different one. This replacement sentence is randomly selected from the top 5 most similar sentences based on cosine similarity.

These steps are represented in pseudo-code by the following algorithm:

Algorithm 6 Sentence Transformer Embedding Anonymization

```
for note in test notes do
    anonymized_note ← ""
    note_sentences ← sentence_tokenize(note)          ▷ Each note as a list of sentences
    for sentence in note_sentences do
        preprocess(sentence)                                                  ▷ Optional
        sentence_embedding ← model.encode(sentence)
        cos_scores ← cosine_similarity(sentence_embedding, embeddings)
        new_sentence ← choice(top(cos_scores, 5))          ▷ 1 from the 5 most similar
        anonymized_note ← anonymized_note + new_sentence
```

This method leverages the semantic richness of pre-trained sentence transformers, which are trained on large amounts of data, to effectively capture and compare the meanings of the different sentences.

5.6 Evaluation Metrics

Evaluating the effectiveness of the methods developed for the de-identification of clinical text is crucial to ensure both the protection of patient privacy and the preservation of data utility. Traditional evaluation metrics, such as precision, recall, and F_1-score, are widely used in various natural

language processing tasks to measure a model's performance by comparing its predictions with the ground-truth labels [49, 67]. These metrics provide a precise and quantitative assessment of how well a model identifies and categorizes specific entities. However, de-identification presents unique challenges that require a nuanced approach to evaluation, particularly when innovative methods such as sentence and word replacement are employed.

Traditional evaluation metrics such as precision, recall, and F_1-score are inadequate for assessing our proposed de-identification approach. In NER-based techniques, these metrics are typically calculated by comparing the model's predictions for each token with the corresponding ground-truth labels. However, our strategy relies on the replacement of text pieces, rendering traditional metrics impractical, as the sensitive entities may still be present in the anonymized text but in a location different from their original positions, which results in the associated labels no longer being relevant. For instance, in the 2014 i2b2/UTHealth de-identification task, the labels for sensitive entities include their starting and ending positions, as well as the corresponding text. Let us consider the sentence, "The patient is John.". Its ground-truth label would be something like [start="16" end="20" text="John"]. Using our sentence replacement strategy for de-identification, one could obtain the sentence "John is the patient.". While the sensitive information remains, the connection between the token and the label is lost. To address this challenge, one might consider using exact string-matching techniques. However, this approach is often insufficient due to its inability to accommodate minor variations in the text. For instance, if a patient's name is "John Doe," exact string-matching would fail to recognize slightly altered forms such as "J. Doe" or "John D." Therefore, even small changes to the sensitive entities can significantly affect the correctness of the de-identification process.

Considering these problems, we employ new metrics that do not depend on the alignment between tokens and labels, allowing for a more accurate assessment of our de-identification method's effectiveness. These evaluation metrics are a conjoint work of researchers at Fraunhofer Portugal AICOS [51]. They can be divided into two categories: anonymization sensitivity metrics and clinical information retention metrics. The first category, whose focus is on the masking of sensitive entities, relies on the usage of Levenshtein Distance (LD). The clinical information retention metrics are based on the usage of a BioBERT [27] model, which has been pre-trained on a hierarchical classification task of ICD-10 code categories.

5.6.1 Anonymization Sensitivity

The LD measures the difference between two strings by counting the minimum number of single-character edits (insertions, deletions, substitutions) needed to transform one string into the other [71]. The smaller the distance between two strings, the more similar they are.

The Levenshtein Ratio (LRa) quantifies similarity based on the LD according to the following expression, where $LD(a,b)$ denotes the LD between strings a and b, and A and B represent their respective lengths:

$$LRa(a,b) = 1 - \frac{LD(a,b)}{\max(A,B)} \tag{5.1}$$

LRa yields a score ranging from 0 to 1: 0 indicates complete dissimilarity between the strings, while 1 signifies they are identical.

Two metrics based on LRa are proposed: the Average Levenshtein Index of Dissimilarity (ALID) and the Levenshtein Recall (LR). These metrics aim to assess the effectiveness of anonymization in situations where token sensitivity is unknown, overcoming limitations associated with string matching. Consider we have a list of n sensitive entities, se, contained in an original note ON of length L. For a specific entity se_i from this list, we begin by determining its length, denoted as e. Subsequently, we slide a window of length e across the anonymized note AN, moving one character at a time. The Levenshtein Similarity Index (LSI) of se_i against AN is computed using the following expression, where w_j denotes the jth window of length e within AN:

$$LSI = \max_{j=1}^{L-e} LRa(se_i, w_j) \tag{5.2}$$

This measure signifies the highest degree of similarity between se_i and the content within AN. Having a list S containing the computed LSIs for each entity in se, the Average Levenshtein Index of Dissimilarity (ALID) is defined as follows, where $\langle S \rangle$ represents the arithmetic mean of S:

$$ALID = (1 - \langle S \rangle) \times 100 \tag{5.3}$$

LR also utilizes the concept of LSI. To compute LR, each LSI in S is compared against a similarity threshold of 0.85, th_s. Entities with an LSI below this threshold are classified as de-identified, while those exceeding the threshold are considered not de-identified. The metric's final value is determined using the conventional recall computation, which divides the count of de-identified entities by the total number of entities.

$$LR@th_s = \frac{\sum_{i=1}^{n} (S_i < th_s)}{n} \times 100 \tag{5.4}$$

Regarding privacy implications, evaluating text anonymization should address additional considerations. For instance, failing to mask direct identifiers - identifiers that uniquely identify an individual - poses greater risks compared to not masking quasi-identifiers - identifiers that do not uniquely identify an individual but can be combined to increase the chances of re-identification. Furthermore, effective anonymization requires all instances of direct identifiers to be masked, not just some. To address these concerns, two new LR-based metrics are introduced: the Levenshtein Recall for Direct Identifiers (LRDI) and the Levenshtein Recall for Quasi-identifiers (LRQI).

