David Pereira da Silva
Orientador: Romuere Rodrigues Veloso e Silva

Reconhecimento de Entidades Nomeadas em Receitas Médicas Manipuladas

Picos - PI
20 de julho de 2023

David Pereira da Silva
Orientador: Romuere Rodrigues Veloso e Silva

Reconhecimento de Entidades Nomeadas em Receitas Médicas Manipuladas

Monografia submetida ao Curso de Bacharelado em Sistemas de Informação como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação. Orientador: Prof. Dr. Romuere Rodrigues Veloso e Silva.

Universidade Federal do Piauí
Campus Senador Helvídio Nunes de Barros
Bacharelado em Sistemas de Informação

Picos - PI
20 de julho de 2023

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca José Albano de Macêdo

S586r Silva, David Pereira da
Reconhecimento de entidades nomeadas em receitas médicas manipuladas [recurso eletrônico] / David Pereira da Silva - 2023.
45 f.

1 Arquivo em PDF
Indexado no catálogo *online* da biblioteca José Albano de Macêdo-CSHNB
Aberto a pesquisadores, com restrições da Biblioteca

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Piauí, Bacharelado em Sistemas de Informação, Picos, 2023.
"Orientador : Prof. Dr. Romuere Rodrigues Veloso e Silva"

1. Linguagem natural – Sistemas de informação. 2. Reconhecimento de entidade nomeada. 3. Receitas médicas manipuladas. 4. Registros eletrônicos de saúde. I. Silva, Romuere Rodrigues Veloso e. II. Título.

CDD 006.35

Emanuele Alves Araújo CRB 3/1290

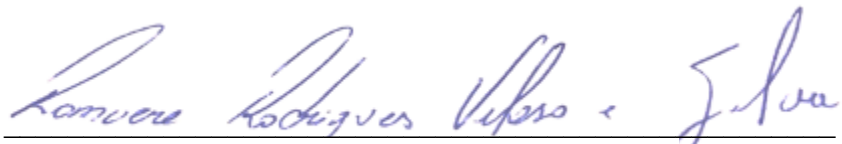
RECONHECIMENTO DE ENTIDADES NOMEADAS EM RECEITAS MÉDICAS MANIPULADAS

DAVID PEREIRA DA SILVA

Monografia aprovada como exigência parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Data de Aprovação

Picos – PI, 14 de agosto de 2023



Prof. Romuere Rodrigues Veloso e Silva

Profa. Deborah Maria Vieira Magalhães

Me. Orrana Lhaynher Veloso de Sousa

Agradecimentos

Desejo expressar minha profunda gratidão a todas as pessoas que estiveram ao meu lado durante esta incrível jornada.

Em primeiro lugar, quero expressar minha profunda gratidão aos meus pais, Maria de Jesus e José Valmir, por seu constante incentivo aos meus estudos, pela dedicação que sempre demonstraram (e continuam demonstrando) em prol de mim, pelo apoio incansável e pelos valiosos conselhos. Mãe e pai, não tenho palavras suficientes para agradecer a vocês.

Minha irmã Aline merece um agradecimento especial por sua presença constante, apoio incondicional, proteção incansável e por tornar esta jornada mais fácil. Meu irmão Daniel por estar sempre ao meu lado com um apoio inabalável, tornando essa caminhada mais tranquila. E ao meu sobrinho Gabriel, agradeço por trazer alegria e risadas sempre que está por perto.

À minha avó, que, embora não possa estar presente neste momento, sei que está feliz por mim. Agradeço profundamente à vovó Nonata por todo o apoio ao longo dos anos.

À minha tia Girlania, por seu interesse constante na minha jornada acadêmica e incentivo incansável. Obrigado, tia Lana.

Aos amigos que considero como irmãos, Amanda e João Marcos, que tornaram minha vida mais leve, proporcionaram risadas inesquecíveis e momentos incríveis, e sempre estiveram disponíveis quando precisei, além do apoio e incentivo constantes.

À Brenda, por ser autenticamente ela mesma.

Aos que iniciaram essa jornada comigo, Carlos Daniel, Mayra e Vinícius, e nos autodenominamos o "quarteto fantástico", obrigado por estarem ao meu lado desde o começo. E a Fernando, que desempenhou um papel crucial nos últimos anos da graduação, obrigado, meu amigo!

Aos meus amigos Eraldo e Ianny, por todos os momentos incríveis que compartilhamos ao longo da graduação. Agradeço pelas risadas, pelos desabafos e pelo apoio mútuo que sempre nos fortaleceu.

Ao Professor Dr. Romuere Rodrigues Veloso e Silva, por sua orientação, ensinamentos valiosos e disponibilidade incansável. Obrigado, Professor Romuere!

À Orrana, que me acompanhou durante a elaboração desta monografia, por sua orientação, paciência e gentileza. Sou profundamente grato a você, Orrana.

A todas as pessoas que gentilmente me ofereceram carona no início da graduação e às bolsas BAE e PIBEX, sem as quais eu provavelmente não teria conseguido.

A todos os professores do curso e a toda a comunidade da UFPI Picos.

A todos que, de alguma forma, contribuíram positivamente para a minha formação, meu sincero agradecimento.

Seja paciente, porque você não tem outra escolha. E seja gentil com as pessoas.
Frank Ocean

Resumo

Este trabalho aborda o desafio de Reconhecimento de Entidade Nomeada em receitas médicas manipuladas com o objetivo de extrair informações importantes e relevantes armazenadas em registros eletrônicos de saúde. Várias técnicas de processamento de linguagem natural e aprendizado de máquina foram usadas e comparadas para atingir esse objetivo. A pesquisa implementa modelos clássicos como campos aleatórios condicionais, em inglês *Conditional Random Fields* (CRF) comumente usados para tarefas NER e arquiteturas de redes neurais, como memória de longo e curto prazo, em inglês *Long Short-Term Memory* (LSTM) e memória longa de curto prazo bidirecional, em inglês *Bidirectional Long Short-Term Memory* (BiLSTM). Além disso, este estudo usou um modelo na sua versão pré-treinada em português de representações bidirecionais de codificadores de transformadores, em inglês *Bidirectional Encoder Representations from Transformers* (BERT), chamado BERTimbau conhecido por sua capacidade de representar palavras com base em uma ampla gama de contextos. A técnica de incorporação de palavras também foi aplicada para ajudar a mapear as palavras para vetores numéricos, permitindo uma compreensão mais profunda dos padrões semânticos das receitas. Os modelos LSTM e BiLSTM, quando combinados com as representações Word2Vec e GloVe, alcançaram um desempenho de 95% em medida F1-*score*. Esse resultado evidencia a eficácia da abordagem na solução do desafio de Reconhecimento de Entidades Nomeadas em receitas médicas manipuladas.

Palavras-chaves: Processamento de Linguagem Natural; Reconhecimento de Entidade Nomeada; Registros Eletrônicos de Saúde; Receitas Médicas Manipuladas.

Abstract

This work explores the challenge of Named Entity Recognition (NER) on manipulated medical prescriptions with the aim of extracting important and relevant information stored in Electronic Health Records (EHRs). We employed and compared various natural language processing and machine learning techniques to achieve this objective. The research implements classical models such as Conditional Random Fields (CRF), commonly used for NER tasks, as well as neural network architectures like Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM). Additionally, we utilized a pre-trained model in its Portuguese version of Bidirectional Encoder Representations from Transformers (BERT), known as BERTimbau, renowned for its ability to represent words based on a wide range of contexts. We applied word embedding techniques to help map words to numerical vectors, enabling a deeper understanding of the semantic patterns within prescriptions. The LSTM and BiLSTM models, when combined with Word2Vec and GloVe representations, achieved 95% of F1-score. This result highlights the effectiveness of the approach in meeting the challenge of recognizing named entities in manipulated medical prescriptions.

Keywords: Natural Language Processing; Named Entity Recognition; Electronic Health Records; Manipulated Prescriptions.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ADE	Adverse Drug Events
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
CBOW	Continuous Bag-of-Word
CRF	Conditional Random Fields
EHR	Electronic Health Record
GRU	Gated Recurrent Unit
GloVe	Global Vectors for Word Representation
LSTM	Long Short-Term Memory
MLM	Masked Language Model
MSEN	Multiple Single-Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NSP	Next Sentence Prediction
PA	Passive Aggressive
RE	Relation Extraction
RNN	Recurrent Neural Network
SEARN	Search-based Structured Prediction
TF-IDF	Term Frequency – Inverse Document Frequency

Sumário

1 Introdução

Registro Eletrônico de Saúde, em inglês *Electronic Health Record* (EHR) dispõem de um acervo de informações importantes que são indispensáveis para o bem-estar geral dos pacientes, como prescrições médicas, anotações clínicas, prontuários, etc. Este acervo contém as informações de prescrição que são instruções mais detalhadas sobre a medicação do paciente (TAO; FILANNINO; UZUNER, 2018). O Conselho Regional de Medicina do Estado de São Paulo afirma que provavelmente o documento mais emitido pelos médicos são receitas clínicas que estão em um formato não estruturado, ou seja, dados que não possuem uma estrutura ou padrão. Geralmente eles estão armazenados em formato de texto, chamado texto livre (CAMILO et al., 2010).

O problema é que cerca de 80% destes dados médicos permanecem não estruturados e não são aproveitados depois de serem criados (CONSULTANT, 2015). Como é difícil lidar com esse tipo de dado no EHR ou na maioria dos sistemas de informações hospitalares, muitas vezes ele é ignorado, não armazenados ou abandonados na maioria dos centros médicos por um longo tempo (Hon S. Pak, 2018). Estes dados em formato não estruturado possuem um valor considerável em termos de informações clínicas, requerendo a extração de conhecimento e adquirindo significado após um processamento adequado (BISTA; RANJAN, 2017). A extração dessas informações pode levar a grandes avanços na pesquisa biomédica, como uma epidemiologia, monitorar a adesão da receita ao tratamento do paciente, farmacovigilância.

Nesse contexto, Reconhecimento de Entidade Nomeada, em inglês *Named Entity Recognition* (NER) tem se mostrado uma técnica promissora para automatizar a extração e identificação precisa de informações contidas em documentos clínicos. NER é um subárea de Processamento de Linguagem Natural, em inglês *Natural Language Processing* (NLP) que visa identificar e classificar entidades específicas em texto, como nomes de pessoas, lugares, organizações, datas e, no caso de documentos clínicos, posologia, nome do medicamento, duração e outras informações contidas/existentes no documento (MOTA CRISTINA; SANTOS, 2007; NADEAU DAVID; SEKINE, 2007).

Extrair informações de documentos médicos é uma tarefa difícil devido à sua natureza não estruturada. Tradicionalmente, essas técnicas exigiam muita engenharia manual de recursos e mapeamento de ontologia, por essas razões a adoção dessas técnicas tem sido limitada (SHICKEL, 2017). A maneira como esses documentos são escritos dificulta a extração de nomes de medicamentos e outras informações de prescrição. Esses documentos são de fácil leitura para quem tem conhecimento médico, mas não se destinam ao processamento por computador. A presença de várias expressões, abreviaturas médicas, erros ortográficos e termos ambíguos que transmitem a mesma informação dificulta a análise automatizada desses documentos (TAO CARSON; FILANNINO, 2017).

Vários métodos estão atualmente disponíveis para extrair informações em diferentes idiomas. No entanto, a nossa investigação centra-se na extração de informação de receitas médicas manipuladas, designadamente tarefas de NER. Embora o inglês seja a língua com mais falantes ativos, é importante desenvolver trabalhos em outras línguas. O português é falado por mais de 250 milhões de falantes nativos e atualmente existe um interesse em processar documentação médica neste idioma. Com este estudo, esperamos avançar nos desafios da NER em receitas médicas manipuladas em português e contribuir para melhorias neste domínio.

