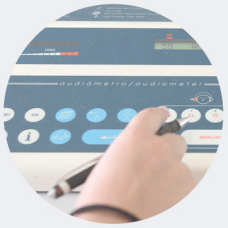
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA BIOMÉDICA






ANONIMIZAÇÃO AUTOMÁTICA DE TEXTO CLÍNICO: UM ESTUDO SOBRE TÉCNICAS EMERGENTES E MÉTODOS DE AVALIAÇÃO

RITA ALEXANDRE PINTO RIBEIRO
Setembro de 2023

P.PORTO

Instituto Superior de Engenharia do Porto

ANONIMIZAÇÃO AUTOMÁTICA DE TEXTO CLÍNICO: UM ESTUDO SOBRE TÉCNICAS EMERGENTES E MÉTODOS DE AVALIAÇÃO

Rita Alexandre Pinto Ribeiro

2023

Instituto Superior de Engenharia do Porto

Departamento de Física

P.PORTO

ANONIMIZAÇÃO AUTOMÁTICA DE TEXTO CLÍNICO: UM ESTUDO SOBRE TÉCNICAS EMERGENTES E MÉTODOS DE AVALIAÇÃO

Rita Alexandre Pinto Ribeiro

Estudante n.º 1210164

Dissertação apresentada ao Instituto Superior de Engenharia do Porto para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Biomédica, realizada sob a orientação da Professora Doutora Maria Goreti Carvalho Marreiros e coorientação do Doutor Vitor Guerra Rolla.

2023

Instituto Superior de Engenharia do Porto

Departamento de Física


isep
Instituto Superior de
Engenharia do Porto


P.PORTO


isep

AGRADECIMENTOS

Gostaria de agradecer, primeiramente, ao Instituto Superior de Engenharia do Porto por me ter acolhido nestes últimos dois anos, contribuindo para a minha formação e desenvolvimento académico e pessoal.

À Professora Doutora Goreti Marreiros agradeço toda a dedicação, empenho e orientação dada durante este processo.

Ao Doutor Vitor Rolla agradeço pela coorientação, pela positividade e incentivo dado ao longo deste período e pela partilha de conhecimento, fulcral para o meu desenvolvimento. Também ao Investigador Bruno Ribeiro tanto pela partilha de conhecimento como pelo encorajamento e dedicação que contribuíram para o sucesso do meu percurso.

Gostaria de estender os meus agradecimentos à Fraunhofer Portugal por me ter recebido de forma tão acolhedora e amigável.

Quero também expressar gratidão à minha família, pais, irmão e avós, pelo apoio incondicional que me proporcionaram e por acreditarem sempre em mim.

Aos meus amigos e namorado pela paciência, apoio e motivação durante esta fase.

página propositadamente em branco

RESUMO

O Processamento de Linguagem Natural (PLN) teve uma evolução explosiva nos últimos 5 anos, principalmente devido ao desenvolvimento e utilização de Modelos de Linguagem baseados em *Deep Learning*, como BERT (*Bidirectional Encoder Representatioms from Transformers*) e GPT (*Generative Pre-trained Transformer*), surgindo assim os LLMs (*Large Language Models*).

A anonimização do texto clínico é uma tarefa crucial para mitigar preocupações de privacidade ao lidar com dados clínicos sensíveis, presentes em Registos Eletrónicos de Saúde e notas clínicas. Vários métodos de PLN podem ser implementados para executar esta tarefa automaticamente, evitando a morosa desidentificação manual do texto.

Uma das maneiras de realizar automaticamente a anonimização de texto clínico é através da técnica de Reconhecimento de Entidade Nomeada (REN) onde um modelo de PLN pode identificar os *tokens* que correspondem a Informações Privadas de Saúde (IPS) num texto, como o nome de um paciente, idade, o nome do hospital, etc. Outra possibilidade é através da utilização da estratégia de substituição por *word embeddings*, que substituem cada palavra de um determinado texto por outras semanticamente relacionadas. No caso de dados clínicos, as informações médicas relevantes devem permanecer inalteradas após a anonimização, o que pode ser avaliado extraindo códigos ICD-10.

Este estudo teve como objetivo comparar o desempenho das técnicas de anonimização baseadas em REN (CRF (*Conditional Random Field*) e Presidio com o modelo spaCy) com as técnicas baseadas em *word embeddings* (Word2Vec e GloVe) para perceber se estas últimas podem ser consideradas uma alternativa mais viável para esta tarefa.

Além disso, foram realizadas experiências em dois contextos linguísticos diferentes: inglês e português. Os resultados deste estudo comparativo entre idiomas diferentes demonstram que, apesar dos escassos dados disponíveis para idiomas de baixo recurso (como o caso do português), grande parte das tendências observadas com os dados ingleses será extensível a outros idiomas.

Para acompanhar este tipo de técnicas emergentes foi necessário desenvolver uma nova métrica – *Levenshtein Recall* (LR) – de forma a ultrapassar os desafios encontrados pelas métricas tradicionais.

Com este estudo conclui-se que os métodos baseados em REN ainda são os mais apropriados para anonimização de texto clínico, ainda que os métodos baseados em *word embeddings* se revelem muito promissores nesta tarefa de PLN, com grande poder de anonimização, mas a custo de grande perda de informação clínica.

PALAVRAS-CHAVE: Anonimização, texto clínico, *word embeddings*, Reconhecimento de Entidades Nomeadas, Processamento de Linguagem Natural

página propositadamente em branco

ABSTRACT

Natural Language Processing (NLP) has had an explosive evolution in the last 5 years, mainly due to the development and use of Language Models based on Deep Learning, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), thus giving rise to LLMs (Large Language Models).

Anonymizing clinical text is crucial to mitigate privacy concerns when dealing with sensitive clinical data in Electronic Health Records and clinical notes. Various NLP methods can be implemented to perform this task automatically, avoiding time-consuming manual text de-identification.

One of the ways to automatically perform clinical text anonymization is through the Named Entity Recognition (NER) technique using the CRF (Conditional Random Field) and Presidio methods with the spaCy model, where an NLP model can identify the tokens, you find Private Health Information (PHI) in text, such as a patient's name, age, hospital name, etc. Another possibility is through the use of word embeddings (such as Word2Vec and GloVe) that replace each word in a given text with other semantically related ones. In the case of clinical data, relevant medical information must remain unchanged after anonymization or can be evaluated by extracting ICD-10 codes.

To keep up with these emerging techniques, it was necessary to develop a new metric – Levenshtein Recall (LR) – to overcome the challenges encountered by traditional metrics.

This study concludes that methods based on NER are still the most suitable for anonymizing clinical text, although methods based on word embeddings are very promising in this NLP task, with great anonymization power but at the cost of a great loss of clinical information.

KEYWORDS: Anonymization, clinical text, word embeddings, Named Entity Recognition, Natural Language Processing

página propositadamente em branco

ÍNDICE

página propositadamente em branco

ÍNDICE DE FIGURAS

página propositadamente em branco

ÍNDICE DE TABELAS

página propositadamente em branco

LISTA DE SIGLAS

Lista de Siglas

BERT	*Bidirectional Encoder Representations from Transformers* (Representações Bidirecionais de Codificadores de Transformadores)
CIL	*Clinical Information Loss* (Perda de Informação Clínica)
CNN	*Convolutional Neural Network* (Rede Neuronal Convolucional)
CRF	*Conditional Random Field* (Campo Aleatório Condicional)
EHR	*Eletronic Health Record* (Registo Eletrónico de Saúde)
ELMo	*Embeddings from Language Models* (Incorporações de Modelos de Linguagem)
GAN	*Generative Adversarial Network* (Rede Generativa Adversarial)
GDPR	*General Data Protection Regulation* (Regulamento Geral de Proteção de Dados)
GPT	*Generative Pre-trained Transformer* (Transformador Pré-treinado Generativo)
HIPAA	*Health Insurance Portability and Accountability Act* (Lei de Portabilidade e Responsabilidade do Seguro de Saúde)
HIDE	*Health Information DE-identification*
i2b2	*Informatics for Integrating Biology and the Bedside*
LSTM	*Long short-term memory* (Memória de Longo Prazo)
LD	*Levenshtein Distance* (Distância de Levenshtein)
LR	*Levenshtein Recall* (Lembrança de Levenshtein)
LRa	*Levenshtein Ratio* (Rácio de Levenshtein)
LSI	*Levenshtein Similarity Index* (Índice de Similaridade de Levenshtein)
LLM	*Large Language Model* (Modelo de Linguagem Grande)
NLP	*Natural Language Processing* (Processamento de Linguagem Natural)
NER	*Named Entity Recognition* (Reconhecimento de Entidade Nomeada)
OLA	*Optimal Lattice Anonymization*
PHI	*Private Health Information* (Informação Privada de Saúde)
REGEX	*Regular Expressions* (Expressões Regulares)
RNN	*Recurrent Neural Network* (Rede Neuronal Recorrente)
SVM	*Support Vector Machine* (Máquina de Vetores de Suporte)

página propositadamente em branco

1. INTRODUÇÃO

Os registos eletrónicos de saúde (EHR) são documentos que contêm anotações médicas em texto livre sobre o estado clínico de pacientes, tornando estes registos em recursos valiosos que poderiam ser usados em pesquisas médicas de grande escala (Friedrich et al., 2020).

A utilização dessas notas clínicas para um âmbito estatístico ou de investigação necessita de um processo de desidentificação e anonimização dos dados dos pacientes de forma a proteger a sua privacidade e confidencialidade (Yang et al., 2019). A desidentificação é uma etapa crucial de limpeza dos EHR e tem como objetivo detetar e remover ou substituir as Informações Privadas de Saúde (PHI) em notas clínicas (Friedrich et al., 2020).

Através de Processamento de Linguagem Natural (NLP) é possível desenvolver sistemas que identificam automaticamente informações pessoais em notas clínicas e as substituem por identificadores anónimos, protegendo assim a privacidade dos pacientes. Esse processo de desidentificação pode ser aplicado em larga escala, permitindo que grandes conjuntos de dados sejam desidentificados com rapidez e eficiência.

O conceito de *word embeddings* que tem demonstrado muito sucesso em variadas tarefas de NLP, apresenta potencial na anonimização das PHI nas notas clínicas, com a intenção de utilidade para fins de pesquisa. Esta técnica será explorada de forma a entender o seu potencial e aumentar a sua eficácia.

Tal como as *word embeddings*, a emergência de novas técnicas, como as baseadas em redes adversariais generativas (GANs) e em *Large Language Models* (LLMs), impulsionam a necessidade de uma melhor avaliação na tarefa de anonimização. Devido à sua diferente forma de atuar para esta tarefa de NLP, as métricas tradicionais como a *recall*, a precisão e a *f1-score* podem já não conseguir determinar o valor real que estas métricas pressupõem, daí a necessidade de exploração de novos métodos de avaliação.

1.1. Enquadramento e pertinência

A desidentificação de notas clínicas é um tema de grande importância na área da saúde. Existe uma grande quantidade de dados sensíveis que são recolhidos e armazenados em hospitais e centros de saúde e a utilização desses dados em pesquisas médicas pode trazer avanços significativos na medicina. Para isso é fundamental garantir a privacidade e confidencialidade dos pacientes (Abdalla et al., 2020; Friedrich et al., 2020).

A desidentificação refere-se à remoção ou substituição de informações que possam identificar os pacientes, como nomes, datas de nascimento e números de identificação pessoal. O objetivo é permitir que as informações sejam compartilhadas com investigadores e outros profissionais de saúde, sem expor os pacientes a riscos de privacidade (Abdalla et al., 2020).

Considerando a crescente preocupação global com a privacidade dos dados em saúde é evidente a necessidade da criação de mecanismos de anonimização, seguindo e cumprindo os regulamentos de cada região como o Regulamento Geral de Proteção de Dados (GDPR), da União Europeia, ou o *Health Insurance Portability and Accountability Act* (HIPAA), dos Estados Unidos.

Um estudo recente de Abdalla et al. apresentou uma nova abordagem que utiliza medidas de proximidade entre *word embeddings*. Os autores propõem substituir cada *token* por outro que seja semanticamente próximo ao original no espaço latente dos *embeddings* para colmatar a negligência presente nos algoritmos de Reconhecimento de Entidades Nomeadas (NER). Esta nova técnica de ofuscação garante a remoção de todos os dados sensíveis (atingindo 100% de *recall*), mas poderá afetar a legibilidade das informações das notas clínicas e resultar em acentuadas perdas de informação, pelo que é necessário explorá-la para aperfeiçoar estes últimos aspetos (Abdalla et al., 2020).

Outras dificuldades encontradas nas tarefas de NLP são a escassa pesquisa e aplicação de idiomas de baixo recurso na área clínica, como é o caso do português. Embora haja avanços constantes e melhorias de desempenho das tarefas de NLP, como em NER, em inglês, o mesmo não acontece para outros idiomas (Schneider et al., 2020).

Além disso, em métodos tradicionais (como as técnicas baseadas em NER), as métricas comumente utilizadas são *recall*, precisão e *f1-score*. Estas métricas pressupõem que cada *token* está associado a um rótulo e é relativamente simples aplicar o cálculo. No entanto, surgem vários desafios quando se recorre a outros métodos de anonimização (como os baseados em *word embeddings* ou os modelos generativos), como a não garantia de que as entidades sensíveis permaneçam na mesma localização na versão original e na anonimizada.

Este foi um estudo proposto e desenvolvido na empresa Fraunhofer Portugal para realizar uma comparação das técnicas baseadas em NER com as técnicas baseadas em *word embeddings* na tarefa de anonimização de texto clínico, explorando a estratégia de substituição de todos os *tokens* da nota clínica original, de Abdalla et al..

A grande contribuição deste estudo, além da análise da comparação de técnicas e métodos de anonimização, foi a proposta de uma nova métrica independente da associação de um *token* com um rótulo e assim ser passível de ser usada para qualquer método de anonimização – a *Levenshtein Recall* (LR).

Em suma, o tema da anonimização é de grande importância e relevância que necessita de contínua exploração de novos métodos e já existentes e métricas mais inovadoras e complexas para avaliação das anonimizações. Apesar dos elevados desempenhos demonstrados na literatura, o problema da anonimização está longe de estar resolvido. É, por isso, urgente a investigação nesta área de forma a atingir a maior privacidade dos pacientes assim como a preservação das informações relevantes para as pesquisas médicas.

1.2. Questão e objetivos de investigação

De acordo com o enquadramento e pertinência anteriormente descritos, neste trabalho pretende-se investigar a eficácia da técnica *word embeddings* como técnica de anonimização de texto clínico. A questão central que conduz esta investigação é: "Serão as técnicas de anonimização baseadas em *word embeddings* alternativas viáveis às técnicas baseadas em Reconhecimento de Entidades Nomeadas? Como poderemos avaliá-las de forma justa e desbloquear o seu potencial?".

Pretende-se testar diferentes modelos de *word embeddings* para se entender como os diferentes algoritmos, assim como os diferentes parâmetros na fase de treino dos modelos dos algoritmos, podem afetar o desempenho final destes e, assim, determinar qual o melhor algoritmo para a tarefa da anonimização. Pretende-se estudar, analisar e avaliar a tarefa de anonimização não só para notas clínicas em inglês como também para outros idiomas de baixo recurso, neste caso o português.

Para a avaliação do sistema de anonimização vão ser exploradas novas métricas que possam avaliar a eficácia e qualidade das técnicas e algoritmos utilizados em termos de privacidade, confidencialidade, qualidade, legibilidade e consequente perda de informação dos dados resultantes das notas clínicas originais.