Let l_{di} denote the length of a list containing direct identifiers from ON. S_{di} represents a list of LSIs computed for each direct identifier, also of length l_{di}. The LRDI metric assumes one of two values: 100, indicating that all occurrences of direct identifiers are anonymized, or 0, indicating otherwise.

$$LRDI@th_s = \left(\forall k \in \{0,1,\ldots,l_{di}-1\},\ S_{di}[k] < th_s\right) \times 100 \tag{5.5}$$

Let l_{qi} represent the length of a list containing the quasi-identifiers from *ON*. The LRQI is calculated similarly to the LR but only takes into account the quasi-identifiers:

$$LRQI@th_s = \frac{\sum_{k=1}^{l_{qi}} (S_k < th_s)}{l_{qi}} \times 100 \tag{5.6}$$

Additionally, we also measure the string-match recall (SMR) to verify if the sensitive entities' text is present on the *AN*. It is counted as a true positive if the entity is not found, and the recall is given by the ratio of true positives to the total number of sensitive entities.

5.6.2 Clinical Information Retention

Maintaining the relevant clinical information intact during the anonymization process is an important aspect of guaranteeing future data utility, so two new metrics are introduced. Their computation utilizes the publicly available BioBERT model, pre-trained on a hierarchical classification task involving ICD-10 code categories, a coding system developed by the World Health Organization to catalog health conditions [70]. The ICD-10 codes are grouped into 157 categories based on their type, e.g., *Cerebrovascular diseases I60-I69*, and the model[9] outputs the confidence for each of those categories being present on the text it receives. By comparing the model's outputs when given a clinical note before and after anonymization, the amount of kept information is estimated.

The first metric employs the Jaccard Similarity Coefficient (JSC) [21]. The BioBERT model outputs are converted to probabilities using a softmax function, followed by applying a threshold of 0.05, th_b. Values above this threshold are set to 1, and those below are set to 0, creating a binary representation of the ICD-10 code categories identified by the BioBERT model in each note. The JSC is then calculated between the binary representations of the note before and after anonymization. Let C_{11} represent the number of classes where both representations have a value of 1, and $C_{01}+C_{10}$ represent the number of classes where the representations differ. The clinical information retention based on the JSC is expressed as follows:

$$JSC@th_b = \frac{C_{11}}{C_{11}+C_{01}+C_{10}} \times 100 \tag{5.7}$$

Additionally, the Normalized Softmax Discounted Cumulative Gain (NSDCG) was used, which is a variant of the widely used NDCG (Normalized Discounted Cumulative Gain) ranking metric [22]. NSDCG operates under the assumption that higher scores indicate closer proximity between original and anonymized logit distributions, thereby measuring the retention of clinical information. Unlike NDCG, NSDCG utilizes a discount factor derived from applying the softmax function

[9] https://huggingface.co/rjac/biobert-ICD10-L3-mimic

to transformer logits, denoted as *sd* (refer to Equation 5.9), instead of the typical logarithmic discount $\log(i+1)$. This discount is conventionally applied to the relevance score *rel*. Consequently, *SDCG* (Softmax Discounted Cumulative Gain) is calculated as follows:

$$SDCG@K = \sum_{i=1}^{K} sd_i \cdot rel_i \tag{5.8}$$

Regarding the discount sd_i, s represents the logits sorted in descending order from the original note. The softmax discount, considering the N ICD-10 classes at position i, is expressed as:

$$sd_i = \frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}} \tag{5.9}$$

The key advantage of using the softmax function is that it allows for more precise weighting of each ICD-10 class logit. Unlike the typical logarithmic discount, which assigns diminishing importance uniformly across all samples and results in weak sensitivity to individual classes, the softmax function maintains variability among the logit outputs. Although this issue could be partially addressed by considering only the top K ranked classes using the K parameter, the variability of logit outputs can still cause problems when using a logarithmic function.

Lastly, rel_i denotes the relevance of the item at position i in the ranked original logits z. Here, z refers to the logits from the original note arranged according to the anonymized note:

$$rel_i = e^{z_i} \tag{5.10}$$

The NSDCG is obtained by dividing the SDCG of the anonymized note by that of the ideal and original note, yielding a percentage value. In our experiments, K was set to 10:

$$NSDCG@K = \frac{SDCG@K}{ISDCG@K} \times 100 \tag{5.11}$$

5.6.3 Summary

In this section, we present and describe the evaluation metrics that will be used in the assessment of our solution. Although traditional metrics like precision and recall can generally be effective in evaluating the performance of the developed solution, there are cases where they might be lacking. Typically, precision is thought of as the data utility metric, and recall is the anonymization rate. But what if a sensitive entity is just replaced by a common abbreviation? A string-matching recall would fail to detect its presence. Regarding precision, we could have two systems with 95%, but one could be (incorrectly) hiding adverbs or determinants, while the other could be (incorrectly) hiding medical terms, which would result in a lower data utility even though their precision is the same.

These proposed metrics aim to extend the evaluation to those cases, and we believe that they can contribute to the better assessment of an automated solution's real performance.

Table 5.3 provides a summarized description of what the anonymization sensitivity (SMR, ALID, LR, LRDI and LRQI) and clinical information retention (JSC and NSDCG) metrics do.

Metric	Summary	Example
SMR	Recall based on string-matching.	*ON* → "The patient *John* assisted on *12/9*", *AN* → "The patient visited" then SMR is 100% as both sensitive strings were removed.
ALID	Compares each sensitive entity in the original text with strings of equal length in the anonymized text using a sliding window. It calculates the average dissimilarity between these segments.	*ON* → "The patient *John*", *AN* → "The patient joined", then the maximum LRa would be obtained for (John, join): LD(John, join) = 2 → LRa(John, join) = 0.5 → LSI = 0.5 → S = [0.5] → ALID = 50%.
LR	Uses a similarity threshold, considering entities below this threshold as de-identified and above as not de-identified. The final value is the ratio of de-identified entities to the total number of sensitive entities.	*ON* → "The patient *John*", *AN* → "The patient joined", then S_i = 0.5, which is lower than 0.85, so it is considered de-identified and LR = 100%.
LRDI	LR for direct identifiers, all should be removed.	-
LRQI	LR for quasi-identifiers.	-
JSC	Calculates the similarity between the sets of ICD-10 code categories identified by the BioBERT model before and after anonymization. By transforming the model outputs into binary representations, the JSC quantifies how many categories are preserved or lost through anonymization.	*ON* → [A: 0.60, B: 0.36, C: 0.04] and *AN* → [B: 0.95, A: 0.03, C: 0.02]. After binarization, *ON* → [A: 1, B: 1, C: 0] and *AN* → [B: 1, A: 0, C: 0]. Then, JSC = $\frac{1}{1+1} \times 100$ → JSC = 50%.
NSDCG	Measures the preservation of the ranking of important ICD-10 code categories after anonymization. Compares the rankings before and after anonymization, providing a score that reflects the degree to which the original ranking is maintained.	*ON* → [A: 0.60, B: 0.36, C: 0.04] and *AN* → [B: 0.95, A: 0.03, C: 0.02]. Then, [e^1=1.82, e^2=1.43, e^3=1.04] → [sd_1=0.4241, sd_2=0.3337, sd_3=0.2422]. SDCG@3 for *AN* = 0.4241×0.36 + 0.3337×0.60 + 0.2422×0.04 → SDCG@3 for *AN* = 0.3626. The ISDCG@3 = 0.4241×0.60 + 0.3337×0.36 + 0.2422×0.04 → ISDCG@3 = 0.3843. Then NSDCG@3 = $\frac{0.3626}{0.3843} \times 100$ → NSDCG@3 = 94.37%.