Neste estudo, realizamos uma avaliação de várias metodologias, incluindo aprendizado de máquina, aprendizado profundo e técnicas de aprendizado de transferência para NER em prescrições médicas em português. Nosso objetivo é usar essas abordagens avançadas para detectar com precisão entidades especificadas em receitas médicas manipuladas.

1.1 Objetivos

O objetivo geral deste trabalho é desenvolver um modelo de Reconhecimento de Entidade Nomeada em português para identificação de entidades em receitas médicas manipuladas.

1.1.1 Objetivos especificos

- Construir uma base de receitas manipuladas rotuladas para a tarefa NER;
- Desenvolver modelos NER com diferentes classificadores para a identificação das entidades nomeadas;
- Avaliar o desempenho dos modelos NER em receitas médicas manipuladas com diferentes classificadores.

1.2 Contribuições

Como contribuições significativas para a área de NLP em saúde destacamos as seguintes:

- A avaliação comparativa de modelos pode fornecer informações valiosas sobre a eficácia e as limitações de cada método, ajudando a orientar a seleção de técnicas apropriadas para futuros projetos de NER em texto clínico;
- Desempenho de modelos tradicionais e modernos, incluímos modelos tradicionais, como CRF, e modelos modernos, como BERT no estudo, o trabalho fornecerá informações sobre como essas técnicas mais atuais se comportam em comparação com

abordagens mais antigas. Isso é particularmente relevante para entender se o uso de modelos mais avançados, como o BERT, justifica o investimento em recursos computacionalmente mais intensivos;

- Impacto de incorporação de palavras, usamos diferentes tipos (Word2Vec Skip-gram, GloVe e BERT) nos modelos que poderiam destacar como diferentes representações semânticas afetam o desempenho da tarefa NER em prescrições médicas. Isso fornecerá informações sobre a importância de incorporação de palavras pré-treinadas em contextos clínicos específicos e sua capacidade de capturar informações relacionadas ao reconhecimento de entidades.

1.3 Estrutura do trabalho

Além deste capítulo introdutório, o restante do monografia está organizada da seguinte forma: o Capítulo 2 apresenta os fundamentos teóricos da metodologia abordada. No Capítulo 3 são apresentados os trabalhos relacionados. No Capítulo 4 é apresentada a metodologia proposta. No Capítulo 5 são discutidos os resultados obtidos com a metodologia proposta. Por fim, no capítulo 6 é apresentada a conclusão e sugestões para trabalhos futuros.

2 Referencial Teórico

Nesta seção serão apresentados os principais conceitos inerentes ao projeto. Estes conceitos são divididos nos seguintes tópicos: (i) Registros Eletrônicos de Saúde, (ii) Reconhecimento de Entidades Nomeadas; (iii) Incorporação de palavras; (iv) Classificadores; e (v) Ajuste fino.

2.1 Registros Eletrônicos de Saúde

Progressos significativos foram feitos nos últimos anos na digitalização de registros médicos e na introdução de registros eletrônicos de saúde. Um EHR é uma plataforma digital que permite aos profissionais médicos armazenar, acessar e gerenciar eletronicamente as informações de saúde do paciente. O EHR é o registro de dados de saúde do paciente onde são armazenadas informações, tais como, diagnósticos, exames e resultados laboratoriais, prescrições, imagens radiológicas, anotações clínicas e mais (BIRKHEAD GUTHRIE S.; KLOMPAS, 2015). Embora tenham como objetivo principal melhorar a eficiência da assistência médica de uma perspectiva operacional, muitos estudos identificaram usos secundários para aplicativos de informações clínicas. Em particular, os dados do paciente contidos nos sistemas EHR são usados para tarefas como extração de informações (MEYSTRE, 2008).

2.2 Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas é uma tarefa importante no campo de NLP que consiste em identificar trechos de texto que mencionam entidades nomeadas e classificá-las em categorias pré-definidas. As categorias gerais de entidades nomeadas são nome de pessoas, organizações (como empresas, organizações governamentais, comitês), local (como cidades, países, rios), expressões temporais e expressões numéricas (como dinheiro, quantidades de outras unidades e porcentagens) (NADEAU DAVID; SEKINE, 2007). Também podem ser definidas categorias especificas de um domínio. Por exemplo, no domínio de receitas médicas, a via de administração, nome do medicamento, frequência que deve tomar a medicação, dosagem, entre outras entidades próprias do domínio.

O NER é comumente modelado como uma tarefa de rotulação de sequência que executa a identificação e classificação de entidades nomeadas. Dada uma sequência de entrada de n *tokens*, que são as palavras do conjunto de dados separadas por espaços (x1, x2, ..., xn), o modelo deve gerar uma sequência de entidades, que são as entidades nomeadas que serão preditas (y1, y2, ..., yn). Cada esquema de marcação define um vocabulário de tags e

restrições de transição de rótulos. O esquema de marcação IOB2, frequentemente utilizado na literatura, define tags B, I, O. A tag B marca o início de uma entidade e a tag I marca *tokens* subsequentes dentro da mesma entidade. A tag O é usada para *tokens* externos que não pertencem a nenhuma entidade. Portanto, no esquema IOB2, cada classe possui suas próprias tags B e I, permitindo que as entidades sejam identificadas e classificadas coletivamente (SOUZA FáBIO; NOGUEIRA, 2019).

Tabela 1 – Exemplo do esquema de marcação IOB2

Token	Tag
Alex	B-PER
está	O
indo	O
para	O
Teresina	B-LOC
Piauí	I-LOC

2.3 Incorporação de Palavras

Segundo MIKOLOV (2013), os sistemas e técnicas de NLP como *One-Hot Encoding*, N-gram, TF-IDF, tratam as palavras como entidades atômicas representadas por índices em um vocabulário, sem considerar o conceito de similaridade entre as palavras. Essa abordagem é suportada por sua simplicidade, robustez e pelo fato de que modelos simples treinados em grandes conjuntos de dados tendem a ter um desempenho melhor do que sistemas complexos treinados em pequenos conjuntos.

Todavia, essas técnicas simples atingem seus limites para muitas tarefas. Dessa forma, existem situações em que a escalabilidade das técnicas básicas não resultará em avanços significativos, requerendo a concentração em técnicas mais avançadas, como as incorporação de palavras, em inglês *words embeddings*. Com os avanços recentes nas técnicas de aprendizado de máquina, tornou-se possível treinar modelos mais complexos em conjuntos de dados maiores, e esses modelos geralmente superam os modelos mais simples (MIKOLOV, 2013). Provavelmente, o conceito mais bem-sucedido é usar representações distribuídas de palavras conhecida como vetores de palavras (MIKOLOV, 2013).

A Incorporação de palavras são vetores de números reais que representam palavras em um espaço n-dimensional, aprendidos baseados em grandes corpora (múltiplos conjuntos de dados) não anotados e que podem capturar conhecimento sintático, semântico e morfológico (HARTMANN, 2017). De acordo com (BARONI MARCO; DINU, 2014) a incorporação de palavras podem ser divididas em duas famílias, modelos de contagem que trabalham com uma matriz de palavras de co-ocorrência, como *Global Vectors* (GloVe) e modelos preditivos que tentam prever palavras vizinhas com base em uma ou mais palavras de contexto, como Word2Vec (HARTMANN, 2017).

2.3.1 Word2Vec

Word2Vec é um modelo preditivo que pode ser implementado de duas formas: *continuous bag-of-words* (CBOW) e *continuous skip-gram* (MIKOLOV, 2013). Os modelos CBOW e *skip-gram* usam pequenas redes neurais para aprender o mapeamento de palavras para um ponto em um espaço vetorial. A diferença entre os métodos é que no CBOW o modelo recebe uma sequência de palavras sem a palavra do meio e tenta prever essa palavra omitida, enquanto no *skip-gram*, o modelo recebe uma palavra e tenta prever suas palavras vizinhas.

Como explicado anteriormente, o CBOW é usado para inferir a palavra do meio de uma sentença baseado nas palavras que a cercam. Para representação, utilizaremos a seguinte notação:

([palavras de contexto], palavra alvo),

No exemplo a seguir usamos uma janela de contexto de tamanho 2. Isso significa que iremos "olhar" apenas uma palavra antes e depois da palavra alvo.

A receita "uso topico minoxidil 5% d pantenol 2% spray qsp 100 ml aplicar 6 borrifadas no couro cabeludo a noite" pode ser representada:

- ([uso, minoxidil], topico)
- ([topico, 5%], minoxidil)
- ([minoxidil, d], 5%)
- ([5%, pantenol], d)
- ([d, 2%], pantenol)
- ([pantenol, spray], 2%)

Com estes dados pretendemos ensinar o modelo a predizer uma palavra alvo, baseada em palavras de contexto. O modelo CBOW é mostrado na Figura 1 onde a camada de entrada recebe os dados pré-processados como acima e a palavras que representa a maior probabilidade da palavra alvo é incluída na saída.

No *skip-gram*, o segundo modelo utilizado no Word2Vec, é o inverso do CBOW, ao invés de tentarmos inferir a palavra central, da palavra central tentaremos inferir suas palavras vizinhas. Utilizaremos a mesma receita "uso topico minoxidil 5% d pantenol 2% spray qsp 100 ml aplicar 6 borrifadas no couro cabeludo a noite" e notação (palavra central, [palavras de contexto]) para exemplo.

- (topico, [uso, minoxidil])
- (minoxidil, [topico, 5%])

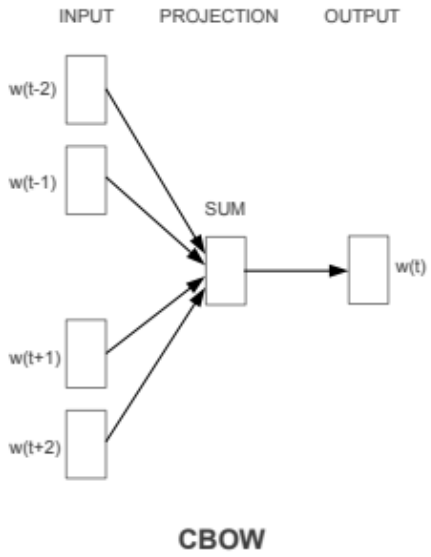


Figura 1 – Representação da arquitetura do modelo CBOW (MIKOLOV, 2013).

- (5%, [minoxidil, d])
- (d, [5%, pantenol])
- (pantenol, [d, 2%])
- (2%, [pantenol, spray])

Na Figura 2 temos a arquitetura do modelo *skip-gram* onde na camada de entrada, temos apenas a palavra alvo, e na saída temos diferentes probabilidades, cada uma contendo palavras de contexto possíveis.

Dois parâmetros são chave para o treinamento de incorporação de palavras Word2Vec: 1) o número de dimensões de representações (geralmente entre 50 e 300, pois equilibra a capacidade de capturar as informações semânticas e a eficiência computacional) e 2) o comprimento da janela de contexto (ou seja, quantas palavras antes e depois da palavra-alvo devem ser usadas como contexto para treinar a incorporação de palavras, geralmente 5 ou 10 palavras, pois janelas menores são focadas em capturar informações da palavra alvo, enquanto as maiores são para capturar um conteúdo mais amplo) (KHATTAK, 2019). Representações de treinamento com mais dimensões geralmente requerem mais dados de treinamento, mas cada dimensão deve capturar algum aspecto do significado, de modo que os representações devem ser grandes o suficiente para distinguir as palavras.

