



UNIVERSIDADE FEDERAL DO PIAUÍ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ORRANA LHAYNHER VELOSO DE SOUSA

ROTULAÇÃO E RECONHECIMENTO DE ENTIDADES CLÍNICAS EM PORTUGUÊS
ATRAVÉS DE MODELOS DE APRENDIZADO PROFUNDO

PICOS
ABRIL DE 2023

ORRANA LHAYNHER VELOSO DE SOUSA

ROTULAÇÃO E RECONHECIMENTO DE ENTIDADES CLÍNICAS EM PORTUGUÊS
ATRAVÉS DE MODELOS DE APRENDIZADO PROFUNDO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da Universidade Federal do Piauí, como um dos pré-requisitos para obtenção do título de Mestre em Engenharia Elétrica. Área de Concentração: Sistemas de Energia Elétrica

Orientadora: Prof.^a Dr.^a Deborah Maria Vieira Magalhães

Co-Orientador: Prof. Dr. Victor Eulalio Sousa Campelo

PICOS

ABRIL DE 2023

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Carlos Castello Branco
Divisão de Representação da Informação

S725r Sousa, Orrana Lhaynher Veloso de.
Rotulação e reconhecimento de entidades clínicas em português
através de modelos de aprendizado profundo / Orrana Lhaynher
Veloso de Sousa. -- 2023.
66 f.

Dissertação (Mestrado) – Universidade Federal do Piauí,
Programa de Pós-Graduação em Engenharia Elétrica, Picos, 2023.
“Orientadora: Prof.^a Dr.^a Deborah Maria Vieira Magalhães”.
“Co-Orientador: Prof. Dr. Victor Eulalio Sousa Campelo

1. Ensemble. 2. Dados clínicos. 3. Reconhecimento de Entidade
Nomeada. 4. Aprendizado Profundo. 5. BERTimbau. 6. Ajuste Fino.
I. Vieira, Deborah Maria Magalhães. II. Campelo, Victor Eulalio
Sousa. III. Título.

CDD 621.3

Bibliotecária: Francisca das Chagas Dias Leite – CRB3/1004



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PIAUÍ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELETRICA
Campus Universitário Ministro Petrônio Portela, Bairro Ininga, Teresina Piauí CEP 64049-550
Telefone (86)3237 1658; Email: ppgee_engeletrica@ufpi.edu.br

ATA DA DEFESA DE DISSERTAÇÃO DE MESTRADO Nº _____

DISSERTAÇÃO PARA OBTENÇÃO DO TÍTULO DE MESTRE EM ENGENHARIA ELÉTRICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA
ÁREA DE CONCENTRAÇÃO: SISTEMAS DE ENERGIA ELÉTRICA
LINHA DE PESQUISA: AUTOMAÇÃO E CONTROLE DE SISTEMAS

No dia 05 de abril de 2023 às 14:00h reuniu-se na sala virtual do Google Meet (<https://meet.google.com/iep-asoo-boz>) a banca examinadora composta pelos pesquisadores indicados a seguir, para examinar a dissertação de mestrado da candidata **Orrana Lhaynher Veloso de Sousa** intitulada: **Rotulação e Reconhecimento de Entidades Clínicas em Português Através de Modelos de Aprendizado Profundo**. O trabalho foi orientado pela Profa. Dra. **Deborah Maria Vieira Magalhães**. Após a apresentação, o candidato foi arguido pelos examinadores que, em seguida à manifestação dos presentes, consideraram o trabalho de pesquisa: (X) Aprovado. () Aprovado com restrições. Revisor indicado para verificação: _____ . () Reprovado.

Nada mais havendo a tratar, a sessão foi encerrada às 16:00, dela sendo lavrada a presente ata, que segue assinada pela Banca Examinadora e pelo Candidato.

O candidato está ciente que a concessão do referido título está condicionada à: (a) satisfação dos requisitos solicitados pela Banca Examinadora; (b) entrega da dissertação em conformidade com as normas exigidas pela UFPI; (c) atendimento ao requisito de publicação estabelecido nas normas do Programa; e (d) entrega da documentação necessária para elaboração do Diploma. A Banca Examinadora determina um **prazo máximo de 30 dias**, considerando os prazos máximos definidos no Regulamento Geral do Programa, para o cumprimento dos requisitos (desconsiderar caso reprovado), sob pena de, não o fazendo, ser desvinculado do Programa sem o Título de Mestre.

Profa. Dra. Deborah Maria Vieira Magalhães –
(UFPI/CSHNB)

(Orientadora)

Deborah Maria Vieira Magalhães

Prof. Dr. Victor Eulálio Sousa Campelo –
(UFPI/DME)

(co-orientador)

Dr. Victor Eulálio Sousa Campelo
Cm: 5559

Prof. Dr. Romuere Rodrigues Veloso e Silva –
(UFPI/CSHNB)

(Avaliador Interno)

Romuere Rodrigues Veloso e Silva

Prof. Dr. Rafael Torres Anchiêta – (IFPI)
(Avaliador Externo)

Rafael Torres Anchiêta

Nome: Orrana Lhaynher Veloso de Sousa
(Assinatura da Candidata)

Orrana Lhaynher Veloso de Sousa

Reservado à Coordenação

DECLARAÇÃO PARA A OBTENÇÃO DO TÍTULO DE MESTRE

A Coordenação do Programa declara que foram cumpridos todos os requisitos exigidos pelo Programa de Pós-Graduação para a obtenção do título de Mestre.

Teresina, ____ de _____ de 20 ____.

Carimbo e Assinatura do(a) Coordenador(a) do Programa


ORRANA LHAYNHER VELOSO DE SOUSA

ROTULAÇÃO E RECONHECIMENTO DE ENTIDADES CLÍNICAS EM PORTUGUÊS
ATRAVÉS DE MODELOS DE APRENDIZADO PROFUNDO


Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) da Universidade Federal do Piauí, como um dos pré-requisitos para obtenção do título de Mestre em Engenharia Elétrica. Área de Concentração: Sistemas de Energia Elétrica

Trabalho Aprovado em: 05/04/2023


BANCA EXAMINADORA

Documento assinado digitalmente
 DEBORAH MARIA VIEIRA MAGALHAES
Data: 05/04/2023 17:38:49-0300
Verifique em <https://validar.iti.gov.br>


Prof.^a Dr.^a Deborah Maria Vieira
Magalhães (Orientadora)
Universidade Federal do Piauí (UFPI)

Documento assinado digitalmente
 VICTOR EULALIO SOUSA CAMPELO
Data: 10/04/2023 10:56:05-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Victor Eulalio Sousa
Campelo (Co-Orientador)
Universidade Federal do Piauí (UFPI)

Documento assinado digitalmente
 ROMUERE RODRIGUES VELOSO E SILVA
Data: 05/04/2023 20:30:13-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Romuere Rodrigues Veloso e Silva
Universidade Federal do Piauí (UFPI)

Documento assinado digitalmente
 RAFAEL TORRES ANCHIETA
Data: 05/04/2023 18:26:12-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Rafael Torres Anchieta
Instituto Federal do Piauí (IFPI)

Dedico este trabalho à minha família e a todos que estiveram comigo nessa interessante jornada.

AGRADECIMENTOS

Agradeço a Deus, que esteve comigo desde o primeiro dia e é a Ele que dedico a concretização deste trabalho.

Agradeço aos meus pais por sempre me incentivarem e acreditarem na minha capacidade de superar os obstáculos que a vida me apresentou. Além disso, agradeço também por demonstrarem muito amor e apoio, o que serviu de alicerce para as minhas realizações.

Agradeço a minha irmã, por todo o apoio e ajuda nos momentos difíceis e por, mesmo com todas as incertezas da vida, nunca duvidar do meu potencial e sempre estar ali como forma de incentivo.

Agradeço aos meus amigos, que me incentivaram e me ajudaram ao longo desta caminhada. Em especial, agradeço a minha amiga Camila, que esteve comigo lado a lado na superação dos desafios encontrados no Mestrado e na vida.

Agradeço aos meus orientadores pelo apoio, correções, incentivo e suporte, sempre tornando essa jornada mais leve e divertida.

Agradeço também ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) e o corpo docente por toda a ajuda e educação dada.

Por fim, agradeço a todos que de forma direta ou indireta contribuíram para a realização deste sonho.

“Todas as conquistas começam com o simples ato de acreditar que elas são possíveis.”

(Autor Desconhecido)

RESUMO

Os sistemas de registro eletrônico de saúde (EHR) têm sido amplamente utilizados na prática médica, o que gerou um grande volume de dados não estruturados contendo abreviaturas, termos ambíguos e erros de digitação. Assim, a classificação automática de dados médicos em categorias clínicas informativas pode reduzir substancialmente o custo dessa tarefa. Além disso, tarefas de Processamento de Linguagem Natural (NLP) e Aprendizado de Máquina (ML), como o Reconhecimento de Entidade Nomeada (NER), têm sido usadas para processar esses dados. Em receitas médicas, por exemplo, é possível a extração de informações úteis para a farmacovigilância e o desenvolvimento de sistemas de apoio a tomada de decisão. Nesse contexto, este trabalho emprega uma metodologia que engloba desde a construção da base de dados até o processamento dos textos clínicos em português. Essa metodologia é dividida em duas etapas. Na etapa inicial, é investigada a utilização de uma *ensemble* de classificação para categorizar textos clínicos nas classes receitas, notas clínicas e solicitações de exames. Para isso, utilizamos diferentes combinações de métodos de vetorização para representar o texto. Em uma das combinações, analisamos o uso do *framework* Snorkel para supervisão fraca. Em seguida, a *ensemble* formada pelos algoritmos Máquina de Vetores de Suporte, Floresta Aleatória e Perceptron Multicamadas realiza a classificação. Na segunda etapa, técnicas de ML e aprendizado profundo (DL) são avaliadas para a extração de entidade clínicas nomeadas de receitas médicas. Cinco combinações de métodos de extração de características com classificadores foram avaliadas: características customizadas com os modelos Perceptron, Multinomial Naive Bayes e Campos Aleatórios Condicionais, as *embeddings* Glove com a rede neural BiLSTM, e uma versão com ajuste fino do BERTimbau. Os resultados alcançados com esta metodologia foram promissores, atingindo uma precisão de 1,00, kappa de 0,99 e F1-score de 1,00 na etapa de classificação, enquanto os modelos de DL obtiveram F1-score de 0,99 na extração das entidades. Assim, concluímos que nossa abordagem permite a classificação automática e precisa do conteúdo de textos clínicos, alcançando melhores resultados de categorização do que as abordagens únicas avaliadas; e o uso de *embeddings* de palavras e modelos de aprendizado profundo produzem melhores resultados no reconhecimento de entidades clínicas em português do que abordagens de ML.

Palavras-chave: Dados Clínicos. Classificação. Ensemble. Reconhecimento de Entidade Nomeada. Aprendizado Profundo. BERTimbau. Ajuste Fino.

ABSTRACT

The medical practice widely uses Electronic Health Record (EHR) systems, which has generated a large volume of unstructured data containing abbreviations, ambiguous terms, and typing errors. Thus, the automatic classification of medical data into informative clinical categories can substantially reduce the cost of this task. Furthermore, researchers use Natural Language Processing (NLP) and Machine Learning (ML) tasks such as Named Entity Recognition (NER) to process this data, for example, the useful information extraction for pharmacovigilance and the development of decision-making support systems from medical prescriptions. This work employs a methodology from database construction to processing clinical texts in Portuguese. This methodology has two stages. In the initial stage, we adopt a classification ensemble to categorize clinical texts in the following classes: prescriptions, clinical notes, and exam requests. To this end, we use different vectorization methods to represent the text, supported by the framework Snorkel for weak supervision. Then, the ensemble formed by the Support Vector Machine, Random Forest, and Multilayer Perceptron algorithms performs the classification. In the second stage, we evaluate ML and Deep learning (DL) techniques for extracting named clinical entities from medical prescriptions. We also evaluate five combinations of feature extraction methods with classifiers: custom features with the Perceptron, Multinomial Naive Bayes, and Conditional Random Fields models, the embeddings Glove with the BiLSTM neural network, and a fine-tuned version of BERTimbau. The results are promising, reaching a precision of 1.00, kappa of 0.99, and F1-score of 1.00 in the classification stage, while the DL models obtained an F1-score of 0.99 for entity extraction. Thus, we conclude that our approach allows the automatic and accurate classification of the content of clinical texts, achieving better categorization results than the single evaluated approaches; and the use of word embeddings and deep learning models produce better results for clinical entity recognition in Portuguese than ML approaches.

Keywords: Clinical Data. Classification. Ensemble. Named Entity Recognition. Deep Learning. BERTimbau. Fine Tuning.

LISTA DE ILUSTRAÇÕES

Figura 1 – Arquitetura do <i>framework</i> Snorkel publicada no artigo (RATNER <i>et al.</i> , 2020).	24
Figura 2 – Arquiteturas dos algoritmos saco-de-palavras contínuo e <i>skip-gram</i> apresentadas no artigo (MIKOLOV; LE; SUTSKEVER, 2013).	28
Figura 3 – Metodologia para classificação de documentos clínicos.	44
Figura 4 – Distribuição de classes com o algoritmo PCA nas <i>embeddings</i> do BERTimbau.	46
Figura 5 – Metodologia de reconhecimento de entidades clínicas em receitas médicas. .	47

LISTA DE TABELAS

Tabela 1 – Exemplo de transformação de texto em unigramas baseado no trabalho (SCHON- LAU; GUENTHER; SUCHOLUTSKY, 2017).	26
Tabela 2 – Exemplo de transformação de texto em bigramas.	26
Tabela 3 – Descrição dos quatro valores usados para calcular as métricas.	32
Tabela 4 – Trabalhos da literatura que realizaram tarefa de classificação ou NER de documentos clínicos.	41
Tabela 5 – Exemplos de amostras pré-processadas por categoria no banco de dados. . .	44
Tabela 6 – Quantidade de amostras rotuladas por classe nos conjuntos de treino e teste.	45
Tabela 7 – Descrição das entidades clínicas do trabalho e exemplos de <i>tokens</i> por entidade.	48
Tabela 8 – Quantidade de <i>tokens</i> rotulados por entidade no conjunto de dados.	49
Tabela 9 – Parâmetros selecionados por modelo.	51
Tabela 10 – Resultados obtidos usando combinações de <i>embeddings</i> com o classificador SVM. Em negrito estão os melhores resultados.	54
Tabela 11 – Resultados obtidos com a predição.	54
Tabela 12 – Distribuição do número de amostras erroneamente classificadas na predição.	55
Tabela 13 – Resultados obtidos na extração de entidades clínicas em receitas médicas. .	56
Tabela 14 – Resultados obtidos segundo as classes da extração de entidades clínicas em receitas médicas.	58

LISTA DE ABREVIATURAS E SIGLAS

EHR	Registro Eletrônico de Saúde
NLP	Processamento de Linguagem Natural
ML	Aprendizado de Máquina
NER	Reconhecimento de Entidade Nomeada
NE	Entidade Nomeada
SVM	Máquina de Vetores de Suporte
RF	Floresta Aleatória
MLP	Perceptron Multicamadas
MNB	Naive Bayes Multinomial
BiLSTM	Memória Bidirecional de Longo-Curto Prazo
DL	Aprendizado Profundo
CRF	Campos Aleatórios Condicionais
EMR	Registro Eletrônico Médico
IE	Extração de Informações
TF-IDF	Frequência do Termo - Frequência Inversa do Documento
Glove	Vetor Global para Representação de Palavras
BERT	Representações de Codificador Bidirecional de Transformadores
BOW	Saco de Palavras
LR	Regressão Logística
CNN	Rede Neural Convolucional
DT	Árvore de Decisão
kNN	k-ésimo Vizinho mais Próximo
NB	Naive Bayes
RT	Árvore Aleatória
TF	Frequência do Termo
TO	Ocorrência do Termo
BO	Ocorrência Binária
SVC	Classificação de Vetores de Suporte