Table 5.3: Summary of the used evaluation metrics.

Each model's performance was tested on the 19,989 notes reserved for the evaluation. Anonymized versions of the clinical notes were produced using the replacement strategies previously described in Sections 5.3, 5.4 and 5.5, which were then evaluated using the mentioned evaluation metrics. The following distribution of MIMIC-III categories was used for the LRDI and LRQI metrics: NAME, CONTACT_NUMBER, ID, and EMAIL were considered direct-identifiers, and LOCATION, DATE, URL, AGE_ABOVE_89, INSTITUTION, and HOLIDAY were considered quasi-identifiers.

Chapter 6

Results and Discussion

In this chapter, we present some experiments and their respective results on the 19,989 clinical notes reserved for evaluation, utilizing the anonymization sensitivity and clinical information retention metrics described in Section 5.6. Different strategies and models will be compared, with a focus on the sentence replacement approach. Additionally, we analyze the temporal cost of our proposed solution, as it is an important aspect of machine learning solutions.

6.1 Anonymization Sensitivity and Clinical Information Retention

Table 6.1 shows the original version of a clinical note and the same note anonymized by each of the strategies and provides insight into some of the results we can expect to find in our experiments. The sentences obtained using the sentence tokenizer are underlined to facilitate comparison of the changes.

Original Note	Shift Update Neurologically pt os intact, MAE to command. Pt is confused at times but reorients easily. Mullins pt is hypertensive, NTG infusing then hydralizine given with effect. NTG to off. Lungs with rales at the bases. Pt encouraged to cough and deep breathe. SA02 improved throughout the day, Fi02 weaned through day. Pt oob to chair with minimal assist. Pt with elevated glucose despite insulin sc, therefore gtt restarted with effect. See flowsheet for details.
Word2Vec Anonymization	Destination updated Construction aphonic arbs patent of MOVES of commends a She was incoherent Real RLSB however inconsistnently Sleepy and Luis Pt was agitated with Ntg infuing Then 2LFluids raises happend 74293486 and Nitro . Photographs of RRR , rhonchi Real right fields , She Encourage on expectorate the shortness transmitted and TRAIL worsened through . onl The Fio2 Switched Through month the He Oob in stand happend small assists and Patient happend increasing unclotted Despite Insulin SubQ and Diahrrea drip stopped w undelying a see careview 95962947 careview of
Doc2Vec Anonymization	Shift Note Neurologically pt is intact, MAE to command. Pt alert but confused at times and reorients easily. Hydralazine given with effect. Weaned off NTG drip. Lungs clear with diminished bases. Pt encouraged to cough and deep breath. PPF STOPPED BRIEFLY, FIO2 INCREASED, LEVO TITRATED UP, AND 500CC BOLUS GIVEN. Pt OOB to chair. Pt was also hyperkalemic with peaked t-waves and was therefore treated with calcium gluconate, insulin and kayexylate. SEE FLOWSHEET FOR DETAILS.
Sentence Transformer Anonymization	Shift Note Neurologically pt is intact, MAE to command. Pt alert but confused at times and reorients easily. Patient received PO anti-hypertensives this AM, IVP hydralizine Q6hrs PRN. NTG OFF. Lungs w/diffuse wheezes & rales @ bases. Pt encouraged to cough and deep breath. A: DS improved this afternoon. Pt oob to chair with one assist tolerated fine. Insulin gtt restarted for poor control of glucose. SEE FLOWSHEET FOR DETAILS.

Table 6.1: Comparison of original note and anonymized versions produced using each strategy.

The first noticeable thing is the lack of readability of the note resulting from the Word2Vec anonymization. As every token is being replaced, the text becomes very incoherent and semantically incorrect. Interestingly, despite every token being substituted by a different one, some words remain the same, just written in a different manner. For example, the words "despite" and "insulin" in the original note were replaced by "Despite" and "Insulin", respectively. However, some medical information is effectively lost. For instance, "confused" gets replaced by "incoherent", and

the information about the lungs completely disappears. Additionally, we can observe that gender agreement is lost, which is also medically relevant. In the sentence with cyan color, the patient is referred to as "she", but in the yellow sentence, we have the pronoun "he". Another significant mistake is the replacement of "improved" with "worsened", which have completely differing meanings. However, it is unclear why this happened, as it could have resulted from a bad performance of the model or simply bad luck when randomly selecting one of the five tokens selected for replacement.

With the sentence replacement strategies, the note remains much more similar to the original version while maintaining its readability. For example, we can observe that the first and second sentences in the anonymized versions are practically equal to the first and second sentences in the original note. In fact, it inadvertently results in the correction of a syntactic error, as "os" is corrected into "is". Contrary to the word replacement strategy, this time, a lot of the relevant medical information is maintained, such as the confused state of the patient, the hypertension and the lung rales.

Figure 6.1 shows the results obtained for our baseline experiment without any text pre-processing. In this experiment, both the Word2Vec and Doc2Vec models produce embeddings with a dimension of 256. The all-MiniLM-L6-v2, S-PubMedBert-MS-MARCO and GIST-large-Embedding-v0 sentence transformer models produce embeddings with 384, 768 and 1024 dimensions, respectively. The results are presented as the average of the scores obtained across the 19,989 notes used for testing.