São considerados como objetivos de investigação os seguintes:

- Observação de qual método obteve melhor desempenho tanto com os dados em inglês como com os dados em português;
- Análise da diferença de resultados quando utilizados dados em inglês e quando utilizados dados em português para treino e teste de diferentes modelos de NLP;
- Exposição das diferenças entre a utilização do método Word2Vec e do método GloVe na anonimização de notas clínicas, em ambos os idiomas;
- Aplicação de métricas inovadoras capazes de ultrapassar desafios impostos por métricas tradicionais para métodos de anonimização mais complexos (*word embeddings* ou métodos generativos);
- Compreensão dos resultados das métricas na alteração de alguns parâmetros definidos;
- Demonstração de qual o tipo de métodos de anonimização (NER ou *word embeddings*) com melhor desempenho na tarefa de anonimização de notas clínicas.

1.3. Opções metodológicas

Nesta secção são expostas as opções metodológicas que foram utilizadas neste estudo. Primeiramente é aprofundado o conceito de NLP e as suas várias vertentes, seguido da técnica de NER e da estratégia de substituição através de *word embeddings*, terminando com a explicação da adoção de novas métricas ao invés das métricas tradicionais.

1.3.1. Métodos de Anonimização

Para a anonimização das notas clínicas foram utilizadas duas técnicas de NLP: o NER e a ofuscação do texto clínico através de *word embeddings*.

Processamento de Linguagem Natural

O NLP dedica-se a compreender a linguagem humana de forma automática e computacional. Este tem uma vasta possibilidade de aplicações em várias áreas como a classificação de documentos, tradução automática, análise de sentimentos, geração de texto e até prever diagnósticos de doenças.

Uma das aplicações largamente mencionada na literatura relacionada com NLP é a anonimização automática de notas clínicas. O NLP é utilizado para identificar as informações pessoais identificáveis e torná-las anónimas.

Para esta tarefa, existem algumas principais abordagens utilizadas em NLP, como:

- **Baseadas em regras (*rule-based*)** (Dehghan et al., 2015; Povlsen et al., 2016) – esta utiliza regras gramaticais e linguísticas para analisar e interpretar texto, como expressões regulares, dicionários de sinónimos, listas de palavras-chave, análise de estrutura e classificação de entidades.
- **Baseadas em estatísticas** (Taira et al., 1999) – utiliza modelos e padrões estatísticos para extrair informação do texto através de supressão de informações pessoais, perturbação aleatória e substituição por frequência, entre outras.
- **Baseadas em dicionário** (Dehghan et al., 2015) – utiliza um conjunto de palavras e frases pré-definidos para analisar o texto, usando técnicas como mapeamento de entidades, substituição por sinónimos, ocultação por ofuscação e remoção de entidades.
- ***Machine Learning*** (Mollaei et al., 2022; Pethani & Dunn, 2023) – utiliza algoritmos de *machine learning* de forma a ensinar o computador a identificar padrões e relações de texto através de técnicas como aprendizagem supervisionada, aprendizagem não supervisionada e *deep learning*.
- **Baseadas em representação de palavras (*word embeddings*)** (Abdalla et al., 2020; Ribeiro et al., 2023) – utiliza modelos de forma a representar as palavras num espaço vetorial, onde existem palavras semelhantes próximas umas das outras. Existem modelos para realizar esta tarefa através de *word embeddings* não-contextuais, como o Word2Vec, o GloVe ou o FastText, e existem modelos de *word embeddings* contextuais como é caso do BERT (*Bidirectional Encoder Representatioms from Transformers*), do ELMo (*Embeddings from Language Models*) ou do GPT (*Generative Pre-trained Transformer*).
- **Métodos generativos** (Chen et al., 2020) – focam-se na criação de dados sintéticos semelhantes a dados reais. Surgiram com o aparecimento dos modelos de *deep learning*. Algumas aplicações destes métodos são na geração de textos, imagens ou áudios de maior qualidade e são considerados métodos importantes em tarefas criativas e de simulação. Existem alguns exemplos como as GANs ou os LLMs.

Reconhecimento de Entidades Nomeadas

O NER é uma técnica utilizada em NLP com o objetivo de identificar e classificar entidades nomeadas em texto, como nomes próprios, datas, locais, entre outros. Esta técnica é baseada em modelos de *machine learning*, de aprendizagem supervisionada ou semi-supervisionada.

No contexto da anonimização de notas clínicas, o NER é usado para identificar e classificar informações sensíveis que carecem de anonimização, como nomes de pacientes, nomes de médicos, datas de nascimento, contactos telefónicos, nomes de hospitais ou clínicas, entre outras.

Geralmente, a técnica de NER é baseada em modelos de aprendizagem supervisionada que são treinados com um conjunto de dados rotulados manualmente para aprender a identificar e classificar as entidades nomeadas. Após o treino deste conjunto de dados, o modelo pode ser utilizado para identificar as entidades nomeadas em novas notas clínicas automaticamente. Alguns exemplos de modelos que podem ser treinados para tarefas de NER são o CRF (*Conditional Random Field*) (Lafferty et al., 2001) e o BERT (Devlin et al., 2019).

Na seguinte Figura 1 está demonstrado o processo geral da técnica de NER.

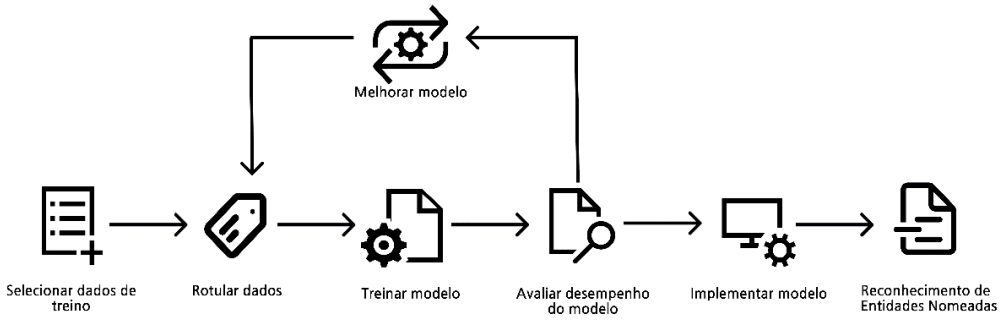


Figura 1- Processo geral da técnica de NER. Adaptado de (Microsoft, 2023).

Estratégia de substituição através de *word embeddings*

Recentemente, Abdalla et al. apresentou uma nova abordagem da técnica de ofuscação para a tarefa da anonimização baseada em *word embeddings* (Abdalla et al., 2020).

Foi proposta a substituição de cada *token* por outro semanticamente próximo presente no espaço de *embeddings*, de forma a conseguir uma representação semelhante das palavras. Desta forma, o objetivo é substituir todo o conteúdo da nota clínica, garantindo a anonimização das informações sensíveis, mas mantendo as informações relevantes e o contexto geral de modo que continue legível (Abdalla et al., 2020).

A Figura 2 seguinte representa alguns exemplos de resultados possíveis utilizando *word embeddings*.

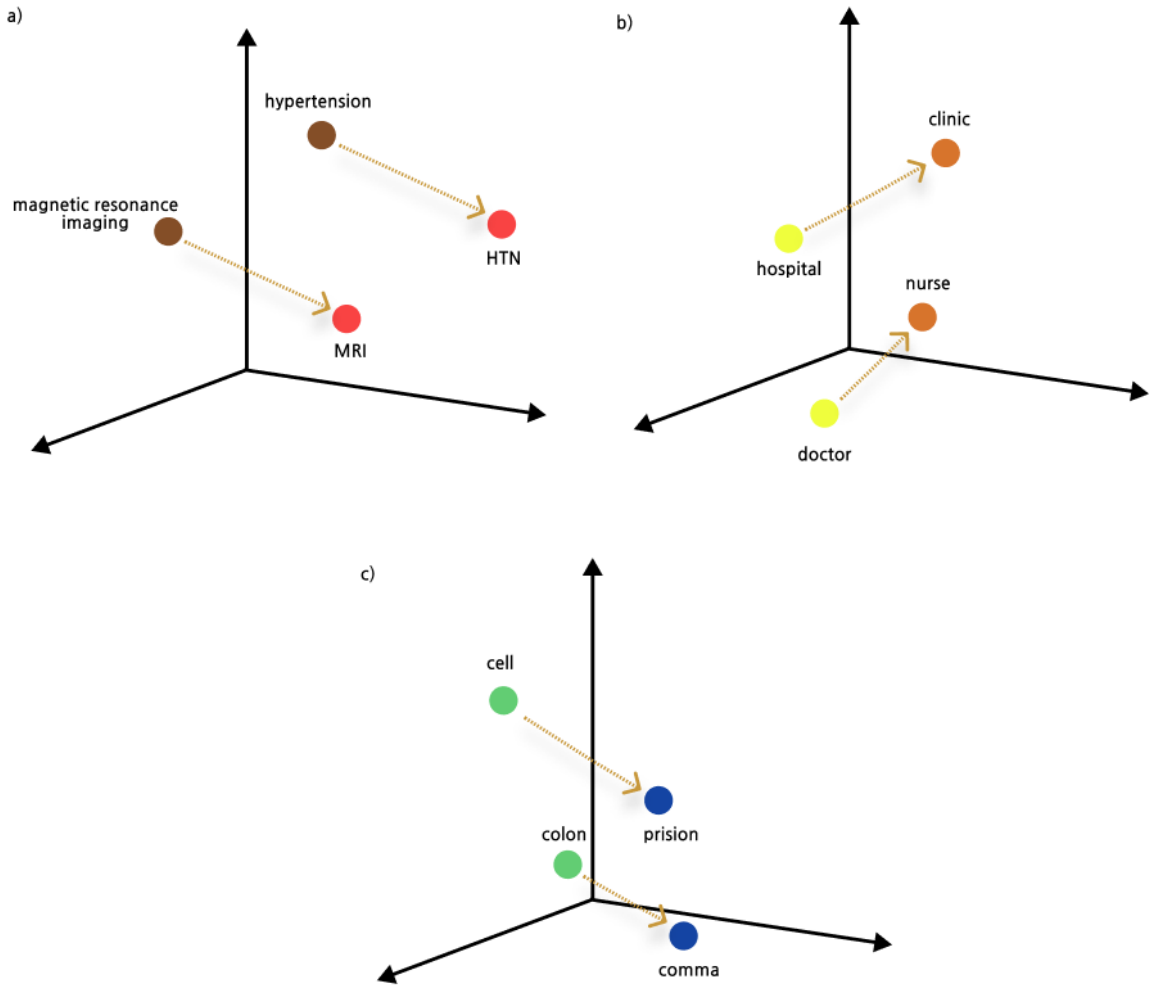


Figura 2- Representações de possíveis exemplos utilizando *word embeddings*. Adaptado de (Anala, 2020).

Na Figura 2 a) encontra-se representado com a cor castanha os *tokens* que poderão surgir na nota clínica original ("*magnetic resonance imaging*" e "*hypertension*") e, utilizando a estratégia de *word embeddings*, surgem os *tokens* representados a vermelho ("MRI" e "HTN") na nota clínica anonimizada que, sendo siglas dos termos originais, poderão fazer parte das palavras mais próximas semanticamente. Na Figura 2 b) o processo é o mesmo. Com a utilização das *word embeddings* é possível corresponder os *tokens* originais ("*hospital*" e "*doctor*") a *tokens* semanticamente próximos ("*clinic*" e "*nurse*", respetivamente) que aparecerão na nota anonimizada. Na Figura 2 c) é possível perceber possíveis resultados das *word embeddings* não-contextuais. Ao não considerar o contexto, dependendo dos dados de treino utilizados para treinar o modelo, os *tokens* "*cell*" e "*colon*" podem vir a ser substituídos por *tokens* do mesmo campo semântico, mas não com contexto clínico, como "*prision*" e "*comma*", respetivamente. Se fosse considerado o contexto poderiam resultar *tokens* como "*cytoplasm*" que corresponderia a "*cell*" e "*colonoscopy*" que corresponderia a "*colon*".

1.3.2. Métricas de Avaliação

Com o surgimento de novas técnicas de anonimização, como os métodos generativos e o uso de *word embeddings* para a anonimização do texto clínico na totalidade, as métricas tradicionais como a *recall*, a precisão e o *f1-score* não têm capacidade para avaliar eficazmente estes métodos, como têm para técnicas baseadas em NER.

Nos métodos baseados em NER um Falso Positivo refere-se a um *token* previsto como sensível que foi substituído ou removido, mas que não fazia parte da lista de entidades sensíveis. Na substituição de todos os *tokens* com *word embeddings* não é verificado este conceito de Falso Positivo, porque todas aos *tokens* (sensíveis e não sensíveis) foram substituídos. Nos métodos generativos nem todos os *tokens* removidos são considerados sensíveis, mas, ainda assim poderão transportar as informações relevantes. Considerá-los como Falsos Positivos levaria a um injusto julgamento de desempenho destes modelos.

É devido a essas limitações que, além de ser considerada a métrica de Perda de Informação Clínica (CIL) (Ribeiro et al., 2023) que permite a avaliação da permanência das informações clínicas relevantes na nota anonimizada, também foi desenvolvida uma nova métrica denominada por *Levenshtein Recall* que irá funcionar como uma *recall* mas adaptada para as técnicas mais complexas que ultimamente têm surgido.

1.4. Apresentação da empresa Fraunhofer Portugal

A Fraunhofer Portugal é uma associação privada sem fins lucrativos fundada pela Fraunhofer-Gesellschaft, a maior organização de investigação aplicada na Europa. Esta desempenha um papel crucial no cenário de inovação e desenvolvimento tecnológico em Portugal.

A Fraunhofer Portugal promove a evolução no conhecimento científico de forma a gerar valor para os seus parceiros. Colabora com universidades e empresas em várias áreas como tecnologias da informação, engenharias de *software*, sistemas de energia e muito mais. Desta forma, a Fraunhofer Portugal tem vindo a demonstrar um impacto significativo no crescimento económico e tecnológico do país, visto que contribui para a criação de soluções inovadoras que, além de beneficiarem a sociedade, impulsionam o setor empresarial.

1.5. Estrutura do trabalho

Este trabalho está dividido em cinco capítulos. Além da Introdução, é também constituído pela Revisão Bibliográfica, pela Metodologia, pelos Resultados e Discussão e, por último, pela Conclusão.

Na Revisão Bibliográfica vão ser abordadas as diferentes estratégias de anonimização que foram sendo utilizadas ao longo do tempo desde as primeiras abordagens como a anonimização manual, a baseada em regras, através de dicionários e métodos estatísticos, até ao momento mais recente em que têm vindo a ser mais explorados os métodos de *Machine Learning*.

Na Metodologia vão ser especificados os recursos utilizados e os métodos aplicados. Para o desenvolvimento deste estudo foi necessário o uso de alguns recursos como a linguagem de programação e suas bibliotecas, datasets disponíveis para utilização e modelos de processamento de linguagem natural.

No capítulo dos Resultados e Discussão serão apresentados e discutidos os resultados obtidos pelos diferentes métodos de anonimização.

E, por fim, no capítulo da Conclusão vão ser expostas as conclusões finais sobre os vários estudos feitos ao longo do trabalho. Vão ser referidos os resultados e reflexões finais de maneira a serem verificados

se os objetivos iniciais foram atingidos e se a questão central do estudo foi adequadamente respondida pelas experiências realizadas.