LDA	Alocação Latente de Dirichlet
SVR	Regressão de Vetor de Suporte
WE	<i>Embeddings</i> de Palavras
KF	Filtragem baseada em Palavras-chave
irAES	Eventos Adversos relacionados ao Sistema Imunológico
TC	Tomografia Computadorizada
RM	Ressonância Magnética
MOA	Mecanismo de Ação da Droga
DA	Doença de Alzheimer
CCI	Índice de Comorbidade de Charlson
MAE	Erro Médio Absoluto
LP	Rótulo <i>Powerset</i>
PCA	Análise de Componentes Principais
IAA	Concordância entre Anotadores
UF	Unidade Federal
CRM	Conselho Regional de Medicina
RTF	Formato <i>Rich Text</i>

SUMÁRIO

1	INTRODUÇÃO	16
1.1	Objetivo Geral	18
1.2	Objetivos Específicos	18
1.3	Contribuições do Trabalho	18
1.4	Publicações	19
<i>1.4.1</i>	<i>Artigos Publicados</i>	19
<i>1.4.2</i>	<i>Artigos Aceitos</i>	19
1.5	Estrutura do trabalho	19
2	REFERENCIAL TEÓRICO	21
2.1	Sistemas de Registro Eletrônico de Saúde	21
2.2	Processamento de Linguagem Natural e a Extração de Informações	21
<i>2.2.1</i>	<i>Reconhecimento de Entidades Nomeadas</i>	22
2.3	Supervisão Fraca	23
2.4	Pré-processamento de dados textuais	23
<i>2.4.1</i>	<i>Unitização e tokenização</i>	24
<i>2.4.2</i>	<i>Padronização e limpeza dos dados</i>	24
<i>2.4.3</i>	<i>Remoção de stopwords</i>	25
<i>2.4.4</i>	<i>Lematização</i>	25
2.5	Extração de Características	25
<i>2.5.1</i>	<i>N_grama</i>	25
<i>2.5.2</i>	<i>Frequência do Termo - Frequência Inversa do Documento</i>	26
<i>2.5.3</i>	<i>Word2Vec</i>	27
<i>2.5.4</i>	<i>Vetor Global para Representação de Palavras</i>	27
2.6	Classificação	28
<i>2.6.1</i>	<i>Máquina de Vetores de Suporte</i>	29
<i>2.6.2</i>	<i>Floresta Aleatória</i>	29
<i>2.6.3</i>	<i>Naive Bayes Multinomial</i>	29
<i>2.6.4</i>	<i>Campos Aleatórios Condicionais</i>	30
<i>2.6.5</i>	<i>Perceptron</i>	30
<i>2.6.6</i>	<i>Perceptron Multicamadas</i>	30

2.6.7	<i>Memória Bidirecional de Longo-Curto Prazo</i>	31
2.7	Representações de Codificador Bidirecional de Transformadores	31
2.8	Ensembles	31
2.9	Métricas de Validação	32
3	TRABALHOS RELACIONADOS	35
3.1	Considerações Finais	42
4	MÉTODO PROPOSTO	43
4.1	Construção do conjunto de dados e classificação de documentos clínicos	43
4.1.1	<i>Aquisição do conjunto de dados</i>	43
4.1.2	<i>Pré-processamento</i>	43
4.1.3	<i>Rotulação</i>	44
4.1.4	<i>Seleção de técnicas de vetorização</i>	45
4.1.5	<i>Ensemble e validação</i>	46
4.2	Reconhecimento de entidades nomeadas clínicas	47
4.2.1	<i>Rotulação para NER</i>	48
4.2.2	<i>Extração de características e abordagens NER</i>	49
5	RESULTADOS E DISCUSSÕES	53
5.1	Resultados da definição dos modelos de vetorização	53
5.2	Resultados da classificação	54
5.3	Reconhecimento das entidades nomeadas clínicas	55
6	CONCLUSÕES E TRABALHOS FUTUROS	59
	REFERÊNCIAS	60

1 INTRODUÇÃO

Os sistemas de registro eletrônico de saúde, do inglês *Electronic Health Record* (EHR), têm sido amplamente utilizados na prática médica. Eles geram um grande volume de dados que servem de insumo para a pesquisa clínica e desenvolvimento de sistemas de auxílio à tomada de decisão (WEI *et al.*, 2020). No tocante aos dados textuais em EHRs, eles são criados geralmente na forma de texto livre e são caracterizados por uma multiplicidade de expressões pelas quais médicos podem relatar uma mesma condição clínica ou procedimento (CABITZA *et al.*, 2019).

Tarefas de processamento de linguagem natural (*Natural Language Processing* - NLP) aliadas às técnicas de Aprendizado de Máquina (*Machine Learning* - ML) são amplamente usadas para processar dados textuais de EHRs (HENRY *et al.*, 2020) com diferentes fins. Por exemplo, a classificação automática dos tipos de documentos médicos, pode ajudar a identificar e organizar esses dados. Dessa forma, eles podem ser usados de forma mais eficiente no uso orientado à pesquisa e análise. Além disso, tal associação pode ser empregada na extração de informações clínicas, treinamento de sistemas de preenchimento automático hospitalar, gerenciamento de medicamentos prescritos com testes de interação medicamentosa, etc (CUI *et al.*, 2019).

Para a extração de informações de EHRs, o Reconhecimento de Entidade Nomeada (*Named Entity Recognition* - NER) é utilizado (HENRY *et al.*, 2020). Ele é uma tarefa da área de Extração de Informação que consiste em identificar e classificar tipos de elementos de informação textual, chamados de entidades nomeadas (*Named Entities* - NE), em categorias semânticas predefinidas (MARRERO *et al.*, 2013; LI *et al.*, 2020). O NER aplicado às receitas pode ser empregado para extrair conceitos e eventos médicos importantes, como nome dos medicamentos, suas dosagens, duração e frequência de administração, forma do medicamento, via e motivo da administração, e quaisquer eventos adversos associados aos medicamentos prescritos.

Apesar das diversas aplicações, processar dados clínicos ainda é uma tarefa desafiadora. Vários registros contêm abreviações, termos ambíguos e erros ortográficos, altos níveis de ruído, esparsidade, vocabulários médicos complexos, medidas e frases gramaticais incorretas (MUJTABA *et al.*, 2019). Assim, os métodos tradicionais de aprendizado de máquina podem não funcionar satisfatoriamente ao lidar com dados complexos, como documentos clínicos e o uso de uma única técnica nem sempre garante um alto nível de precisão (DONG *et al.*, 2020; HOSNI *et al.*, 2019).

Destacamos que a criação de dados de treinamento rotulados no domínio médico requer muito esforço, principalmente devido à: i) falta de corpora clínicos publicamente disponíveis por motivos de privacidade e ii) exigência de conhecimento médico para anotar textos clínicos com exatidão (WANG *et al.*, 2019). Há uma carência de trabalhos que utilizam dados clínicos em português para este fim. A maior parte dos estudos na literatura tem como base a língua inglesa devido ao número maior de iniciativas de disponibilização para desafios de processamento de linguagem natural, como o compartilhamento de conjuntos de dados como o Monitoramento Inteligente Multiparâmetro em Terapia Intensiva (MIMIC) (LEE *et al.*, 2011; JOHNSON *et al.*, 2016).

Nesse contexto, este trabalho desenvolve uma metodologia para processamento de textos clínicos em português dividida em duas etapas. Na etapa inicial, é investigada a utilização de uma *ensemble* de classificação para categorizar textos clínicos em prescrições, notas clínicas e solicitações de exames. Para isso, utilizamos diferentes combinações de *embeddings* para representar o texto. Em uma das combinações, analisamos o uso do *framework* Snorkel (RATNER *et al.*, 2020) para supervisão fraca. Em seguida, a *ensemble* formada pelos algoritmos Máquina de Vetores de Suporte (*Support Vector Machines* - SVM) (BURGES, 1998), Floresta Aleatória (*Random Forest* - RF) (SWAIN; HAUSKA, 1977) e Perceptron Multicamadas (*Multilayer Perceptron* - MLP) (HORNIK; STINCHCOMBE; WHITE, 1989) realiza a classificação. Na segunda etapa, técnicas de ML e aprendizado profundo (*Deep Learning* - DL) são avaliadas para a extração de entidade clínicas nomeadas de receitas médicas. Sete tipos de entidades nomeadas foram utilizadas: apresentação do medicamento, dosagem, frequência e quantidade de uso, nome do medicamento, duração do tratamento e via de administração. Cinco combinações de métodos de extração de características com classificadores foram avaliadas: características customizadas com os modelos Perceptron (FREUND; SCHAPIRE, 1998), Naive Bayes Multinomial (KIBRIYA *et al.*, 2005) e Campos Aleatórios Condicionais (*Conditional Random Fields* - CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001), as *embeddings* da Glove (PENNINGTON; SOCHER; MANNING, 2014) com a rede neural BiLSTM (SCHUSTER; PALIWAL, 1997), e uma versão com ajuste fino do BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020). Para a realização dessa metodologia de processamento de textos, uma base de documentos clínicos foi construído nesse trabalho.

1.1 Objetivo Geral

O objetivo deste trabalho é desenvolver uma metodologia que vai desde a aquisição de documentos clínicos até o reconhecimento de entidades nomeadas em receitas médicas.

1.2 Objetivos Específicos

Para que o objetivo geral seja plenamente alcançado, foram definidos os seguintes objetivos específicos:

- Realizar um levantamento do estado da arte sobre conjunto de dados textuais clínicos em português, assim como trabalhos que utilizem bases nessa língua;
- Construir uma base de dados de documentos clínicos em português nas classes receitas, notas clínicas e solicitações de exames;
- Avaliar o uso do *framework* de supervisão fraca Snorkel;
- Avaliar abordagens de aprendizado de máquina e aprendizado profundo para a classificação de textos clínicos, assim como a ensemble constituída por eles;
- Avaliar e comparar abordagens de aprendizado de máquina e aprendizado profundo para o reconhecimento de entidades clínicas nomeadas.

1.3 Contribuições do Trabalho

As contribuições desse trabalho se dividem em duas etapas: aquisição e rotulação de textos clínicos em português para fins de classificação e reconhecimento de entidades nomeadas em receitas médicas. Na primeira categoria, as contribuições foram:

1. A construção e disponibilização de uma base de dados de documentos clínicos em português rotulados nas classes: prescrições, notas clínicas e solicitações de exames¹;
2. A análise da utilização e, conseqüentemente, o impacto do *framework* Snorkel nos resultados com a criação de rótulos no conjunto de treinamento; e
3. A proposta de uma *ensemble* para a classificação de documentos clínicos em português nas classes de interesse;

Já na segunda categoria, as contribuições feitas foram:

1. A construção e disponibilização de uma base de dados de receitas médicas em português

¹ O código-fonte e a base de dados do nosso estudo está disponível publicamente em https://github.com/pavicutpi/ISDA_Clinical_Text.

rotulada para a tarefa de NER; e

2. A comparação de metodologias de NER clínico com o uso de modelos de ML e DL, em conjunto com a técnica de transferência de aprendizado.

1.4 Publicações

Durante o mestrado, os seguintes artigos foram produzidos:

1.4.1 Artigos Publicados

- Orrana Sousa, Deborah Magalhães, Pablo Vieira e Romuere Silva. Deep Learning in Image Analysis for COVID-19 Diagnosis: a Survey. **IEEE Latin America Transactions, 2021. (Qualis A4)** (SOUSA *et al.*, 2021).
- Pablo Vieira, Orrana Sousa, Deborah Magalhães, Ricardo Rabêlo e Romuere Silva. Detecting pulmonary diseases using deep features in X-ray images. **Pattern Recognition, 2021. (Qualis A1)** (VIEIRA *et al.*, 2021).
- Natural Language Processing for Clinical Data Classification. **ISys - Brazilian Journal of Information Systems, 2022. (Qualis B1)** (SOUSA *et al.*, 2022).

1.4.2 Artigos Aceitos

- Orrana Sousa, David Silva, Victor Campelo, Romuere Silva e Deborah Magalhães. Ensemble of Classifiers for Multilabel Clinical Text Categorization in Portuguese. **22nd International Conference on Intelligent Systems Design and Applications (ISDA), 2023. (Qualis B1)**.

1.5 Estrutura do trabalho

O restante deste trabalho está organizado da seguinte forma:

- Capítulo 2 - Referencial Teórico: os conceitos necessários para entendimento do problema e das técnicas empregadas na produção do trabalho desenvolvido;
- Capítulo 3 – Trabalhos Relacionados: apresentamos um levantamento do estado da arte das obras relacionadas ao tema proposto;
- Capítulo 4 – Método Proposto: metodologia proposta nesse trabalho, detalhando todas as etapas do processo;

- Capítulo 5 – Resultados e Discussões: os resultados obtidos com a metodologia deste trabalho são mostrados, seguidos de uma breve discussão sobre os mesmos;
- Capítulo 6 - Conclusões: as considerações finais do trabalho, abordando os tópicos principais. Ao final, apresentamos possíveis sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo, são discutidos conceitos e técnicas necessárias para a compreensão da metodologia proposta. Como seções, temos: Sistemas de Registro Eletrônico de Saúde 2.1; Processamento de Linguagem Natural e Extração de Informações 2.2; Supervisão Fraca 2.3; Pré-processamento de Dados Textuais 2.4; Extração de Características 2.5; Classificação 2.6; Representações de Codificador Bidirecional de Transformadores 2.7; *Ensembles* 2.8; e Métricas de Validação 2.9.

2.1 Sistemas de Registro Eletrônico de Saúde

O registro eletrônico de saúde (EHR) é geralmente definido como um repositório de informações mantidas eletronicamente sobre o estado e os cuidados de saúde dos pacientes, por exemplo, na forma de registros médicos eletrônicos (*Electronic Medical Record* – EMRs) (SHI *et al.*, 2020; TANG; MCDONALD, 2006). Um dos principais benefícios desses sistemas é a disponibilidade de grandes volumes de dados, que podem ser usados para facilitar a análise de dados e o aprendizado de máquina, por exemplo, para informar outros esforços de pesquisa médica, como a previsão de doenças (SHI *et al.*, 2020).

Os sistemas EHR são frequentemente citados por seu potencial de reduzir erros médicos por meio de suporte à decisão, como interações medicamentosas adversas, e também têm o potencial de fornecer outros benefícios, como reduzir custos com medicamentos e disponibilizar dados de histórico médico durante cuidado de emergência (CHAUDHRY *et al.*, 2006; SCHILLINGER *et al.*, 2003; STIELL *et al.*, 2003). Eles têm o potencial de transformar o sistema de saúde de uma indústria baseada principalmente em papel para uma que utiliza informações clínicas e outras informações para ajudar os provedores a oferecer atendimento de maior qualidade a seus pacientes (MENACHEMI; COLLUM, 2011).