2.3.2 GloVe

Vetores Globais para Representação de Palavras, em inglês *Global Vectors for Word Representation* mais conhecida como GloVe é um algoritmo de aprendizado não supervi-

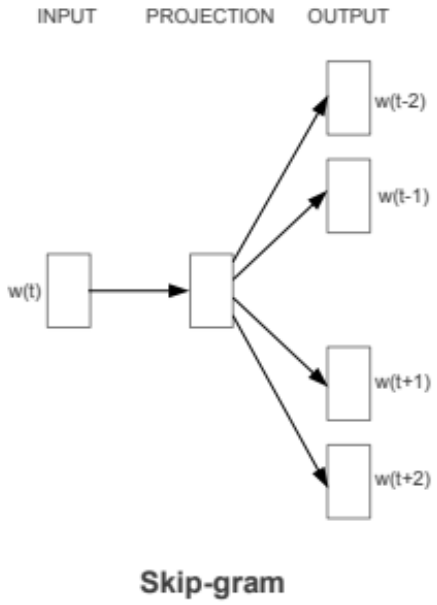


Figura 2 – Representação da arquitetura *skip-gram* (MIKOLOV, 2013).

Probabilidade e Razão	k = sólido	k = gás	k = água	k = moda
P(k|gelo)	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
P(k|vapor)	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
P(k|gelo)/P(k|vapor)	8.9	8.5×10^{-2}	1.36	0.96

Tabela 2 – Exemplo de taxas de probabilidade de uma combinação de palavras (PENNINGTON JEFFREY; SOCHER, 2014).

sionado proposto por (PENNINGTON JEFFREY; SOCHER, 2014) que obtêm representações vetoriais para palavras que capturam as relações semânticas e sintáticas entre as palavras com base em sua coocorrência. O treinamento é realizado em estatísticas globais agregadas de coocorrência palavra-palavra de um corpus, e as representações resultantes exibem subestruturas lineares interessantes do espaço vetorial de palavras. Essa técnica obteve resultados de última geração para tarefas de analogias sintáticas e semânticas.

GloVe combina o melhor do Word2Vec e matrizes de co-ocorrência. Ele aprende representações vetoriais usando informações globais de co-ocorrência de palavras. Isso significa considerar a frequência com que as palavras coocorrem no corpus, evitando a necessidade de calcular uma matriz de coocorrência densa potencialmente computacionalmente intensiva. A Tabela 2 mostra a probabilidade de co-ocorrência para palavras-alvo gelo e vapor utilizando um conjunto de palavras de contexto, e o resultado do calculo de probabilidade entre as palavras gelo e vapor com o contexto em questão, apresenta a relação entre elas e o contexto.

Como sólido é uma palavra usada com maior frequência no contexto de gelo, a razão P(gelo)/P(vapor) é alta, enquanto quando à palavra gás que está mais relacionada com

vapor a razão deve apresentar um resultado mais baixo. Para palavras que estão relacionadas com as duas, como água, palavras que não estão relacionadas com nenhuma das duas, como moda, a razão resultante do cálculo deve ser próxima de 1.

O processo de treinamento da GloVe começa com a construção de uma matriz de coocorrência que registra a frequência de coocorrência de cada palavra no corpus. Os pesos do modelo são então ajustados usando uma função objetivo destinada a minimizar a diferença do produto escalar das representações vetoriais das palavras. O resultado final é uma representação vetorial densa de cada palavra com sua proximidade no espaço vetorial refletindo a semelhança semântica ou sintática entre as palavras.

2.4 Classificadores

Os classificadores são uma abordagem popular e eficaz para realizar tarefas NER. Os classificadores usados em NER atribuem rótulos a palavras ou fragmentos de texto que indicam a qual categoria a entidade pertence. Esses classificadores são normalmente treinados usando técnicas de aprendizado supervisionado, onde um conjunto de dados rotulado é fornecido como entrada para o modelo. Existem várias abordagens e algoritmos de classificação que podem ser utilizados para NER. Alguns dos classificadores mais comuns incluem: classificadores baseados em aprendizado supervisionado (como, campos aleatórios condicionais), classificadores baseados em aprendizado profundo (como, Rede Neural Recorrente) e mais recentemente os classificadores baseados em modelos pré-treinados (como, representações bidirecionais de codificadores de transformadores).

2.4.1 Campos Aleatórios Condicionais

O modelo Campos Aleatórios Condicionais, em inglês *Conditional Random Fields* comumente chamado de CRF foi proposto por (LAFFERTY JOHN; MCCALLUM, 2001a) e é usado para prever dados sequenciais através da utilização de informações contextuais de rótulos anteriores, aumentando assim a quantidade de informações que o modelo possui para fazer uma boa previsão. É particularmente eficaz para problemas de aprendizado supervisionado onde os dados estão organizados em sequências, como análise de sentimentos, identificação de entidades nomeadas e segmentação de fala.

O CRF é um modelo estatístico sequencial que considera o contexto das palavras vizinhas para tomar decisões sobre a rotulação das entidades. Diferentemente de outros modelos, o CRF pode utilizar características arbitrárias dos dados para inferir as rotulações corretas. A maneira básica como um CRF funciona é definir um conjunto de características relacionadas à tarefa em mãos. Esses recursos podem ser extraídos de palavras, posições relativas de palavras, rótulos anteriores, etc. Cada instância da sequência é representada por um vetor de recursos.

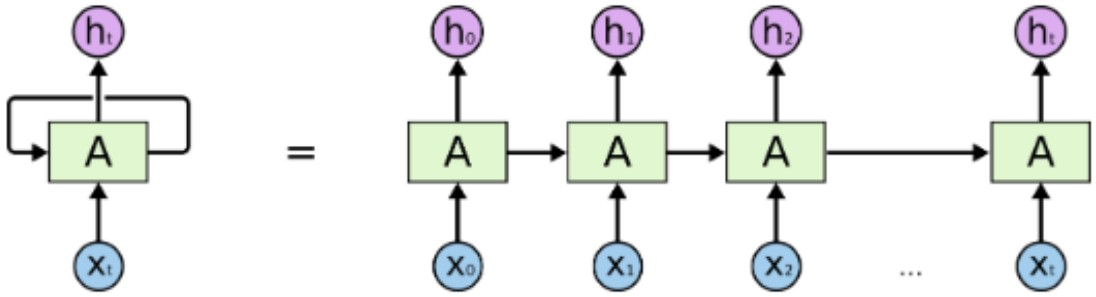


Figura 3 – Arquitetura básica de uma RNN (OLAH, 2015b).

O modelo CRF aprende como atribuir pesos a esses recursos durante a fase de treinamento. Esses pesos são usados para calcular a probabilidade condicional do rótulo para os recursos observados. Como a função de probabilidade é modelada usando uma distribuição exponencial, o CRF pode capturar relacionamentos complexos entre recursos e rótulos.

2.4.2 Memória de Longo e Curto Prazo

Uma rede neural recorrente (RNN) em aprendizado profundo é uma rede neural artificial que usa *links* reversos que permitem que os nós se conectem a outros nós em camadas anteriores ou a si mesmos para formar *loops* direcionados. Então essas arquiteturas possuem uma função de armazenamento onde os valores passados de um neurônio junto com a entrada da camada anterior representam a entrada do neurônio/camada. Assim, os valores de saída passados de um neurônio são determinados por suas entradas passadas e afetam sua saída atual. Na Figura 3, X_0 para X_t representa entradas em diferentes etapas de tempo, onde X_0, X1, X2... representam 1, 2, 3, ... entradas e X_t é a entrada atual.

As RNNs são frequentemente usadas para reconhecimento de padrões onde os resultados anteriores influenciam os resultados atuais, como dados de séries temporais e processamento de linguagem natural. Ainda assim, uma RNN simples como a arquitetura acima sofre com o problema de gradiente de fuga. Isso significa que a rede pode se lembrar apenas de entradas recentes e esquecer rapidamente as entradas de longo prazo. Para resolver este problema, uma variante de RNN conhecida como rede de memória de longo e curto prazo, em inglês *Long Short-Term Memory* (LSTM) foi introduzida por (HOCHREITER SEPP; SCHMIDHUBER, 1997).

Uma LSTM consiste em células de memória que podem armazenar informações por longos períodos de tempo e portas de entrada, saída e esquecimento que controlam o fluxo de informações dentro e fora das células de memória. As arquiteturas LSTM podem aprender a cada passo de tempo quais informações devem ser retidas, esquecidas ou atu-

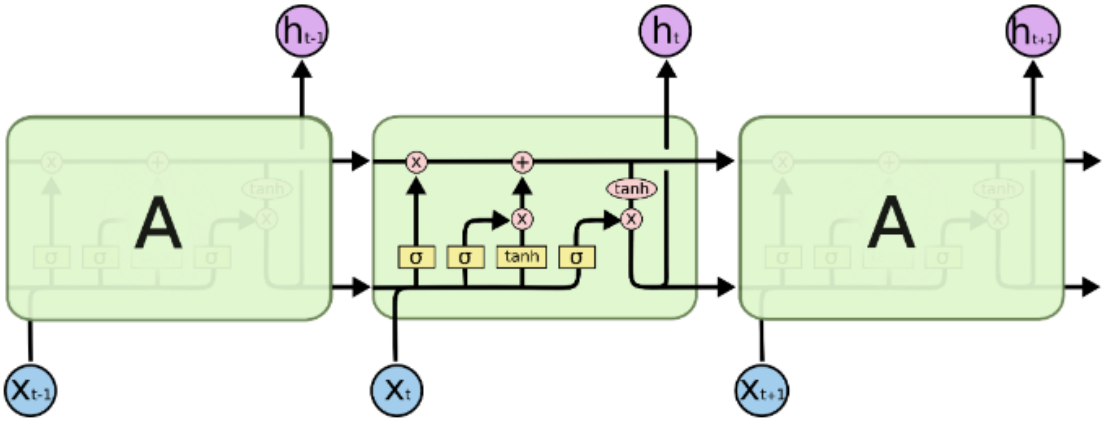


Figura 4 – Arquitetura de uma rede LSTM (OLAH, 2015b).

alizadas, tornando-as particularmente eficazes na modelagem de dependências de longo prazo (GREFF, 2016). Na Figura 4 a LSTM atinge capacidade de armazenamento de longo prazo com uma nova arquitetura. Como mencionado anteriormente, há três portas que cada uma das quais afeta o estado da célula de uma maneira diferente:

- *Portão de esquecimento*: determina quais informações o portão de esquecimento deve apagar da célula de memória;
- *Portão de entrada*: determina quais novas informações a porta de entrada adiciona à célula de memória;
- *Portão de saída*: determina quais informações a porta de saída apresenta como saída da célula de memória.

Essas portas normalmente produzem valores entre 0 e 1 e utilizam uma função de ativação sigmoide que descreve o nível de ativação da porta (onde 0 porta fechada e 1 porta aberta) (SANTANA, 2017). As LSTMs são muito populares devido à sua eficácia e são a forma predominante de RNN para fins práticos, especialmente ao lidar com dados de sequências e séries temporais.

2.4.3 Memória Bidirecional de Longo e Curto Prazo

A Memória Bidirecional de Longo e Curto Prazo, em inglês *Bidirectional Long Short-Term Memory* (BiLSTM) a entrada flui em duas direções, tornando uma BiLSTM diferente da LSTM. Com a LSTM, só é possível fazer a entrada fluir em uma direção, para trás ou para frente. No entanto, em bidirecional, é possível fazer a entrada fluir em ambas as direções para preservar as informações futuras e passadas (WANG, 2019).