2. REVISÃO BIBLIOGRÁFICA

Para ser possível determinar uma resposta para a questão de investigação "Serão as técnicas de anonimização baseadas em *word embeddings* alternativas viáveis às técnicas baseadas em Reconhecimento de Entidades Nomeadas? Como poderemos avaliá-las de forma justa e desbloquear o seu potencial?" de forma pertinente e adequada é necessário investigar o estado-da-arte e entender as diferentes estratégias de anonimização que foram sendo utilizadas ao longo do tempo desde as primeiras abordagens como a anonimização manual, a baseada em regras, através de dicionários e métodos estatísticos, até ao momento mais recente em que têm vindo a ser mais explorados os métodos de *Machine Learning*.

Para a revisão bibliográfica deste tema foi necessária pesquisa e recolha de informação científica. Para isso, foram utilizadas bases de dados como a ScienceDirect, a Springer, a IEEE Xplore e a Google Scholar. As palavras-chave para pesquisa foram: "*anonymization*", "*clinical text anonymization*", "*Named Entity Recognition*", "NER", "*word embeddings*", "*Natural Language Processing*" e "NLP".

Do resultado de artigos encontrados foram considerados todos os escritos em inglês, excluindo todos os artigos publicados anteriormente a 1995. Após a aplicação destes critérios de inclusão e exclusão foram selecionados aqueles que demonstraram relevância e pertinência para o presente estudo.

2.1. Primeiras abordagens

Os primeiros trabalhos desenvolvidos para desidentificação de informações sensíveis em notas clínicas consistiam na identificação e substituição de palavras sensíveis através da pesquisa em dicionários, regras manuais, ou algoritmos simples baseados em padrões e métodos estatísticos (Taira et al., 2002). A grande falha destes métodos é a sua baixa capacidade de generalização, uma vez que são muito específicos para cada conjunto de notas partilhadas.

Estas primeiras abordagens conduziram a uma investigação mais profunda nesta área levando à exploração e utilização de técnicas como o NLP para a tarefa de anonimização, mais recentemente.

2.1.1. Manual

A desidentificação pode ser realizada manualmente por anotadores humanos, criando rótulos e classificando as PHI. Esta estratégia primordial apresenta vários obstáculos muito limitativos como: a restrição nos profissionais com permissão de acesso livre às notas clínicas com os dados dos pacientes, o custo do serviço (é uma tarefa que tem tanto de morosa como de dispendiosa) (Dernoncourt et al., 2017) e, tal como apresentado por Neamatullah et al., os resultados de sensibilidade da tarefa variam de profissional para profissional, atingindo valores entre 0.63 e 0.94 (Neamatullah et al., 2008). Além disso, a abordagem manual é impraticável quando o conjunto de notas clínicas é extenso (South et al., 2014).

Apesar de todas as desvantagens que a abordagem manual apresenta, esta pode ser vantajosa quando utilizada em combinação com outras técnicas de anonimização. Normalmente esta abordagem é usada para validar os resultados de um sistema automático (Gupta et al., 2004; Uzuner et al., 2007) ou utilizada em sistemas semi-automáticos em que é realizada uma pré-desidentificação automática em que o sistema permite ao revisor humano modificar, excluir ou adicionar anotações aos resultados obtidos (South et al., 2014; Stenetorp et al., 2012). Esta continua a ser considerada como a alternativa mais confiável por parte das instituições clínicas, visto que ainda se verifica uma grande resistência na partilha de dados devido à dificuldade na tarefa de anonimização.

2.1.2. Baseada em regras

Na abordagem baseada em regras são utilizadas diversas normas para identificar entidades nomeadas num texto. Essas regras, inicialmente, eram definidas por especialistas manualmente. Algumas das regras que são tipicamente usadas nesta abordagem é a anonimização do que vem posterior ao prefixo "Dr." ou após "Mr." ou "Mrs.".

Sweeney et al. criou o Sistema Scrub que usa algoritmos de deteção baseado em regras para identificar PHIs e substituí-las por códigos ou informações generalizadas. Este é um artigo muito influente, publicado em 1996, na área da privacidade e segurança de dados em saúde que alavancou a exploração de técnicas para a tarefa de anonimização (Sweeney, 1996).

Esta é uma técnica relativamente simples e que pode ser útil em alguns contextos. Uma das vantagens consideradas por Meystre et al. é a forma célere e fácil no caso de necessidade de adicionar regras ou expressões regulares para melhorar o desempenho (Meystre et al., 2010).

Ainda assim, é improvável que um sistema seja composto apenas pela técnica baseada em regras. Primeiramente, são necessárias múltiplas regras e dificilmente são suficientes para a eficiência na identificação de todas as entidades nomeadas. Múltiplas regras também poderão ser um problema visto que pode haver conflito de interação entre elas (Taira et al., 1999). Outro problema da abordagem baseada em regras, quando usada sozinha, é a ambiguidade a que esta está sujeita: tem um bom desempenho para PHI inequívocos, mas péssimo desempenho quando se confronta com dados ambíguos (Catelli et al., 2021). Além disso, as regras criadas para um dado conjunto de dados precisam de ser ajustadas para cada novo conjunto de dados, as regras podem não conseguir ser generalizáveis, não são sensíveis a mudanças de idioma, variações nas palavras, erros tipográficos ou algumas abreviações e podem não ter em conta o contexto da palavra (Dernoncourt et al., 2017; Meystre et al., 2010).

2.1.3. Dicionários

Para a anonimização de texto clínico, a abordagem de dicionários é uma técnica utilizada que substitui PHIs de documentos médicos por códigos ou outras informações de forma a não ser possível identificar o paciente em questão.

O uso de dicionários pode ser abordado de duas formas. Pode ser utilizado um dicionário com um conjunto de dados e ir verificando a existência de cada um deles no texto a ser anonimizado ou utilizar um dicionário/glossário médico e aplicar a estratégia de manter apenas os termos clínicos, removendo tudo o resto. Ambas as estratégias apresentam limitações, visto que na primeira abordagem é praticamente impossível um dicionário conter todas as variações de nomes, contactos telefónicos, moradas, etc., e na segunda abordagem também é praticamente impossível que o dicionário/glossário contenha todos os termos clínicos existentes, agravando o facto que poderá tornar-se de difícil legibilidade ao remover todas as outras palavras (Tveit et al., 2004).

Apesar disso, estas são abordagens simples e eficazes quando o conjunto de PHIs que necessitam de anonimização é restrito. Poderão também ser úteis em situações em que seja necessário preservar a estrutura e o formato do texto a anonimizar. Tal como a abordagem baseada em regras, a utilização de dicionários também é considerado um método fácil e rápido aquando da necessidade de melhorar o desempenho, acrescentando termos ao dicionário (Meystre et al., 2010).

No entanto, esta técnica é comumente utilizada em combinação com outras técnicas de anonimização, como o caso da abordagem baseada em regras (Wang et al., 2020). Tal acontece porque a utilização de

dicionários não é totalmente eficaz quando há uma grande quantidade de conjuntos de dados, sendo difícil impedir a re-identificação de pacientes.

2.1.4. Métodos estatísticos

Os métodos estatísticos, tal como o nome indica, envolvem a aplicação de conceitos e ferramentas estatísticas com o objetivo de proteger a privacidade dos dados dos pacientes. Estes ganharam muita atenção devido a algumas lacunas basilares em métodos simbólicos, como a abordagem baseada em regras e a técnica de dicionários (Taira et al., 1999).

Existem diferentes métodos estatísticos como: a supressão de dados que envolve a remoção de informações identificáveis de uma nota clínica; a substituição de dados que substitui as informações sensíveis por valores aleatórios ou por palavras mais generalizadas; e a perturbação dos dados que envolve a alteração de dados originais para impedir que estes identifiquem o paciente ou que possam ser reconstruídos para a sua forma original.

Para se aplicar métodos estatísticos é necessário um processo que inclui várias etapas, como especificado por Taira et al., em 1999. São essas etapas: analisador estrutural, analisador lexical, analisador de dependências, interpretador semântico e o processador de discurso (Taira et al., 1999).

O método *k-anonymity*, apesar de não ser um método estatístico puro, está relacionado com o tratamento de dados estatísticos de forma a garantir a anonimização e privacidade dos dados. Num determinado conjunto de dados, o *k-anonymity* tem o objetivo de pelo menos "k" indivíduos possuam características semelhantes, de forma a tornar difícil a identificação de uma pessoa em específico. Este método foi inicialmente desenvolvido em 1998 por Samarati e Sweeney que, após várias aplicações práticas conseguiram perceber que ao utilizar as técnicas de generalização e supressão neste método a perda de informação não era crítica, apesar das suas limitações iniciais (como dedicar estudo a entender o tamanho mais adequado para o "k") (Samarati & Sweeney, 1998).

Ao longo dos anos têm sido desenvolvidos muitos algoritmos de *k-anonymity* que são utilizados em vários estudos. El Emam et al. desenvolveram um algoritmo de *k-anonymity* denominado por OLA (*Optimal Lattice Anonymization*). Este algoritmo determina qual o melhor nó numa determinada rede de forma a obter uma perda mínima de informação. Através de estratégias de generalização, o OLA implementa uma procura binária iniciando pelo nó ótimo encontrado na rede formada (El Emam et al., 2009). Apesar do *k-anonymity* ser um dos algoritmos mais conhecidos em questões de privacidade, a sua aplicação a dados de texto livre é um desafio, visto que este foi desenvolvido para dados tabulares.

2.2. *Machine Learning*

Sendo uma subárea da inteligência artificial, o campo de *machine learning* é, segundo Tom M. Mitchell: "o estudo de algoritmos que permitem que programas de computador melhorem automaticamente através da experiência" (Mitchell, 1997).

Existem diferentes técnicas de *machine learning* que podem ser utilizadas em diferentes tipos de tarefas. Entre elas salientam-se as técnicas de aprendizagem supervisionada, aprendizagem não supervisionada e *deep learning*.

Na tarefa de anonimização de notas clínicas, as técnicas de *machine learning* são utilizadas para automatizar o processo. Assim, o processo de ocultação das informações sensíveis presentes nas notas clínicas é automático, permitindo aos investigadores que usem esses registos para fins de pesquisa sem comprometer a privacidade dos pacientes.

A maioria dos sistemas de anonimização têm sido baseados na técnica de aprendizagem supervisionada, identificando as palavras que são PHI e classificando-as como tal. Esta técnica de *machine learning*, por ser supervisionada, requer um conjunto de dados de treino em que todas as PHIs são rotuladas manualmente. Alguns exemplos de modelos de *machine learning* tradicional que podem ser usados para a tarefa de NER de entidades protegidas são: CRF, SVM (*Support Vector Machine*), Árvores de Decisão e Entropia Máxima (Meystre et al., 2010; Yang et al., 2019).

As principais vantagens de *machine learning* nesta tarefa centram-se na facilidade de aprendizagem do reconhecimento de padrões de PHI, mesmo sendo complexos, têm uma melhor capacidade de generalização aquando da entrada de novos dados e mantêm a velocidade de processamento ao longo do tempo. No entanto, a necessidade de grandes quantidades de dados para treino pode ser uma desvantagem da técnica, além de que poderá ser difícil descobrir a origem de algum erro de desidentificação, em que acrescentar mais dados para treino poderá ser uma solução como também poderá não influenciar na correção desse erro (Meystre et al., 2010; Yang et al., 2019).

Conditional Random Fields

Os CRF são um tipo de modelo probabilístico, mais concretamente um modelo de aprendizagem supervisionada, utilizado em tarefas de *machine learning*.

Este modelo apresenta uma capacidade de combinação de propriedades como: o treino de modelos discriminativamente para a segmentação e determinação de rótulos, o treino e descodificação eficientes utilizando programação dinâmica e a estimativa de parâmetros para encontrar o ideal (Lafferty et al., 2001).

No caso da tarefa de anonimização de notas clínicas, o modelo é treinado com um conjunto de notas clínicas anonimizadas manualmente para identificar padrões nas mesmas que indicam a presença de informações sensíveis e que precisam de ser anonimizadas. Posteriormente, depois de treinado, este modelo poderá ser usado para anonimizar novas notas clínicas automaticamente.

Ao longo dos anos, diversos trabalhos desenvolvidos por investigadores na área do NLP estabeleceram estes modelos como um ponto de referência importante na tarefa de anonimização automática de texto clínico. Alguns exemplos destes trabalhos são aqueles apresentados nos desafios de desidentificação da i2b2 (*Informatics for Integrating Biology and the Bedside*) de 2006 e 2014 (Stubbs et al., 2015; Uzuner et al., 2007).

Aramaki et al. participou no desafio de 2006 da i2b2 e, com um sistema que utilizou um modelo CRF, foram considerados um dos sistemas com melhor desempenho na tarefa de anonimização. Este sistema conseguiu alcançar resultados de precisão, *recall* e *f1-score* superiores a 94% para a deteção geral de PHI, enquanto para identificação de categorias individuais *f1-score* apresentou valores a partir dos 70% (Aramaki et al., 2006).

Gardner et al. desenvolveram um sistema denominado por HIDE (*Health Information DE-identification*). A extração de PHIs, sendo considerada um problema de NER, foi realizada através de CRFs. O valor relatado para a precisão geral do sistema HIDE foi de 98,2% (Gardner & Xiong, 2008).

No desafio de i2b2 de 2014 os dois melhores sistemas de desidentificação foram desenvolvidos com um modelo de CRFs. O dois sistemas são baseados em CRFs mas aliados a outras técnicas. O melhor sistema do desafio combinou o modelo de CRFs com uma abordagem baseada em regras, expressões regulares e dicionários no pós-processamento. O segundo sistema utilizou também a abordagem baseada em regras para a identificação de PHIs padrão (telefone, FAX e número de registo médico). Outro trabalho desenvolvido com esta técnica também combinou com a abordagem baseada em regras no pós-

processamento e alcançou o segundo melhor desempenho na tarefa de CEGS N-GRID de 2016 (Yang et al., 2019).

Microsoft Presidio

Microsoft Presidio[1] (Mendels & Balter, n.d.) é uma ferramenta que pretende garantir que as informações sensíveis sejam identificadas e, seguidamente, anonimizadas num dado texto, removendo-as ou substituindo-as. É, portanto, dividido em duas partes: o analisador e o anonimizador. O analisador pode utilizar várias técnicas diferentes, como é o caso do NER (utilizando um modelo de NLP – spaCy[2]), REGEX (expressões regulares) ou baseada em regras. O anonimizador anonimiza as entidades sensíveis detetadas pelo analisador. Este último processo, por padrão, substitui no texto anonimizado cada entidade sensível pelo seu tipo de entidade correspondente, por exemplo: "Maria" no texto original é seria substituído por "<NOME>" no texto anonimizado.

Este algoritmo da Microsoft Presidio já demonstrou resultados promissores em alguns trabalhos, como a identificar e anonimizar informações sensíveis em notas clínicas (Ribeiro et al., 2023), ou noutras aplicações como identificar os vários formatos de números de segurança social, por exemplo, em e-mails (Friebely, 2022).