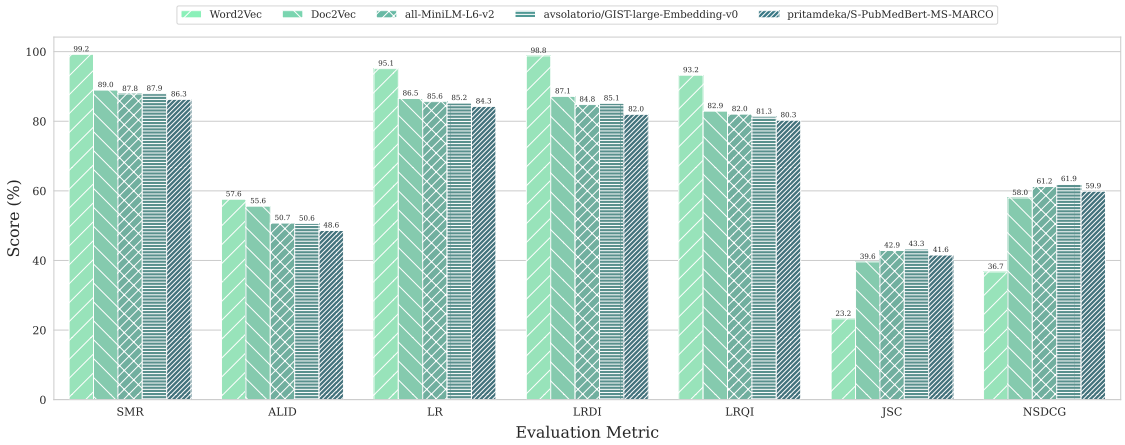


Figure 6.1: Performance results obtained by each model on the different evaluation metrics, without text pre-processing.

We can observe that the word replacement strategy yields better results on all of the anonymization sensitivity metrics (SMR, ALID, LR, LRDI, and LRQI) but performs worse regarding the clinical information retention ones (JSC and NSDCG). This outcome is expected due to the fundamental differences in how the anonymization is conducted. For example, when anonymizing the clinical note "The patient's name is John Doe", the word replacement strategy replaces each word

in the sentence individually. In contrast, sentence replacement might substitute the entire sentence with another that contains common elements, such as "John" or "Doe", thereby negatively impacting the performance of anonymization metrics.

The superior performance of sentence replacement in information retention metrics can be explained similarly. If a medical condition appears in the clinical note to be anonymized, word replacement will result in that term being removed entirely. However, sentence replacement might yield a substitute sentence that still contains the name of the medical condition, thus preserving more clinical information.

Interestingly, the word replacement approach did not achieve a 100% score in any recall-based metrics. This can be attributed to sensitive entities sometimes appearing as a part of non-sensitive entities, which reflects on the metrics as they are based on Levenshtein Distance.

ALID is the anonymization metric with the lowest results for every strategy and model, as it is naturally demanding because it is based on the Levenshtein Ratio, which is only 0 if every character is different between two words. ALID is based on the complement of this ratio, so it would only be 100% if, for every sensitive entity, all the strings with their lengths differed in all characters. As there are always small similarities, even if just one character or two, it already impacts the result.

Regarding anonymization sensitivity metrics, there is no significant difference in performance between the Doc2Vec and Sentence-Transformer models. Notably, Doc2Vec, the model producing vectors with the fewest dimensions, slightly outperformed those generating higher-dimensional vectors. This is likely because higher-dimensional vectors capture a wider range of semantic and syntactic attributes, which may not all be relevant for anonymization purposes. For the anonymization itself, what matters is the replacement of one sentence with another and the semantic similarity aspect, which is the motive behind creating a sentence embedding space, does not come into play.

In contrast, the Sentence-Transformer models perform better than the Doc2Vec model on the information retention metrics. The dimensionality of the vectors likely influences these results, as information retention relies on the similarity between the original and anonymized notes. The avsolatorio/GIST-large-Embedding-v0 model achieves the best performance in both metrics. This result is expected since this model produces vectors with the highest number of dimensions, better capturing useful features for sentence similarity. Additionally, this pre-trained model is among the best performers on the BIOSSES benchmark. Conversely, the lowest performance of the pritamdeka/S-PubMedBert-MS-MARCO model among the three sentence transformer models suggests that the PubMed text it was trained on differs significantly from the clinical text in the MIMIC-III database.

Figure 6.2 illustrates the performance of the same models with the same parameters, but the following pre-processing steps are applied for the embedding generation: lowercasing, removal of non-alphanumeric characters and replacement of consecutive white spaces with a single one.

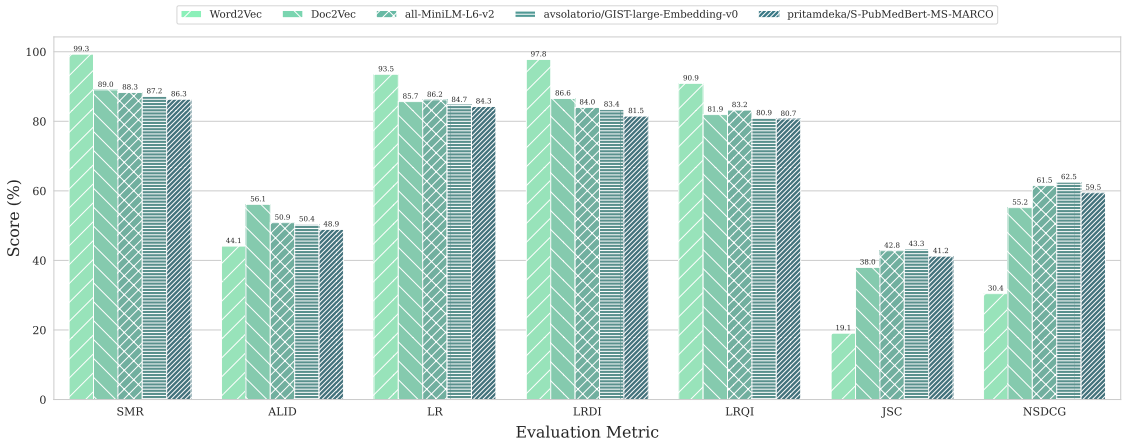


Figure 6.2: Performance results obtained by each model on the different evaluation metrics, without text pre-processing, with lowercasing, removal of non-alphanumeric characters and removal of consecutive white spaces.

On the information retention part, Word2Vec has a drop of 4.2% and 6.3% in the JSC and NSCDG scores, respectively. As we observed in Table 6.1, this word replacement approach benefited from the same word being stored in different variants or with small alterations in the vector space, e.g., "insulin" and "INSULIN" or "Fi02" and "Fio2". With the text pre-processing steps we are applying, namely lowercasing, these variants no longer happen, so during the replacement process, they have to be replaced with an actual different word, resulting in the loss of medical information.