A BiLSTM adiciona mais uma camada LSTM, que inverte a direção do fluxo de informações. Isso significa que a sequência de entrada flui para trás na camada LSTM adicional,

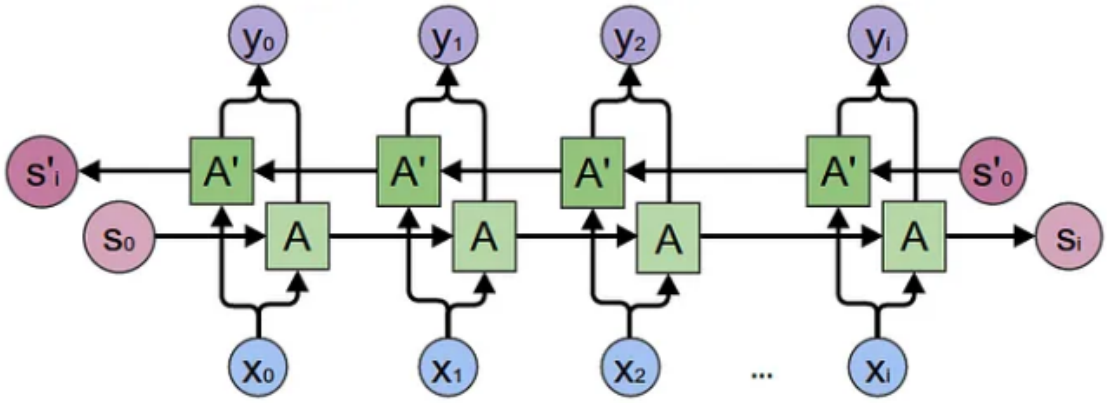


Figura 5 – Arquitetura de uma rede BiLSTM (OLAH, 2015a).

seguida pela agregação das saídas de ambas as camadas LSTM de várias maneiras, como média, soma, multiplicação ou concatenação. Na Figura 5 é mostrado uma arquitetura de uma rede BiLSTM, onde ambas as redes emitem suas saídas individuais com base nas informações do passado-presente e do presente futuro a cada passo de tempo.

Este tipo de arquitetura tem muitas vantagens em problemas do mundo real, especialmente em NLP. A principal conclusão é que cada componente de uma sequência de entrada possui informações do passado e do presente. Com isso dito, a BiLSTM pode produzir uma saída mais significativa, especialmente no caso de construção de modelos de linguagem, uma vez que as palavras em um bloco de texto geralmente são conectadas de duas maneiras - com palavras anteriores e palavras futuras. Portanto, o modelo BiLSTM é benéfico em muitas tarefas de processamento de linguagem natural, como classificação de sentenças, tradução e reconhecimento de entidades.

2.4.4 Representações Bidirecionais de Codificadores de Transformadores

Representações Bidirecionais de Codificadores de Transformadores, em inglês *Bidirectional Encoder Representations from Transformers* (BERT) foi desenvolvido por pesquisadores do Google AI Language (DEVLIN et al., 2019) que inovou ao apresentar resultados de ponta em uma ampla variedade de tarefas de NLP, incluindo resposta a perguntas, inferência de linguagem natural e outras. O BERT é baseado em *Transformers*, um modelo de aprendizado profundo no qual cada elemento de saída está conectado a cada elemento de entrada, e as ponderações entre eles são calculadas dinamicamente com base em sua conexão.

O BERT foi projetado para ler em ambas as direções ao mesmo tempo. Essa capacidade, possibilitada pela introdução dos *Transformers*, é conhecida como bidirecionalidade. Usando esse recurso bidirecional, o BERT é pré-treinado em duas tarefas de NLP diferen-

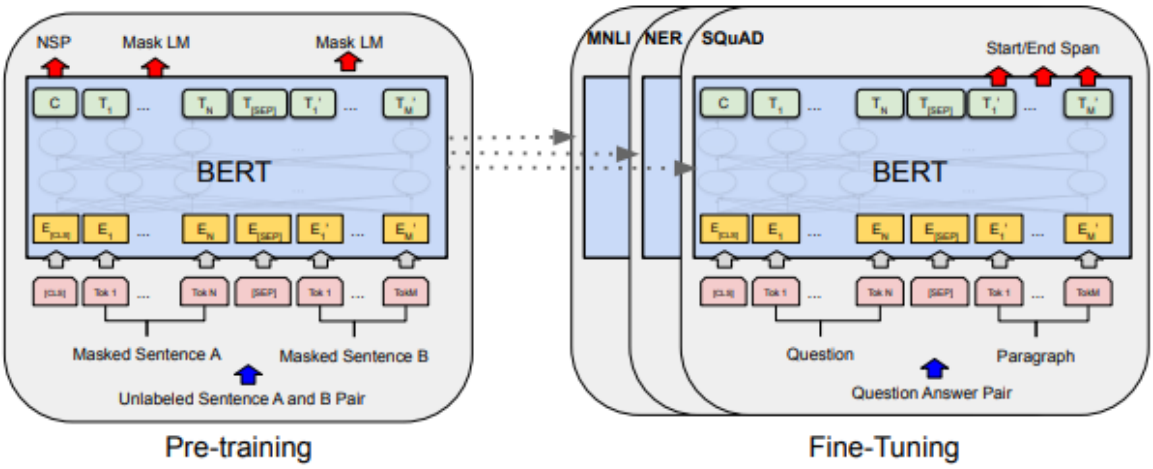


Figura 6 – Arquitetura para tarefa de perguntas e respostas (DEVLIN et al., 2019).

tes, mas relacionadas: Modelagem de linguagem mascarada, em inglês *Masked Language Model* (MLM) e Previsão da próxima frase, em inglês *Next Sentence Prediction* (NSP). Na MLM o objetivo do treinamento é ocultar uma palavra em uma frase e, em seguida, fazer com que o programa preveja qual palavra foi ocultada (mascarada) com base no contexto da palavra oculta. Já na NSP o objetivo do treinamento é fazer com que o programa preveja se duas sentenças fornecidas têm uma conexão lógica e sequencial ou se sua relação é simplesmente aleatória (DEVLIN et al., 2019).

Há duas maneiras de implementá-lo: pré-treinamento e ajuste fino. Durante o pré-treinamento, o modelo é treinado em dados não rotulados. Para ajuste fino, o modelo BERT é primeiro inicializado com parâmetros pré-treinados e todos os parâmetros são otimizados usando dados rotulados de tarefas *downstream*. Cada tarefa *downstream* tem um modelo de ajuste fino separado, mas eles são inicializados com os mesmos parâmetros pré-treinados (DEVLIN et al., 2019). Na Figura 6 um exemplo de uma arquitetura de perguntas e respostas descrevendo a execução.

O BERT também permite a extração de incorporação de palavras, além do fato de que possui vantagens sobre modelos tradicionais como o Word2Vec. Isso ocorre porque cada palavra no Word2Vec tem uma representação fixa, mas o BERT gera uma representação da palavra independentemente do contexto em que a palavra aparece, que é dinamicamente influenciado pelas palavras circundantes. Por exemplo, considere as duas frases a seguir.

- João foi ao banco sacar dinheiro;
- João sentou no banco.

Em ambas as sentenças, Word2Vec produz as mesmas incorporação de palavras para a palavra “banco”, mas o BERT tem representações diferentes para “banco” para cada

sentença. Além de capturar diferenças óbvias, como ambiguidade, as incorporação de palavras são sensíveis ao contexto também capturando outras formas de informação que levam a representações mais precisas de recursos, o que, por sua vez, melhora o desempenho do modelo.

2.5 Ajuste Fino

O ajuste fino, em inglês *Fine-tune* é um conceito usado no aprendizado de máquina e em uma ampla gama de tarefas e refere-se ao processo de ajustar ou melhorar um modelo pré-treinado para executar uma tarefa específica. Geralmente, os modelos pré-treinados são treinados em grandes quantidades de dados comuns, como texto da web, imagens de várias fontes, *Wikipédia*. Ao aplicar o ajuste fino, o modelo pré-treinado é ajustado para um conjunto de dados mais específico e relevante para a tarefa de interesse. Isso é feito continuando a treinar o modelo com dados adicionais rotulados ou anotados para tarefas específicas. A concepção por trás disso é que um modelo pré-treinado já internalize características abrangentes do domínio do problema, enquanto o ajuste fino pode personalizar o modelo para incorporar características mais específicas da tarefa em foco.

O ajuste fino treina apenas uma pequena parte do modelo (geralmente a classificação ou a última camada) e deixa as camadas anteriores inalteradas. Dessa forma, o modelo pré-treinado retém o conhecimento prévio e pode se adaptar a dados mais específicos. O ajuste fino é amplamente utilizado em vários campos, como processamento de linguagem natural, visão computacional e reconhecimento de fala, e é especialmente útil quando modelos pré-treinados estão disponíveis e podem ser ajustados para uma tarefa específica.

3 Trabalhos Relacionados

Nesta seção, exploramos estudos e pesquisas relevantes que fornecem uma base sólida para o contexto deste trabalho. Pesquisas anteriores abordaram questões semelhantes nessa área, fornecendo informações valiosas e estabelecendo as bases para nossa pesquisa. Esta revisão da literatura serve para posicionar nossa pesquisa dentro do contexto existente, destacando as lacunas de conhecimento e apontando a contribuição de nossa abordagem para este campo.

CHAPMAN (2019) propuseram um sistema de NER e RE para detecção de eventos adversos a medicamentos com o conjunto de notas clínicas do hospital da Faculdade de Medicina da Universidade de Massachusetts (UMASS). Na etapa de NER utilizaram o CRF para rotular cada *token* identificado no documento clínico em uma das entidades de interesse: nome do medicamento, frequência, dosagem, duração, via, gravidade, indicação, evento adverso a medicamento, SSLIF (outros sinais, sintomas e doenças, em inglês *Other Signs, Symptoms and Diseases*). Duas abordagens foram usadas para incorporação de palavras neste sistema.: (1) um conjunto foi treinado com o CBOW de fontes públicas e quase 100.000 notas de EHR e (2) outro conjunto foi treinado com o *skip-gram* sem nenhum dado de EHR. No pré-processamento foi feito uso de tokenização, marcação de parte da fala, em inglês *part of speech* (POS), detecção de nomes de medicamentos conhecidos usando recursos da MedEx (XU, 2010) e vetores de atributos. Optaram por utilizar o CRF, devido às restrições de tempo e recursos e por ser um algoritmo computacionalmente menos intensivo e focado na engenharia de recursos. Obtiveram um medida F1 médio de 0,88% no sistema NER.

(KIM YOUNGJUN; MEYSTRE, 2020) propôs um método de extração de medicamentos e informações relacionadas a partir de textos clínicos, utilizando um conjunto de modelos combinados (*ensemble method*) que consiste em combinar os resultados de múltiplos modelos de aprendizado de máquina para melhorar a precisão e o desempenho geral. Os modelos foram CRF (LAFFERTY JOHN; MCCALLUM, 2001b), CRFext, SEARN (DAUMÉ HAL; LANGFORD, 2009) e BiLSTM (SCHUSTER MIKE; PALIWAL, 1997) com CRF. Foram extraídas nove entidades: medicamento (nome), quantidade, duração, via, forma, ADE, dosagem, motivo (para prescrição) e frequência. O pré-processamento do texto envolveu tokenização, lematização, marcação de parte da fala. Cada modelo é treinado usando uma abordagem diferente e então eles são combinados usando um método de voto majoritário, onde a decisão final é tomada com base nas previsões da maioria dos modelos. Alcançaram 0,92% em medida F1. A utilização de um método *ensemble* possui complexidade e custo computacional alto necessitando mais tempo de treinamento em comparação com modelos individuais.