2.2 Processamento de Linguagem Natural e a Extração de Informações

O processamento de linguagem natural (*Natural Language Processing* - NLP) surgiu na década de 1950 como a interseção da inteligência artificial e da linguística, com o objetivo de fazer com que os computadores executem tarefas úteis envolvendo a linguagem humana, como permitir a comunicação homem-máquina, melhorar a comunicação humano-humano ou simplesmente fazer processamento útil de texto ou fala (JURAFSKY, 2000). Dentre as subáreas

desenvolvidas na NLP, temos: tradução de linguagem natural, recuperação de informações, extração de informações, resumo de textos, resposta às perguntas, modelagem de tópicos, mineração de opinião, dentre outras.

Na extração de informações (*Information Extraction - IE*), o objetivo é tornar explícita a estrutura semântica do texto, já que textos são referidos como dados não estruturados, o que dificulta a busca ou análise das informações contidas neles. Assim, a IE é o processo de análise de textos e identificação de menções de entidades e relacionamentos semanticamente definidos dentro deles (GRISHMAN, 2015). Esses relacionamentos podem então ser registrados em um banco de dados para pesquisar relacionamentos específicos ou inferir informações adicionais dos fatos explicitamente declarados.

A IE é melhor empregada em aplicações onde o volume de dados textuais a serem estudados simplesmente sobrecarrega o leitor e as relações semânticas de interesse aparecem com alta frequência. Por exemplo, a literatura médica e biomédica está crescendo a uma taxa de mais de 500.000 artigos por ano, e hospitais e consultórios médicos geram grandes volumes de registros médicos eletrônicos a serem revisados em cada consulta ou internação do paciente (GRISHMAN, 2015).

2.2.1 Reconhecimento de Entidades Nomeadas

O reconhecimento de entidades nomeadas (*Named Entity Recognition - NER*) é uma tarefa da IE que visa reconhecer menções de designadores rígidos de texto pertencentes a tipos semânticos predefinidos (NADEAU; SEKINE, 2007). O NER não apenas atua como uma ferramenta autônoma para a IE, mas também desempenha um papel essencial em uma variedade de aplicações de NLP.

Como principal elemento do NER temos as entidades nomeadas (*Named Entity - NE*), termo usado pela primeira vez na sexta Conferência de Entendimento de Mensagens (MUC-6) (GRISHMAN; SUNDHEIM, 1996). As NE são palavras ou frases que são nomeadas ou categorizadas em um determinado tópico, como nomes de organizações, pessoas e locais em domínio geral, e nomes de genes, proteínas, drogas e doenças no domínio biomédico. Elas geralmente carregam informações importantes em uma frase que servem como alvos importantes para a maioria dos sistemas de processamento de linguagem (MOHIT, 2014).

Quanto às técnicas aplicadas no NER, existem quatro correntes principais: abordagens baseadas em regras, que não precisam de dados anotados, pois dependem de regras

artesanais; abordagens de aprendizado não supervisionado, que dependem de algoritmos não supervisionados sem exemplos de treinamento rotulados à mão; abordagens de aprendizado supervisionado baseadas em características, que dependem de algoritmos de aprendizado supervisionado com cuidadosa engenharia de características; e abordagens baseadas em aprendizado profundo, que descobrem automaticamente as representações necessárias para a classificação e/ou detecção da entrada bruta.

2.3 Supervisão Fraca

A supervisão fraca é uma abordagem simples e adaptável que utiliza conjuntos de treinamento fracamente rotulados criados de forma programática (WANG *et al.*, 2019). Assim, são criados métodos de aprendizado de máquina que aproveitam rótulos imperfeitos para treinamento e, portanto, não dependem de corpus de treinamento em grande escala rotulados manualmente.

Dentre os métodos de supervisão fraca, temos o Snorkel. Ele é um *framework* para criar dados de treinamento sem usar rotulagem manual. Ele permite que os usuários especifiquem genericamente várias fontes de supervisão programática fraca, como regras e padrões sobre texto, que podem variar em precisão e cobertura e ser arbitrariamente correlacionados (RATNER *et al.*, 2020).

O *pipeline* Snorkel segue três estágios principais, como apresentado na Figura 1. Primeiro, os usuários escrevem funções de rotulagem que podem utilizar várias fontes de supervisão fracas, como padrões, heurísticas, bases de conhecimento externas, etc. Essas funções de rotulagem coletam amostras de dados não rotulados e produzem um rótulo ou abstêm-se. Depois, o Snorkel aprende automaticamente um modelo generativo que estima a precisão das funções de rotulagem e então repondera e combina os rótulos para produzir um conjunto de rótulos de treinamento probabilístico, o que permite estimar suas precisões e correlações. Por fim, um modelo discriminativo final arbitrário é então treinado com esses rótulos e usado como um classificador final (BACH *et al.*, 2019).

2.4 Pré-processamento de dados textuais

O pré-processamento de texto é tradicionalmente uma etapa importante no NLP, recebendo uma entrada de texto bruto e retornando *tokens* "limpos". *Tokens* são conceituados

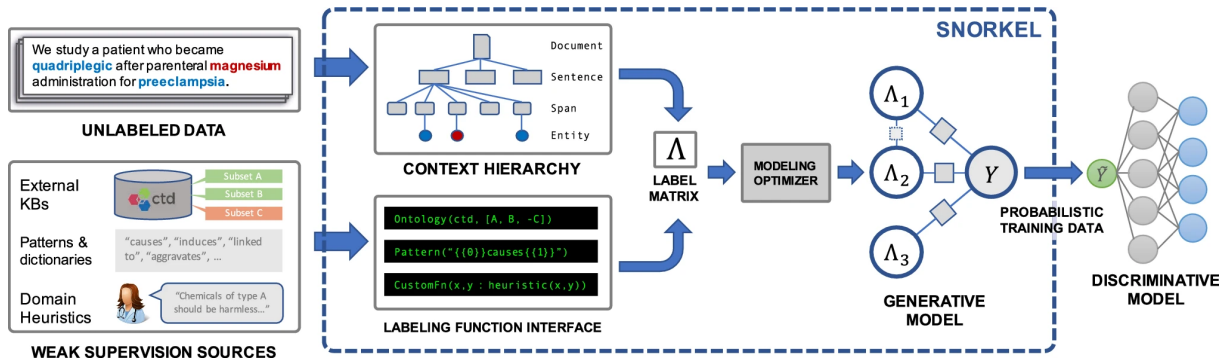


Figura 1 – Arquitetura do *framework* Snorkel publicada no artigo (RATNER *et al.*, 2020).

como palavras únicas ou grupos de palavras que são computados por sua frequência e servem como características da análise (ANANDARAJAN *et al.*, 2019). Assim, o pré-processamento inclui as seguintes etapas: (1) unitização e tokenização, (2) padronização e limpeza dos dados, (3) remoção de *stopwords* e (4) lematização.

2.4.1 Unitização e tokenização

Essa etapa consiste na escolha da unidade de texto a ser analisada e a separação do texto com base na unidade de análise. Esta unidade pode ser uma palavra, um agrupamento de palavras ou uma frase. Dessa forma, após selecionada a unidade de análise, o procedimento de tokenização é realizado, segmentando o texto nessas unidades. Geralmente, a segmentação é realizada considerando apenas caracteres alfabéticos ou alfanuméricos que são delimitados por caracteres não alfanuméricos (por exemplo, pontuação, espaço em branco, etc) (UYSAL; GUNAL, 2014).

2.4.2 Padronização e limpeza dos dados

A etapa de padronização e limpeza dos dados transforma os termos encontrados nos documentos de forma que possam ser analisados comparavelmente. Por exemplo, "caractere, caractere" e "Caractere" não devem ser considerados itens separados só porque um tem vírgula e o outro inicia com um C maiúsculo. A padronização e limpeza evita que essa possibilidade ocorra (ANANDARAJAN *et al.*, 2019). Assim, é inicialmente realizado a conversão do formato do texto para minúsculo. Depois, a remoção de números, pontuação, espaços extras, e caracteres especiais pode ser feita, levando sempre em consideração o domínio e a versão final esperada do texto após essa etapa.

2.4.3 *Remoção de stopwords*

Stopwords são palavras comumente encontradas em textos sem dependência de um tópico específico, por exemplo, conjunções, preposições, artigos, etc (UYSAL; GUNAL, 2014). Portanto, são consideradas palavras irrelevantes usadas com frequência, que não agregam valor à análise. Sendo assim, elas são removidas já que servem a um propósito gramatical, mas fornecem pouca informação em termos de conteúdo. Elas são específicas do idioma de estudo, por isso, algoritmos de remoção de *stopwords* são desenvolvidos baseados em listas de palavras de paradas da língua de interesse.

2.4.4 *Lematização*

A lematização envolve a redução de palavras até sua palavra raiz, incorporando informações sobre a classe gramatical do termo (YATSKO, 2011). Esse método combina palavras que contêm a mesma raiz em um único *token* para reduzir o número de *tokens* exclusivos no conjunto de análise, pois palavras com uma raiz comum geralmente compartilham um significado semelhante. Porém, há exceções para as raízes das palavras que compartilham o mesmo significado. Entretanto, a redução adicional na complexidade muitas vezes compensa a categorização incorreta de algumas palavras (UYSAL; GUNAL, 2014).

2.5 **Extração de Características**

A extração de características é uma das etapas mais importantes no processamento de textos, já que permite a geração e extração de representações numéricas do texto. Os dados são representados por um número fixo de características, que podem ser binárias, categóricas ou contínuas, criando assim uma abstração sobre o texto de interesse (WANG; SU; YU, 2020). Ela pode ser realizada através do levantamento de características ortográficas e linguísticas ou o uso de algoritmos estatísticos.

2.5.1 *N_grama*

O *N_grama* é modelo de linguagem usado para prever o próximo item em uma sequência textual, onde uma sequência contígua de n itens representa uma amostra de texto, com $1, 2, 3, \dots, n$ itens para esta representação (BROWN *et al.*, 1992). Em geral, *n_gramas* fornecem

Tabela 1 – Exemplo de transformação de texto em unigramas baseado no trabalho (SCHONLAU; GUENTHER; SUCHOLUTSKY, 2017).

Frase	Unigramas						
	Eu	Amo	Vezes	Céu	20	Praia	Lindo
Eu amo a praia	1	1	0	0	0	1	0
Hoje o céu está lindo	0	0	0	1	0	0	1
Ele repetiu a mesma coisa 20 vezes	0	0	1	0	1	0	0

Tabela 2 – Exemplo de transformação de texto em bigramas.

Frase	Bigramas						
	Eu amo	O céu	20 vezes	A praia	Está lindo	Ele repetiu	Mesma coisa
Eu amo a praia	1	0	0	1	0	0	0
Hoje o céu está lindo	0	1	0	0	1	0	0
Ele repetiu a mesma coisa 20 vezes	0	0	1	0	0	1	1

processamento independente de idioma, rastreamento de informações lexicais e contextuais e comportamento mais robusto na presença de diferentes tipos de erros textuais, já que os erros afetam apenas um número limitado de n -gramas (CAVNAR; TRENKLE *et al.*, 1994). As Tabelas 1 e 2 apresentam exemplos de representação através de unigramas e bigramas.

2.5.2 Frequência do Termo - Frequência Inversa do Documento

A medida estatística Frequência do Termo - Frequência Inversa do Documento (*Term Frequency - Inverse Document Frequency* - TF-IDF) é uma abordagem comumente usada para indicar a importância de uma palavra em um documento para uma coleção de documentos ou corpus (YUN-TAO; LING; YONG-CHENG, 2005). Ele captura a relevância entre palavras específicas, documentos de texto e categorias. Entretanto, essa abordagem possui suas próprias desvantagens, já que se o corpus de texto for grande, esse método pode gerar vetores de características com um grande número de dimensões, o que pode aumentar as chances de superajuste do modelo de classificação (DZISEVIČ; ŠEŠOK, 2019).

Esse método de ponderação de palavras no corpus funciona da seguinte forma: dada uma coleção de documentos D , uma palavra w e um documento individual $d \in D$, calculamos:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right), \quad (2.1)$$

onde $w_{i,j}$ é o peso do termo i no documento j , N é o número de documentos no corpus, $tf_{i,j}$ é a frequência de termo do termo i no documento j e df_i é a frequência de documento do termo i no corpus.

Assim, a ideia do TF-IDF é que as palavras em um documento podem ser divididas em duas classes: as palavras que são comuns, como artigos e preposições, e as palavras que são relevantes, definidas por terem um valor de TF-IDF maior que as palavras comuns (RAMOS *et al.*, 2003).

2.5.3 *Word2Vec*

Dado o alto número de palavras distintas em um texto, o número de colunas construídas utilizando um esquema tradicional de representação pode se tornar muito alto, com cada coluna correspondendo a uma palavra, e cada valor na coluna especificando se a palavra existe no texto correspondente à linha ou não. O Word2vec ajuda a melhor representar os dados, com palavras semelhantes entre si tendo vetores semelhantes, enquanto palavras não semelhantes entre si tendo vetores diferentes (AYYADEVARA; AYYADEVARA, 2018). Assim, este modelo pode detectar palavras sinônimas ou sugerir palavras adicionais para uma frase após treinada.

O algoritmo Word2Vec usa um modelo de rede neural, o saco-de-palavras contínuo ou o *skip-gram*, para aprender associações de palavras de um grande corpus de texto (MIKOLOV *et al.*, 2013). O algoritmo saco-de-palavras contínuo prediz a palavra dado seu contexto, enquanto o *skip-gram* prediz o contexto dado uma palavra, como mostrado na Figura 2. Ao alimentar o corpus de texto em um desses dois modelos de aprendizado, o Word2Vec finalmente gera os vetores de palavras. Nesse processo, o Word2Vec primeiro constrói um vocabulário a partir do corpus de texto de treinamento e aprende as representações vetoriais de cada palavra. Além disso, o Word2Vec tem a capacidade de calcular a distância do cosseno entre cada palavra.

2.5.4 *Vetor Global para Representação de Palavras*

O Vetor Global para Representação de Palavras (*Global Vector for Word Representation - Glove*) (PENNINGTON; SOCHER; MANNING, 2014) é um modelo de representação de palavras conhecido por sua capacidade de mapear palavras em um espaço significativo, onde a distância entre elas está relacionada à sua similaridade semântica. Ele usa um algoritmo não supervisionado para aprender a representação vetorial por meio da redução de dimensionalidade na matriz de co-ocorrência, utilizando estatísticas locais (informações de contexto local das palavras) incorporadas às estatísticas globais (co-ocorrência de palavras) para obter os vetores.