Word embeddings

Word embeddings são representações vetoriais de palavras num espaço latente em que cada dimensão representa um conceito abstrato. A representação vetorial permite que as palavras sejam usadas como entrada em modelos de *machine learning*, como redes neuronais. Esta é uma representação muito utilizada em problemas de NLP, visto que mapeia as palavras num espaço vetorial baseado nas suas relações semânticas e sintáticas (Pandey et al., 2022). Os métodos de geração de *word embeddings* podem ser divididos nos que apontam à geração de *word embeddings* não contextuais, e nos que apontam à geração de *word embeddings* contextuais. A principal diferença entre esses dois tipos de *word embeddings* está na consideração do contexto das palavras. Enquanto as não-contextuais apenas captam um único significado da palavra (por exemplo: "banco", neste tipo de *embeddings*, apenas representaria um assento ou uma entidade bancária, nunca os dois), as *word embeddings* contextuais têm a capacidade de distinguir o significado da palavra consoante o contexto em que ela se encontra no texto.

No caso das *word embeddings* não-contextuais temos os exemplos do Word2Vec e do GloVe que utilizam esta estratégia desenvolvida por Abdalla et al.. Estas técnicas (Word2Vec e Glove) embora não consigam relacionar os diferentes significados que uma palavra pode ter, como mencionado anteriormente, ainda são consideradas técnicas importantes para identificar sinónimos e relações semânticas (Sogancioglu et al., 2021; Zhang et al., 2020).

Word2Vec[3] (Mikolov et al., 2013) é um algoritmo de representação de vetores que aprende as representações vetoriais das palavras através do treino de redes neuronais. Ao receber um corpus de entrada constrói um vocabulário a partir desses dados e, após isso, gera um vetor para cada palavra como resultado de saída.

[1] https://github.com/microsoft/presidio
[2] https://github.com/explosion/spaCy
[3] https://code.google.com/archive/p/word2vec/

GloVe[4] (Pennington et al., 2014) é um algoritmo que também tem como objetivo representar palavras através de vetores. Em contraste com o Word2Vec, o treino deste algoritmo é através de co-ocorrência estatística de palavras num determinado corpus de texto, captando as relações semânticas e sintáticas entre as palavras com base nessa co-ocorrência de palavras no texto.

Por outro lado, as *word embeddings* contextuais, ao contrário das não-contextuais, são representações vetoriais que têm em consideração os diferentes contextos das palavras. Isto acontece porque os vetores são calculados avaliando as palavras vizinhas, conseguindo identificar os diferentes significados. São exemplos de modelos que utilizam *word embeddings* contextuais os modelos de *deep learning* BERT (Devlin et al., 2019), GPT-3 (Radford et al., 2018) e ELMo (Peters et al., 2018). Aprender o contexto em que as palavras se inserem é considerada uma mais-valia quando se pretende analisar vários cenários com a mesma palavra e quando se pretende resolver ambiguidades (Soganciogl u et al., 2021; Zhang et al., 2020).

Recentemente, Abdalla et al. desenvolveram uma abordagem de *word embedding* que consiste em substituir cada *token* na nota clínica por outro *token* aleatório dentro do mesmo campo semântico (N *tokens* vizinhos, excluindo ele próprio) na mesma posição da nota original. Visto que a substituição é realizada a 100%, garante uma confiança de anonimização porque os dados da nota anonimizada não existem no conjunto de dados originais (Abdalla et al., 2020).

Apesar desta estratégia conseguir manter o contexto aproximado do texto a nível lexical, a legibilidade é, ainda assim, um dos grandes problemas. É necessário haver um equilíbrio no grau de ofuscação (N) porque um N muito pequeno pode tornar fácil a reconstrução da nota original e um N demasiado grande faria com que certas alternativas de *tokens* tivessem uma relação semântica muito afastada com o original, podendo perder-se nestes casos uma grande quantidade de informação. No estudo de Abdalla et al. foi verificado um melhor resultado, de acordo com a correlação de Pearson realizada, quando utilizaram um N=5 (Abdalla et al., 2020).

Com a substituição de todas as palavras a *recall* demonstrou-se perfeita, atingindo os 100%, a custo da baixa precisão.

De forma a aumentar a segurança e a dificuldade de reconstrução, os autores sugerem variar o N ao longo da nota clínica. Para aumentar a legibilidade da nota anonimizada sugerem a exploração do uso desta abordagem de *word embeddings* em combinação com outros modelos de NLP para anonimização de texto clínico, como o uso de dicionários (embora alertem para o aumento do risco) (Abdalla et al., 2020).

Deep Learning

O *deep learning* é uma subcategoria de *machine learning*. No entanto, o *deep learning* é normalmente projetado para lidar com tarefas que envolvem grandes quantidades de dados e a partir de dados brutos. Este método é composto por modelos matemáticos denominados por redes neuronais profundas que se inspiram no funcionamento do cérebro humano. Essas redes neuronais apresentam várias camadas, cada uma com as suas características, funções e parâmetros que, ao combinarem conseguem alcançar um nível mais profundo na realização de tarefas de aprendizagem em comparação com as técnicas de *machine learning* tradicionais.

O *deep learning* demonstrou-se com bom desempenho em aplicações como processamento de imagem, reconhecimento de fala e tradução de texto (Wu et al., 2015).

[4] https://github.com/stanfordnlp/GloVe

No âmbito da tarefa de anonimização de texto clínico, os métodos de *deep learning* são utilizados para identificar e remover as informações sensíveis, mantendo as informações não sensíveis e relevantes para análise e investigação.

Existem vários tipos de arquiteturas de *deep learning* que podem ser usadas com o objetivo de anonimizar texto clínico, como: redes neuronais convolucionais (CNNs), redes neuronais recorrentes (RNNs), redes adversariais generativas (GANs), estratégia de memória de longo prazo (LSTM), modelos de aprendizagem por reforços, transformadores ou *autoencoders*.

Nos últimos anos, os modelos de *deep learning* mais utilizados são os transformadores. Estes têm sido aplicados a tarefas de anonimização e NER em contexto clínico, demonstrando bom desempenho (Yang et al., 2019).

- ***Long short-term memory***

LSTM é uma técnica de *deep learning* e um tipo de RNN. De acordo com o estado-da-arte, LSTM é o mais utilizado e explorado tipo de RNN (Lample et al., 2016).

As RNNs foram desenvolvidas para lidar com dados sequenciais, como o texto. A grande vantagem das RNNs é o facto de estas terem uma "memória interna". Têm uma arquitetura de rede que é composta por *loops*, o que permite que beneficiem das informações da entrada anterior para gerar os dados de saída (Wu et al., 2018). Assim sendo, a implementação de LSTM, de forma a conseguir a melhor previsão, mistura as várias informações anteriormente adquiridas (Lample et al., 2016). Esta característica é particularmente útil em tarefas de NLP, como a anonimização de notas clínicas, onde é importante ter em conta o contexto das palavras. Esta técnica tem a capacidade de, além de identificar as informações sensíveis que precisam ser anonimizadas, preservar o contexto e a semântica da nota original.

A LSTM pode ser utilizada sozinha, mas, de forma a melhorar a qualidade da desidentificação e a preservação da privacidade dos pacientes, pode ser combinada com outras técnicas de *deep learning* como as GANs ou as CNNs, assim como com a CRF.

Em termos de desempenho, a LSTM tem vindo a demonstrar um bom desempenho na tarefa de desidentificação. Liu et al. comprovaram com resultados que os sistemas em que utilizaram LSTM alcançaram melhores resultados de *f1-score* do que os sistemas que foram baseados apenas em CRF, neste caso. Tanto no desafio de 2014 (i2b2) como no de 2016 (N-GRID) os resultados de f1-score foram superiores nos métodos que utilizaram LSTM. No método BI-LSTM conseguiram um valor de 94,29% e no BI-LSTM-FEA de 94,51% enquanto o CRF obteve 92,58%, no desafio de 2014. Em 2016, também se obteve valores superiores com uma diferença entre 0,75%-1,61% em comparação com o método baseado em CRF (Liu et al., 2017).

Quando utilizadas as duas técnicas num sistema híbrido LSTM-CRF também conseguem alcançar um bom desempenho, como demonstrado nos trabalhos realizados por Huang et al. e Lample et al. (Huang et al., 2015; Lample et al., 2016).

- **BERT**

BERT é baseado na arquitetura de transformadores, que é um tipo de arquitetura de *deep learning* altamente relevante em aplicações de NLP. É um modelo de linguagem natural pré-treinado num grande corpus de texto que consiste em prever palavras ausentes numa dada frase com base no seu contexto (Devlin et al., 2019).

O BERT foi desenvolvido pela Google em 2018 e é considerado dos modelos de linguagem mais avançados e influentes na comunidade de inteligência artificial, atualmente. Demonstrou ser de grande interesse em tarefas de NLP, como classificação de texto ou NER. É reconhecido pelas suas capacidades de entendimento de contexto e nuances do idioma, o que o torna muito eficaz, incluindo na anonimização de notas clínicas. Como ele pode ser usado para representar palavras de uma sequência de dados de entrada numa representação vetorial rica em contexto, o BERT demonstra ter muita utilidade em tarefas de anonimização (Lee et al., 2020).

Alsentzer et al. demonstrou que ao treinar o BERT com dados num domínio mais específico, como com dados clínicos, estes modelos são capazes de mostrar desempenhos superiores nas tarefas de NLP, como a desidentificação. Neste estudo concluíram que em cinco datasets, dois deles (i2b2 2006 e i2b2 2014) demonstraram um *f1-score* mais elevado (94.8 e 93.0, respetivamente) quando utilizado o modelo BioBERT, enquanto com os outros datasets (MedLI, i2b2 2010 e i2b2 2012) o melhor *f1-score* foi obtido pelo BioBERT afinado clinicamente através de notas clínicas (Alsentzer et al., 2019).

3. METODOLOGIA

Nesta secção vão ser especificados os recursos utilizados e os métodos aplicados. Para o desenvolvimento deste estudo foi necessário o uso de alguns recursos como a linguagem de programação e suas bibliotecas, datasets disponíveis para utilização e modelos de processamento de linguagem natural. Estes recursos foram essenciais para o processo de criação e treino de modelos que foram utilizados para a anonimização de notas clínicas, em dois contextos linguísticos distintos: o inglês e o português.

3.1. Recursos utilizados

Nesta secção são apresentados os recursos utilizados para a realização deste estudo. Primeiramente é apresentada e justificada a linguagem de programação utilizada e, de seguida, são indicados os diferentes datasets que foram usados para teste e treino dos diferentes modelos.

Linguagem de Programação

Neste estudo utilizou-se a linguagem de programação Python.

Python é uma linguagem de programação muito utilizada na área de *Data Science* e *Machine Learning*. Isto deve-se ao facto de apresentar características ideais para a análise de dados e a construção de modelos, como a simplicidade e a legibilidade. Além disso, são várias as bibliotecas disponíveis, e que foram utilizadas no desenvolvimento deste trabalho, como o pandas[5] (para análise e manipulação de dados), NumPy[6] (utilizado para cálculo vetorial e matricial), scikit-learn[7] (comumente usado para a construção de modelos de *machine learning*), NLTK[8] (útil para tarefas de pré-processamento de texto) e Matplotlib[9] (para efeitos de visualização). Estas são bibliotecas importantes para as etapas de pré-processamento, criação e treino de modelos, visualização e aplicação de algoritmos de *Machine Learning*, fundamentais para o desenvolvimento deste estudo.

Datasets

Foram utilizados quatro datasets diferentes (dois em inglês e dois em português) para treino e teste dos modelos a ser avaliados.

Os datasets em inglês utilizados para este estudo são provenientes das seguintes fontes:

- **MIMIC-III** (*Medical Information Mart for Intensive Care*) (Johnson et al., 2016): um banco de dados que alberga notas clínicas e informações dos pacientes internados em unidades de cuidados intensivos, especialmente relatórios de alta dos pacientes. Deste conjunto de dados foram utilizadas 54652 notas clínicas para treinar os modelos de *word embeddings* e do CRF, em inglês.
- **i2b2** (Stubbs et al., 2015; Uzuner et al., 2007): conjuntos de dados fornecidos para as competições realizadas em 2006 e 2014 que contêm registos clínicos não estruturados, como notas e relatórios médicos. Destes dados foram utilizadas 514 notas clínicas do

[5] https://github.com/pandas-dev/pandas
[6] https://github.com/numpy/numpy
[7] https://github.com/scikit-learn/scikit-learn
[8] https://www.nltk.org/
[9] https://matplotlib.org/stable/

dataset de 2014 para teste (quantidade de notas clínicas para teste fornecidas pela i2b2 no desafio de 2014) de todos os modelos a serem avaliados neste estudo.

Os datasets para treino e teste de modelos em português foram disponibilizados por:

- **SemClinBr** (Oliveira et al., 2022): é um corpus que contém 1000 notas clínicas em português do Brasil. Estas notas foram utilizadas para treinar os modelos de *word embeddings* em português.
- **Notas clínicas de um departamento hospitalar de cardiologia:** é um conjunto de 100 notas clínicas, de pacientes submetidos a cirurgia cardiotorácica, fornecidas por um hospital de Portugal previamente anotadas manualmente por investigadores da Fraunhofer Portugal. Para a sua utilização foi necessário recorrer a uma substituição das anotações por informação falsa, utilizando o a biblioteca de Python *Faker*[10], que permite gerar dados falsos incluindo informação pessoal fictícia.

3.2. Utilização de modelos de NLP

Para a anonimização de notas clínicas foram utilizados modelos pré-treinados (CRF e ferramenta Presidio com modelo spaCy) e modelos de *word embeddings* criados de raiz (Word2Vec e GloVe).

O CRF e o Presidio foram aplicados de acordo com uma configuração de NER para substituir as PHI pelo nome de categorias já pré-definidas pelos modelos, como os nomes, números de telemóvel, moradas, localidades, datas ou nomes de hospitais, mantendo a restante estrutura e conteúdo da nota clínica anonimizada igual à original.

Os modelos de *word embeddings* Word2Vec e GloVe foram aplicados de acordo com a estratégia proposta por Abdalla et al. de forma a obter uma nota clínica anonimizada em que todas as palavras (representadas por um vetor cada uma) do texto original são substituídas por outras palavras que apresentam representação vetorial similar à original, verificando-se uma relação entre palavras e semântica o mais próxima possível.

Para testar e avaliar a eficácia da anonimização de notas clínicas em inglês foram utilizados todos os algoritmos descritos acima, enquanto para a anonimização de notas clínicas em português só foram testados os dois modelos de *word embeddings* e a ferramenta do Presidio com o modelo spaCy, visto que o dataset SemClinBR ao não ter anotações de entidades sensíveis, não foi possível treinar NER.

3.2.1. Anonimização de dataset inglês

Para desempenhar a tarefa de anonimização de notas clínicas em inglês foram considerados os quatro modelos já referidos anteriormente.

Modelos de NER

O CRF e o Presidio + spaCy são modelos do estado-da-arte que têm vindo a demonstrar bons resultados nesta tarefa específica de NLP para NER. O modelo de CRF ganhou destaque nos desafios da i2b2

[10] https://fakerjs.dev/

tornando-se um modelo de referência em tarefas de anonimização e o Presidio tem-se demonstrado interessante para estudar e aprofundar as suas capacidades.

Primeiramente, utilizou-se um modelo pré-treinado do CRF com o conjunto de dados do MIMIC-III.

Outro método que foi testado foi a ferramenta Presidio com o modelo spaCy, com vocabulário em inglês. O spaCy é uma biblioteca de Python que fornece modelos de linguagem pré-treinados no conjunto de dados OntoNotes 5 (Weischedel et al., 2013).