Em (SCHNEIDER, 2020) realizaram o ajuste fino em três modelos baseados no BERT

em corpora clínicos e biomédicos em português, i) BioBERTpt(clin) um modelo com os dados clínicos, a partir das narrativas de hospitais brasileiros; ii) BioBERTpt(bio) um modelo contendo dados biomédicos com base em resumos de artigos científicos; e iii) BioBERTpt(all) uma versão completa, ou seja, usando dados clínicos e biomédicos. Efetuaram dois experimentos NER em relação ao corpora, i) no primeiro experimento, usaram o SemClinBr (OLIVEIRA, 2022), um corpus semanticamente anotado para NER clínico em português, contendo 1.000 notas clínicas rotuladas com o esquema de marcação IOB2; ii) no segundo experimento, executaram os modelos no pequeno conjunto de dados CLINpt (LOPES FáBIO; TEIXEIRA, 2019) com o formato IOBES com 281 descrições de casos clínicos de neurologia. Em relação aos dois corpus, no SemClinBr o BioBERTpt(all) obteve melhor desempenho com medida F1 de 0,60%, já no CLINpt o BioBERTpt(clin) teve o melhor desempenho alcançando 0,92% em medida F1.

RAMACHANDRAN R; ARUTCHELVAN (2021) propuseram um método híbrido para NER em documentos biomédicos que combina abordagens baseadas em regras e baseadas em aprendizado de máquina. Utilizaram dados advindos da *web* como PubMed, medlineplus, WebMD e FDA (*Food and Drug Administration*) que são sites de informações de saúde, utilizando uma *API* em Python para buscar esses documento biomédicos. Para validação do modelo, foi realizado o desenvolvimento de um dicionário e intervenções humanas. O dicionário possui três entidades: sintomas, via de administração e formas de dosagem, além de extraírem também nome do medicamento e nome da doença. Na intervenção humana, o especialista de domínio pode identificar a entidade correta e atualizar as sentenças anotadas. O modelo chamado de Hybrid-NER (hNER) foi construído com base na CNN-LSTM. A abordagem proposta alcançou média de 0,73% em medida F1 nas cinco entidades e possui melhor desempenho quando um especialista de domínio intervem na validação.

Em (BÁEZ, 2022), os autores desenvolveram um sistema automatizado para a extração de entidades aninhadas em textos de encaminhamentos clínicos escritos em espanhol, as entidades aninhadas referem-se a entidades que estão contidas ou relacionadas hierarquicamente umas com as outras em um texto, ou seja, são entidades que estão dentro de outras entidades. O corpus utilizado inclui encaminhamentos não identificados da lista de espera em hospitais públicos chilenos com duas classes (58,6% médicos e 41,4% odontológicos) ele foi anotado manualmente com dez tipos de entidades, seis atributos (abreviações, doença, medicação, achado, parte do corpo, membro da família do paciente e procedimento) e pares de relações com relevância clínica. Eles utilizaram o esquema de marcação IOB2 para anotação do corpus e propuseram uma abordagem de Múltiplas Entidades Individuais, em inglês *Multiple Single-entity* NER (MSEN) que consiste no treinamento independente de vários modelos (BiLSTM + CRF), um para cada tipo de entidade. O modelo MSEN foi implementado no *framework Flair* (AKBIK, 2019). Destacamos como uma limitação a falta de investigação das entidades aninhadas, o que resultou em erros

na classificação nas entidades.

Tabela 3 – Trabalhos relacionados

Autor	Método	Incorporação de palavras	Domínio	Idioma
(CHAPMAN, 2019)	CRF	CBOW e *Skip-gram*	Notas clínicas	Inglês
(KIM YOUNG-JUN; MEYSTRE, 2020)	CRF, CRFext, SEARN e Bi-LSTM	GloVe	ADEs	Inglês
(SCHNEIDER, 2020)	BioBERTpt(clin), BioBERTpt(bio) e BERTpt(all)	X	Notas clínicas	Português
(RAMACHANDRAN R; ARUTCHELVAN, 2021)	CNN e LSTM	X	Notas clínicas	Inglês
(BÁEZ, 2022)	BiLSTM + CRF	Flair e BERT	Encaminhamentos	Espanhol

A Tabela 3 apresenta uma visão geral das tecnologias utilizadas nos trabalhos relacionados, organizadas cronologicamente de 2019 a 2022. Uma das principais características distintivas deste trabalho é a abordagem de quatro modelos distintos (CRF, LSTM, BiLSTM e BERTimbau) para a tarefa NER em prescrições médicas manipuladas. Em comparação com outros trabalhos apresentados usando métodos específicos. Isso permite comparar o desempenho de modelos tradicionais, como o CRF, com abordagens mais recentes, como o BERT, proporcionando uma visão mais abrangente dos avanços tecnológicos e possíveis benefícios do uso dos modelos mais avançados.

Além disso, a utilização de três tipos de representações (Word2Vec, GloVe e BERTimbau) também é um diferencial deste trabalho. Embora outros estudos possam ter reduzido a uma ou duas representações, este estudo explorou diferentes representações semânticas, permitindo uma análise mais profunda de como diferentes representações podem afetar o desempenho do modelo NER em um ambiente clínico. O conjunto de dados utilizado também é notável, pois consiste em textos em língua portuguesa relacionados a receitas médicas manipuladas. Portanto, a característica mais distintiva deste estudo, em comparação com outros, é a utilização desse conjunto de dados específico em português voltado para receitas médicas manipuladas.

4 Metodologia Proposta

Este capítulo apresenta os passos necessários para a execução deste trabalho. A seguir, descrevemos o conjunto de dados utilizado, seguido pelas etapas de pré-processamento. Em seguida, apresentamos a seleção das incorporação de palavras, de acordo com a literatura. Posteriormente, descrevemos os classificadores implementados. Após isso, detalhamos a preparação dos dados para cada modelo. Em seguida, apresentamos os hiperparâmetros de cada modelo. Por fim, descrevemos o processo de treinamento e as métricas escolhidas para avaliar o desempenho dos modelos.

4.1 Conjunto de Dados

O conjunto de dados utilizado neste trabalho foi extraído do trabalho de (SOUSA, 2022) que disponibilizou um conjunto de dados que compreende três categorias: receitas, anotações clínicas e solicitações de exames. O conjunto de dados foi produzido por médicos durante consultas presenciais. Foram coletadas 3.000 amostras de 10 de maio de 2010 a 11 de agosto de 2021.

No conjunto de dados original foi disponibilizado 1000 amostras para a classe de receitas, onde tem duas classes de receitas, as manipuladas e as industrializadas. Para uma melhor compreensão das duas classes, exemplo das duas categorias abaixo:

Exemplo de uma amostra de uma receita manipulada: “uso topico ac retinoico 0.05% ac kojico 3% ac fitico 3% desonida 0.05% alfa bisabolol 1% gel qsp 30g passar na face a noite e lavar pela manha”

Exemplo de uma amostra de uma receita industrializada: “uso oral alektos tomar um comp via oral uma vez ao dia por 15 dias”

Nos exemplos fornecidos, fica evidente que as receitas manipuladas detalham medicamentos feitos sob medida para atender às particularidades de cada paciente, ao passo que as receitas industrializadas se aplicam a medicamentos já prontos e disponíveis comercialmente no mercado.

Para este trabalho foi realizado a extração de 500 amostras de receitas manipuladas do conjunto de dados original. Na Tabela 4 são apresentadas as estatísticas descritivas, sendo elas (i) quantidade de receitas; (ii) quantidade de *tokens*; (iii) quantidade média de palavras do conjunto de dados.

Tabela 4 – Estatísticas descritivas do conjunto de dados

Quantidade de receitas	500
Quantidade de tokens	19754
Quantidade média de palavras	33

4.2 Pré-processamento

Após a extração das amostras do conjunto de dados disponibilizado, foi realizado a limpeza dos dados. O conjunto de dados disponibilizado não possui acentuação e os dados estão todos em minúsculos, então não se fez necessário essas etapas, ao invés disso, foram removidos caracteres especiais como: (”#’*-=@[]_|‘^ :) pois no conjunto de dados existem múltiplos medicamentos prescritos por receita, separados por “:”. A remoção destes caracteres foi realizada para fins de limpeza e padronização do conjunto de dados. Após isso, a tokenização das receitas foi realizada. Neste processo, as amostras são divididas em palavras individuais, chamadas de *tokens*. Por exemplo, na seguinte receita “tacrolimus 0.03 % creme qsp 30 g passar na mancha 2 vezes ao dia”, os *tokens* seriam: [“tacrolimus”, “0.03”, “%”, “creme”, “qsp”, “30”, “g”, “passar”, “na”, “mancha”, “2”, “vezes”, “ao”, “dia”].

A rotulação de dados, também conhecida como anotação de dados, refere-se ao processo de atribuir rótulos ou entidades a um conjuntos de dados. O processo para essa etapa foi a seguinte: i) Definição das entidades; ii) Atribuição dos rótulos; iii) Revisão. Foi definido seis entidades e a entidade O utilizando o esquema de marcação IOB2, a Tabela 5 apresenta todas elas.

Tabela 5 – Categorias de entidades definidas.

Entidade	Descrição
VIA	Via de administração do medicamento
MED	Nome do medicamento
DOS	Dosagem do medicamento
QTD	Quantidade total do medicamento
FREQ	Frequência de uso
DUR	Duração de uso
O	*Tokens* que não pertencem a nenhuma entidade

Na Figura 7 é possível ver a distribuição das entidades. Nas receitas manipuladas, a maioria das informações se concentra em duas principais entidades: “DOS” e “MED”. Na entidade “MED”, são detalhados os princípios ativos, que constituem os medicamentos personalizados criados para atender às necessidades individuais de cada paciente. Já na entidade “DOS”, são especificadas a dosagem personalizada dos princípios ativos, expressas em miligramas (mg), mililitros (ml) ou porcentagem (%).