O modelo pode ser estimado da seguinte forma: dado um corpus grande, as estatísticas de coocorrência de palavras são coletadas no formato de matriz de coocorrência de palavras

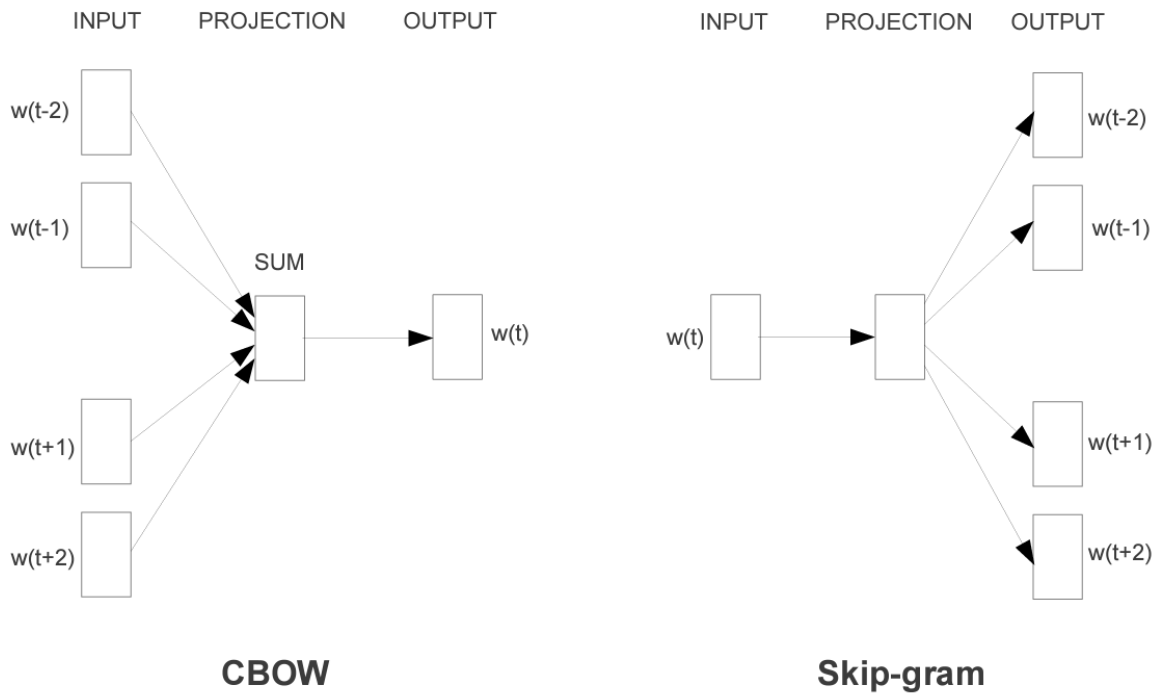


Figura 2 – Arquiteturas dos algoritmos saco-de-palavras contínuo e *skip-gram* apresentadas no artigo (MIKOLOV; LE; SUTSKEVER, 2013).

X . Cada elemento X_{ij} dessa matriz representa a contagem de quantas vezes a palavra i aparece no contexto da palavra j . O contexto de uma palavra é definido como um número predefinido de palavras antes e depois dela.

Um dos benefícios do GloVe em relação a outros esquemas de incorporação de palavras como o Word2Vec é que ele não captura apenas as informações de contexto local das palavras (estatísticas locais), mas também captura a co-ocorrência de palavras, também conhecidas como estatísticas globais em um corpus, para obter os vetores de palavras. Além disso, ele permite a implementação paralela, o que facilita o treinamento em um grande corpus e também combina as melhores características de duas famílias de modelos; os métodos de janela de conteúdo local e a fatoração de matriz global.

2.6 Classificação

A classificação é uma tarefa em que dados são categorizados de acordo com um grupo de rótulos predefinidos. Em textos, por exemplo, a classificação automática envolve a atribuição de classes a documentos textuais, usando uma técnica de ML (DALAL; ZAVERI, 2011). Ela é feita geralmente com base em palavras ou características significativas extraídas do documento de texto. Como as classes são predefinidas, é uma tarefa de aprendizado de máquina

supervisionado.

A definição do problema de classificação é: dado um conjunto de registros de treinamento $D = X_1, \dots, X_n$, de modo que cada registro seja rotulado com um valor de classe extraído de um conjunto de k valores discretos diferentes indexados por $1 \dots k$ (AGGARWAL; ZHAI, 2012). Os dados de treinamento são usados para construir um modelo de classificação, que relaciona os recursos no registro subjacente a um dos rótulos de classe. Para uma determinada instância de teste em que a classe é desconhecida, o modelo de treinamento é usado para prever um rótulo de classe para essa instância.

2.6.1 Máquina de Vetores de Suporte

As Máquinas de Vetores de Suporte (*Support Vector Machine* - SVMs) são uma família de métodos de ML, originalmente introduzidos para o problema de classificação e posteriormente generalizados para várias outras situações. Esse modelo aprende funções discriminantes de duas classes a partir de um conjunto de exemplos de treinamento (MAMMONE; TURCHI; CRISTIANINI, 2009). O método combina ideias da teoria de aprendizado estatístico e otimização convexa, para encontrar um limite adequado no espaço de dados para separar as duas classes de pontos.

2.6.2 Floresta Aleatória

O modelo Floresta Aleatória (*Random Forest* - RF) é uma ferramenta popular de ML baseada em árvores que é altamente adaptável a dados, e é capaz de levar em conta a correlação, bem como as interações entre as características. Elas são uma combinação de preditores de árvores, de modo que cada árvore depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores da floresta (BREIMAN, 2001).

2.6.3 Naive Bayes Multinomial

O classificador Naive Bayes Multinomial (*Multinomial Naive Bayes* - MNB) é uma variante do classificador Naive Bayes (NB) (RISH *et al.*, 2001) usado para dados distribuídos multinomialmente como os encontrados na classificação de texto (XU; LI; WANG, 2017). É frequentemente usado como linha de base porque é rápido e fácil de implementar. Além disso,

com pré-processamento apropriado, é competitivo com métodos mais avançados.

2.6.4 Campos Aleatórios Condicionais

O algoritmo Campos Aleatórios Condicionais (*Conditional Random Fields* - CRF) é um modelo de estrutura probabilística de aprendizado supervisionado usado para rotular e dividir dados de estrutura de sequência (SUTTON; MCCALLUM *et al.*, 2012), considerado na literatura como uma das abordagens mais eficazes para NER. Ele é uma forma de modelo gráfico não direcionado que define uma única distribuição log-linear sobre sequências de rótulos, dada uma sequência de observação particular.

2.6.5 Perceptron

O Perceptron (FREUND; SCHAPIRE, 1998) é um tipo de rede neural de camada única constituída por um único neurônio, inexistindo-se qualquer tipo de realimentação de valores produzidos pelo seu único neurônio (ZAREI; BOZORG-HADDAD; NIKOO, 2022). Ele pode ser visto como o tipo mais simples de rede neural *feedforward*: um classificador linear. Como classificador binário, o perceptron mapeia a entrada x (um vetor de valor real) para um valor de saída $f(x)$ (um valor binário simples através da matriz

$$f(x) = \begin{cases} 1, & \text{se } w \cdot x + b \geq 0, \\ 0, & \text{otherwise} \end{cases} \quad (2.2)$$

2.6.6 Perceptron Multicamadas

Um Perceptron Multicamadas (*Multilayer Perceptron* - MLP) consiste num sistema de neurônios interconectados de forma simples, podendo ser formada por uma ou mais camadas ocultas (GARDNER; DORLING, 1998). É composta de pelo menos três camadas de neurônios totalmente conectados por pesos, sendo uma camada de entrada (que recebe os dados), uma camada oculta (que permite lidar com problemas não lineares) e uma camada de saída (onde ocorre a predição). Os neurônios utilizam uma função de ativação não linear. Ao se adicionar camadas nessas estruturas, elas se tornam mais profundas e complexas, sendo assim capazes de fazer e aprender relações mais intrincadas. Essas redes utilizam a técnica de aprendizado supervisionado de retro-propagação para treinamento (RUMELHART; HINTON; WILLIAMS, 1986).

2.6.7 Memória Bidirecional de Longo-Curto Prazo

A Memória Bidirecional de Longo-Curto Prazo (*Bidirectional Long-Short Term Memory* - BiLSTM) é uma rede neural de processamento de sequência que consiste em duas LSTMs: uma recebendo a entrada na direção direta e a outra na direção inversa, aumentando efetivamente a quantidade de informações disponíveis para a rede e melhorando o contexto disponível para o algoritmo (CHIU; NICHOLS, 2016).

2.7 Representações de Codificador Bidirecional de Transformadores

Representações de Codificador Bidirecional de Transformadores (*Bidirectional Encoder Representations from Transformers* - BERT) é um modelo de aprendizado profundo pré-treinado baseado em transformadores, que são modelos DL que adotam o mecanismo de auto-atenção, ponderando diferencialmente o significado de cada parte dos dados de entrada (DEVLIN *et al.*, 2018). O BERT é projetado para NLP, oferecendo *embeddings* contextuais de palavras (WEHNERT *et al.*, 2022). É um modelo bidirecional; ou seja, considera simultaneamente o contexto esquerdo e direito e torna mais fácil discernir o contexto das palavras com base nas palavras que as cercam.

O BERT foi treinado com dados em inglês da Wikipedia e do Book Corpus (DEVLIN *et al.*, 2018), com duas versões: a *base* e a *large*. Depois, outros modelos foram disponibilizados em outras línguas, sendo um deles o BERTimbau. Ele é o primeiro modelo BERT que foi treinado para português brasileiro. Para isso, uma enorme base de dados foi utilizada, a *Brazilian Web as Corpus* (FILHO *et al.*, 2018), que contém 2,68 bilhões de *tokens* de 3,53 milhões de documentos de páginas brasileiras (SOUZA; NOGUEIRA; LOTUFO, 2020). Foram criadas duas versões do BERTimbau. Na primeira, BERTimbau *base*, os pesos foram inicializados com a rede Multilingual BERT base, uma versão BERT treinada para 107 idiomas (PIRES; SCHLINGER; GARRETTE, 2019). A segunda versão chama-se BERTimbau *large*; em que os pesos foram inicializados com a rede original BERT na versão *large*.

2.8 Ensembles

Uma *ensemble* consiste em um conjunto de classificadores treinados individualmente cujas previsões são combinadas de alguma forma para formar a previsão final (SAINI; GHOSH, 2017). Na *ensemble*, a previsão de cada classificador pode ser considerada um voto para uma

determinada classe. Portanto, o resultado da *ensemble* geralmente é com base em diferentes estratégias de votação que dependem diretamente do tipo de saída obtida.

Na estratégia de votação por média simples, cada classificador produz um vetor dimensional com tamanho igual ao número de classes. Cada coluna desse vetor representa o suporte para a hipótese de que a amostra x submetida à classificação pertence à classe i . Sem perda de generalidade, podemos assumir que cada saída do classificador pertence ao conjunto $\{0,1\}$ (KUNCHEVA, 2014).

Este tipo de votação usa o vetor de confiança nos rótulos sugeridos e estimativas de probabilidade posteriores para as classes (KUNCHEVA, 2014). Cada instância do conjunto de dados de teste tem suas previsões médias calculadas, sendo o rótulo final a classe com a maior média. Essa abordagem é definida na Equação 2.3.

$$\mu_j(x) = \frac{1}{L} \sum_{i=1}^L d_{i,j}(x), \quad (2.3)$$

onde o rótulo da amostra x recebe o índice do maior $\mu_j(x)$ e $d_{i,j}(x)$ é o vetor de probabilidades.

2.9 Métricas de Validação

A avaliação dos modelos é parte fundamental na construção de uma metodologia eficaz, pois somente a avaliação pode dizer o quão bom é o modelo na generalização. A avaliação das diferentes etapas do trabalho foi feita através do uso de métricas acurácia (Equação 2.4), kappa (Equação 2.5), precisão (Equação 2.6), recall (Equação 2.7), F1-score (Equação 2.8) e a área sob a curva ROC (AUC). Os valores apresentados na Tabela 3 constituem o cálculo de quatro dessas métricas.

Tabela 3 – Descrição dos quatro valores usados para calcular as métricas.

Valor	Descrição
Verdadeiro positivo (TP)	O modelo prevê corretamente a classe positiva
Verdadeiro negativo (TN)	O modelo prevê corretamente a classe negativa
Falso positivo (FP)	O modelo prevê incorretamente a classe positiva
Falso negativo (FN)	O modelo prevê incorretamente a classe negativa

O cálculo da acurácia (BARATLOO *et al.*, 2015) é apresentado na Equação 2.4:

$$Acuracia = \frac{TP + TN}{TP + TN + FP + FN}. \quad (2.4)$$

onde TP , TN , FP e FN correspondem a Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo, respectivamente. Quanto mais próximo de 1, melhor o resultado da precisão.

A métrica kappa de Cohen (COHEN, 1960) indica como os resultados escolhidos superaram a jogada aleatoriamente de acordo com a frequência de cada classe. Os valores do índice são categorizados como: ruim ($\kappa \leq 0.2$), regular ($0.21 \leq \kappa \leq 0.4$), bom ($0.41 \leq \kappa \leq 0.6$), muito bom ($0.61 \leq kappa \leq 0,8$) e excelente ($\kappa \geq 0,81$). O cálculo do Kappa é apresentado na Equação 2.5:

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \quad (2.5)$$

onde p_o é a concordância observada e p_e é a concordância esperada.

A métrica precisão (OLSON; DELEN, 2008) calcula a capacidade do classificador de não rotular uma amostra negativa como positiva. Quanto mais próximo de 1, melhor o resultado. O cálculo da precisão é apresentado na Equação 2.6:

$$Precisao = \frac{TP}{TP + FP}. \quad (2.6)$$

O recall (OLSON; DELEN, 2008) calcula a capacidade do classificador de encontrar todas as amostras positivas. O melhor desempenho é alcançado quando o valor se aproxima de 1. O cálculo do recall é apresentado na Equação 2.7:

$$Recall = \frac{TP}{TP + FN}. \quad (2.7)$$

O F1-score corresponde à média harmônica entre precisão e recall. O valor mais alto possível da pontuação F1 é 1, indicando precisão e recuperação perfeitas, e o valor mais baixo possível é 0 se a precisão ou sensibilidade for zero. O cálculo do F1-score é apresentado na Equação 2.8:

$$F1 - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \quad (2.8)$$

Finalmente, usamos a área sob a curva ROC (AUC) (HANLEY; MCNEIL, 1982) métrica. O melhor desempenho é obtido quando o valor de AUC se aproxima de 1 (equivalente a 100%).

3 TRABALHOS RELACIONADOS

Este capítulo apresenta um levantamento do estado da arte sobre o emprego de tarefas de NLP em bases de textos relacionados à saúde, especialmente de dados clínicos. São apresentados e discutidos trabalhos em que as tarefas de classificação automática e NER foram realizadas, em sua maioria, em textos em português. Na Tabela 4, as principais características desses estudos são destacadas.

Em Cusick *et al.* (2021), os autores usaram um método fracamente supervisionado para detectar ideação suicida a partir de notas clínicas não estruturadas em EHRs. Este método é uma abordagem de NLP baseada em regras para rotular notas de treinamento e validação. Eles testaram as seguintes representações de palavras: saco de palavras (ZHANG; JIN; ZHOU, 2010), bigrama (BROWN *et al.*, 1992), TF-IDF (YUN-TAO; LING; YONG-CHENG, 2005) e Word2Vec (CHURCH, 2017). Com este conjunto de dados, vários modelos de aprendizado de máquina foram treinados - Regressão Logística (LR) (WRIGHT, 1995), Máquina de Vetores de Suporte (SVM) (BURGES, 1998), Naïve Bayes (NB) (RISH *et al.*, 2001) e rede neural convolucional (CNN) (ALBAWI; MOHAMMED; AL-ZAWI, 2017). O modelo CNN superou todos os outros métodos, alcançando uma precisão geral de 94% e um F1-score de 82% em documentos com ideação suicida “atual”.