Modelos de *word embeddings*

O Word2Vec e o GloVe são técnicas de estado-da-arte que, apesar das diferenças no processo da escolha das palavras, têm ambas a capacidade de identificar sinónimos e palavras vizinhas, tendo o objetivo de se manter a proximidade semântica entre a palavra original e a palavra substituta (anonimizada), para que haja a menor perda de informação possível. Estes modelos foram criados de raiz e, como tal, foi necessário um pré-processamento que antecedeu ao treino destes modelos.

O pré-processamento é uma etapa crucial para tarefas de NLP. Neste contexto, o pré-processamento do conjunto de dados fornecido é essencial devido à liberdade de escrita destas notas. As notas clínicas são registos médicos não estruturados que podem seguir normas diferentes de escrita de relatórios entre as diferentes instituições e, além disso, cada clínico tem a sua própria forma de escrita, gerando conteúdos altamente variáveis.

Para pré-processamento do dataset foi necessário recorrer a:

- **Tokenização**: um processador que divide o texto de entrada em *tokens* e frases, para que a sua anotação seja mais simples e possa ocorrer a um nível mais simplificado.
- **Lematização**: reduz as palavras à sua forma raiz, o que ajuda a tratar as diferentes formas da mesma palavra como equivalentes.
- **Conversão para letras minúsculas**: converte todas as letras para minúsculas, evitando a diferenciação entre palavras iguais que apenas diferem em letras maiúsculas e minúsculas.

Para estas etapas de pré-processamento foi utilizada uma biblioteca de NLP para Python, com aplicação em muitos idiomas, denominada por Stanza (Qi et al., 2020).

Após o pré-processamento de todas as notas clínicas, foram iniciados os treinos de ambos os modelos com essas mesmas 54652 notas clínicas de treino pré-processadas, originalmente fornecidas pelo dataset MIMIC-III.

Para iniciar a tarefa da anonimização dos dados de teste foi considerado um grau de ofuscação (N) de 5. A opção de utilizar este valor foi devido ao estudo de Abdalla et al. que apresentou resultados da correlação de Pearson para três graus de ofuscação (N=3, 5 e 7). O que demonstrou melhor desempenho em termos de manter a legibilidade, mas, ao mesmo tempo, não ser de fácil reconstrução da nota original foi com a utilização do N=5 (Abdalla et al., 2020). Assim considerou-se os cinco *tokens* vizinhos mais próximos de cada *token* da nota original.

Os quatro algoritmos foram testados com as 514 notas clínicas disponibilizadas para teste do dataset da i2b2 de 2014.

3.2.2. Anonimização de dataset português

Para desempenhar a tarefa de anonimização de notas clínicas em português foram considerados três dos modelos já referidos anteriormente (Presidio, Word2Vec e GloVe).

Modelo NER

Como já existem modelos pré-treinados do spaCy para português, decidiu-se estudar a eficácia da ferramenta Presidio para outro idioma (de baixo recurso) – o português.

Modelos de *word embeddings*

Para aplicar o Word2Vec e o GloVe a um conjunto de notas em português, foi necessário criar estes modelos de raiz e treiná-los com dados em português, após um pré-processamento desses dados.

Tal como para o dataset em inglês, recorreu-se à biblioteca Stanza para o pré-processamento dos dados, mas, desta vez, definida para o idioma português, para se proceder à tokenização, à lematização e à conversão de todas as palavras para letras minúsculas.

O dataset utilizado para treino das *word embeddings* em ambos os casos, e que sofreu o pré-processamento acima descrito, foi o SemClinBr. Este é um corpus consideravelmente inferior em termos de quantidade de notas clínicas, comparando com o dataset inglês. Adicionalmente, é importante salientar que o idioma do SemClinBR é português do Brasil, enquanto o conjunto de dados de teste provém do português de Portugal, cedido por um estabelecimento hospitalar de Portugal.

Para este conjunto de dados de teste foi necessário realizar a anotação manual das PHI e, subsequentemente, recorreu-se à ferramenta denominada "*Faker*" para garantir que as notas clínicas de teste não contivessem qualquer informação sensível autêntica.

Também para a anonimização em português foi utilizado um valor de N=5, considerando os cinco *tokens* vizinhos mais próximos semanticamente de cada *token* da nota original.

3.3. Métricas

Para técnicas baseadas em NER, a *recall*, precisão e *f1-score* são métricas comumente usadas para avaliação da anonimização. Como cada *token* está associado a uma dada categoria previamente anotada (como "NOME" ou "LOCALIDADE", por exemplo) e a posição do *token* da nota original é igual à posição do *token* anonimizado, então é simples uma comparação entre o *token* original e o *token* resultante na nota anonimizada para concluir a sua eficiência, uma vez que toda a restante nota permanece igual.

Apesar destas métricas serem eficazes em técnicas NER, quando utilizados outros tipos de métodos de anonimização, como *word embeddings* ou métodos baseados em modelos generativos, é necessário ultrapassar alguns desafios com recurso a outras abordagens de avaliação. Visto que, no caso das *word embeddings*, todas as palavras do texto são substituídas, a aplicabilidade das métricas tradicionais neste contexto é questionável, como irá ser explicado de seguida. Assim como através dos modelos generativos, como os LLMs, deixa de ser possível garantir que a localização das informações sensíveis permaneçam inalteradas para a versão anonimizada da nota, visto que estes são imprevisíveis. Estes modelos poderão alterar completamente o todo o texto através de sumarização ou reorganização de ideias.

Desta forma, é proposta uma nova métrica para complementar a avaliação da anonimização: a *Levenshtein Recall* (LR).

Esta métrica é baseada na *Levenshtein Distance* (LD). A LD é utilizada para medir quantitativamente a diferença entre duas *strings*. Através desse processo obtém-se o número mínimo de operações (inserções, exclusões ou substituições de caracteres) necessárias para transformar uma *string* noutra. Quanto maior a LD entre duas *strings*, mais diferentes elas são entre si (Haldar & Mukhopadhyay, 2011).

Além destas, foi também considerada a métrica de Perda de Informação Clínica (CIL), uma métrica específica para contexto clínico e que permite também avaliar a qualidade e utilidade da nota anonimizada (Ribeiro et al., 2023).

3.3.1. *Recall*

A *Recall*, no contexto de avaliação da anonimização de notas clínicas, é a métrica responsável por identificar se as informações sensíveis no texto original foram corretamente anonimizadas no texto anonimizado.

Este é um cálculo tradicional de grande importância para se analisar o seu resultado, visto que ajuda a perceber a qualidade da anonimização e o nível de proteção de dados dos pacientes, auxiliando na escolha do melhor método de anonimização, nesta perspetiva.

A equação que descreve o cálculo para a *Recall* é a representada na Equação 1.

$$Recall = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos}$$ Equação 1

Tradicionalmente, no contexto de técnicas baseadas em NER (como o caso do CRF e do spaCy pelo Presidio), a *recall* é calculada com base nas previsões dos modelos na tarefa de NER. Quando estes preveem um *token* como entidade sensível e confere ser sensível, então considera-se um Verdadeiro Positivo. Se preveem um *token* como entidade não sensível e for considerada sensível, então considera-se um Falso Negativo. Isto é possível no caso do NER porque é possível corresponder o *token* original ao *token* previsto.

Nas técnicas emergentes (como as baseadas em *word embeddings* ou LLMs) não é possível realizar essa correspondência porque são introduzidas novas palavras, logo não é garantido que não correspondam a informação sensível. Por esse motivo, a *recall* tradicional não é considerada eficaz para estas técnicas.

O *pattern-matching* é um método que poderá ser utilizado para avaliar estas técnicas emergentes, mas não é considerado ideal, apresentando várias desvantagens para este contexto da anonimização.

Considera-se *pattern-matching* quando se associa dois *tokens* ou expressões iguais. Ou seja, quando uma expressão que está no texto original continua a aparecer no texto anonimizado é considerada um Falso Negativo porque, mesmo depois do texto ser anonimizado, a expressão continua presente. Os Verdadeiros Positivos são considerados todas as expressões que estão na lista de anotações, mas não no texto anonimizado. Logo, foi corretamente anonimizado. A desvantagem é que a expressão tem de ser coincidente a 100%, senão já não é considerada e, consequentemente, é considerada anonimizada. Portanto, a aplicação do *pattern-matching* nos métodos mais inovadores é questionável porque a expressão poderá continuar no texto anonimizado, mas sob uma forma mais encurtada. Por exemplo: se na nota clínica original estiver presente o nome "Ana P. Silva" e na nota clínica anonimizada estiver "Ana Silva", a *Recall* tradicional, pelo *pattern-matching*, vai considerar a informação sensível como anonimizada, ainda

que o nome continue presente na nota anonimizada, mas sob uma forma encurtada, que pode acontecer quando a tarefa é realizada por métodos generativos.

Para contornar estes problemas, é proposta uma nova métrica que não depende das posições dos *tokens* nas notas nem da correspondência a 100% de cada *token* da nota original com os da nota anonimizada, apenas verificam o *index* de similaridade entre *tokens* em toda a extensão da nota, através do conceito de LD.

3.3.2. *Levenshtein Ratio*

A *Levenshtein Ratio* (LRa) é uma medida de similaridade baseada na LD. Esta métrica fornece um valor entre 0 e 1, sendo que o 0 significa que as duas *strings* são completamente diferentes e o 1 que são exatamente iguais.

A Equação 2 expressa o cálculo para a LRa, onde **LD (a, b)** é a LD entre duas *strings* **a** e **b**, e **A** e **B** são os respetivos comprimentos dessas *strings*.

$$LRa(a,b) = 1 - \frac{LD\ (a,b)}{max\ (A,B)} \qquad \text{Equação 2}$$

3.3.3. *Levenshtein Recall*

Esta métrica destina-se a verificar a eficácia da anonimização quando não há informações sobre a categoria de todos os *tokens* (como no caso das *word embeddings* ou dos modelos generativos), sejam eles informação sensível ou não. Além disso, o objetivo desta métrica é também contornar o problema encontrado do *pattern-matching* para este tipo de anonimizações.

Para conseguir ultrapassar este desafio é proposto uma estratégia através de uma *sliding window*. Isto é:

- Considera-se uma lista de comprimento **l** que é composta pelas entidades sensíveis, **se**, de uma nota original, **ON** , de comprimento **L**.
- Calcula-se o comprimento, **e**, de uma determinada entidade sensível dessa lista, $\mathbf{se_i}$.
- Utiliza-se esse comprimento **e** para a *sliding window* que vai percorrer toda a nota anonimizada, **AN**, com o passo de um caractere de cada vez.
- Calcula-se o **LRa** entre cada *sliding window* e $\mathbf{se_i}$.

O *Levenshtein Similarity Index* (LSI) consiste no valor da máxima semelhança encontrada entre $\mathbf{se_i}$ da **ON** e o conteúdo da **AN**, e é dado pela seguinte Equação 3:

$$LSI = max_{j=1}^{L-e}\ LRa(se_i, w_j) \qquad \text{Equação 3}$$

Este processo é realizado para cada uma das entidades sensíveis contidas nessa lista, resultando um LSI associado a cada uma dessas entidades, formando uma lista, **S**, dos LSIs medidos.

Com isto, se um *token* da lista de entidades sensíveis da nota original estiver presente na nota anonimizada, então o índice de similaridade será máximo e compreende-se que é a palavra exata.

A métrica LR é proposta com base na LRa e no conceito LSI.

Para calcular a LR, cada um valores de LSI da lista **S** são comparados a um *threshold*, $\mathbf{th_s}$, de similaridade (valor ajustável com as necessidades).

Para compreender se a entidade sensível foi corretamente anonimizada:

- Se LSI > $\mathbf{th_s}$ - entidade considerada não anonimizada.
- Se LSI < $\mathbf{th_s}$ - entidade considerada anonimizada.

O valor final da métrica de LR é dado segundo o cálculo tradicional de *Recall*, dividindo o número de entidades sensíveis anonimizadas pelo número de entidades sensíveis total, como demonstrado na Equação 4:

$$LR = \frac{\sum_{i=1}^{l}(S_i < th_s)}{l} \times 100 \qquad \text{Equação 4}$$

3.3.4. Perda de Informação Clínica

A métrica de CIL (Ribeiro et al., 2023), específica para contextos clínicos, permite avaliar a perda de informação clínica na nota anonimizada em que é aproveitado um modelo pré-treinado do BioBERT (Lee et al., 2020) treinado numa tarefa hierárquica de classificação de códigos ICD-10 (Classificação Internacional de Doenças, 10ª revisão). Estes códigos são utilizados para classificar e codificar doenças, condições de saúde, lesões e causas de morte, fornecendo uma estrutura-padrão de forma a categorizar e registar informações relacionadas com a saúde. O objetivo é identificar as N categorias de códigos de ICD-10 (de entre os 157 possíveis) mais frequentes tanto da nota clínica original como da nota clínica anonimizada.

A perda de informação, tal como representada na Equação 5, é estimada ao dividir o número de categorias presentes simultaneamente em ambas as listas de predição pelo número N de categorias principais consideradas. A lista dos N principais códigos previstos na nota clínica original é denominada na equação por $\boldsymbol{y_{orig}}$ e a lista dos N principais códigos previstos na nota clínica anonimizada é denominada por $\boldsymbol{y_{anon}}$.

$$CIL = \left(\frac{1 - \sum_{i=0}^{n}(y_{anon} \in \ y_{orig}\ )}{N}\right) \times 100 \qquad \text{Equação 5}$$

4. RESULTADOS E DISCUSSÃO

Depois de aplicadas as metodologias de pré-processamento do texto e implementados os diferentes modelos de anonimização, os conjuntos de dados foram analisados e avaliados através das métricas propostas. Neste capítulo serão apresentados e discutidos os resultados obtidos pelos diferentes métodos de anonimização.

4.1. Apresentação de resultados

Para obter os resultados e conseguir analisar e avaliar os métodos de anonimização utilizados foram aplicadas as três métricas propostas e descritas anteriormente.

De referir que o *threshold* (utilizado pela métrica LR) inicialmente foi definido como 0.7. Ou seja, as entidades sensíveis foram consideradas anonimizadas se o LSI de cada entidade sensível for menor do que o *threshold* definido (neste caso 0.7). Isto quer dizer que o *token* na nota anonimizada é suficientemente diferente do *token* da nota original para ser considerado anonimizado. Considerando o exemplo dado anteriormente: "Ana P. Silva" e "Ana Silva" resultam num rácio de 0.86 (> 0.7), pelo que, desta forma, esta entidade sensível não é considerada anonimizada. Mas "*asthma*" e "*bronchitis*" já apresentam um rácio de 0.13 (< 0.7), pelo que é considerada anonimizada, ainda que se encontrem relacionadas com um campo semântico próximo.

O *threshold* foi submetido a algumas experiências iniciais para perceber qual o valor que apresentava um melhor balanço entre garantir a anonimização, mas não considerar palavras diferentes como sendo a mesma. Assim, 0.7 foi o valor escolhido (de acordo com os resultados obtidos na secção 4.1.4).

4.1.1. Modelos treinados com dataset inglês

Numa primeira fase, os diferentes métodos discutidos no capítulo da Metodologia foram testados contra o dataset inglês correspondente ao conjunto de teste do desafio de desidentificação da i2b2 de 2014. Estas técnicas foram comparadas em termos da *recall*, calculada através de um método de *pattern-matching*, da *recall* de Levenshtein, e ainda da CIL. Através destas métricas é possível avaliar não só a eficácia do processo de anonimização, mas também o efeito do mesmo na retenção ou perda de informação relevante.