4.3 Seleção de incorporação de palavras

Na revisão da literatura, foram analisados estudos sobre incorporação de palavras, que é uma técnica fundamental de processamento de linguagem natural. Essa técnica captura as relações semânticas e contextuais entre as palavras, permitindo que os modelos de aprendizado de máquina e aprendizado profundo entendam melhor o significado das

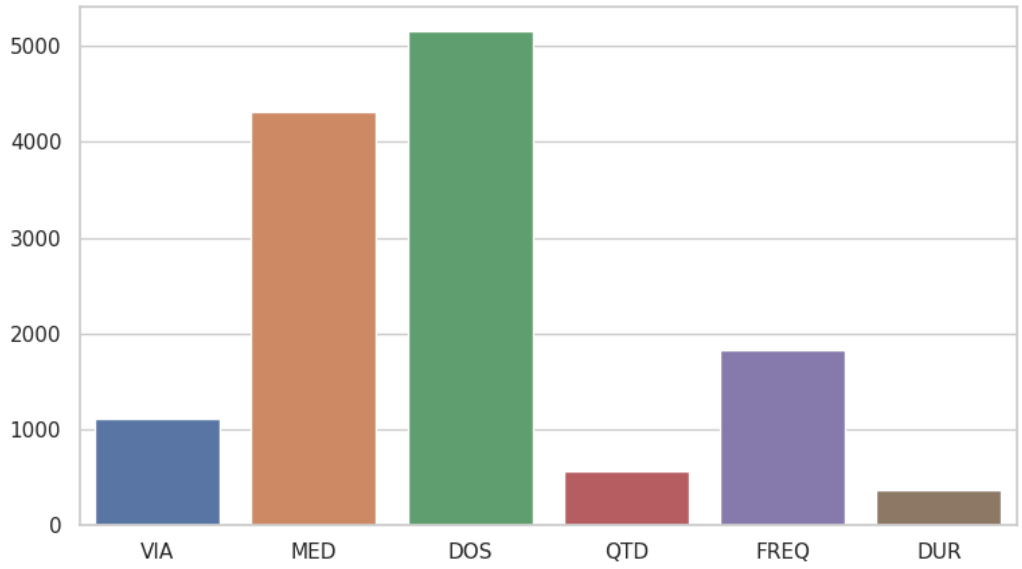


Figura 7 – Distribuição das entidades de NER.

palavras no texto. Durante a pesquisa, foram buscados artigos acadêmicos e artigos relacionados que tratassem de diferentes abordagens e usos, as mais amplamente utilizadas são Word2Vec (CHAPMAN, 2019), (ZHOU, 2023), GloVe (KIM YOUNGJUN; MEYSTRE, 2020), (ÇETINDAĞ CAN; YAZICIOĞLU, 2023) e mais recentemente, BERT (BÁEZ, 2022), (ZHOU, 2023).

Neste trabalho, os seguintes parâmetros foram adotados para cada algoritmo de incorporação de palavras: i) Para Word2Vec, escolhemos a versão de 100 dimensões *Skip-gram*; ii) Em GloVe foi utilizada a versão de 100 dimensões; iii) Para o BERT, escolhemos uma versão pré-treinada em português chamada BERTimbau que realizamos a extração das representações. As seleções foram feitas levando em consideração as necessidades e especificidades da pesquisa em questão, com o objetivo de obter representações eficientes e coerentes. As representações são extraídas do repositório NILC-Embeddings (HARTMANN, 2017), especialmente criado para a Língua Portuguesa, garantindo assim uma abordagem adequada ao contexto do trabalho.

4.4 Classificadores implementados

Nesta estudo, foram implementados quatro modelos para NER: CRF, duas RNNs (LSTM e BiLSTM) e o BERT na abordagem ajuste fino. O CRF foi implementado utilizando a biblioteca sklearn-crfsuite. As RNNs foram implementadas em Python utilizando a biblioteca de código aberto *Tensorflow*. Para o modelo BERT utilizou-se a arquitetura BERT base, com o modelo pré-treinado em português BERTimbau (SOUZA FáBIO; NO-

GUEIRA, 2020), também, para a linguagem Python com a biblioteca *Transformers*.

A biblioteca sklearn-crfsuite (KOROBOV, 2015) fornece uma interface de estilo scikit-learn simples e familiar para treinar e avaliar modelos CRF. É baseado no pacote crfsuite, uma biblioteca C++ eficiente para CRF. Com o sklearn-crfsuite, é possível aproveitar muitos dos recursos do scikit-learn, incluindo: capacidade de usar pipelines, realizar busca de hiperparâmetros e aproveitar outras ferramentas fornecidas pelo scikit-learn.

TensorFlow (LLC, 2015) é uma popular biblioteca de machine learning e deep learning de código aberto desenvolvida pelo Google. Suporta diversas arquiteturas de RNN, como LSTM e GRU. Estas são variantes de RNNs destinadas a resolver o problema comum do desaparecimento de gradientes para sequências longas. A arquitetura básica para RNN envolve a construção de uma sequência de células RNN. Cada célula é responsável por processar as entradas na sequência e atualizar o estado interno.

O BERT é um modelo de linguagem pré-treinado desenvolvido pelo Google, que utiliza a arquitetura *Transformer*, e tem sido amplamente utilizado para tarefas de NLP, como classificação de texto, perguntas e respostas, entre outros. Na abordagem de ajuste fino, adotamos a implementação do BERT com a versão pré-treinada em português Bertimbau, para classificação de palavras da *HuggingFace* na versão baseada em Pytorch, devido ao fato de que nosso conjunto de dados está em português.

4.5 Preparação dos dados

Os modelos possuem entradas de dados distintas, portanto, os conjuntos foram vetorizados de acordo com cada entrada dos modelos.

4.5.1 Entrada dos dados do CRF

Na etapa de treinamento, dois vetores são considerados como entradas para o CRF. Primeiro, um vetor contendo os rótulos, e segundo, um vetor de atributos. Esse vetor de atributos visa caracterizar todas as palavras do corpus selecionadas para o processo. Neste trabalho foi utilizado as seguintes atributos baseadas em (SOUZA et al., 2019):

- bigram
- trigram
- *Token* sem vogais
- Tamanho do *token*
- Quantidade de vogais
- Número de letras do *token*

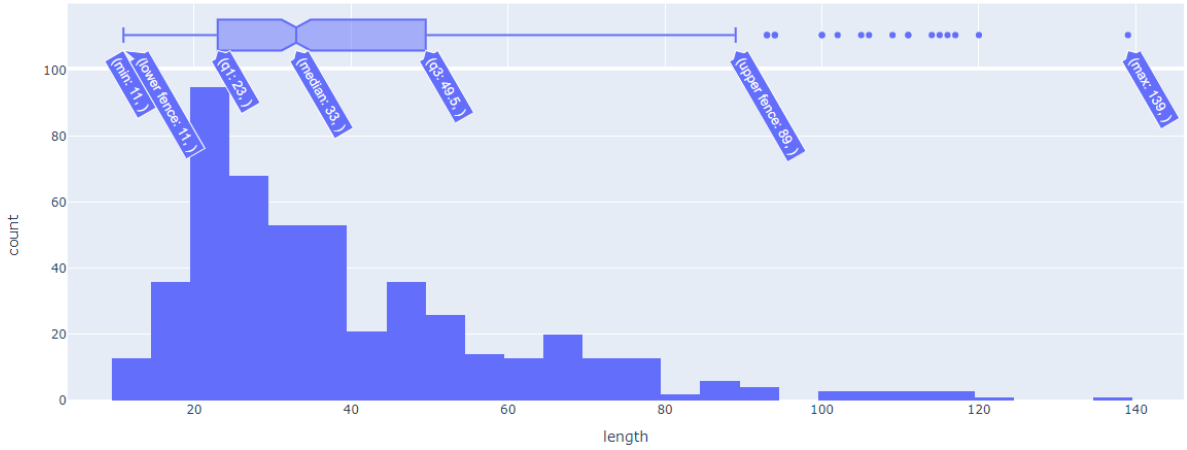


Figura 8 – Estatística descritiva das amostras

- Todos os caracteres são dígitos
- Número máximo de vogais consecutivas
- Número máximo de consoantes consecutivas
- *Token* possui mais de duas consoantes consecutivas

4.5.2 Entrada dos dados das RNNs

A vetorização de texto é a conversão de textos em uma forma numérica. Isso foi feito usando um dicionário, uma maneira simples de representar palavras individuais como vetores numéricos. Cada palavra é representada por um vetor cujas dimensões correspondem às palavras únicas no dicionário. Após isso, foi realizado o preenchimento das sequências para ajustar o comprimento das sequências, para que todas tenham o mesmo comprimento. De acordo com a Figura 8, o tamanho máximo escolhido de 120, por ser um valor próximo a maior sequência do conjunto de dados. As Tabelas 6 e 7 mostram a representação das amostras antes e depois deste processo.

Tabela 6 – Representação dos dados antes da vetorização

	tokens	tag
1	uso topico 1. syovea 0.5...	B-VIA I-VIA O B-MED B-DOS...

Tabela 7 – Representação dos dados após a vetorização

	tokens	tag
1	544, 183, 154, 450 287...	6, 3, 5, 1, 2...

4.5.3 Entrada dos dados no BERTimbau

A primeira etapa no BERT é executar o pré-processamento do *token* no *WordPiece* (WU YONGHUI, 2016), isso divide palavras longas em subpalavras menores com base no vocabulário, onde "##" indica que a subpalavra faz parte de uma palavra maior usando o tokenizador da biblioteca *Transformers*. Ao processar o texto, o tokenizador verifica se cada palavra está no vocabulário. Nesse caso, a palavra é mantida como um *token* exclusivo. Caso contrário, o *token* divide a palavra em subpalavras e verifica se essas subpalavras estão no vocabulário. As subpalavras presentes no vocabulário são mantidas como *tokens* individuais, enquanto as subpalavras que não estão presentes são divididas em subpalavras menores até que todas sejam incluídas no vocabulário. Além disso, dois rótulos são adicionados: [CLS] é usado como um indicador para marcar o início da sequência e [SEP] é usado como um separador para marcar o final da sequência. Depois disso, o redimensionamento da sequência é realizado, pois o *WordPiece* desajusta o tamanho das sequências, ele é reajustado adicionando a tag [PAD] para o preenchimento de sequências para que todas tenham o mesmo comprimento.

Tabela 8 – Representação dos dados antes do WordPiece

	tokens
1	'uso', 'topico', '1.', 'minoxidil', '5'...

Tabela 9 – Representação dos dados após o WordPiece

	tokens
1	'[CLS]', 'uso', 'top', '##ico', '1', '.', 'min', '##ox', '##idi', '##l', '5'...

4.6 Hiperparâmetros dos modelos

Na abordagem CRF, foram selecionados os parâmetros (o algoritmo *Passive Aggressive (PA)*, c = 1 e pa_type = 1), além de alguns atributos (*token* sem vogais, quantidade de vogais, número de letras do *token*, todos os caracteres são dígitos, número máximo de vogais consecutivas, número máximo de consoantes consecutivas, *token* possui mais de duas consoantes consecutivas) implementadas por (SOUZA et al., 2019) que apresentaram resultados satisfatórios em seu trabalho. No caso das RNNs, seguimos a arquitetura de (DANDALA BHARATH; JOOPUDI, 2019) com decaimento de peso de 0,4, taxa de aprendizado de 0,02, regularização de 1e-7 e quantidade de camadas ocultas de 150 e utilizando um máximo de 20 épocas para treinamento. Já no modelo BERT, adotamos alguns hiperparâmetros indicados por (SCHNEIDER, 2020) incluindo o uso do otimizador AdamW, fator de decaimento de 0,01, tamanho do lote de 4, taxa de aprendizado de 3e-5, 10 épocas treinamento e passos de aquecimento como 0,1.

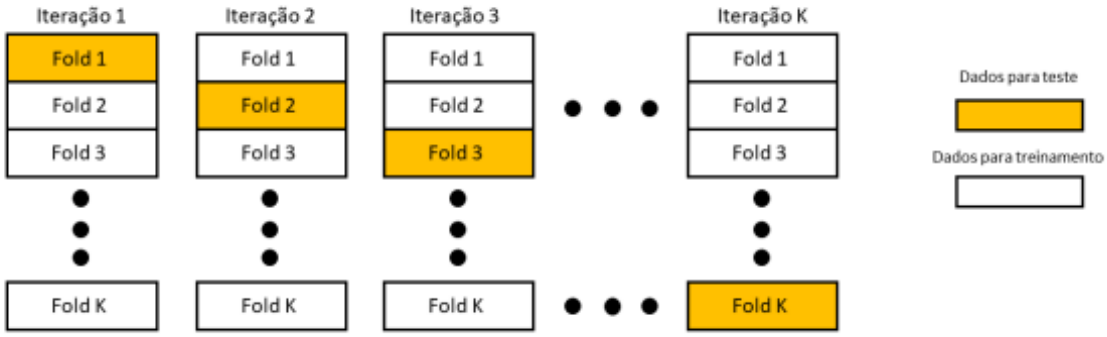


Figura 9 – Processo da validação cruzada k-fold (FARIAS THIAGO S; ROSSI,).