The performance of the sentence embedding models remains fairly the same, with small gains or losses in performance. The best-performing sentence transformer, GIST-large-Embedding-v0, obtains a new highest value for the NSDCG metric, with a small increase of 0.6%, not significant enough to draw conclusions about the influence of pre-processing. The reason that was previously described to justify the decrease in performance for the Word2Vec model regarding the information retention metrics can also happen in the sentence replacement approach but on a smaller scale, as we can have the same sentence written in lower-case or upper-case. The fact that we are applying relatively soft pre-processing steps, compared to stemming, lemmatization or stop-word removal, for example, can also explain the discreet difference in performance. For instance, the removal of consecutive white spaces has no effect on the semantics of the sentences.

We will now only focus on the sentence embedding models, as the sentence replacement strategy is the central point of this work. Figure 6.3 shows the results obtained by each model on the notes of type Nursing/other, which is the type of notes with the biggest presence in the training and testing sets.

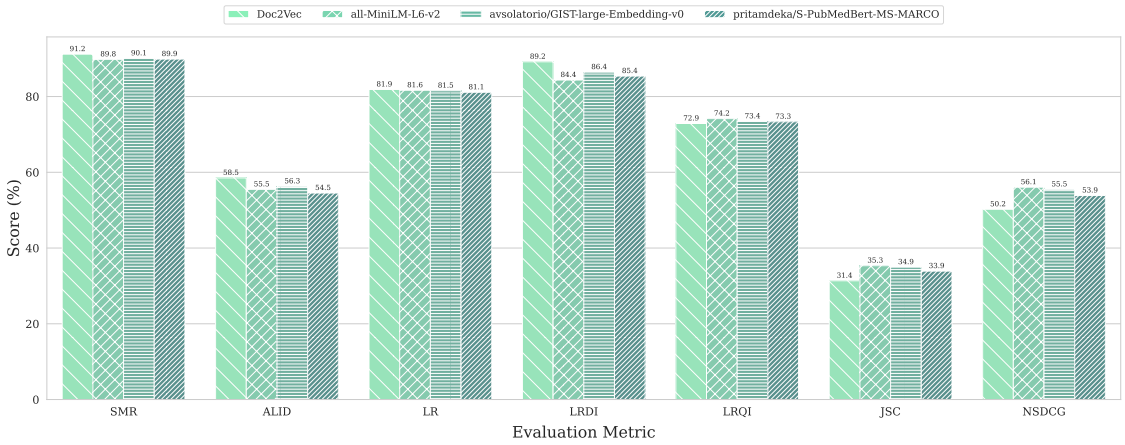


Figure 6.3: Performance results obtained by each model on the different evaluation metrics for the Nursing/other note type.

Compared to the averaged results, on the ALID metric, we can observe an increase of 5.0 to 6.0% for the sentence transformer models. On the other hand, LR scores decrease between 4.5 and 3.2%. On the information retention metrics, JSC and NSCCG, there are notable decreases, with the GIST-large-Embedding-v0 transformer, the best-performing model for these metrics in the general results, suffering drops of 8.4 and 7.0% in performance. In fact, for this note type, the top performer regarding information retention was the all-MiniLM-L6-v2 model. The inconsistent results across this note type, and possibly across the others, could come from the fact that we have one global embedding space with sentences from all note types. Dividing the embedding space with sub-spaces clustered by note type or medical department might benefit the search for similar sentences and would also lower the time cost of the anonymization process.

Table 6.2 contains the performance obtained on the clinical information retention metrics by the different models, including the Doc2Vec model with a different number of dimensions on the embeddings. It allows us to compare its performance with the sentence transformer counterparts regarding the embedding's dimensionality. The all-MiniLM-L6-v2, S-PubMedBert-MS-MARCO and GIST-large-Embedding-v0 models produce embeddings with 384, 768 and 1024 dimensions, respectively.

	JSC (%)	**NSDCG (%)**
Doc2Vec-256	38.00	55.20
Doc2Vec-384	37.82	55.02
all-MiniLM-L6-v2	42.80	61.50
Doc2Vec-768	37.82	54.88
S-PubMedBert-MS-MARCO	41.20	59.50
Doc2Vec-1024	38.30	55.10
GIST-large-Embedding-v0	43.30	62.50

Table 6.2: Performance of various models on the clinical information retention metrics.

We can observe that increasing the number of dimensions for the Doc2Vec model did not result in achieving the performance of the best sentence transformer model regarding any of the two metrics. In fact, there is barely any difference when changing the number of dimensions from 256 to 384, 768 or 1024. We can also see that the Doc2Vec model always obtains worse results than the sentence transformer with the same number of dimensions, suggesting that this lower performance originates from their different architectures, which is expected as the sentence transformers are a more recent approach and for the Doc2Vec training we use the lighter architecture that only leverages the sentence vector of a sentence and not its word vectors.

In summary, while word replacement is effective for anonymization, sentence replacement better retains critical clinical information. However, it is important to notice that the anonymization metrics are not as relevant to this work as the information retention ones, as our strategy is based on the premise that the data used to generate the embeddings contains no real sensitive and personal information, and so neither will the anonymized version. Even so, we feel it is important to show and discuss the results of the anonymization sensitivity metrics, even if just to validate our solution. For example, if the results obtained by the sentence embedding models were much lower than the ones obtained by the word embedding model, there would be reasons to doubt the effectiveness of our solution and the truthfulness of our assumption. Still, it is possible that the anonymization results are being worsened due to the fact that we use the same data for the vector space generation and testing, as the fake entity generator may have produced fake entities that overlap between these two sets, therefore impacting the process of finding similar sentences. For example, on a test note, we may have a sentence containing a fake name, and during the anonymization process, it can be replaced by one containing the same name, which will impact the results, as the annotation contains the fake name. This would be less likely to happen if we had a test set with real sensitive information, as the overlap of entities would be smaller. Even so, if a real sensitive entity happened to occur in a sentence contained in the replacement group, it would merely be a coincidence.

6.2 Time Cost Analysis

The trade-off between the model's performance and its temporal cost is very present in machine and deep learning architectures. Therefore, in addition to assessing the model's performance on the proposed evaluation metrics for anonymization sensitivity and clinical information retention, we also analyze the time cost factor regarding the three main aspects of our solution: embedding space generation, inference and evaluation. All the experiments were performed on a NVIDIA A16-8C GPU with 8 GiB of memory.