É de salientar que, ao contrário das métricas da *recall*, onde altos valores estão associados a uma maior performance, a CIL é interpretada de forma inversa, sendo que valores elevados desta medida denunciam uma indesejável alta taxa de perda de informação.

Através da Figura 3 percebe-se que na *Recall* os métodos de *word embeddings* têm vantagem acentuada em relação aos métodos que utilizam NER. Isto deve-se ao facto de todos os *tokens* da nota original serem substituídos outros semanticamente similares e, portanto, a *Recall* apresenta-se muito próxima de 1 (completamente anonimizada). Pela mesma razão, a métrica que mede a perda de informação clínica na nota anonimizada apresenta valores bastante mais elevados nos métodos de *word embeddings* porque, enquanto o Presidio e o CRF apenas substituem ou removem a informação sensível, o Word2Vec e o GloVe alteram todas as palavras do texto por outras semanticamente semelhantes, estando esta tarefa também dependente do dataset utilizado para treino dos modelos.

Analisando os valores de *recall* por *pattern-matching* e a LR, verifica-se não ter havido variações significativas entre as duas medidas. Ainda assim, é possível verificar uma ligeira descida na LR para

certos métodos, o que pode indicar a permanência de informação sensível que, apesar de não se apresentar exatamente na sua forma original, poderá colocar em risco a privacidade dos envolvidos

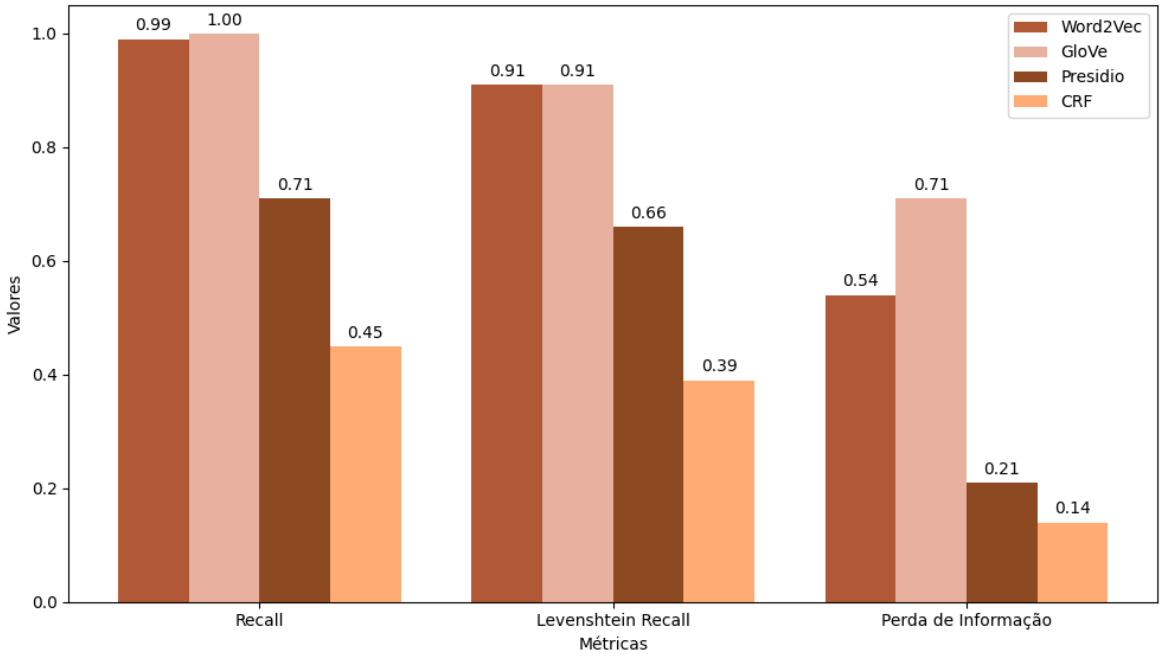


Figura 3- Desempenho de Cada Método na Tarefa de Anonimização - EN

De uma forma global, mantendo um equilíbrio entre uma anonimização eficaz das entidades sensíveis e uma perda de informação o mais reduzida possível, e de acordo com a Tabela 1 é possível perceber que o Word2Vec tem melhor desempenho geral, alcançando 0.99 de *Recall* e 0.54 de CIL, apenas perdendo na perda de informação para os métodos por NER. Assim, se for priorizada a menor perda de informação então o Presidio e o CRF levam vantagem, sendo que o Presidio apresenta um melhor equilíbrio entre anonimização e perda de informação, com valores de 0.71 e 0.21, respetivamente.

Tabela 1 – Desempenho Geral dos Métodos de Anonimização - EN

	Recall	*Levenshtein Recall*	Perda de Informação
Word2Vec	0.99	0.91	0.54
GloVe	1	0.91	0.71
Presidio	0.71	0.66	0.21
CRF	0.45	0.39	0.14

4.1.2. Modelos treinados com dataset português

Em relação aos modelos treinados e testados com datasets em português, os resultados obtidos são os que podemos observar na Figura 4 e na Tabela 2 seguintes.

As tendências observadas foram semelhantes aos obtidos com os modelos treinados e testados com os datasets em inglês.

A *Recall* tem melhor desempenho nos métodos de *word embeddings,* mas mais uma vez se verifica que estes métodos têm também maior perda de informação na nota anonimizada. Ainda que a diferença seja pequena, o GloVe apresenta um valor mais baixo de perda de informação em comparação ao Word2Vec, sendo que a diferença de ambos para a menor perda de informação utilizando o Presidio é muito pronunciada, como também já se tinha verificado nos resultados obtidos no dataset em inglês.

Na LR consegue perceber-se uma queda ténue dos valores dos dois primeiros métodos, sendo maior no Word2Vec.

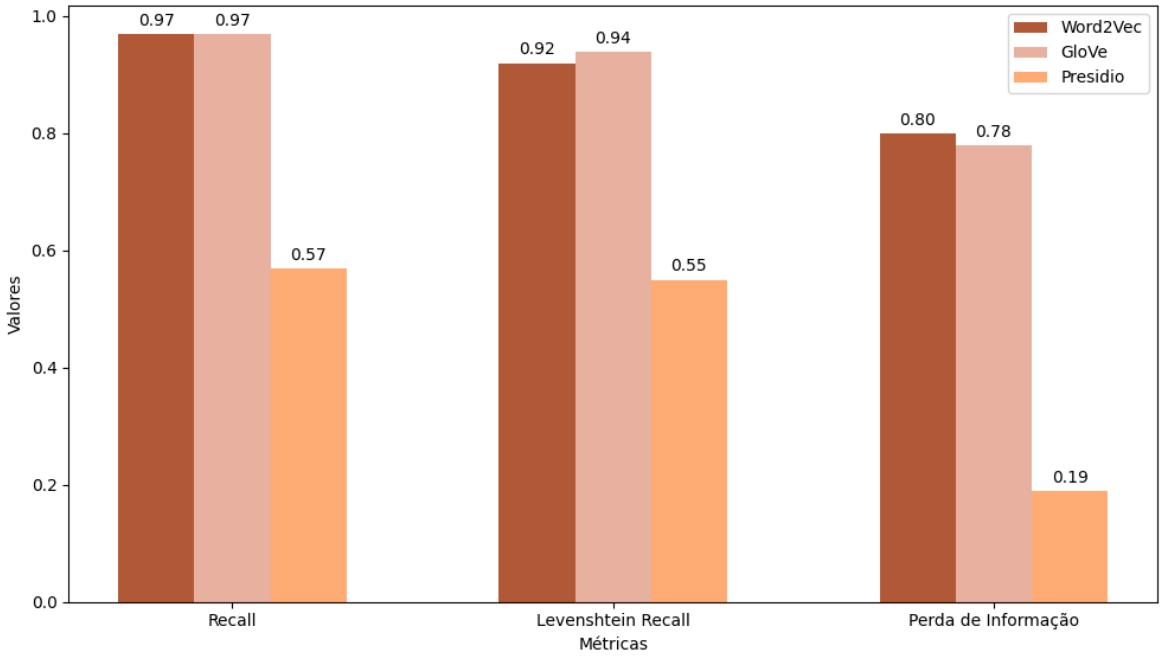


Figura 4- Desempenho de Cada Método na Tarefa de Anonimização - PT

Nos modelos treinados com dados em português a distinção de melhor desempenho geral é do GloVe, ainda que o Word2Vec se apresente muito próximo dos valores deste, como observado na Tabela 2.

É de destacar que a perda de informação continua a ser um critério diferenciador dos modelos de NER, conseguindo atingir valores muito mais baixos (0.19) em comparação aos métodos por *word embeddings* (0.8 e 0.78 do Word2Vec e do GloVe, respetivamente).

Tabela 2 – Desempenho Geral dos Métodos de Anonimização - PT

	Recall	Levenshtein Recall	Perda de Informação
Word2Vec	0.97	0.92	0.8
GloVe	0.97	0.94	0.78
Presidio	0.57	0.55	0.19

4.1.3. Comparações entre modelos treinados com dataset inglês e dataset português

Para analisar e comparar os resultados obtidos para ambos os idiomas testados foram comparados todos os resultados obtidos para os métodos testados nos dois contextos linguísticos.

Quanto à *Recall* conseguimos verificar pela Figura 5 que esta é mais elevada para os modelos em inglês e a maior diferença entre os dois idiomas é na anonimização através do Presidio.

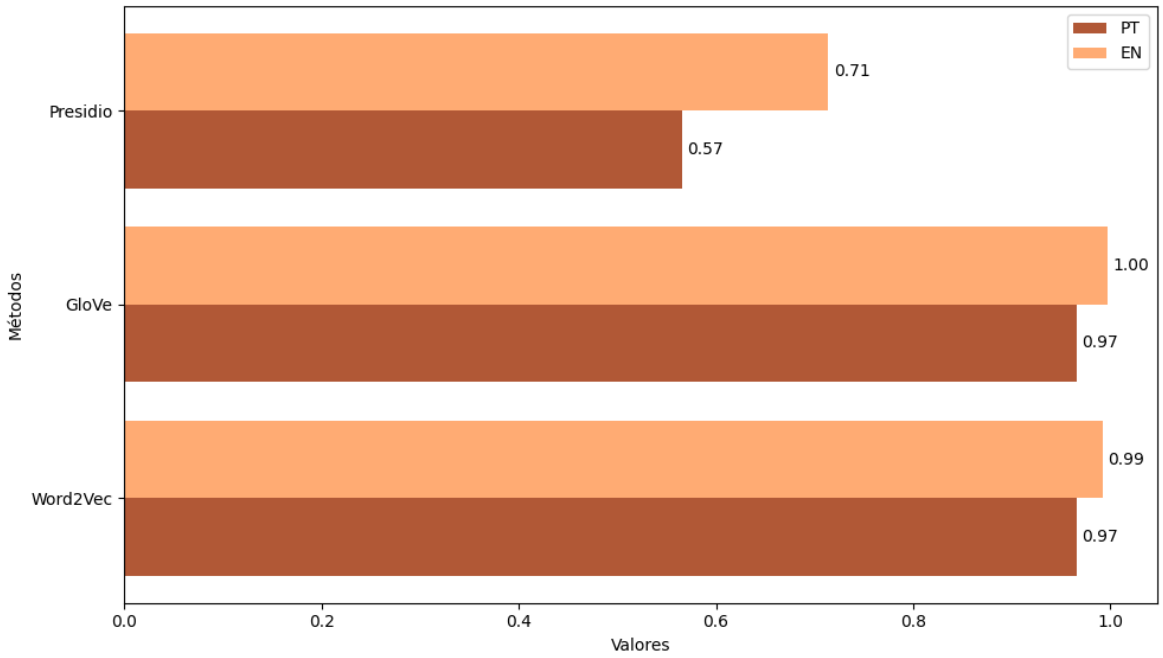


Figura 5- *Recall* Alcançada em Diferentes Contextos Linguísticos

Através da Figura 6 apresentada de seguida, ao comparar a LR entre ambos os idiomas nos métodos Word2Vec e GloVe percebe-se que a diferença não é significativa, ainda que se note um melhor desempenho pelos modelos com dados portugueses. Os valores da LR do Presidio são iguais à *Recall* tradicional, já analisados anteriormente.

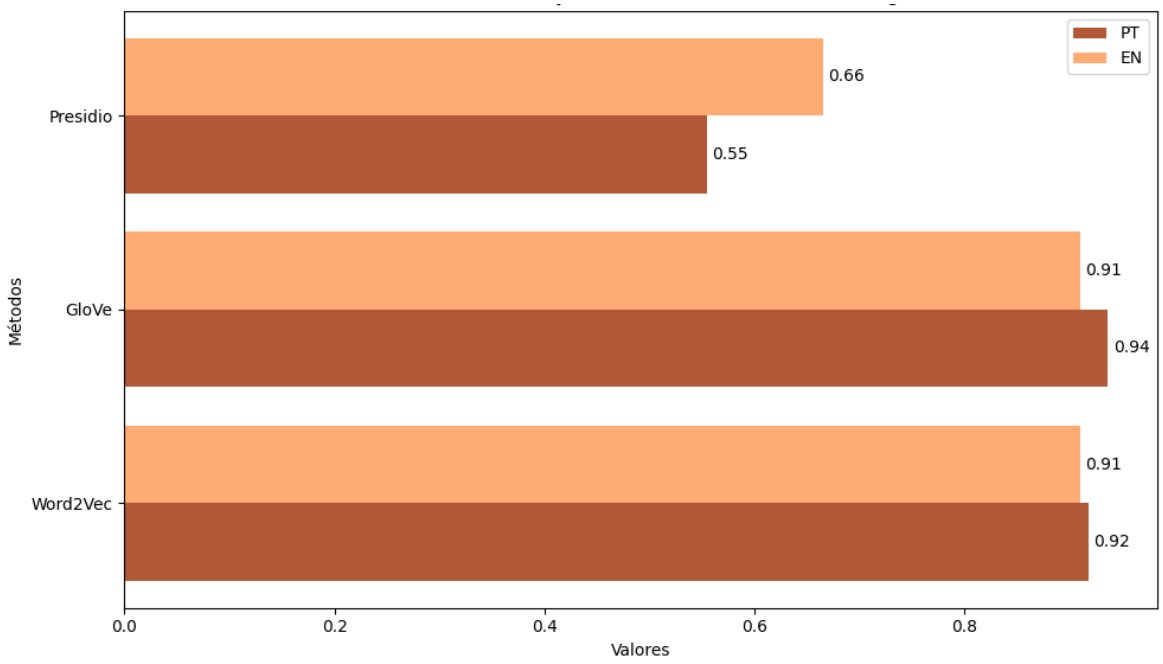


Figura 6- *Levenshtein Recall* Alcançada em Diferentes Contextos Linguísticos

No que diz respeito à perda de informação (Figura 7), nos métodos de *word embeddings*, denota-se uma clara distinção entre os modelos treinados em inglês e os modelos treinados em português, sendo que há maior perda de informação no idioma português. O contrário acontece com o Presidio, em que a perda de informação é maior com os dados em inglês, ainda que a diferença não seja grande.