4.7 Treinamento e Métricas

Foi usada a validação cruzada k-fold, em inglês *K-fold cross-validation* para a validação do modelo, essa é uma técnica usada para avaliar a eficácia do modelo e estimar sua capacidade de generalizar para dados não vistos. Isso reduz os problemas de superajuste e permite estimativas mais precisas do desempenho do modelo. O conjunto de dados é dividido em K subconjuntos (ou *folds*) de tamanho igual e o modelo é treinado K vezes, de modo que cada iteração se ajuste a uma combinação diferente de *folds* de treinamento e seja testada em diferentes *folds*. O resultado final é obtido pela média da métrica de avaliação para K iterações. O valor de 10 foi escolhido para K por ser o valor mais utilizado na literatura (KARAPETIAN, 2023), (DARJI HARSHIL; MITROVIĆ, 2023), (PATIL NITA; PATIL, 2020), (BÁEZ, 2022). A Figura 9 mostra o processo desta técnica.

Ao desenvolver um modelo NER, é necessária uma etapa de avaliação para entender a qualidade do modelo. Esta avaliação é baseada em diferentes métricas que demonstram desempenho do modelo. As métricas mais utilizadas para avaliar modelos são *precisão*, *revocação* e *medida F1*. A base dessas métricas são baseadas nos conceitos de Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo. A tabela 10 ilustra isso (considere o valor 1 como uma previsão positiva).

Tabela 10 – Exemplos de Verdadeiro/Falso Positivo e Negativo

Classe prevista	Classe real	Tipo	Descrição
1	1	Verdadeiro positivo	Previu positivo corretamente
0	0	Verdadeiro negativo	Previu negativo corretamente
1	0	Falso positivo	Previu positivo incorretamente
0	1	Falso negativo	Previu negativo incorretamente

A *precisão* é a proporção de verdadeiros positivos sobre a soma de falsos positivos e verdadeiros negativos. É uma métrica útil e mostra que, dentre aqueles previstos como positivos, quão precisa foi a previsão. De todas as classificações de classe positivo que

o modelo realizou, quantas estão realmente corretas. A precisão é descrita pela seguinte equação:

$$precisão = \frac{VP}{VP + FP}, \tag{4.1}$$

onde VP e FP representam verdadeiro positivo e falso positivo, respectivamente.

O *revocação* é a proporção de resultados previstos corretamente para todas as previsões. É o número de verdadeiros positivos dividido pelo número de falsos negativos, entre as classificações de classe positivo como valor esperado, quantas estão corretas. O revocação é descrito pela seguinte equação:

$$revocação = \frac{VP}{VP + FN}, \tag{4.2}$$

onde VP e FN representam verdadeiro positivo e falso negativo, respectivamente.

medida F1 é a média harmônica entre a precisão e a revocação, que é definida como:

$$medidaF1 = \frac{2 * precisão * revocação}{precisão + revocação} \tag{4.3}$$

5 Resultados

Neste capítulo, são apresentados os resultados obtidos pelos quatro classificadores discutidos anteriormente no Capítulo 2.

Os experimentos foram realizados com os classificadores CRF, as RNNs com as três abordagens de incorporação de palavras (Word2Vec *Skip-gram*, GloVe e as representações do BERT) e o BERTimbau. Na Tabela 11 são apresentados os resultados obtidos com base nas métricas precisão, revocação e medida F1, considerando todo o conjunto de dados usando a técnica de validação cruzada k-fold com o tamanho de 10 *folds*.

Tabela 11 – Resultados obtidos com diferentes metodologias de NER

Modelo	Precisão	Revocação	Medida F1
CRF	0,92	0,91	0,91
LSTM + Word2Vec	0,95	0,95	0,95
LSTM + GloVe	0,95	0,95	0,95
LSTM + Representações BERT	0,94	0,95	0,94
BiLSTM + Word2Vec	0,95	0,95	0,95
BiLSTM + GloVe	0,96	0,95	0,95
BiLSTM + Representações BERT	0,95	0,95	0,95
BERTimbau	0,92	0,93	0,93

Em NER no domínio de texto clínico, valores maiores que 90% para as medidas de precisão, revocação e medida F1 podem ser considerados excelentes. De fato, todos os modelos analisados tiveram ótimos desempenhos, com pontuações acima de 90% em todos os indicadores. Pode-se observar que os cenários que usam RNNs com representações têm melhor desempenho em comparação com outras abordagens, pois foram projetadas especificamente para lidar com dados sequenciais e com a adição da incorporação de palavras que capturam as relações semânticas entre as palavras, produzem excelentes resultados. Considerando os resultados das RNNs, o uso da LSTM e BiLSTM com a incorporação de palavras GloVe e Word2Vec tiveram o melhor desempenho com pontuações de 95% em revocação e medida F1. O BERTimbau teve desempenho inferior as RNNs que usam representações. Já o CRF, embora tenha obtido um resultado mais baixo do que outros modelos, é importante considerar que é uma abordagem mais simples e antiga. Ainda assim, o CRF oferece a vantagem de ser menos custoso.

Na Tabela 11 podemos visualizar os resultados dos modelos com o uso de incorporação de palavras. Foi constatado que as três apresentaram resultados promissores em prescrições médicas manipuladas. A análise mostra que as diferenças de desempenho entre elas são relativamente pequenas, com resultados muito próximos entre si. Em particular, o uso da *embedding* GloVe se destaca por apresentar resultados ligeiramente superiores aos outros métodos, podendo indicar que sua representação semântica e qualidade na captura

de relacionamentos de palavras são particularmente relevantes para o contexto particular de prescrições médicas. No entanto, é importante considerar que as diferenças de desempenho entre as combinações são mínimas, sugerindo que todas são opções viáveis para a tarefa de NER em prescrições médicas.

Nas Tabelas 12 e 13 são apresentados os resultados detalhados por entidade com base nas métricas precisão, revocação e medida F1, considerando todas as abordagens dos modelos. Para facilitar o entendimento, decidimos agrupar as entidades que possuem tag I, visto que é uma continuação da tag B. É nítido observar que as entidades com os melhores resultados são VIA devido a padronização dos tokens e DOS por ser a entidade com o maior número de amostras presentes no conjunto de dados. No modelo CRF, as entidades QTD e DUR obtiveram o pior resultado entre todas as entidades avaliadas, por serem as duas entidades com o menor número de amostras, ocasionando assim em poucos dados na validação com a validação cruzada k-fold dificultando o aprendizado neste modelo.

Detalhando as entidades, VIA é a entidade com a melhor avaliação alcançando 0,99% nas RNNs na maioria das métricas. MED, é a entidade que possui a maior diversidade de dados, apresenta resultados semelhantes na maioria dos modelos, atingindo 93% em sete modelos, com exceção do CRF. A entidade DOS, como mostra na Figura 7 é a entidade com a maior quantidade de amostras no conjunto de dados, apresenta resultados excelentes, alcançando mais de 98% nas RNNs em todas as métricas. FREQ apresenta bons resultados em todos os modelos, chegando a 92% na maior parte das métricas e 86% em revocação no CRF. QTD e DUR obtiveram o pior resultado, no modelo CRF, mas nos outros sete modelos apresentaram resultados excelentes com pontuação entre 90% e 95%, evidenciando assim, que a utilização de técnicas mais avançadas como as RNNs e incorporação de palavras são mais eficazes no reconhecimento de entidades nomeadas, enquanto no BERTimbau as duas entidades alcançaram 100% em todas as métricas.

Nos modelos de RNNs, todas as entidades obtiveram excelentes resultados acima de 90% ou mais em todas as métricas. Para o modelo de BiLSTM + GloVe que obteve o melhor desempenho geral, as entidades VIA e DOS alcançaram mais de 98% para a maioria das métricas avaliadas, exceto para a precisão da entidade DOS, enquanto a entidade DUR obteve o pior desempenho. Além disso, verificou-se que o modelo BERTimbau, embora superando o CRF, registrou apenas uma pequena diferença de 1-2 décimos nas métricas de avaliação. Essa pequena vantagem do BERTimbau sobre o CRF pode ser significativa, pois a capacidade de generalização das entidades nomeadas em prescrições médicas manipuladas no BERTimbau apresenta um desempenho melhor. Essa capacidade de generalizar é especialmente valiosa em ambientes clínicos, onde a variedade e a complexidade dos termos podem ser altas, portanto, é um sinal positivo para outras tarefas relacionadas a NLP na área da saúde, onde a identificação precisa de informações-chave é essencial.

Tabela 12 – Resultados obtidos de acordo com cada classe individualmente.

Entidade	**Precisão**	**Revocação**	**Medida F1**
CRF			
VIA	0,90	0,90	0,90
MED	0,82	0,84	0,81
DOS	0,93	0,95	0,94
QTD	0,55	0,55	0,55
FREQ	0,90	0,86	0,87
DUR	0,50	0,50	0,50
LSTM + Word2Vec			
VIA	0,99	0,99	0,99
MED	0,93	0,93	0,93
DOS	0,98	0,98	0,98
QTD	0,92	0,92	0,92
FREQ	0,93	0,91	0,92
DUR	0,92	0,94	0,93
LSTM + GloVe			
VIA	0,99	0,99	0,99
MED	0,93	0,93	0,93
DOS	0,98	0,98	0,98
QTD	0,93	0,91	0,92
FREQ	0,93	0,92	0,92
DUR	0,92	0,93	0,92
LSTM + Representações BERT			
VIA	0,99	0,99	0,99
MED	0,93	0,92	0,92
DOS	0,98	0,98	0,98
QTD	0,93	0,91	0,92
FREQ	0,93	0,91	0,92
DUR	0,91	0,91	0,91

Tabela 13 – Resultados obtidos de acordo com cada classe individualmente pt2.

Entidade	**Precisão**	**Revocação**	**Medida F1**
BiLSTM + Word2Vec			
VIA	0,99	0,98	0,99
MED	0,93	0,93	0,93
DOS	0,98	0,99	0,99
QTD	0,95	0,95	0,95
FREQ	0,92	0,93	0,92
DUR	0,93	0,93	0,93
BiLSTM + GloVe			
VIA	0,99	0,99	0,99
MED	0,93	0,94	0,93
DOS	0,98	0,99	0,99
QTD	0,95	0,94	0,94
FREQ	0,93	0,93	0,93
DUR	0,92	0,93	0,92
BiLSTM + Representações BERT			
VIA	0,99	0,99	0,99
MED	0,93	0,92	0,92
DOS	0,98	0,98	0,98
QTD	0,95	0,92	0,94
FREQ	0,93	0,92	0,92
DUR	0,92	0,93	0,92
BERTimbau			
VIA	0,98	0,95	0,96
MED	0,89	0,90	0,90
DOS	0,98	0,99	0,98
QTD	1,00	1,00	1,00
FREQ	0,87	0,95	0,91
DUR	1,00	1,00	1,00

6 Conclusão

O reconhecimento de entidades nomeadas em prescrições médicas é uma área de grande interesse e progresso no campo da inteligência artificial e processamento de linguagem natural na área da saúde. O reconhecimento de entidade nomeada (NER) é uma tarefa fundamental na extração de informações relevantes do texto, permitindo a identificação de entidades específicas, como nome do medicamento, via de administração, frequência de uso, duração, entre outros, a partir de textos não estruturados.