Table 6.3 displays the hours taken for the embedding space generation and for the inference process. During the embedding space generation, the models create an embedding space containing 644,052 sentences. During inference, the sentence replacement process is executed for the 19,989 notes reserved for evaluation, using the model and the embedding space to find similar sentences. We compare the Doc2Vec model with different dimensions to the best-performing model

on the information retention metrics.

	Doc2Vec				GIST-large-Embedding-v0
Dimensions	256	384	768	1024	1024
Generation (hours)	0.53	0.59	0.79	0.87	1.96
Inference (hours)	2.46	3.11	4.66	5.53	112.73

Table 6.3: Embedding space generation and inference time results (hours) for different models and different number of dimensions.

We can clearly observe that the Doc2Vec model is much faster at both the embedding space generation and inference. With a 300% increase in the number of dimensions, from 256 to 1024, there is only an increase of approximately 64.15% in the time taken to generate the embedding space. Regarding inference time, when going from vectors with 256 dimensions to vectors with 1024 dimensions, the time cost nearly doubled. However, the most impressive value is the time it took to perform the substitution process using the avsolatorio/GIST-large-Embedding-v0 model, with nearly 5 days being necessary.

As for the Doc2Vec model, we can not say that the increase in time cost (due to the increase in the number of dimensions) results in a better performance. It was observed in Section 6.1 that there is no noticeable difference between a Doc2Vec model with 256 or 1024-dimensional vectors in the clinical information retention metrics, which are the ones that mostly rely on the sentence similarity aspect. In contrast, this trade-off is observed for the sentence transformer model, as it takes longer to generate the embedding space and perform the anonymization, but its performance is better.

6.3 Summary

Most clinical text anonymization systems rely on NER. Although many have achieved great performances, with precision and recall of around 90%, they aren't really used by healthcare institutions. Even if a NER model had a precision and a recall of 100%, there is no guarantee that performance would be maintained when being applied to new data.

In this chapter, we compare two text anonymization strategies based on the substitution of words or sentences that guarantee the removal of sensitive information. However, it comes at the cost of readability and information retention, performing worse in this aspect compared to traditional NER models, as these only replace the tokens they identify as sensitive.

We hope that the presentation and discussion of our results bring insights into the research community and help in the development of more robust clinical text anonymization solutions.

Chapter 7

Conclusion

De-identification of clinical text is crucial for the secure and ethical sharing of health data, as it ensures patient privacy and compliance with legal regulations such as HIPAA and GDPR. By removing or anonymizing PII, de-identification facilitates the use of clinical data in research, public health monitoring, and policy development without compromising individual privacy. This process not only mitigates the risk of data breaches and misuse but also promotes a broader collaboration among healthcare providers, researchers, and policymakers, fostering advancements in medical research and improving public health outcomes.

In this work, we present a comparison between two different and novel techniques for the anonymization of clinical notes that guarantee the removal of all sensitive information within the text, word, and sentence replacement, with a focus on the latter. Five different models were tested and evaluated across several evaluation metrics aimed at anonymization sensitivity and clinical information retention. The discussed results indicate that both replacement techniques have their unique strengths and are viable alternatives to the traditional NER approaches when the removal of sensitive information is a priority over data usefulness, as the latter are never capable of detecting all the sensitive information. The sentence embeddings-based substitution method also proved to outperform the word embedding-based substitution regarding clinical information retention, which was one of the underlying motivations for the realization of this work. Additionally, state-of-the-art sentence transformers performed the best on the information retention metrics, highlighting their capacity to capture similarity between sentences. The best sentence embedding model obtains gains from 20 to 32% over the tested word embedding model on clinical information retention metrics, answering our second research question.

Answering our first research question, we understand that there is no perfect solution, as they all rely on a trade-off between sensitive information removal and future data utility, and achieving 100% performance on both seems unrealistic. While our solution ensures anonymity, the clinical information retention results vary from 40% to 60%, which might not be enough to justify its usage. Additionally, our information retention metrics are based on a single task of ICD-10 code category classification. This is an obvious limitation of our approach, but there are steps that can be taken to improve it.

7.1 Limitations

While the two compared strategies ensure the removal of sensitive information by replacing every word or sentence with similar counterparts from a de-identified dataset, it does come with notable limitations, particularly concerning readability and data usefulness. There is no guarantee that the anonymized version of the note will be semantically or syntactically correct, leading to potential disruptions in the coherence of the clinical document. Consequently, the agreement on attributes such as gender, age group, and person may not be consistently maintained throughout the new clinical note.

One significant drawback of the word replacement approach is the loss of relevant medical terms. If a crucial medical term appears in the original clinical notes, it is guaranteed not to appear in the anonymized version, as every word is replaced. This issue is mitigated with the sentence replacement approach, which shows better performance in clinical information retention metrics. However, if the clinical note contains a sentence with a medical term not present in any sentence of the replacement group, that term will also be permanently lost, compromising the retention of essential medical information.

Furthermore, we assume that the retention of relevant clinical information can be fully assessed by an ICD-10 code category classification task. While this provides a reasonable overview of the information lost during anonymization, it is not exhaustive. Other evaluation strategies could yield deeper insights. For instance, training models for tasks such as clinical NER, diagnosis factuality or medical progress on both original and anonymized data can provide a more detailed assessment. By comparing the performance of these models, we can better evaluate how well the anonymized data retains critical information necessary for various clinical applications, thus offering a more comprehensive assessment of the de-identification method's effectiveness.

Another potential limitation is the use of the same database for both embedding generation and anonymization evaluation. This longstanding issue in text anonymization research arises because solutions are often tailored to specific datasets or note types, with no guarantee of consistent performance across different scenarios. Using the same type and structure of clinical notes throughout our process may facilitate the identification of similar words or sentences, potentially inflating our results regarding information retention. The performance observed in these experiments might be lower if evaluated on a different dataset, where finding similar words or sentences would be more challenging.

Lastly, the reliance on sentence embeddings themselves introduces some limitations. While embeddings capture semantic similarity, they may not always preserve the nuanced meanings crucial in clinical contexts. The embedding-based approach might miss subtle yet important distinctions in medical terminology and context, affecting the overall accuracy and reliability of the de-identification process.

In summary, while our approach shows promise in removing sensitive information from clinical texts, these limitations highlight the need for further research to enhance the readability, accuracy, and applicability of anonymized clinical documents across diverse datasets and real-world

scenarios.