O trabalho Gupta *et al.* (2021) teve como objetivo avaliar uma abordagem de fenotipagem baseada em ML para identificar pacientes com eventos adversos relacionados ao sistema imunológico (irAEs) a partir de anotações clínicas. Os dados utilizados foram do sistema MedStar Health - Baltimore-Washington (EUA). Eles compararam o desempenho de diferentes modelos de ML: LR, SVM, Random Forest (RF) (SWAIN; HAUSKA, 1977), CNN e Memória Bidirecional Longa de Curto Prazo (BiLSTM) (ZHOU *et al.*, 2016). Tais modelos foram associados a diferentes métodos de representação de características (TF-IDF e BioWordVec (ZHANG *et al.*, 2019)) e redução de texto (filtragem baseada em palavras-chave - KF). Como resultado, a CNN com incorporações BioWordVec e KF obteve o melhor desempenho com um F1-score de 75,2% e AUC-ROC de 78,2%. Além disso, eles também avaliaram o desempenho dos modelos na detecção dos três principais irAEs no conjunto de dados: toxicidades relacionadas à pele, toxicidades endócrinas e colite. Os resultados indicam que o desempenho dos classificadores para prever os três principais subtipos de irAEs é tipicamente inferior. No entanto, o desempenho comparativo dos modelos com e sem KF é semelhante aos resultados obtidos para classificar qualquer irAE. Para classificar as toxicidades relacionadas à pele, a CNN com KF obteve o

melhor F1-score de 65,1%. O BiLSTM com KF e parâmetro de classe de peso foi o melhor modelo para detectar toxicidade endócrina com um F1-score de 58,8%. Para colite, CNN com KF obteve os melhores resultados e pesos de classe com um F1-score de 70,9%.

Em López-Úbeda *et al.* (2021), modelos de classificação ML foram utilizados em um conjunto com dados de pacientes presentes em laudos radiológicos para atribuir protocolos de procedimentos de imagens médicas em exames de Tomografia Computadorizada (TC) e Ressonância Magnética (RM). Eles compararam vários modelos, incluindo métodos tradicionais como SVM, RF, BiLSTM, CNN e técnicas de aprendizado de transferência (BETO (CAÑETE *et al.*, 2020) - BERT treinado em um conjunto de dados espanhol e XLM-17 (CONNEAU *et al.*, 2019) - BERT treinado em 17 línguas). TF-IDF + unigrama foram utilizados para a representação numérica dos dados. O SVM apresentou melhores resultados em ambos os conjuntos de dados, com um F1-score de 76% e uma precisão de 92% para o conjunto TC e um F1-score de 76% e uma precisão de 87% para o RM. Os piores desempenhos foram das redes neurais.

Os autores em Kamar *et al.* (2022) automatizaram a análise e classificação do mecanismo de ação da droga (MOA) em textos de tratamento ambulatorial da doença de Alzheimer (DA). No total, o conjunto de dados contém 233 registros rotulados. Os autores pré-processaram os dados e então usaram o TF-IDF para produzir a matriz de representação de texto. Para encontrar o melhor classificador para esse fim, eles exploraram diferentes algoritmos de ML, como RF, XGBoost (CHEN; GUESTRIN, 2016), LR, SVM e Árvores de Decisão (DT) (SAFAVIAN; LANDGREBE, 1991). Este último apresentou os resultados mais promissores, com 95% de precisão, 100% de sensibilidade e um F1-score de 92%.

No trabalho Li *et al.* (2022), os autores treinaram modelos ML (LR, SVM e RF), bem como diferentes ensembles desses modelos, para detectar rupturas do menisco medial ou lateral em relatórios de RM de texto livre. Eles também avaliaram a capacidade de generalização desses algoritmos treinados exclusivamente, desde relatórios de ressonância magnética até relatórios de artroscopia. Os modelos mostraram alto desempenho de validação cruzada para detectar ruptura meniscal em relatórios de ressonância magnética do joelho (menisco medial F1-score - 93% 94%, menisco lateral F1-score - 86% 88%). O desempenho desses algoritmos em laudos de artroscopia foi semelhante, apesar de nunca terem sido treinados com esse tipo de dado. Porém, foram superiores ao conjunto dos modelos (F1-score do menisco medial - 97%, F1-escore do menisco lateral - 99%).

No artigo Kumar *et al.* (2020), o objetivo foi desenvolver um sistema de classificação

para o reconhecimento de 16 comorbidades a partir de registros clínicos de pacientes. Para isso, foram utilizadas abordagens clássicas de ML: SVM, k-Nearest Neighbors (kNN) (PETERSON, 2009), NB, RF, Árvore Aleatória (RT) (ALDOUS, 1993), J-48 (RAJPUT *et al.*, 2011) e J-Rip (RAJPUT *et al.*, 2011)) e aprendizagem profunda (BiLSTM). Como *embeddings*, foram utilizadas as representações de saco de palavras, TF-IDF, Word2Vec, GloVe (STURMAN; ZELTZER, 1994), FastText (JOULIN *et al.*, 2016) e Universal Sentence Encoder (CER *et al.*, 2018). Os autores combinaram abordagens clássicas de ML e aprendizado profundo com métodos de representação de recursos e métodos de seleção de recursos para definir a melhor combinação para o problema. Como resultado, o SVM com TF-IDF obteve um F1-score de 99,26% e a ensemble um F1-score de 99,27%. Embora o desempenho da ensemble seja apenas ligeiramente melhor do que o SVM, os autores concluíram que a eficácia dos modelos do conjunto pode ser relevante por seus altos valores de média e baixo desvio padrão.

Os trabalhos citados acima realizaram a classificação de conjuntos de dados em inglês e espanhol com finalidades diversas, desde a classificação de ideação suicida até a identificação de lesões a partir de imagens médicas. Por outro lado, Sousa *et al.* (2012) mostrou que os métodos de recuperação de informação e classificação de texto com NB alcançam bons desempenhos na categorização do conteúdo da web de saúde em português brasileiro para fornecer as melhores informações ao público leigo. O conjunto de dados contém 3.702 páginas da web, divididas em 19 categorias de um diretório público, incluindo acidentes, cuidados pessoais, medicina preventiva e odontologia. Comparou-se o desempenho de 4 métodos de representação de textos por vetores de peso: frequência de termo (TF) (SALTON; BUCKLEY, 1988), TF-IDF, ocorrência de termo (TO) (SALTON; BUCKLEY, 1988) e ocorrência binária (BO) (SALTON; BUCKLEY, 1988). O método de representação TO combinado com o classificador NB apresentou os resultados mais promissores com *recall* de 0,91 e 0,98 para a primeira e quinta posições do ranking de relevância da categoria.

Assim como no trabalho anterior, Silva *et al.* (2019) analisou descrições de textos em português brasileiro. No entanto, as 15.479 descrições referem-se a cirurgias e registros pós-operatórios para prever e detectar infecções em centros cirúrgicos para garantir a segurança do paciente. A estratégia de mineração de texto envolveu as etapas de normalização, tokenização, lematização e vetorização. Em seguida, os autores utilizaram unigramas, bigramas e trigramas na etapa de seleção de atributos e os algoritmos TF e TF-IDF para a construção do vocabulário. Os seguintes algoritmos de aprendizado de máquina realizaram a previsão: Linear SVC, LR, MNB,

Nearest Centroid (MOSCHITTI, 2003), RF, Descida do Gradiente Estocástico (KETKAR, 2017) e Support Vector Classification (SVC). A Descida do Gradiente Estocástico e LR obtiveram os melhores resultados de previsão com 79,7% ROC-AUC e 80,6% ROC-AUC, respectivamente.

Os autores em Santos *et al.* (2018) estimaram automaticamente o índice de comorbidade de Charlson (CCI), usado para prever a mortalidade de pacientes com comorbidades. Os autores usaram um conjunto de dados com 1,5 milhão de notas clínicas de 48.900 registros de admissão. Eles usaram métodos de extração de recursos de texto: TF-IDF (unigrama e multigrama), representação de tópico (Latent Dirichlet Allocation - LDA) (BLEI; NG; JORDAN, 2003) e Word2Vec. Com base nas características extraídas, os autores usaram os seguintes métodos para estimar o CCI: RF, kNN, Regressão Linear de Vetores de Suporte e Redes Neurais em duas abordagens com camadas densas e convolucionais. Os resultados mostraram que o Word2Vec associado a uma rede neural com camadas densas supera as demais combinações, atingindo um erro médio absoluto (MAE) de 0,51. Os autores também apontam o RF + unigrama como uma alternativa adequada para conjuntos de dados menores (MAE=1,41).

Com relação a trabalhos que realizar a tarefa de NER, Lopes, Teixeira e Oliveira (2019) avaliaram o uso do algoritmo Conditional Random Fields (CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001) para a extração de entidades de 281 textos clínicos de neurologia em português. Na etapa de pré-processamento, foi realizada a tokenização das sentenças, a atribuição da tag POS, assim como o lemma. Na rotulação, foram utilizadas as entidades caracterização, teste, evolução, genética, sítio anatômico, negação, observações adicionais, doença, resultados, data hora (datetime), terapêutico, valor, e via de administração. Consideraram uma janela de contexto de 5 *tokens* (*token* atual, dois anteriores e dois seguintes) para o qual características ortográficas foram extraídas e depois selecionaram as características que seriam utilizadas no modelo CRF final. Como resultados, obtiveram micro F1-score médio de 80.5% para avaliação relaxada e 73.1% para avaliação rigorosa. Na primeira avaliação, a tag predita é comparada com a tag rotulada, e na segunda, todos os *tokens* de uma sentença devem corresponder as tags rotuladas.

Em Lopes, Teixeira e Oliveira (2020), os autores compararam os modelos CRF, BiLSTM-CRF (LAMPLE *et al.*, 2016) e BiLSTM-CRF com conexões de aprendizado residual. Foram adicionados 20 textos de Serviço de Neurologia do Centro Hospitalar e Universitário de Coimbra (CHUC), mesmo conjunto de dados do trabalho anterior (LOPES; TEIXEIRA; OLIVEIRA, 2019). Além disso, foram comparados os desempenhos dos modelos BiLSTM-CRF

usando modelos de *embeddings* de palavras (WE) treinados com texto clínico e treinados com textos de linguagem geral. Para treinamento do modelo de WE de domínio clínico, coletaram 3377 textos clínicos de todos os volumes da revista Sinapse, publicados entre 2001 e 2018. O modelo de *embedding* FastText (WU; MANBER, 1992) pré-treinado em português foi utilizado para comparação com o modelo de domínio. A BiLSTM-CRF com aprendizado residual alcançou os melhores micro F1-scores com 82.9% e 74.9% para avaliação relaxada e rigorosa em textos extraídos da revista médica. Para textos coletados no hospital, a rede neural alcançou F1-score de 71.2% e 61.8%, respectivamente.

No trabalho Peters *et al.* (2022), um corpus anotado semanticamente em português foi desenvolvido usando texto clínico de várias especialidades médicas, tipos de documentos e instituições. Esse corpus, nomeado como SemClinBr, possui 1000 notas clínicas, rotuladas com 65.117 entidades e 11.263 relações. Além disso, os autores também apresentaram um esquema e uma ferramenta de rotulação com foco em um recurso de sugestão de anotação. A pontuação média de concordância dos anotadores variou de 71%, aplicando correspondência estrita, à 92%, considerando uma correspondência relaxada, aceitando sobreposições parciais e tipos semânticos hierarquicamente relacionados.

Os autores em Schneider *et al.* (2020) desenvolveram um modelo de NER com o uso de uma rede neural contextual profunda para o português, denominado BioBERTpt. Como base de dados do trabalho, coletaram 2.100.546 notas clínicas de hospitais brasileiros de 2002 a 2018, e selecionaram títulos e resumos de artigos científicos portugueses publicados na Pubmed e na Scielo. Na etapa de pré-processamento, dividiram as notas clínicas e os resumos em sentenças e tokenizaram com o tokenizador padrão BERT. Eles realizaram o fine-tuning do modelo BERT multilingual, gerando as redes BioBERTpt(all), BioBERTpt(clin) e BioBERTpt(bio). Para avaliar o desempenho, realizaram experimentos NER em dois corpora anotados, os conjuntos de dados (PETERS *et al.*, 2022) e (LOPES; TEIXEIRA; OLIVEIRA, 2019). Depois, compararam os resultados com o CRF, BiLSTM-CRF e modelos BERT existentes: BERT multilingual uncased, BERT multilingual cased, e Portuguese BERT (BERTimbau) nas versões base e large. O modelo BioBERTpt(all) alcançou 60.8%, 60.7% e 60.4% nas métricas precisão, recall e F1-score no primeiro conjunto de dados, enquanto o BioBERTpt(clin) obteve 91.7%, 93.5% e 92.6%, respectivamente, no segundo conjunto de dados. O modelo BERTimbau base alcançou resultados superiores que a versão large, com 59.5%, 58.7% e 58.5% no primeiro conjunto de dados, e 91%, 92.2% e 91.6% no segundo conjunto de dados.

Em Souza *et al.* (2020), os autores utilizaram modelos baseados na rede BERT e técnicas de Label Powerset (LP), que transforma um problema multilabel em um problema multiclasse, para fazer o reconhecimento de entidades nomeadas clínicas. O corpus utilizado foi o apresentado no trabalho (PETERS *et al.*, 2022). Os modelos usados para comparação e seleção da melhor abordagem foram BERT multilingual uncased, BERT multilingual cased, BioBERTpt(all), BioBERTpt(clin), BioBERTpt(bio), Portuguese BERT (BERTimbau) nas versões large e base, e CRF, como modelo baseline. Nos resultados, BioBERTpt(clin) obteve o melhor desempenho com 75.2%, 75.3% e 75.2% nas métricas precisão, recall e micro F1-score, respectivamente.

Tabela 4 – Trabalhos da literatura que realizaram tarefa de classificação ou NER de documentos clínicos.