O uso de NER em prescrições médicas apresenta desafios especiais, pois esses documentos podem conter texto não padronizado, abreviações, erros de digitação e contexto específico da área médica. No entanto, os avanços na tecnologia melhoraram a capacidade de reconhecer entidades com mais precisão e eficácia.

Este estudo investigou o uso de vários classificadores baseados em aprendizado de máquina, redes neurais recorrentes e aprendizado de transferência para reconhecer entidades nomeadas em prescrições médicas manipuladas. Para isso, foram desenvolvidos quatro modelos: CRF, LSTM e BiLSTM (incluindo as incorporação de palavras Word2Vec, GloVe e BERT), e uma versão pré-treinada baseada em português do BERT-base, mais precisamente o BERTimbau. Todos os modelos foram validados com o conjunto de dados adquirido e os resultados foram analisados com base nas medidas de precisão, revocação e pontuação F1.

Com base nos resultados experimentais obtidos, pode-se concluir que os modelo de RNNs apresentaram o melhor desempenho, sendo a BiLSTM com a utilização da GloVe, que obteve os resultados mais promissores. É importante ressaltar que há espaço para melhorias, como a exploração de outros hiperparâmetros, como alterar o número de camadas ocultas, taxa de aprendizado ou considerar utilizar outras arquiteturas presentes na literatura, como o BioBERT e BiLSTM+CRF, que podem potencialmente aprimorar ainda mais o desempenho. Como limitações, apontamos o tamanho do conjunto de dados, desbalanceamento das entidades, e de que este estudo foi conduzido sob a análise de apenas um conjunto de dados. Para a validação completa dos modelos desenvolvidos, seria interessante validar a gama de modelos com outros conjuntos de dados, visto que a escrita de receitas médicas podem variar.

Em conclusão, este estudo traz uma contribuição significativa para o avanço da identificação de entidades nomeadas em prescrições médicas. A partir dos resultados obtidos, é possível vislumbrar um potencial promissor para aplicações práticas no meio médico. No entanto, para alcançar a robustez necessária e lidar com os desafios do mundo real, pesquisas futuras devem se concentrar em abordar as limitações identificadas e explorar métodos, técnicas e novos conjuntos de dados para obter um reconhecimento de entidade mais preciso e eficiente em ambientes clínicos.

Referências

AKBIK, A. e. a. Flair: An easy-to-use framework for state-of-the-art nlp. In: *In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*. [S.l.: s.n.], 2019. p. 54–59. Citado na página 28.

BARONI MARCO; DINU, G. K. G. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *Association for Computational Linguistics*, v. 1, p. 238–247, 2014. Citado na página 17.

BIRKHEAD GUTHRIE S.; KLOMPAS, M. S. N. R. Uses of electronic health records for public health surveillance to advance public health. *Annual review of public health*, v. 36, p. 345–359, 2015. Citado na página 16.

BISTA, R.; RANJAN, A. A new approach to extract meaningful clinical information from medical notes. In: IEEE. *2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*. [S.l.], 2017. p. 1–8. Citado na página 13.

BÁEZ, P. e. a. Automatic extraction of nested entities in clinical referrals in spanish. *ACM Transactions on Computing for Healthcare (HEALTH)*, v. 3, n. 3, p. 1–22, 2022. Citado 4 vezes nas páginas 28, 29, 32 e 36.

CAMILO, C. O. et al. Uma metodologia para mineração de regras de associação usando ontologias para integração de dados estruturados e não-estruturados. Universidade Federal de Goiás, 2010. Citado na página 13.

CHAPMAN, A. B. e. a. Detecting adverse drug events with rapidly trained classification models. *Drug safety*, v. 42, p. 147–156, 2019. Citado 3 vezes nas páginas 27, 29 e 32.

CONSULTANT, H. I. T. Why unstructured data holds the key to intelligent healthcare systems [internet]. *HIT Consultant*, 2015. Citado na página 13.

DANDALA BHARATH; JOOPUDI, V. D. M. Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. *Drug safety*, v. 42, p. 135–146, 2019. Citado na página 35.

DARJI HARSHIL; MITROVIĆ, J. G. M. German bert model for legal named entity recognition. *arXiv*, 2023. Citado na página 36.

DAUMÉ HAL; LANGFORD, J. M. D. Search-based structured prediction. *Machine learning*, v. 75, p. 297–325, 2009. Citado na página 27.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>. Citado 3 vezes nas páginas 8, 24 e 25.

FARIAS THIAGO S; ROSSI, R. G. D. S. M. G. Estudo comparativo de arquiteturas de redes neurais em análise de sentimentos. Citado 2 vezes nas páginas 8 e 36.

GREFF, K. e. a. Lstm: A search space odyssey. *IEEE*, v. 28, n. 10, p. 2222–2232, 2016. Citado na página 23.

HARTMANN, N. e. a. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv*, 2017. Citado 2 vezes nas páginas 17 e 32.

HOCHREITER SEPP; SCHMIDHUBER, J. Long short-term memory. *Neural computation*, v. 9, n. 8, p. 1735–1780, 1997. Citado na página 22.

Hon S. Pak. *Unstructured Data in Healthcare*. 2018. Disponível em: <https://artificial-intelligence.healthcaretechoutlook.com/cxoinsights/unstructured-data-in-healthcare-nid-506.html>. Acesso em: 09 de março 2023. Citado na página 13.

KARAPETIAN, K. e. a. Supervised relation extraction between suicide-related entities and drugs: development and usability study of an annotated pubmed corpus. *Journal of medical internet research*, v. 25, 2023. Citado na página 36.

KHATTAK, F. K. e. a. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics*, v. 100, 2019. Citado na página 19.

KIM YOUNGJUN; MEYSTRE, S. M. Ensemble method–based extraction of medication and related information from clinical texts. *Journal of the American Medical Informatics Association*, v. 27, n. 1, p. 31–38, 2020. Citado 3 vezes nas páginas 27, 29 e 32.

KOROBOV, M. *sklearn-crfsuite*. 2015. <https://github.com/TeamHG-Memex/sklearn-crfsuite>. Acesso em: 10 de junho 2023. Citado na página 33.

LAFFERTY JOHN; MCCALLUM, A. P. F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. Citado na página 21.

LAFFERTY JOHN; MCCALLUM, A. P. F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. Citado na página 27.

LLC, G. *TensorFlow*. 2015. <https://github.com/tensorflow/tensorflow>. Acesso em: 15 de maio 2023. Citado na página 33.

LOPES FáBIO; TEIXEIRA, C. O. H. G. Contributions to clinical named entity recognition in portuguese. In: *In: Proceedings of the 18th BioNLP Workshop and Shared Task*. [S.l.: s.n.], 2019. p. 223–233. Citado na página 28.

MEYSTRE, S. M. e. a. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, v. 17, n. 01, p. 128–144, 2008. Citado na página 16.

MIKOLOV, T. e. a. Efficient estimation of word representations in vector space. *arXiv*, 2013. Citado 5 vezes nas páginas 8, 17, 18, 19 e 20.

MOTA CRISTINA; SANTOS, D. R. E. Avaliação de reconhecimento de entidades mencionadas: princípio de arem. p. 161–175, 2007. Citado na página 13.

NADEAU DAVID; SEKINE, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, v. 30, n. 1, p. 3–26, 2007. Citado 2 vezes nas páginas 13 e 16.

OLAH, C. *Neural Networks, Types, and Functional Programming*. 2015. Disponível em: <https://colah.github.io/posts/2015-09-NN-Types-FP/>. Acesso em: 19 de julho 2023. Citado 2 vezes nas páginas 8 e 24.

OLAH, C. *Understanding LSTM Networks*. 2015. Disponível em: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Acesso em: 18 de julho 2023. Citado 3 vezes nas páginas 8, 22 e 23.

OLIVEIRA, L. E. S. e. e. a. Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, v. 13, n. 1, p. 13, 2022. Citado na página 28.

PATIL NITA; PATIL, A. P. B. V. Named entity recognition using conditional random fields. *Procedia Computer Science*, v. 167, p. 1181–1188, 2020. Citado na página 36.

PENNINGTON JEFFREY; SOCHER, R. M. C. D. Glove: Global vectors for word representation. *EMNLP*, p. 1532–1543, 2014. Citado 2 vezes nas páginas 9 e 20.

RAMACHANDRAN R; ARUTCHELVAN, K. Named entity recognition on bio-medical literature documents using hybrid based approach. *Journal of Ambient Intelligence and Humanized Computing*, p. 1–10, 2021. Citado 2 vezes nas páginas 28 e 29.

SANTANA, L. M. Q. d. Aplicação de redes neurais recorrentes no reconhecimento automático da fala em ambientes com ruídos. 2017. Citado na página 23.

SCHNEIDER, E. T. R. e. a. Biobertpt-a portuguese neural language model for clinical named entity recognition. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. [S.l.: s.n.], 2020. p. 65–72. Citado 3 vezes nas páginas 27, 29 e 35.

SCHUSTER MIKE; PALIWAL, K. K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, v. 45, n. 11, p. 2673–2681, 1997. Citado na página 27.

SHICKEL, B. e. a. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, v. 22, n. 22, p. 1589–1604, 2017. Citado na página 13.

SOUSA, O. L. V. e. a. Ensemble of classifiers for multilabel clinical text categorization in portuguese. In: *In: International Conference on Intelligent Systems Design and Applications*. [S.l.]: Springer Nature Switzerland, 2022. p. 42–51. Citado na página 30.

SOUZA FáBIO; NOGUEIRA, R. L. R. Portuguese named entity recognition using bert-crf. *arXiv*, 2019. Citado na página 17.

SOUZA FáBIO; NOGUEIRA, R. L. R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. [S.l.: s.n.], 2020. Citado na página 33.

SOUZA, J. V. de et al. Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups. In: *Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2019. p. 318–323. ISSN 2763-8952. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/6269>. Citado 2 vezes nas páginas 33 e 35.

TAO, C.; FILANNINO, M.; UZUNER, Ö. Fable: A semi-supervised prescription information extraction system. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *AMIA Annual Symposium proceedings*. [S.l.], 2018. v. 2018, p. 1534. Citado na página 13.

TAO CARSON; FILANNINO, M. U. Prescription extraction using crfs and word embeddings. *Journal of biomedical informatics*, v. 72, p. 60–66, 2017. Citado na página 13.

WANG, S. e. a. Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. *International Journal of Electrical Power Energy Systems*, v. 109, p. 470–479, 2019. Citado na página 23.

WU YONGHUI, e. a. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv*, 2016. Citado na página 35.

XU, H. e. a. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, v. 17, n. 1, p. 19–29, 2010. Citado na página 27.

ZHOU, M. e. a. Ensemble transfer learning on augmented domain resources for oncological named entity recognition in chinese clinical records. *IEEE Access*, 2023. Citado na página 32.

ÇETINDAĞ CAN; YAZICIOĞLU, B. K. A. Named-entity recognition in turkish legal texts. *Natural Language Engineering*, v. 29, n. 3, p. 615–642, 2023. Citado na página 32.

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA "JOSÉ ALBANO DE MACEDO"

Identificação do Tipo de Documento

( ) Tese

( ) Dissertação

(X) Monografia

( ) Artigo

Eu, **David Pereira da Silva**, autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação "**Reconhecimento de Entidades Nomeadas em Receitas Médicas Manipuladas"** de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI 14 de Agosto de 2023.

David Pereira da Silva

Assinatura