7.2 Future Work

This research opens several avenues for future exploration, particularly in assessing performance across different languages and integrating evaluation metrics during the anonymization process.

One important direction is to extend our anonymization method to various languages. While this study focused on English clinical texts, future work should investigate the effectiveness of sentence embeddings in multilingual contexts. As most of the data available for clinical and medical NLP tasks is in English, this would likely involve a translation module since it is not possible to obtain an embedding space for every language. For example, if we wanted to anonymize a clinical note written in Portuguese but our embedding space was composed of English sentences, a possible strategy would be to translate the original note into English, then perform the anonymization step, and finally translate it back into Portuguese.

Another promising area for future research is the integration of evaluation metrics into the anonymization process itself rather than applying them solely at the end to assess performance. By incorporating real-time feedback mechanisms, the anonymization process can be dynamically adjusted to preserve critical clinical information better. For instance, instead of finding the replacement sentences considering only the cosine similarity, the NSDCG metric could also be considered during that process. Imagine we wanted to replace a sentence, S_A, and had two candidate sentences, S_B and S_C. The cosine similarity between S_A and S_B and S_C is 0.9 and 0.8, respectively. However, imagine that the NSDCG for S_B was 0.2, while for S_C it was 0.6. It could make sense to sacrifice the 0.1 drop in similarity for the 0.4 increase in information retention.

Exploring these aspects will contribute to the development of a more robust and versatile de-identification technique, ensuring that clinical data remains both secure and useful for a wider range of applications.

References

[1] Mohamed Abdalla, Moustafa Abdalla, Frank Rudzicz, and Graeme Hirst. Using word embeddings to improve the privacy of clinical notes. *Journal of the American Medical Informatics Association*, 27(6):901–907, 05 2020.

[2] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey, 2023.

[3] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[4] Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. Automatic deidentification by using sentence features and label consistency. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 01 2006.

[5] Bruce A Beckwith, Rajeshwarri Mahaadevan, Ulysses J Balis, and Frank Kuo. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making*, 6(1):12, March 2006.

[6] Hanna Berg, Aron Henriksson, and Hercules Dalianis. The impact of de-identification on downstream named entity recognition in clinical text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online, November 2020. Association for Computational Linguistics.

[7] William Blacoe and Mirella Lapata. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[8] Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. Crosslingual named entity recognition for clinical de-identification applied to a covid-19 italian data set. *Applied Soft Computing*, 97:106779, 2020.

[9] Rosario Catelli, Francesco Gargiulo, Valentina Casola, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. A novel covid-19 data set and an effective deep learning approach for the de-identification of italian medical records. *IEEE Access*, 9:19097–19110, 2021.

[10] Raphaël Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis. Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *J Med Internet Res*, 21(5):e13484, May 2019.

[11] Committee on Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences Policy, and Institute of Medicine. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Appendix B, Concepts and methods for De-identifying clinical trial data*. National Academies Press, Washington, D.C., DC, April 2015.

[12] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606, 12 2016.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] David A. Dorr, William Phillips, Shobha Phansalkar, Shannon A. Sims, and John F. Hurdle. Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of Information in Medicine*, 45:246 – 252, 2006.

[15] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[16] Óscar Ferrández, Brett South, Shuying Shen, F Friedlin, Matthew Samore, and Stephane Meystre. Evaluating current automatic de-identification methods with veteran's health administration clinical documents. *BMC Medical Research Methodology*, 12:109, 07 2012.

[17] Óscar Ferrández, Brett South, Shuying Shen, F Friedlin, Matthew Samore, and Stephane Meystre. Generalizability and comparison of automatic clinical text de-identification methods and resources. *AMIA Annual Symposium Proceedings*, 2012:199–208, 11 2012.

[18] F Jeff Friedlin and Clement J McDonald. A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, 15(5):601–610, September 2008.

[19] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

[20] Jigna J. Hathaliya and Sudeep Tanwar. An exhaustive survey on security and privacy issues in healthcare 4.0. *Computer Communications*, 153:311–335, 2020.

[21] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37:547–579, 1901.

[22] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[23] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 2016.

[24] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. A survey of word embeddings for clinical text. *Journal*

of Biomedical Informatics, 100:100057, 2019. Articles initially published in Journal of Biomedical Informatics: X 1-4, 2019.

[25] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[26] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[27] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[28] Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online, August 2021. Association for Computational Linguistics.

[29] Qi Liu, Matt J. Kusner, and Phil Blunsom. A survey on contextual embeddings, 2020.

[30] Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75:S34–S42, 2017. Supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

[31] Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, Fang Zeng, Lichao Sun, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. Deid-gpt: Zero-shot medical text de-identification by gpt-4, 2023.

[32] Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. Evaluating the impact of text de-identification on downstream NLP tasks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 10–16, Tórshavn, Faroe Islands, May 2023. University of Tartu Library.

[33] Martin Malmsten, Love Börjeson, and Chris Haffenden. Playing with words at the national library of sweden – making a swedish bert, 2020.

[34] Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@SEPLN*, 2019.

[35] Christopher Meaney, Wali Hakimpour, Sumeet Kalia, and Rahim Moineddin. A comparative evaluation of transformer models for de-identification of clinical text data. 2022.

[36] Stephane Meystre, F Friedlin, Brett South, Shuying Shen, and Matthew Samore. Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Medical Research Methodology*, 10:70, 08 2010.

[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[38] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[39] Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. In *Conference and Labs of the Evaluation Forum*, 2020.

[40] Ishna Neamatullah, Margaret M Douglass, Li-Wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. Automated de-identification of free-text medical records. *BMC Med. Inform. Decis. Mak.*, 8(1):32, July 2008.

[41] Jihad S. Obeid, Paul M. Heider, Erin R. Weeda, Andrew J. Matuskowitz, Christine M. Carr, Kevin Gagnon, Tami Crawford, and Stephane M. Meystre. Impact of de-identification on clinical text classification using traditional and deep learning classifiers. *Studies in health technology and informatics*, 264:283–287, August 2019.

[42] U.S. Department of Health and Human Services Office for Civil Rights. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard, 2022. [Online; Accessed: January 6, 2024].

[43] U.S. Department of Health and Human Services Office for Civil Rights. The hipaa privacy rule. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html, 2022. [Online; Accessed: January 6, 2024].