Trabalho	Tarefa	Tipo de documento	Vetorização	Modelos	Idioma
Cusick <i>et al.</i> (2021)	CLA	Notas clínicas para detecção de ida- ação suicida	BOW, bigrama, TF-IDF e Word2Vec	LR, SVM, Naive Bayes e CNN	Inglês
Gupta <i>et al.</i> (2021)	CLA	Notas clínicas para identificar even- tos adversos relacionados ao sistema imunológico (irAEs)	TF-IDF e BioWordVec	LR, SVM, RF, CNN e BiLSTM	Inglês
López-Úbeda <i>et al.</i> (2021)	CLA	Relatórios radiológicos para atribuir protocolos de procedimentos de ima- gens médicas	Unigrama + TF-IDF	SVM, RF, BiLSTM, CNN, BETO e XLM-17	Espanhol
Kambar <i>et al.</i> (2022)	CLA	Textos de tratamento de Alzheimer para classificar o mecanismo de ação da droga (MOA)	TF-IDF	RF, XGBoost, LR, SVM e DT	Inglês
Li <i>et al.</i> (2022)	CLA	Relatórios de ressonância magnética para detectar ruptura do menisco me- dial ou lateral	BOW + n_grama (1, 2 e 3)	LR, SVM, RF e diferentes ensembles des- ses modelos	Inglês
Kumar <i>et al.</i> (2020)	CLA	Registros clínicos para reconheci- mento de 16 comorbidades	BOW, TF-IDF, Word2Vec, GloVe, fast- Text e Universal Sentence Encoder	SVM, kNN, NB, RF, RT, J-48, J-Rip, e BiLSTM	Inglês
Sousa <i>et al.</i> (2012)	CLA	Páginas web de saúde divididas em 19 categorias	TF, TF-IDF, TO e BO	NB	Português
Silva <i>et al.</i> (2019)	CLA	Descrições de cirurgias e registros pós-operatórios para previsão de in- fecções	N_grama (1, 2 e 3) + TF e TF-IDF	Linear-SVC, LR, MNB, Nearest Centroid, RF, Stochastic Gradient Descent e SVC	Português
Santos <i>et al.</i> (2018)	CLA	Notas clínicas para prever o índice de comorbidade de Charlson (CCI)	Unigrama e multigrama + TF-IDF, LDA e Word2Vec	RF, kNN, Linear SVR e redes neurais em duas abordagens com camadas densas e convolucionais	Português
Lopes, Tei- xeira e Oliveira (2019)*	NER	Textos clínicos de neurologia para a extração de entidades	-	CRF	Português
Lopes, Tei- xeira e Oliveira (2020)	NER	Textos clínicos de neurologia e tex- tos de serviço de neurologia do Cen- tro Hospitalar e Universitário de Coimbra (CHUC)	FastText e treinamento de modelos de WE	CRF, BiLSTM-CRF e BiLSTM-CRF com conexões de aprendizado residual	Português
Peters <i>et al.</i> (2022)*	NER	Notas clínicas rotuladas semantica- mente em 65.117 entidades e 11.263 relações	-	-	Português
Schneider <i>et al.</i> (2020)	NER	2.100.546 notas clínicas de hospitais brasileiros, e títulos e resumos de ar- tigos científicos portugueses publi- cados na Pubmed e na Scielo	-	CRF, BiLSTM-CRF, BERT multilingual uncased, BERT multilingual cased, BER- Timbau base, BERTimbau large e Bio- BERTpt (modelo proposto)	Português
Souza <i>et al.</i> (2020)	NER	Corpus do trabalho (PETERS <i>et al.</i> , 2022)	-	BERT multilingual uncased, BERT mul- tilingual cased, BioBERTpt(all), Bio- BERTpt(clin), BioBERTpt(bio), BERTIm- bau large, BERTimbau base, e CRF	Português

CLA: Classificação de dados; NER: Reconhecimento de entidade nomeada; *: Trabalhos que disponibilizaram o conjunto de documentos clínicos

3.1 Considerações Finais

O levantamento de estudos do estado da arte teve como principal foco o entendimento de como estavam sendo realizadas pesquisas no campo de processamento de textos clínicos nas tarefas de classificação e NER, em especial em trabalhos que utilizassem conjuntos de dados em português. Inicialmente, realizamos um levantamento sobre estudos que realizaram a classificação desses textos e concluímos que abordagens profundas não alcançaram resultados muito melhores do que as técnicas clássicas, e que era relevante a avaliação de diferentes tipos de algoritmos de vetorização para essa tarefa. Depois, avaliamos artigos sobre o reconhecimento de entidades clínicas. Neles, observamos que as entidades extraídas possuíam uma norma coletiva quanto à sua definição, como o uso das entidades droga química, procedimento, quantitativo e achados que não são específicos de um determinado tipo de documento. Ou seja, as entidades eram determinadas através de elementos significativos e comuns entre diferentes documentos clínicos, como receitas, notas clínicas, notas de admissão e alta, etc.

Com esses artigos com NER, surgiu também o interesse na comparação de modelos de aprendizado de máquina com modelos de aprendizado profundo para NER, já que a maioria dos modelos utilizados nos trabalhos relacionados eram redes neurais, com exceção do CRF. Além disso, os autores alcançaram bons resultados com o emprego da técnica de ajuste fino em modelos BERT. Por fim, um ponto evidenciado pelos estudos elencados é a falta de corpora disponíveis. Dentre todos os trabalhos em português com textos clínicos, apenas dois foram disponibilizados. Assim, essas características observadas foram utilizadas para a construção da metodologia proposta nesse trabalho.

4 MÉTODO PROPOSTO

Este capítulo apresenta as etapas realizadas para a construção da base de dados de documentos clínicos, a classificação desses textos e o reconhecimento das entidades clínicas. A metodologia proposta está dividida em duas etapas: a construção da base de dados e classificação dos documentos clínicos que a compõe; e o reconhecimento de entidades clínicas em receitas médicas, conforme apresentado nas seções a seguir.

4.1 Construção do conjunto de dados e classificação de documentos clínicos

Esta seção apresenta a metodologia de classificação de textos clínicos e descreve sua aplicação. Ela possui seis etapas: aquisição da base de dados, pré-processamento dos textos, rotulação, seleção das abordagens de vetorização, treinamento da *ensemble* e predição dos rótulos e validação dos resultados. A Figura 3 ilustra tais componentes.

4.1.1 Aquisição do conjunto de dados

O banco de dados utilizado neste trabalho é composto por prescrições, encaminhamentos, atestados, relatórios, anotações clínicas e solicitações de exames que médicos produziram durante consultas presenciais. Foram 3.000 amostras coletadas no período de 10 de maio de 2010 a 11 de agosto de 2021. O texto dessas amostras está no formato *rich text* (RTF). Foram consideradas três classes para esta pesquisa: prescrições, anotações clínicas (laudos, encaminhamentos, atestados e prontuários) e requisições de exames. Escolhemos essas classes devido ao interesse futuro em extrair informações específicas dessas categorias usando o reconhecimento de entidades nomeadas (NER).

4.1.2 Pré-processamento

Antes de vetorizar o texto, devemos pré-processar os dados para aumentar a qualidade de sua representação. A maioria dos textos médicos estava no formato RTF, então realizamos a conversão dos arquivos para strings com a biblioteca `striprtf`¹. Em seguida, realizamos a retirada de acentos, excesso de pontuação e troca de vírgula por ponto em decimais. Depois, removemos todos os espaços em branco restantes da conversão para *string*: `\t`, `\n` e `\r`. Finalmente, convertemos todo o texto em minúsculo. A Tabela 5 apresenta os exemplos do conjunto de

¹ <https://github.com/joshy/striprtf>

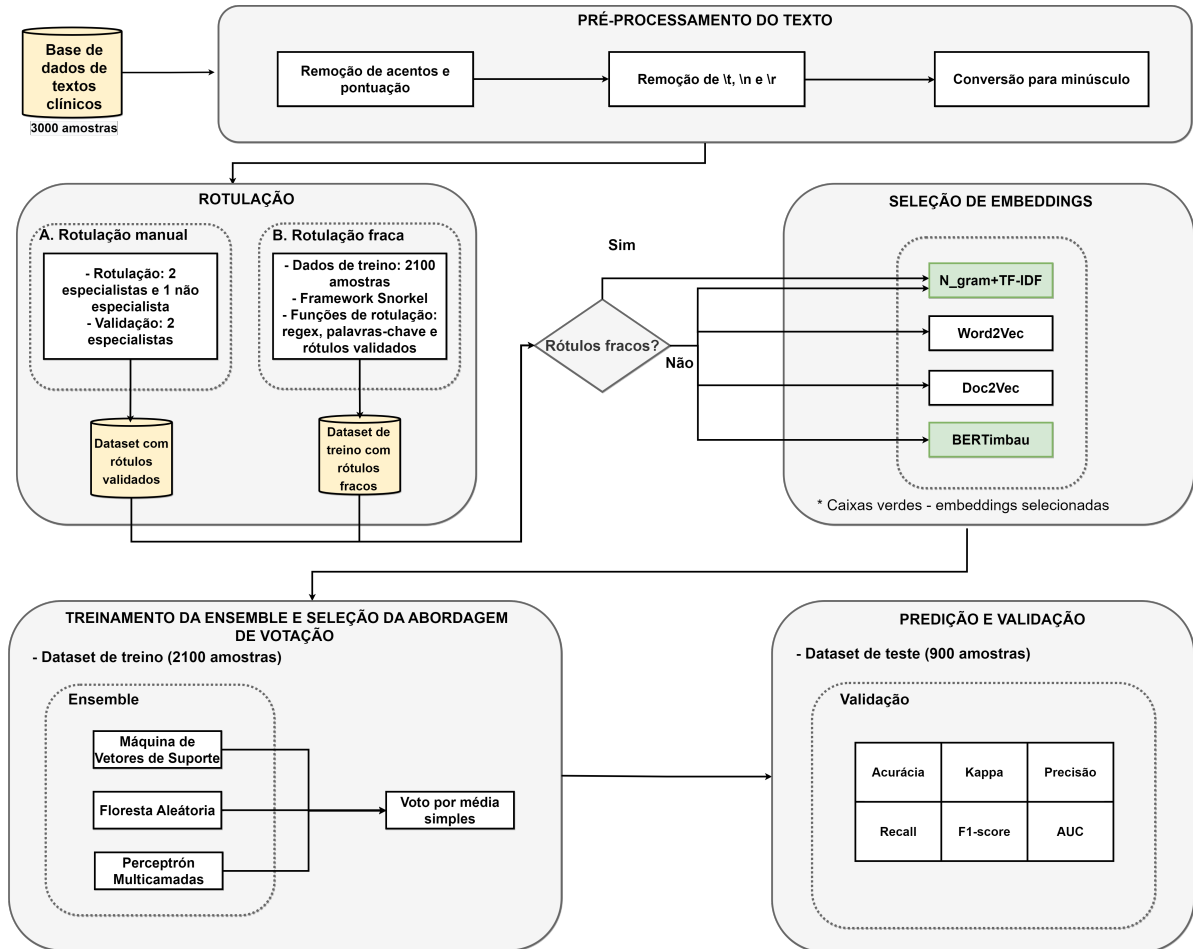


Figura 3 – Metodologia para classificação de documentos clínicos.

Tabela 5 – Exemplos de amostras pré-processadas por categoria no banco de dados.

Categoria	Texto pré-processado
Receitas	uso oral vertex 1 cx tomar um comp via oral a noite por 2 meses
Notas clínicas	ao dentista, encaminhado a paciente com suspeita de bruxismo para avaliação.
Solicitação de exames	solicitação de exame: solicito rm do joelho direito

dados.

4.1.3 Rotulação

A primeira rotulação foi feita manualmente por dois especialistas (farmacêuticos) e um não especialista. Depois de rotular todas as 3.000 amostras (1.000 amostras por tipo de documento), os dois especialistas validaram os rótulos manuais. Em seguida, dividimos o conjunto de dados em conjuntos de treinamento e teste, com 2.100 e 900 amostras, respectivamente. Na Tabela 6, a quantidade de amostras por classe dos conjuntos de treino e teste é apresentada. Como os dados clínicos são sensíveis e há necessidade de conhecimento médico para rotulá-los,

Tabela 6 – Quantidade de amostras rotuladas por classe nos conjuntos de treino e teste.

Classe	Conjunto de treino	Conjunto de teste
Receitas	700	300
Notas clínicas	700	300
Solicitações de exames	700	300

o uso de supervisão fraca tornou-se relevante. Assim, usamos o *framework* Snorkel para criar um segundo conjunto de treinamento com rótulos fracos.

Funções de rotulagem usando regex e palavras-chave foram criadas e metade das amostras do conjunto de dados de treinamento com os rótulos validados pelo especialista foram usados como entrada do *framework*. Como palavras-chave foram utilizados termos frequentes encontrados nos documentos. Em receitas, por exemplo, foram identificadas as palavras-chave: 'uso', 'formula', 'oral', 'topico', 'otologico', 'interno', 'manipulado', 'manipular', 'aplicar', 'tomar', 'via', etc. Já com relação aos regexs, eles foram gerados de acordo com a estrutura textual dos diferentes tipos de documentos, como '(solicito [a-zA-Z]*)' para solicitações de exames. Depois disso, combinamos os rótulos fracos previstos com o conjunto de dados de treinamento. No final, tivemos dois conjuntos de dados de treinamento, um com rótulos de verdade e outro com rótulos fracos.

Utilizamos a Análise de Componentes Principais (*Principal Component Analysis* - PCA) (WOLD; ESBENSEN; GELADI, 1987) para visualizar o conjunto de dados com a representação BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020). A Figura 4 mostra a densidade dos agrupamentos. A classe prescrição possui um agrupamento mais denso do que as classes notas clínicas e solicitação de exames. Além disso, há uma sobreposição mais significativa entre os dados das anotações clínicas e das aulas de solicitação de exames, enquanto os dados da prescrição permanecem em um espaço de representação mais distante. Além disso, existem alguns valores discrepantes entre os agrupamentos, principalmente na classe de notas clínicas.

4.1.4 Seleção de técnicas de vetorização

Para aumentar a variedade de representações das amostras de trabalho, selecionamos duas abordagens para a vetorização do texto. Inicialmente, escolhemos quatro métodos que foram testados e, a partir deles, foram selecionadas as duas abordagens. Dentre as técnicas utilizadas para esta etapa, temos o BERTimbau, a versão pré-treinada em português do Brasil do modelo BERT; Word2Vec e sua extensão para representação de documentos, Doc2Vec (KIM *et*

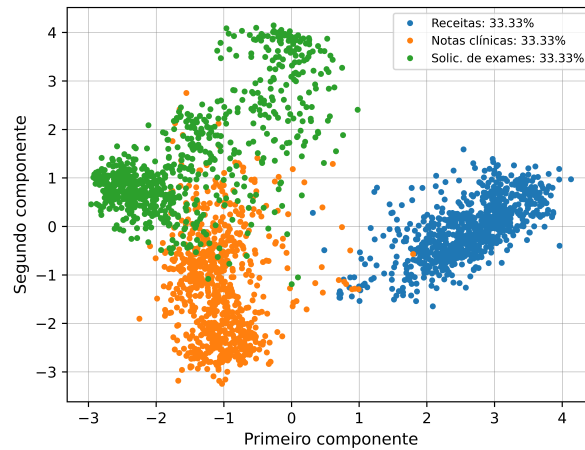


Figura 4 – Distribuição de classes com o algoritmo PCA nas *embeddings* do BERTimbau.

al., 2019) e N_gram juntamente com TF-IDF (YUN-TAO; LING; YONG-CHENG, 2005); com o conjunto de dados criado pelo Snorkel e o conjunto de dados com rótulos verdadeiros.

Usamos a representação com o unigrama e bigrama, já que são os tipos mais utilizados de n_gramas na literatura. Essa vetorização resultou em uma matriz de valores numéricos de 1.400×5.188 correspondente à representação N_gram + TF-IDF. No BERTimbau, foi utilizada a versão base ('bert-base-português-case'), resultando em uma matriz de saída de 1.400×768 . Treinamos os dois últimos modelos nas amostras para a etapa de seleção do algoritmo de vetorização. No Word2Vec, o modelo utilizado foi o bag-of-words contínuo, e o resultado tem dimensão de 1.400×100 , enquanto o Doc2Vec produziu uma representação de 1.400×128 . Nas abordagens de vetorização, usamos 1.400 amostras do conjunto de dados de treinamento. Depois, usamos a saída desses quatro métodos como entrada para o classificador SVM e empregamos a validação cruzada com *k-fold* igual a 10. A partir dos resultados obtidos, os dois métodos selecionados foram a junção de N_gram com o TF-IDF, com a representação dos conjuntos de dados com rótulos verdadeiros e fracos, e BERTimbau.

4.1.5 Ensemble e validação

Uma *ensemble* de classificação para documentos clínicos foi desenvolvida, consistindo em um conjunto de classificadores treinados individualmente cujas previsões são combinadas de alguma forma para formar a previsão final. Utilizamos a votação por média simples, ou seja, a média das probabilidades de certeza de classificação dos modelos.