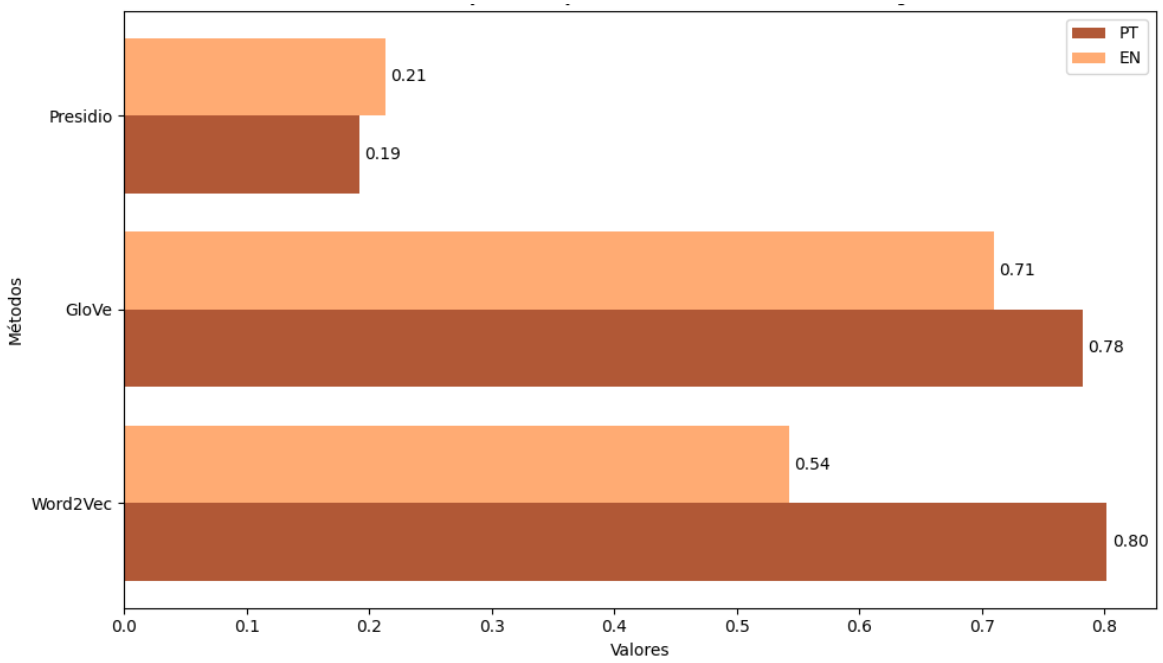


Figura 7- Perda de Informação Alcançada em Diferentes Contextos Linguísticos

4.1.4. Alteração de parâmetros

Para perceber a influência da escolha do valor de *threshold* nos valores obtidos pela *Levenshtein Recall* foram realizadas experiências com diferentes valores de *threshold*: 0.6, 0.7, 0.8 e 0.9, para ambos os idiomas.

Como esperado, e como demonstrado pelas Figura 8 e Figura 9 (para o idioma inglês e para o idioma português, respetivamente), quanto maior for o valor de *threshold*, ou seja, quanto maior o grau de semelhança entre o(s) *token(s)* da nota clínica original e da nota clínica anonimizada, maior é o resultado obtido pela métrica *Levenshtein recall*, porque considera menos variabilidade para um dado *token*.

Como exemplo, o método Word2Vec para o idioma inglês varia o seu resultado da métrica *Levenshtein Recall* de 0.74 (para um *threshold* de 0.6) para 0.99 (para um *threshold* de 0.9). O método Presidio demonstra uma subida mais acentuada nos valores de *Levenshtein Recall* com os dados em inglês do que em português, que apresenta uma subido dos valores mais ligeira.

Observando os resultados de ambos os gráficos podemos afirmar que, de uma forma geral, a maior subida de valores do *Levensthein Recall* é do *threshold* de 0.6 para o *threshold* de 0.7, sendo que os valores seguintes aumentam, mas mais subtilmente. Esta tendência verificou-se em todos os métodos testados, o que parece sugerir que, existem entidades que, não sendo exatamente iguais às entidades sensíveis originais, contém parte do seu conteúdo, e poderão por tanto constituir ameaças à privacidade.

O verdadeiro potencial da métrica LR só poderá ser alcançado determinando o *threshold* ideal que, por um lado, é sensível à existência de variações ligeiras de entidades sensíveis que possam colocar em risco a privacidade, e, por outro lado, não condena a existência de palavras idênticas, mas sem qualquer tipo de informação sensível.

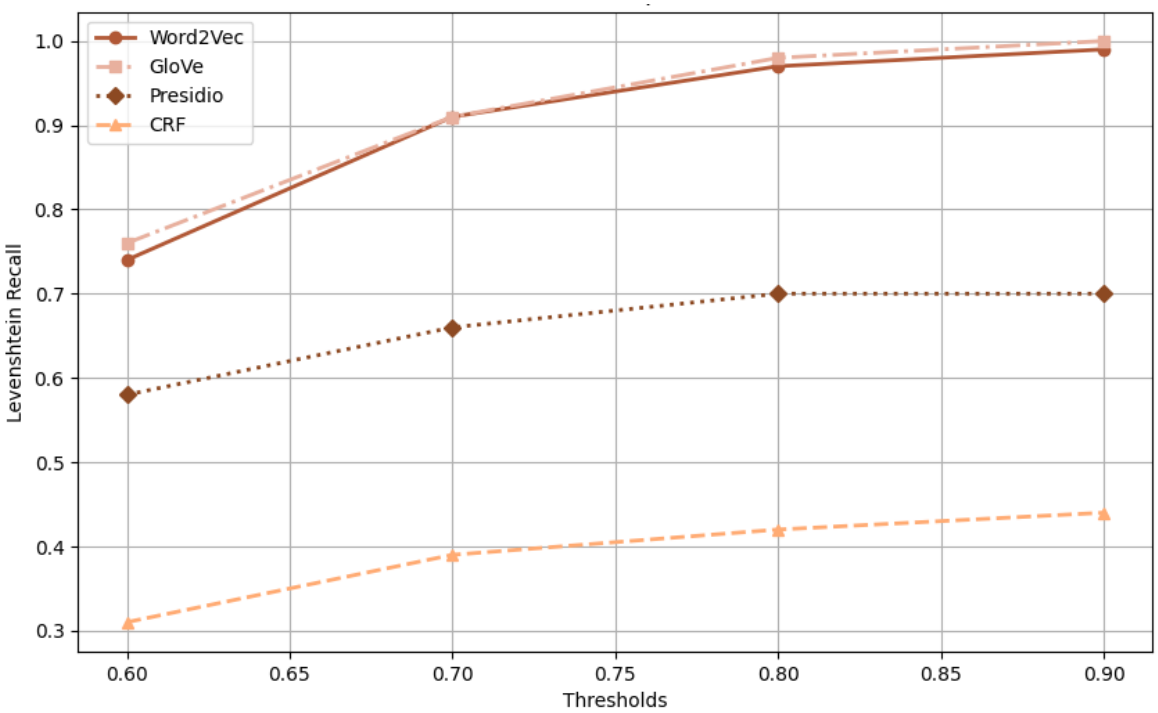


Figura 8- *Levenshtein Recall* vs. *Thresholds* para Diferentes Métodos - EN

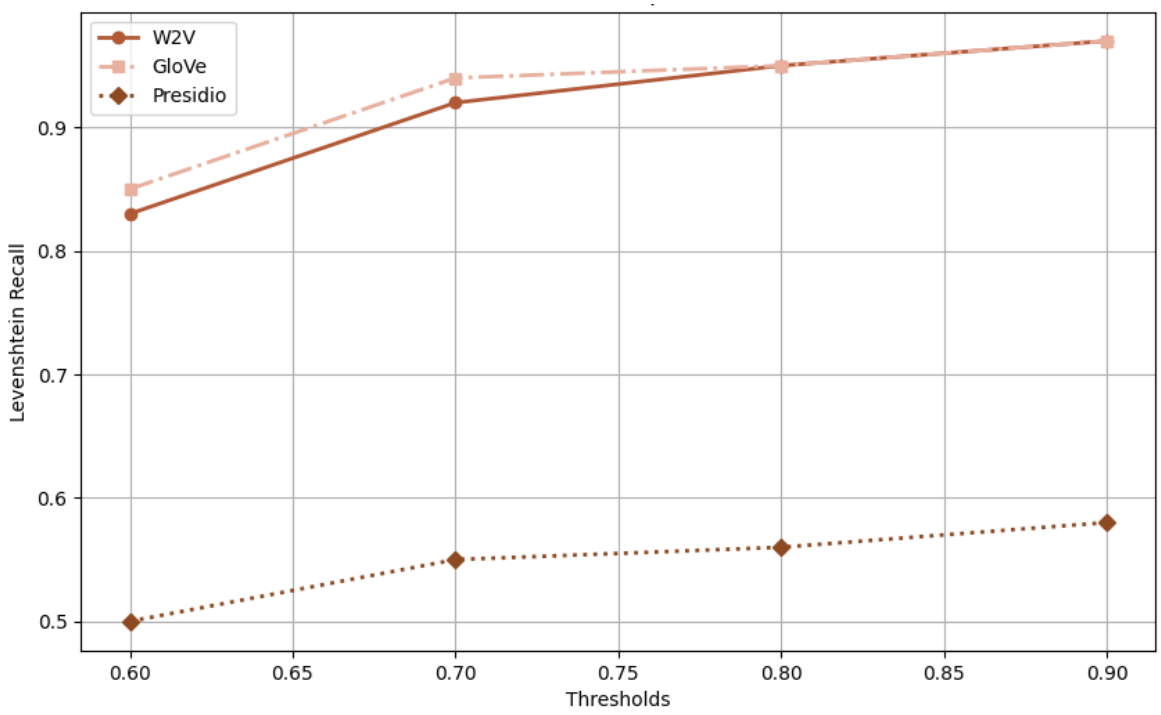


Figura 9- *Levenshtein Recall* vs. *Thresholds* para Diferentes Métodos - PT

4.2. Discussão de resultados

Para a discussão dos resultados foram analisados os gráficos e tabelas, de forma relacional, para que sejam claras as diferenças entre estratégias, entre métodos e entre os diferentes contextos linguísticos.

No caso da métrica que foi desenvolvida, é importante referir a sua importância e contribuição na avaliação de métodos mais inovadores (como é o caso dos *word embeddings* e dos métodos generativos). A LR consegue colmatar as falhas da *Recall* tradicional, conseguindo identificar tentativas de anonimização incompletas e detetar se a entidade sensível estiver presente no texto anonimizado, mas noutra posição diferente do texto original, para além de que o seu cálculo não é baseado em pares *token*-previsão, produzidos apenas por soluções baseadas em NER.

Os valores obtidos na *Recall* e na LR demonstram que os métodos de *word embeddings* (Word2Vec e GloVe) são os que apresentam melhor desempenho em termos de remoção e/ou substituição de informação sensível para a nota anonimizada, podendo atingir os 100% de *recall*. Estes resultados eram os esperados, tal como descrito no artigo de Abdalla et al., uma vez que este tipo de método vai substituir todo o conteúdo da nota original, garantindo a anonimização de todas as entidades sensíveis (Abdalla et al., 2020). No entanto, como podemos verificar nos gráficos anteriores, a *recall*, apesar de estar muito próxima de 100%, apenas tal se verifica num caso. Isto acontece porque, por exemplo, certos nomes são constituídos por apenas duas letras e poderão aparecer nas notas anonimizadas como substituição de outros *tokens* originais. Os métodos de *recall* por *pattern-matching* e a *recall* de Levenshtein detetam essas ocorrências de nomes com duas letras noutros contextos erradamente, gerando Falsos Positivos. Considera, de forma errada, como não anonimizado.

Em termos de perda de informação da nota clínica original para a nota clínica anonimizada, por ambos os idiomas, foi verificado que os métodos baseados em NER são mais eficazes, apresentando valores muito abaixo dos valores obtidos através dos baseados em *word embeddings*, o que se traduz numa maior capacidade em reter informação. Este resultado demonstra que, apesar das novas estratégias baseadas

em *embeddings* serem extremamente promissoras, garantindo uma confiança no processo de anonimização com o qual as tradicionais não podem competir, apresentam ainda graves lacunas no que diz respeito à retenção de informação. É de ressalvar também que os modelos de *word embeddings* usados foram treinados em datasets limitados, e é de esperar que estes resultados melhorem se forem utilizados modelos mais poderosos, treinados num maior volume de dados clínicos de alta qualidade, como aqueles usados para treinar LLMs clínicos.

A diferença de resultados entre os dois contextos linguísticos testados podem ser justificadas por vários fatores. Um dos grandes fatores diferenciadores é a quantidade de notas disponíveis para treino e teste dos modelos. A língua portuguesa é considerada um idioma de baixo recurso devido à escassez de dados para uma exploração mais profunda da capacidade dos modelos neste contexto.

É de notar que a maior perda de informação ocorreu aquando da utilização de métodos baseados em *word embeddings*, e mais expressivo nas experiências em português. Isto pode ser explicado pelo facto de que, na estratégia de *word embeddings* o mais importante é a diversidade de palavras aprendidas no treino do modelo. Quanto maior o número de palavras e contextos diferentes o modelo tiver treinado, melhor desempenho ele vai demonstrar. A eficácia em reter a informação relevante vai aumentar, visto que vai ser ter mais vocabulário para arranjar sinónimos para os *tokens* e, assim, preservar o seu conteúdo clínico. Portanto, como o dataset português é muito menor e, consequentemente, menos diverso do que o dataset inglês utilizado nestes estudos é de esperar que se traduza numa maior perda de informação das notas anonimizadas portuguesas.

Por outro lado, observando os resultados da perda de informação utilizando o Presidio, podemos perceber que a maior perda de informação se destaca com o dataset inglês. Em NER há perda de informação se uma entidade não sensível for considerado como sensível erradamente. Portanto, neste caso, quanto menor a diversidade dos dados de treino, melhor poderá ser o desempenho dos modelos porque, ao não haver muita variabilidade de palavras este modelo fica mais estável e é mais fácil encontrar entidades sensíveis corretamente. Considerando que as notas de teste em português são mais constantes (todas recolhidas no mesmo contexto clínico, em pacientes com características semelhantes e submetidos a cirurgia cardiotorácica) é possível que essa consistência a que ocorram menos Falsos Positivos e, portanto, menos perda de informação.

Ainda assim, foi possível verificar que a grande parte das tendências observadas se verificaram em ambos os contextos linguísticos, sugerindo que o sucesso verificado para o idioma inglês será extensível a outros idiomas, garantindo as mesmas condições.

O impacto da diferença de recursos disponíveis foi particularmente visível nos resultados apresentados para a perda de informação clínica, onde os modelos de *word embeddings* treinados nos escassos dados em português foram incapazes de produzir palavras parónimas em número suficiente para garantir uma proximidade semântica entre os conteúdos das notas originais e anonimizadas elevada o suficiente para reter a maioria da informação clínica relevante. Apesar destas técnicas apresentarem elevadas perdas de informação, foi possível verificar que estas foram consideravelmente inferiores no contexto inglês, potenciadas muito provavelmente por uma maior quantidade de termos clínicos disponíveis para treino.

Em termos das medidas da *recall*, o contexto inglês denotou melhores resultados, como seria de esperar, com exceção das medidas de LR para os métodos baseados em *word embeddings*. Este resultado poderá ser explicado pelas características dos datasets em estudo. O SemClinBR, usado para treinar estes modelos, não apresenta qualquer tipo de informação sensível, já que foi manualmente anonimizado. Por outro lado, as entidades sensíveis presentes no set de teste em Português foram geradas pela biblioteca *Faker*, constituindo por isso conjuntos de palavras mais particulares. Pelo contrário, os modelos em inglês foram treinados em notas clínicas da MIMIC III, que contém uma maior quantidade de termos que se poderão assemelhar a informação sensível, e testados contra o dataset de teste do desafio de 2014

da i2b2, que contém entidades mais desafiantes tais como iniciais de nomes, que poderão ter corrompido o cálculo da *Levenshtein Recall*. Assim sendo, estes resultados poderão ter sido fruto não de uma melhor performance dos métodos em português, mas sim das diferenças inerentes dos diferentes datasets usados.