[44] Council of the European Union. Art. 4 gdpr - definitions. https://gdpr-info.eu/art-4-gdpr/. [Online; Accessed: January 7, 2024].

[45] Council of the European Union. Recital 26 - not applicable to anonymous data. https://gdpr-info.eu/recitals/no-26/. [Online; Accessed: January 6, 2024].

[46] Babita Pandey, Devendra Kumar Pandey, Brijendra Pratap Mishra, and Wasiur Rhmann. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions. *Journal of King Saud University - Computer and Information Sciences*, 34(8, Part A):5083–5099, 2022.

[47] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[48] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[49] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101, December 2022.

[50] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[51] David Pissarra, Isabel Curioso, João Alveira, Duarte Pereira, Bruno Ribeiro, Tomás Souper, Vasco Gomes, André V. Carreiro, and Vitor Rolla. Unlocking the potential of large language models for clinical text anonymization: A comparative study, 2024.

[52] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[53] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[54] Bruno Ribeiro, Vitor Rolla, and Ricardo Santos. INCOGNITUS: A toolbox for automated clinical notes anonymization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 187–194, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[55] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, 2018.

[56] Amber Stubbs, Michele Filannino, and Özlem Uzuner. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of Biomedical Informatics*, 75:S4–S18, 2017. Supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.

[57] Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of Biomedical Informatics*, 58:S11–S19, 2015. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

[58] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *J. of Biomedical Informatics*, 58(S):S20–S29, dec 2015.

[59] European Data Protection Supervisor. Health. https://edps.europa.eu/data-protection/our-work/subjects/health_en. [Online; Accessed: January 6, 2024].

[60] Latanya Sweeney. Replacing personally-identifying information in medical records, the scrub system. *Proceedings : a conference of the American Medical Informatics Association. AMIA Fall Symposium*, pages 333–7, 1996.

[61] György Szarvas, Richárd Farkas, and Róbert Busa-Fekete. State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework. *Journal of the American Medical Informatics Association*, 14(5):574–580, 09 2007.

[62] Head of Technology Thomas Zerdick and Privacy. Pseudonymous data: processing personal data while mitigating risks. https://edps.europa.eu/press-publications/press-news/blog/pseudonymous-data-processing-personal-data-while-mitigating_en, 2021. [Online; Accessed: January 7, 2024].

[63] Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[64] United States Congress. Health insurance portability and accountability act of 1996 (HIPAA), 1996.

[65] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 09 2007.

[66] Thomas Vakili and Hercules Dalianis. Utility preservation of clinical text after de-identification. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 383–388, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[67] Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France, June 2022. European Language Resources Association.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[69] Ben Wellner, Matt Huyck, Scott Mardis, John Aberdeen, Alex Morgan, Leonid Peshkin, Alex Yeh, Janet Hitzeman, and Lynette Hirschman. Rapidly Retargetable Approaches to De-identification in Medical Records. *Journal of the American Medical Informatics Association*, 14(5):564–573, 09 2007.

[70] WHO. *ICD-10: international statistical classification of diseases and related health problems: tenth revision*. World Health Organization, 2004.

[71] Wikipedia contributors. Levenshtein distance — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=1223925517, 2024. [Online; accessed 16-June-2024].

[72] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 06 2018.

[73] Hui Yang and Jonathan M. Garibaldi. Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, 58:S30–S38, 2015. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

[74] Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R. Hogan, and Yonghui Wu. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(S5):232, December 2019.

[75] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: Quantifying biases in clinical contextual word embeddings, 2020.

[76] Özlem Uzuner, Tawanda C. Sibanda, Yuan Luo, and Peter Szolovits. A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine*, 42(1):13–35, 2008.

Appendix A

Dataset Samples

<RECORD ID="1">
<TEXT>
<PHI TYPE="ID">101126659</PHI>
<PHI TYPE="HOSPITAL">MGH</PHI>
<PHI TYPE="DATE">10/29</PHI>/1997 12:00:00 AM
CARCINOMA OF THE COLON .
Unsigned
DIS
Report Status :
Unsigned
Please do not go above this box important format codes are contained .
DISCHARGE SUMMARY
<PHI TYPE="ID">FMT51 DS</PHI>
DISCHARGE SUMMARY NAME :
<PHI TYPE="PATIENT">SLOAN , CHARLES E</PHI>
UNIT NUMBER :
<PHI TYPE="ID">358-51-76</PHI>
ADMISSION DATE :
<PHI TYPE="DATE">10/29</PHI>/1997
DISCHARGE DATE :
<PHI TYPE="DATE">11/02</PHI>/1997
PRINCIPAL DIAGNOSIS :
Carcinoma of the colon .
ASSOCIATED DIAGNOSIS :
Urinary tract infection , and cirrhosis of the liver .
HISTORY OF PRESENT ILLNESS :
The patient is an 80-year-old male , who had a history of colon cancer in the past , resected approximately ten years prior to admission , history of heavy alcohol use , who presented with a two week history of poor PO intake , weight loss , and was noted to have acute on chronic Hepatitis by chemistries and question of pyelonephritis .
</TEXT>
</RECORD>

Figure A.1: Sample discharge summary excerpt from the 2006 i2b2 de-identification corpus using XML representation (obtained from [65]).

Record date: <DATE>2074-04-05</DATE>

<PHI TYPE="HOSPITAL">EMMANUEL HOME</HOSPITAL> EMERGENCY DEPT VISIT

<PATIENT>JACOB,LARRY</PATIENT>
<MEDICALRECORD>910-66-83-7</MEDICALRECORD>
VISIT DATE: <DATE>04/05/74</DATE>

This patient was seen, interviewed and examined by myself as well as Dr. <DOCTOR>Naylor</DOCTOR> whose I have reviewed and whose findings I have confirmed.

HISTORY OF PRESENTING COMPLAINT: This is a <AGE>53</AGE>-year-old male who
[...]
Follow-up appointment scheduled for <DATE>Apr 16th</DATE>

<IDNUM>QC920/47122</IDNUM>

<DOCTOR>ISABELLA COOK</DOCTOR>, M.D.
<USERNAME>IC39</USERNAME>
D:<DATE>04/05/74</DATE>
T:<DATE>04/06/74</DATE>
Dictated by: <DOCTOR>ISABELLA COOK</DOCTOR>, M.D.
<USERNAME>IC39</USERNAME>

Figure A.2: Sample of clinical note of the 2014 i2b2/UTHealth de-identification corpus using XML representation (obtained from [58]).