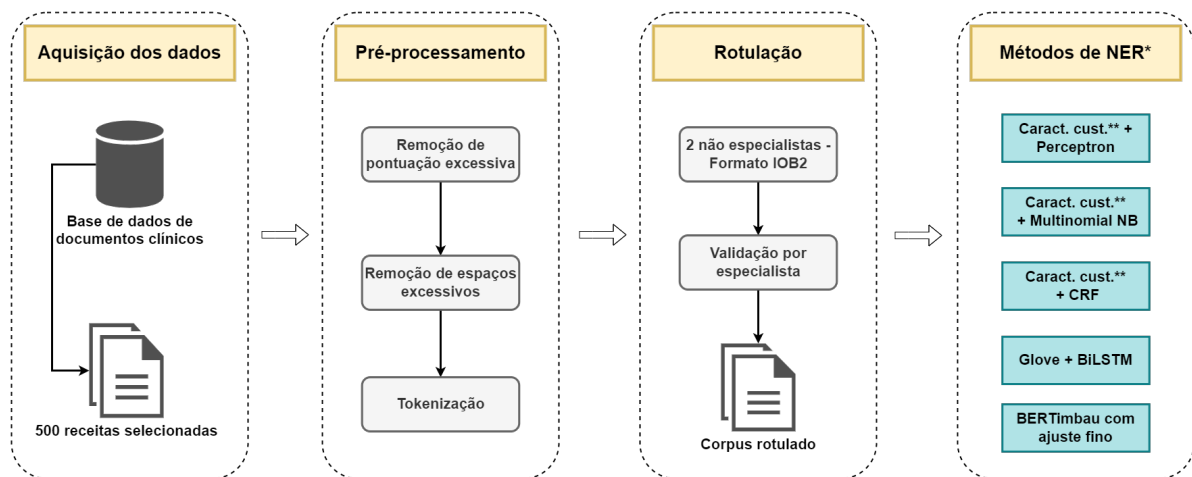
Treinamos o conjunto usando 2.100 amostras com os seguintes classificadores: *Random Forest* (RF) (SWAIN; HAUSKA, 1977), *Multilayer Perceptron* (MLP) (HORNIK;

STINCHCOMBE; WHITE, 1989) e *Support Vector Machine* (SVM) (BURGES, 1998). Eles foram selecionados porque, de acordo com a literatura, são amplamente utilizados na classificação de documentos clínicos. Cada classificador teve como entrada as três *embeddings* definidos na etapa anterior, a representação do conjunto de dados de treinamento com rótulos ground-truth e fracos com N_gram+TF-IDF, com dimensão 2.100×7.442 , e a representação feita pelo BERTimbau, com dimensão de 2.100×768 .

No SVM, definimos o parâmetro de regularização como 1,0 e a função de base radial como kernel. No RF, definimos o número de árvores para 100 e não houve profundidade máxima da árvore. No MLP, usamos 100 neurônios na camada oculta. Adotamos a abordagem de validação cruzada com 10 dobras para criar os nove conjuntos de rótulos a serem usados no conjunto. Depois de combinar os rótulos, usamos a *ensemble* para prever as 900 amostras do conjunto de dados de teste. A avaliação do desempenho das diferentes etapas do trabalho foi feita por meio do uso das métricas acurácia (Acc), kappa (Kap), precisão (Prec), recall (Rec), F1-score (F1) e área sob a curva ROC (AUC).

4.2 Reconhecimento de entidades nomeadas clínicas

Essa seção apresenta as etapas realizadas para a extração das entidades nomeadas. Inicialmente, foram selecionadas amostras de receitas clínicas na etapa de seleção do conjunto de dados para a tarefa NER. Depois, foram tokenizadas e rotuladas manualmente. Por fim, o corpus anotado foi usado para a comparação de cinco métodos de reconhecimento de entidades. A Figura 5 fornece uma ilustração de todas essas etapas.



* Reconhecimento de entidade nomeada
 ** Características customizadas

Figura 5 – Metodologia de reconhecimento de entidades clínicas em receitas médicas.

Tabela 7 – Descrição das entidades clínicas do trabalho e exemplos de *tokens* por entidade.

Entidade	Descrição da entidade	Tokens
APR	Forma de apresentação do medicamento	Creme
DOS	Dosagem do medicamento prescrito	150 mg
DUR	Duração do tratamento	5 dias
FREQ	Frequência de uso	1x/dia
MED	Nome do medicamento	Allegra
QTD	Quantidade de uso durante o tratamento	1 caixa ou 1 comprimido
VIA	Via de administração	Uso tópico

4.2.1 Rotulação para NER

Para a etapa de NER, foram selecionadas 500 receitas médicas do conjunto de dados rotulado na Seção 4.1. Depois dessa seleção, foi feita a tokenização das sentenças. Após isso, sete categorias de entidades foram escolhidas por um especialista para a rotulação do conjunto de dados, sendo elas: via de administração, nome do medicamento, apresentação, dosagem, quantidade e frequência de uso, e duração do tratamento. A definição das entidades participantes na rotulação deu-se pelo levantamento dos elementos caracterizantes de receitas clínicas. Na Tabela 7, uma descrição das entidades deste trabalho é apresentada conforme exemplos de *tokens* encontrados no conjunto de dados.

Depois da definição das entidades clínicas, o formato *Inside-Outside-Beginning 2* (IOB2) (SANG; VEENSTRA, 1999) foi escolhido como esquema de rotulação de *tokens*. Essa técnica permite identificar *tokens* no início (B) ou dentro (I) de uma entidade nomeada, ou *tokens* não participantes (O). Depois disso, dois não especialistas rotularam as amostras e um especialista validou as discordâncias. Para avaliar a confiabilidade da rotulação, baseado no trabalho (PETERS *et al.*, 2022), o cálculo da concordância entre anotadores (*Inter-Annotator Agreement* - IAA) foi realizado, com o uso da Equação 4.1. Ao final, a distribuição de *tokens* dos sete tipos de entidades incluídas no conjunto de dados foi calculada e é mostrada na Tabela 8.

$$IAA = \frac{matches}{matches + non_matches}, \quad (4.1)$$

onde *matches* é a quantidade de amostras rotuladas igualmente entre os rotuladores e *non_matches* é a quantidade de divergências.

A confiabilidade da rotulação se dá pela demonstração da validade do esquema de rotulação, já que se a rotulação não for consistente, então algum dos rotuladores está errado ou o esquema de anotação é inadequado para os dados (ARTSTEIN; POESIO, 2008). Na área médica, Landis e Koch (1977) propõem que os valores de IAA sejam categorizados como

Tabela 8 – Quantidade de *tokens* rotulados por entidade no conjunto de dados.

Entidade	Quantidade no conjunto de dados
APR	147
DOS	987
DUR	1497
FREQ	2936
MED	1625
QTD	2796
VIA	1901
All	19882

$0,41 \leq IAA \leq 0,6$ (moderado), $0,61 \leq IAA \leq 0,8$ (substancial) e $IAA \leq 0,81$ (quase perfeito). Neste trabalho, obtivemos um IAA de 0,896, evidenciando a consistência do corpus e que os rotuladores tiveram um entendimento semelhante das diretrizes de anotação, obtendo assim um desempenho significativo.

4.2.2 *Extração de características e abordagens NER*

Para a extração de características, duas técnicas foram usadas: características customizadas e as *embeddings* criadas pelo modelo Glove. A seleção dessas técnicas ocorreu pelo interesse em se avaliar o uso tanto de características calculadas conforme a estrutura do texto quanto àquelas geradas a partir de modelos de aprendizado profundo que consideram o contexto dos *tokens* para a vetorização. Na primeira abordagem, baseado nos trabalhos (LOPES; TEIXEIRA; OLIVEIRA, 2019; SOUZA *et al.*, 2019; LOPES; TEIXEIRA; OLIVEIRA, 2020), foram utilizados os elementos ortográficos e morfológicos, assim como linguísticos para caracterização do texto. A lista de características customizadas é apresentada a seguir.

- Ortográfica e morfológica
 - O *token* é um sinal de pontuação;
 - O *token* tem apenas caracteres ASCII;
 - O *token* tem apenas caracteres minúsculos;
 - O *token* tem apenas caracteres maiúsculos;
 - O *token* tem apenas caracteres alfabéticos;
 - O *token* é numérico;
 - O *token* é alfanumérico;
 - O *token* começa com um caractere maiúsculo;
 - O *token* termina em "a". Isso ocorre para substantivos femininos regulares do

português;

- O *token* termina em "s". Isso ocorre para substantivos plurais regulares do português;
- A forma do *token*, convertendo caracteres maiúsculos em “A”, minúsculos em “a”, números em “#” e pontuação em “-”;
- Comprimento do *token*;
- Prefixos e sufixos do *token* usando uma janela de 5 caracteres para ambos os afixos;
- A palavra tem mais de 2 consoantes consecutivas;
- Número máximo de consoantes consecutivas;
- A palavra está no início da frase;
- Todas as palavras na frase são maiúsculas;
- Todas as palavras na frase são minúsculas;
- A palavra tem acento;
- Número de vogais da palavra;
- A palavra não tem vogais;
- Número máximo de vogais consecutivas;
- A palavra está no final da frase.

- **Linguística**

- Tag de parte do discurso criada com a biblioteca spacy;
- Lema do *token* também criado com a biblioteca spacy.

Para o reconhecimento das entidades nomeadas, utilizamos os modelos Perceptron, MNB, CRF, BiLSTM e realizamos o ajuste fino do BERTimbau para as entidades clínicas do conjunto de dados. Nas duas primeiras abordagens, com os modelos Perceptron e MNB, a matriz resultante da etapa de extração de características customizadas foi convertida para um dicionário e todas as strings foram mapeadas utilizando a codificação binária *one-hot*. Uma busca de hiperparâmetros foi realizada para os dois classificadores, com o Perceptron sendo inicializado com o número máximo de iterações igual a 50, e a MNB com o parâmetro de suavização aditivo de 0.01. Na Tabela 9, os hiperparâmetros selecionados para cada modelo são apresentados.

No método com o modelo CRF, a etapa de vetorização após a extração de características não foi necessária, já que o modelo implementa a própria vetorização, com as características sendo formatadas em um dicionário. Assim como nos dois primeiros modelos, uma busca de hiperparâmetros foi realizada e os coeficientes de regularização L1 e L2 foram definidos como 0,09362135735093031 e 0,29134934387516137, respectivamente, e o número máximo

Tabela 9 – Parâmetros selecionados por modelo.

Parâmetro	Valor
Perceptron	
Número máximo de iterações	50
Naïve Bayes Multinomial	
Alpha	0,01
CRF	
Algoritmo de treinamento	L-BFGS
Coefficiente de regularização L1	0,09362135735093031
Coefficiente de regularização L2	0,2913493438751613
Número máximo de iterações	100
BiLSTM	
Número de épocas	5
Tamanho do lote	1
Otimizador	Adam
Número de neurônios	352
Taxa de <i>dropout</i>	0,1
Algoritmo de ativação	Softmax
BERTimbau com ajuste fino	
Número de épocas	10
Tamanho do lote	4
Taxa de aprendizado	3e-5
Weight decay	0,01

de iterações como 100. O algoritmo de treinamento escolhido foi o método L-BFGS (LIU; NOCEDAL, 1989).

Com relação ao quarto método, utilizamos a versão pré-treinada em português com dimensão 100 do modelo Glove ², baseado no trabalho (SILVA; OLIVEIRA, 2022). A escolha da dimensão desse modelo se deu pelo interesse em se obter um vetor de representação que capturasse informações, mas que não fosse o mais complexo computacionalmente dentre a lista disponibilizada. Após isso, criamos o vocabulário do trabalho com base nos *tokens* únicos encontrados no conjunto de dados, com tamanho final de 1524, e depois geramos a matriz de representação através desse vocabulário, obtendo uma matriz de dimensão de 1524×100. Um *padding* de 238, tamanho da maior sentença, foi aplicado em todas as receitas. Uma busca por hiperparâmetros também foi realizada com o *framework* KerasTuner (MANASWI; MANASWI, 2018), sendo eles o número de neurônios na camada BiLSTM, a taxa de *dropout*, o algoritmo de otimização, o número de épocas e o tamanho do lote. Assim, definimos a arquitetura da rede BiLSTM. A camada inicial foi a de *embedding* com a dimensão de entrada de 1524, sendo o tamanho do vocabulário, e inicializada com os pesos da matriz de representação. A camada

² <http://www.nilc.icmc.usp.br/embeddings>

seguinte foi a Bidirecional LSTM, com número de neurônios de 352 e *dropout* de 0,1. A última camada foi a *Dense*, com o algoritmo de ativação softmax e a dimensionalidade da saída de 15, número de entidades do trabalho. Os valores escolhidos foram 1 para o tamanho do lote e 5 para as épocas.

Por fim, para o ajuste fino, a versão pré-treinada em português do BERT foi selecionada, o BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020). Esse modelo de linguagem foi escolhido com o intuito de avaliar o uso de uma rede com ajuste fino pré-treinada em português para a tarefa de NER clínico e analisar como essa rede se comportaria em comparação com outras abordagens de extração. Utilizamos o 'bert-base-portuguese-case', tanto como tokenizador quanto como modelo de classificação de *tokens*. O *padding* das receitas também foi de acordo com o tamanho da maior sentença. Implementamos o ajuste fino e executamos o processo de treinamento usando os módulos e a API Trainer da biblioteca HuggingFace Transformer(WOLF *et al.*, 2020), que é otimizada e oferece uma ampla gama de opções de treinamento e recursos integrados. Os hiperparâmetros termo de regularização, também conhecido como decaimento de peso (*weight decay*), assim como o número de épocas, a taxa de aprendizado e o tamanho do lote foram definidos de acordo com os valores utilizados no trabalho (SCHNEIDER *et al.*, 2020).

5 RESULTADOS E DISCUSSÕES

Neste capítulo, apresentamos e discutimos os resultados obtidos na definição dos componentes da *ensemble* proposta e na classificação da base de textos clínicos, assim como a rotulação do conjunto de dados para NER e a comparação dos modelos utilizados para o reconhecimento de entidades clínicas.

5.1 Resultados da definição dos modelos de vetorização

A Tabela 10 apresenta os resultados obtidos na avaliação de diferentes técnicas de vetorização com a SVM. As *embeddings* produzidas pelo BERTimbau e o N_gram+TF-IDF com rótulos verdadeiros e fracos tiveram os melhores desempenhos. O Word2Vec também obteve resultados com valores acima de 0.90 nas métricas, porém seus resultados ainda foram inferiores às abordagens já citadas. Das abordagens avaliadas, o Doc2Vec teve o pior desempenho.

Dois fatores explicam os resultados obtidos pelo Doc2Vec: (1) este modelo geralmente requer documentos mais longos e informativos, e (2) para obter resultados mais conclusivos, o Doc2Vec necessita de um vocabulário rico e diversificado, ou seja, com maior multiplicidade de *tokens* únicos. Nosso conjunto de dados possui, em sua maioria, documentos curtos e uma extensa repetição de *tokens* nas amostras. Essa repetição ocorre por meio do uso já difundido de uma estrutura padrão para a criação desses textos clínicos. Assim, Word2Vec utiliza associação de palavras e tende a ter melhores resultados com o conjunto de dados proposto do que modelos para representação de documentos.