Além disso, o facto de o dataset de treino ser em português do Brasil e o dataset de teste ser em português de Portugal pode ter levado a alguma discrepância nas palavras, uma vez que existem diferenças linguísticas entre as duas vertentes da língua portuguesa.

Comparando as metodologias baseadas em *word embeddings*, o Word2Vec demonstrou um melhor desempenho na tarefa de anonimização quando treinado e testado com datasets em inglês, enquanto o GloVe foi melhor que o Word2Vec em contexto português, embora esta última diferença não seja significativa, apresentando ambos os algoritmos resultados muito próximos. Estes resultados sugerem que as diferenças nas duas metodologias de criação das representações vetoriais não têm impacto significativo na tarefa final de anonimização, constituindo ambas uma alternativa viável para este fim.

Apesar dos métodos baseados em *word embeddings* serem métodos promissores devido ao seu alto poder de anonimização (muito acima dos métodos que utilizam NER), é claro que estes têm ainda um longo caminho a percorrer e precisam de ser alvo de mais exploração, uma vez que a perda de informação das notas clínicas anonimizadas é um fator determinante na avaliação do sucesso de uma técnica automática de anonimização textual. Para que continuem a ser úteis para pesquisa e investigação ou estatística médica é necessário que haja capacidade de retenção de informação relevante na nota anonimizada. Desta forma, as técnicas baseadas em NER continuam a ser as mais apropriadas, tendo em conta todos os fatores que avaliam o processo de anonimização.

No entanto, é necessário continuar a melhorar e investigar estes novos métodos ou outros emergentes como é o caso da utilização das *word embeddings* contextuais ou das LLMs, que são um método de anonimização generativa e poderão ser uma forte alternativa para qualquer tarefa de NLP.

A nova métrica – *Levensthtein Recall* – que foi proposta permite avaliar todos estes novos métodos emergentes, impulsionando a comunidade científica a explorá-los e a propor novos também novas métricas de avaliação para esses métodos e para o próprio processo de anonimização.

5. CONCLUSÃO

Nesta secção vão ser expostas as conclusões finais sobre os vários estudos feitos ao longo do trabalho.

Como forma de conclusão vão ser referidos os resultados e reflexões finais de maneira a serem verificados se os objetivos iniciais foram atingidos e se a questão central do estudo foi adequadamente respondida pelas experiências realizadas.

5.1. Conclusões finais

No presente estudo foram explorados técnicas de anonimização emergentes, assim como novos métodos de avaliação com capacidade de avaliar eficazmente essas novas técnicas. Foram também comparados resultados entre duas estratégias de anonimização, NER ou utilização de *word embeddings*, de forma a entender quais os métodos com melhor desempenho.

Quando foram avaliados os métodos com o conjunto de dados em inglês, o Word2Vec foi o que se demonstrou melhor na tarefa de anonimização, mas a custo da perda de informação, que se demonstrou elevada. O Presidio + spaCy foi o método que se revelou com melhor equilíbrio entre desempenho da anonimização e menor perda de informação. Avaliando os métodos utilizados para o conjunto de dados em português o GloVe demonstrou-se ligeiramente melhor que o Word2Vec, mas, no entanto, a perda de informação é muito pronunciada em ambos os métodos baseados em *word embeddings*, quando comparados com o Presidio + spaCy.

Comparando os métodos baseados em *word embeddings* (Word2Vec e GloVe) para os dois contextos linguísticos (inglês e português) os resultados da métrica de LR são muito semelhantes em ambos os idiomas. O que os distingue é o resultado da métrica da perda de informação que é mais predominante no GloVe quando utilizado no dataset inglês, ou seja, há maior perda de informação.

O surgimento de novos métodos de anonimização, como os baseados em *word embeddings* ou os métodos generativos (como as LLMs), é necessário recorrer a outro tipo de métricas que sejam capazes de avaliar eficazmente a tarefa de anonimização. Por esse motivo a métrica *Levenshtein Recall* foi desenvolvida para ultrapassar os desafios das métricas tradicionais impostos pelos métodos emergentes.

Apesar desta nova métrica conseguir colmatar algumas falhas na avaliação da anonimização pelos métodos tradicionais ainda é necessária uma exploração mais aprofundada deste cálculo para, por exemplo, determinar o valor mais adequado do *threshold*.

Pode afirmar-se que, ainda que os métodos baseados em *word embeddings* se demonstrem promissores e demonstrem um grande poder de anonimização, a grande perda de informação faz com que os métodos de NER continuem a liderar os métodos com melhor desempenho nesta tarefa de anonimização.

Assim, conclui-se que as técnicas de anonimização emergentes poderão vir a ser alternativas viáveis no futuro se o seu poder de anonimização for cada vez mais explorado e se forem investigando estratégias para minimizar a perda de informação crítica que estas novas técnicas acarretam. Para desbloquear este potencial das *word embeddings* e métodos generativos são também necessárias métricas que acompanhem a evolução destas novas estratégias de anonimização, como a *Levenshtein Recall*.

5.2. Limitações e investigação futura

Algumas das limitações encontradas ao longo deste trabalho estão relacionadas com os datasets utilizados. Os datasets em idiomas diferentes apresentam também características muito distintas, o que pode ter impactado os resultados e as conclusões que foram retiradas. A escassez de dados em português disponíveis para estudo foi considerada uma limitação para este trabalho, visto que uma recolha maior de dados teria sido mais benéfica para a análise da capacidade dos modelos desenvolvidos.

Futuramente, seria interessante promover a criação de métodos que permitam perceber as semelhanças e diferenças entre datasets para que seja possível utilizar datasets em idiomas diferentes, mas com características semelhantes para que haja uma comparação mais justa. Com isso, seria necessário realizar uma nova avaliação dos modelos.

Ao longo deste trabalho foram desenvolvidas métricas para auxiliar na avaliação de modelos mais inovadores, como o caso das *word embeddings* e dos métodos generativos. Estas métricas merecem futura exploração e mais experiências para entender a sua influência nos resultados obtidos. O *threshold* utilizado no cálculo da LR não tem qualquer sustentação teórica, o que é uma desvantagem, porque não há comprovação de que é o valor mais adequado de *threshold*.

Outro aspetos para uma possível investigação futura seriam a realização da anonimização através da exploração de *promptings* a LLMs e o uso de métodos baseados em *word embeddings* contextuais para a anonimização total de texto clínico.

REFERÊNCIAS BIBLIOGRÁFICAS

Abdalla, M., Abdalla, M., Rudzicz, F., & Hirst, G. (2020). Using word embeddings to improve the privacy of clinical notes. *Journal of the American Medical Informatics Association*, *27*(6), 901–907. https://doi.org/10.1093/jamia/ocaa038

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). *Publicly Available Clinical BERT Embeddings*. http://arxiv.org/abs/1904.03323

Anala, S. (2020). *A Guide to Word Embedding*. Towards Data Science. https://towardsdatascience.com/a-guide-to-word-embeddings-8a23817ab60f

Aramaki, E., Imai, T., Miyo, K., & Ohe, K. (2006). Automatic deidentification by using sentence features and label consistency. *I2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, *April*, 10–11. http://luululu.com/profile/paper/2006-i2b2/i2b2-deid.pdf

Catelli, R., Gargiulo, F., Casola, V., De Pietro, G., Fujita, H., & Esposito, M. (2021). 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. https://doi.org/10.1109/ACCESS.2021.3054479

Chen, J., Wu, Y., Jia, C., Zheng, H., & Huang, G. (2020). Customizable text generation via conditional text generative adversarial network. *Neurocomputing*, *416*(2019), 125–135. https://doi.org/10.1016/j.neucom.2018.12.092

Dehghan, A., Kovacevic, A., Karystianis, G., Keane, J. A., & Nenadic, G. (2015). Combining knowledge- and data-driven methods for de-identification of clinical narratives. *Journal of Biomedical Informatics*, *58*, S53–S59. https://doi.org/10.1016/j.jbi.2015.06.029

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, *24*(3), 596–606. https://doi.org/10.1093/jamia/ocw156

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., Corriveau, J. P., Walker, M., Chowdhury, S., Vaillancourt, R., Roffey, T., & Bottomley, J. (2009). A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. *Journal of the American Medical Informatics Association*, *16*(5), 670–682. https://doi.org/10.1197/jamia.M3144

Friebely, A. (2022). *Efficacy of Microsoft Presidio*. *May*.

Friedrich, M., Köhn, A., Wiedemann, G., & Biemann, C. (2020). Adversarial learning of privacy-preserving text representations for de-identification of medical records. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 5829–5839. https://doi.org/10.18653/v1/p19-1584

Gardner, J., & Xiong, L. (2008). HIDE: An integrated system for health information DE-identification. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, 254–259. https://doi.org/10.1109/CBMS.2008.129

Gupta, D., Saul, M., & Gilbertson, J. (2004). Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. *American Journal of Clinical Pathology*, *121*(2), 176–186. https://doi.org/10.1309/E6K33GBPE5C27FYU

Haldar, R., & Mukhopadhyay, D. (2011). *Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach*. *Ld*. http://arxiv.org/abs/1101.1232

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. http://arxiv.org/abs/1508.01991

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(1), 160035. https://doi.org/10.1038/sdata.2016.35

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Morgan Kaufmann Publishers Inc.*, *11*(1), 1–84. https://doi.org/10.5555/645530.655813

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 260–270. https://doi.org/10.18653/v1/n16-1030

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Liu, Z., Tang, B., Wang, X., & Chen, Q. (2017). De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, *75*, S34–S42. https://doi.org/10.1016/j.jbi.2017.05.023

Mendels, O., & Balter, A. (n.d.). *Presidio: Context aware, pluggable and customizable data protection and de-identification sdk for text and images*. https://github.com/microsoft/presidio

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: A review of recent research. *BMC Medical Research Methodology*, *10*(August). https://doi.org/10.1186/1471-2288-10-70

Microsoft. (2023). *Documentação da Linguagem de IA do Azure*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Mitchell, T. M. (1997). Machine Learning. In *McGraw-Hill Science/Engineering/Math*. https://doi.org/10.1007/978-3-031-17922-8_9

Mollaei, N., Cepeda, C., Rodrigues, J., & Gamboa, H. (2022). *Biomedical Text Mining: Applicability of Machine Learning-based Natural Language Processing in Medical Database*. *March*, 159–166. https://doi.org/10.5220/0010819500003123

Neamatullah, I., Douglass, M. M., Lehman, L. W. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., & Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, *8*, 1–17. https://doi.org/10.1186/1472-6947-8-32

Oliveira, L. E. S. e., Peters, A. C., da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., & Moro, C. M. C. (2022). SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. *Journal of Biomedical Semantics*, *13*(1), 1–19. https://doi.org/10.1186/s13326-022-00269-1

Pandey, B., Kumar, D., Pratap, B., & Rhmann, W. (2022). 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), 5083–5099. https://doi.org/10.1016/j.jksuci.2021.01.007

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018).

Deep contextualized word representations. *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 2227–2237. https://doi.org/10.18653/v1/n18-1202

Pethani, F., & Dunn, A. G. (2023). Natural language processing for clinical notes in dentistry: A systematic review. *Journal of Biomedical Informatics*, *138*(January). https://doi.org/10.1016/j.jbi.2023.104282

Povlsen, C., Jongejan, B., Hansen, D. H., & Simonsen, B. K. (2016). Anonymization of Court Orders. *Iberian Conference on Information Systems and Technologies, CISTI*, *2016-July*, 1–4. https://doi.org/10.1109/CISTI.2016.7521611

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Ribeiro, B., Rolla, V., & Santos, R. (2023). INCOGNITUS: A Toolbox for Automated Clinical Notes Anonymization. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, 187–194.

Samarati, P., & Sweeney, L. (1998). Protecting Privacy When Disclosing Information: K Anonymity and its Enforcement through Suppression. *International Journal of Computing Algorithm*, *001*(001), 19–22. https://doi.org/10.20894/ijcoa.101.001.001.004

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e, Copara, J., Gumiel, Y. B., Oliveira, L. F. A. de, Paraiso, E. C., Teodoro, D., & Barra, C. M. C. M. (2020). *BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition*. 65–72. https://doi.org/10.18653/v1/2020.clinicalnlp-1.7

Sogancioglu, G., Mijsters Amar Van Uden, F., & Peperzak, J. (2021). Gender bias in (non)-contextual clinical word embeddings for stereotypical medical categories; Gender bias in (non)-contextual clinical word embeddings for stereotypical medical categories. In *Proceedings of Utrecht University (INFOMHCML'2021)* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/xxxxxxx

South, B. R., Mowery, D., Suo, Y., Leng, J., Ferrández, Ó., Meystre, S. M., & Chapman, W. W. (2014). Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. *Journal of Biomedical Informatics*, *50*, 162–172. https://doi.org/10.1016/j.jbi.2014.05.002

Stenetorp, P., Pyysalo, S., & Topi, G. (2012). *brat: a Web-based Tool for NLP-Assisted Text Annotation - ACL Anthology*. *Figure 1*, 102–107. https://aclanthology.org/E12-2021/

Stubbs, A., Kotfila, C., & Uzuner, Ö. (2015). 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. https://doi.org/10.1016/j.jbi.2015.06.007

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system. *Proceedings : A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*, 333–337.

Taira, R. K., Bui, A. A. T., & Kangarloo, H. (2002). Identification of patient name references within medical documents using semantic selectional restrictions. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 757–761.

Taira, R. K., Soderland, S. G., & Hospital, C. (1999). *A Statistical Natural Language Processor for Medical Reports*. 970–974.

Tveit, A., Edsberg, O., Røst, T., & Faxvaag, A. (2004). Anonymization of general practioner's patient records'. *Proceedings of the HelsIT*, *7489*.

http://scholar.google.no/scholar?start=10&q=amund+tveit&hl=no&as_sdt=0,5#0

Uzuner, Ö., Luo, Y., & Szolovits, P. (2007). Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association*, *14*(5), 550–563. https://doi.org/10.1197/jamia.M2444

Wang, D., Su, J., & Yu, H. (2020). Feature extraction and analysis of natural language processing for deep learning english language. *IEEE Access*, *8*, 46335–46345. https://doi.org/10.1109/ACCESS.2020.2974101

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R., & Houston, A. (2013). *OntoNotes*. 11–12. https://doi.org/10.3115/1620950.1620956

Wu, Y., Jiang, M., Lei, J., & Xu, H. (2015). Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. *Studies in Health Technology and Informatics*, *216*, 624–628. https://doi.org/10.3233/978-1-61499-564-7-624

Wu, Y., Yang, X., Bian, J., Guo, Y., Xu, H., & Hogan, W. (2018). Combine Factual Medical Knowledge and Distributed Word Representation to Improve Clinical Named Entity Recognition. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2018*, 1110–1117.

Yang, X., Lyu, T., Li, Q., Lee, C. Y., Bian, J., Hogan, W. R., & Wu, Y. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, *19*(Suppl 5), 1–9. https://doi.org/10.1186/s12911-019-0935-4

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). Hurtful words. *ACM CHIL 2020 - Proceedings of the 2020 ACM Conference on Health, Inference, and Learning*, 110–120. https://doi.org/10.1145/3368555.3384448