As representações do conjunto de dados com rótulos verdadeiros com os métodos BERTimbau e N_gram+TF-IDF obtiveram os mesmos valores nas métricas. O BERTimbau, um modelo pré-treinado em português brasileiro e um método de embutir palavras que retorna diferentes vetores para o mesmo termo, dependendo do contexto de uso, pode criar *embeddings* fortemente contextuais. Assim como N_gram+TF-IDF pode gerar associações de palavras através da sequência de *tokens* utilizados, mantendo o contexto de cada amostra. Esses modelos produzem representações contextuais e encontram padrões nas amostras que facilitam a discriminação entre as classes. Com relação aos resultados com o *framework* Snorkel, por se tratar de uma metodologia em que foram utilizados rótulos fracos, esses resultados são animadores, pois são numericamente comparáveis a abordagens treinadas apenas pelo conjunto de dados com rótulos de verdade de campo, obtendo resultados como 0.96 ± 0.03 em acurácia, precisão, recall

Tabela 10 – Resultados obtidos usando combinações de *embeddings* com o classificador SVM. Em negrito estão os melhores resultados.

Embedding	Acc	Kap	Prec	Rec	F1	AUC
BERTimbau	0.99±0.01	0.98±0.02	0.99±0.01	0.99±0.01	0.99±0.01	1.00±0.00
Doc2Vec	0.73±0.02	0.59±0.03	0.74±0.02	0.73±0.02	0.73±0.02	0.91±0.02
N_gram+TF-IDF	0.99±0.01	0.98±0.02	0.99±0.01	0.99±0.01	0.99±0.01	1.00±0.00
Snorkel+N_gram+TF-IDF	0.96±0.03	0.94±0.04	0.96±0.03	0.96±0.03	0.96±0.03	0.99±0.01
Word2Vec	0.94±0.02	0.91±0.03	0.94±0.02	0.94±0.02	0.94±0.02	1.00±0.00

Tabela 11 – Resultados obtidos com a predição.

Classificador	<i>embeddings</i>	Acc	Kap	Prec	Rec	F1	AUC
RF	N_grama+TF-IDF	0.99	0.99	0.99	0.99	0.99	1.00
	Snorkel+N_grama+TF-IDF	0.97	0.95	0.97	0.97	0.97	1.00
	BERTimbau	0.98	0.96	0.98	0.98	0.98	1.00
MLP	N_grama+TF-IDF	0.99	0.99	0.99	0.99	0.99	1.00
	Snorkel + N_grama + TF-IDF	0.97	0.95	0.97	0.97	0.97	1.00
	BERTimbau	0.99	0.99	0.99	0.99	0.99	1.00
SVM	N_grama+TF-IDF	0.99	0.99	0.99	0.99	0.99	1.00
	Snorkel + N_grama + TF-IDF	0.99	0.98	0.99	0.99	0.99	1.00
	BERTimbau	0.98	0.96	0.98	0.98	0.98	1.00
<i>Ensemble proposta</i>		0.99	0.99	0.99	0.99	0.99	-

e pontuação F1, 0.94 ± 0.04 em kappa e 0.99 ± 0.01 em AUC.

5.2 Resultados da classificação

Tabela 11 apresenta os resultados de classificação obtidos com o conjunto de dados de teste. Há significativa similaridade nos resultados da classificação, independentemente da abordagem utilizada. Com uma estruturação particular já amplamente utilizada em sua criação, a categorização do conjunto de dados não se torna uma tarefa complexa. Amostras de prescrição, por exemplo, seguem um padrão formativo com características como via de administração, dosagem e nome do medicamento. As solicitações de exames apresentam informações como o nome do procedimento ou exame e a área ou local do procedimento. Essa padronização, aliada ao contexto de cada amostra mantido pelos *embeddings*, torna a discretização entre as classes altamente eficiente.

A Tabela 12 compara o número de amostras erroneamente classificadas por categoria obtidas com o *ensemble* e os classificadores únicos que o constituem. Numericamente, a abordagem N_gram+TF-IDF utilizando o classificador MLP teve o menor número de erros, embora o uso desse método de vetorização produza os melhores desempenhos independentemente

Tabela 12 – Distribuição do número de amostras erroneamente classificadas na predição.

Classificador	<i>embeddings</i>	Receitas	Notas	Exames	Total
RF	N_gram+TF-IDF	1	1	5	07
	Snorkel+N_gram+TF-IDF	2	18	7	27
	BERTimbau	1	13	8	22
MLP	N_gram+TF-IDF	0	2	4	06
	Snorkel+N_gram+TF-IDF	1	23	6	30
	BERTimbau	1	4	3	08
SVM	N_gram+TF-IDF	0	1	6	07
	Snorkel+N_gram+TF-IDF	0	5	5	10
	BERTimbau	1	13	8	22
<i>Ensemble</i>		0	4	4	08

do classificador utilizado. Outras abordagens de classificação também obtiveram menos erros, como BERTimbau com a MLP e o conjunto de dados fracamente rotulado com N_gram+TF-IDF com a SVM.

A classe de receitas é a que oferece menor dificuldade em termos de discretização em comparação com as outras categorias (vide Fig. 4). A classe de solicitação de exames obteve uma padronização de erros independente da metodologia. Na classe de notas clínicas, os classificadores treinados com o Snorkel+N_gram+TF-IDF tiveram os piores desempenhos, exceto com SVM; a rotulagem fraca justifica esse comportamento. O Snorkel teve dificuldade em discretizar esta classe, pois havia amostras com elementos ambíguos ou que não ofereciam informações suficientes para serem identificadas. Porém, o Snorkel obteve desempenho semelhante com o conjunto de dados de rótulos validados, provando sua eficácia na criação de conjuntos de dados de treinamento fracamente rotulados.

É importante também mencionar o uso de uma *ensemble* para a classificação de documentos clínicos. As *ensembles* produzem um desempenho mais preciso nos resultados do que classificadores únicos, pois as previsões de diferentes fontes são mescladas, reduzindo assim a propagação ou dispersão das previsões e do desempenho do modelo. Desta forma, o uso de *ensembles* produz modelos com maior robustez ou confiabilidade em desempenho médio.

5.3 Reconhecimento das entidades nomeadas clínicas

Todos os resultados mostrados nessa seção foram obtidos usando *k-fold* igual a 10, ou seja, os valores são suas respectivas médias e desvios-padrão. Na Tabela 13, são apresentados os resultados dos experimentos com as abordagens de classificação. Observa-se que os modelos

Tabela 13 – Resultados obtidos na extração de entidades clínicas em receitas médicas.

Modelo	Acurácia	Precisão	Recall	F1-score
Custom features + Perceptron	0.70±0.145	0.59±0.136	0.58±0.106	0.52±0.135
Custom features + Naive Bayes Multinomial	0.92±0.019	0.87±0.044	0.92±0.019	0.88±0.034
Custom features + CRF	0.98±0.004	0.98±0.010	0.96±0.011	0.97±0.009
Glove + BiLSTM	0.99±0.001	0.99±0.013	0.99±0.011	0.99±0.012
Finetuned BERTimbau	0.99±0.003	0.99±0.011	0.99±0.010	0.99±0.011

de aprendizado profundo e o CRF superaram as outras duas abordagens, com o pior desempenho sendo obtido pela Perceptron com a extração das características customizadas.

Embora os resultados demonstrem que os métodos de aprendizado de máquina são eficazes, eles dependem fortemente do corpus, tanto com relação ao tamanho quanto a seleção de características para a identificação de padrões. Por outro lado, os métodos baseados em aprendizado profundo podem incorporar as informações de contexto e evitar o custoso trabalho de seleção de características, já que eles são eficazes na aprendizagem automática de representações úteis. Com relação ao modelo CRF, método dominante na resolução de problemas NER antes da prevalência das redes neurais, é importante apontar os excelentes resultados obtidos e o custo computacional menor comparado a outras abordagens de NER (AMARAL; VIEIRA, 2014), produzindo resultados acima de 0,95 em todas as métricas de avaliação.

Outro ponto importante é quanto ao uso de características customizadas conforme a estrutura textual das receitas e o uso de *embeddings* criados por modelos de linguagem. Por ser uma metodologia de extração de características facilmente implementada e com um custo computacional menor, a primeira abordagem é uma opção adequada para um método *baseline*, entretanto é necessária uma etapa de engenharia de características e conhecimento de domínio para a correta seleção de características a serem utilizadas na metodologia. Já as abordagens com modelos de linguagem, mesmo não sendo pré-treinadas no domínio, mas apenas no idioma do conjunto de dados, extraem automaticamente as características mais relevantes e úteis dos dados, representando-os conforme as semelhanças e relações entre as palavras. Assim, eles conseguem produzir vetorizações contextuais dos *tokens* e diminuir a complexidade da representação.

Na Tabela 14, são apresentados os resultados nas métricas de avaliação de acordo com as entidades clínicas. VIA foi a entidade que obteve os maiores valores nas métricas em comparação com as outras entidades, independente da abordagem utilizada. Em contrapartida, a entidade APR foi a que teve o pior desempenho nessa mesma comparação. Esses resultados ocorreram devido à quantidade e o formato das amostras encontradas no conjunto de dados.

APR é a classe com o menor número de amostras e também a entidade com a maior discrepância quanto ao entendimento na rotulação. Já VIA é a terceira classe com a maior quantidade de amostras e, diferentemente das duas primeiras (FREQ e QTD), possui grande padronização quanto ao formato em que são constituídos os *tokens*.

Como observado na Tabela 13, a Perceptron teve resultados inferiores nas entidades em relação aos demais modelos. Isso se deu por ser um modelo linear, o que faz com que haja erros na classificação de um conjunto de dados em que as classes se sobrepõem e compartilham vocabulário semelhante. Já o BERTimbau com ajuste fino, no geral, superou as outras metodologias em todas as entidades, com apenas MED e QTD tendo desvios-padrão superiores no F1-score que as metodologias CRF e BiLSTM. Esse tipo de modelo de linguagem produz representações contextuais fortes através de padrões linguísticos, o que facilita a discriminação entre classes. Além disso, o uso da técnica de ajuste fino permitiu que o modelo alcançasse ótimo desempenho mesmo com o uso de um conjunto de dados pequeno, devido aos pesos pré-treinados do modelo na tarefa original.

Tabela 14 – Resultados obtidos segundo as classes da extração de entidades clínicas em receitas médicas.

Modelo	Entidade	Precisão	Recall	F1-score
Custom features + Perceptron	APR	0.04±0.065	0.42±0.420	0.07±0.095
	DOS	0.75±0.336	0.62±0.316	0.65±0.320
	DUR	0.57±0.399	0.62±0.411	0.59±0.392
	FREQ	0.64±0.414	0.42±0.303	0.47±0.317
	MED	0.46±0.397	0.37±0.405	0.36±0.344
	QTD	0.61±0.287	0.76±0.323	0.63±0.250
	VIA	0.90±0.217	0.71±0.225	0.74±0.224
Custom features + MNB	APR	0.64±0.266	0.95±0.218	0.76±0.237
	DOS	0.89±0.048	0.93±0.042	0.91±0.033
	DUR	0.83±0.159	0.92±0.086	0.86±0.116
	FREQ	0.92±0.039	0.84±0.117	0.87±0.077
	MED	0.90±0.140	0.95±0.045	0.92±0.095
	QTD	0.83±0.151	0.93±0.064	0.87±0.099
	VIA	0.99±0.007	0.99±0.008	0.99±0.005
Custom features + CRF	APR	0.92±0.226	0.81±0.232	0.86±0.221
	DOS	0.98±0.022	0.97±0.039	0.97±0.022
	DUR	0.99±0.022	0.98±0.017	0.98±0.010
	FREQ	0.97±0.016	0.97±0.017	0.97±0.013
	MED	0.95±0.037	0.93±0.051	0.94±0.037
	QTD	0.98±0.016	0.99±0.016	0.99±0.011
	VIA	0.99±0.005	0.99±0.007	0.99±0.004
Glove + BiLSTM	APR	0.91±0.113	0.91±0.124	0.90±0.101
	DOS	0.99±0.007	0.98±0.040	0.99±0.024
	DUR	0.98±0.017	0.99±0.013	0.98±0.012
	FREQ	0.98±0.024	0.99±0.015	0.98±0.019
	MED	0.99±0.010	0.99±0.007	0.99±0.008
	QTD	0.98±0.026	0.99±0.012	0.99±0.019
	VIA	0.99±0.005	0.99±0.004	0.99±0.004
Finetuned BERTimbau	APR	0.95±0.082	0.98±0.060	0.97±0.068
	DOS	1.00±0.000	0.99±0.036	0.99±0.019
	DUR	0.99±0.005	0.99±0.005	0.99±0.005
	FREQ	0.99±0.005	0.99±0.005	0.99±0.005
	MED	0.99±0.012	0.99±0.010	0.99±0.011
	QTD	0.99±0.030	0.99±0.011	0.99±0.020
	VIA	0.99±0.003	0.99±0.004	0.99±0.002

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, uma base de documentos clínicos em português brasileiro foi construída. A partir dessa base, foram propostos uma *ensemble* de classificação e a comparação de abordagens de reconhecimento de entidade nomeadas. Através da metodologia de classificação, concluímos que a estratégia de supervisão fraca pode treinar modelos de aprendizado de máquina com alta eficiência na rotulagem de textos clínicos.

No tocante a comparação de métodos NER, nossos experimentos mostram que as abordagens de aprendizado profundo obtêm resultados superiores às abordagens de aprendizado de máquina. Com esses resultados, concluímos que o uso de *embeddings* de palavras e modelos de DL têm melhor desempenho do que o uso de características customizadas e classificadores de aprendizado de máquina, evidenciando que a aprendizagem por transferência com modelos de domínio pode beneficiar tarefas clínicas, de forma estatisticamente significativa.

Como limitações deste trabalho, apontamos o tamanho do conjunto de dados com 3000 amostras rotuladas e sua alta padronização em termos de formatos de amostra por classe. Ou seja, existe um padrão evidente nos textos de acordo com a categoria de documento analisada. Assim, como trabalhos futuros, adicionaremos amostras que não contenham os elementos típicos das classes e dividiremos as classes em outras categorias de interesse, como prescrições de industrializados e manipulados, encaminhamentos, atestados médicos, entre outros. Além disso, avaliaremos a extração de relação entre entidades para a realização de testes de interações medicamentosas.

REFERÊNCIAS

- AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. **Mining text data**, Springer, p. 163–222, 2012.
- ALBAWI, S.; MOHAMMED, T. A.; AL-ZAWI, S. Understanding of a convolutional neural network. In: IEEE. **2017 international conference on engineering and technology (ICET)**. [S.l.], 2017. p. 1–6.
- ALDOUS, D. The continuum random tree iii. **The Annals of Probability**, JSTOR, p. 248–289, 1993.
- AMARAL, D. O. F. do; VIEIRA, R. Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. **Linguamática**, v. 6, n. 1, p. 41–49, 2014.
- ANANDARAJAN, M. *et al.* Text preprocessing. **Practical text analytics: Maximizing the value of text data**, Springer, p. 45–59, 2019.
- ARTSTEIN, R.; POESIO, M. Inter-coder agreement for computational linguistics. **Computational linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 34, n. 4, p. 555–596, 2008.
- AYYADEVARA, V. K.; AYYADEVARA, V. K. Word2vec. **Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R**, Springer, p. 167–178, 2018.
- BACH, S. H. *et al.* Snorkel drybell: A case study in deploying weak supervision at industrial scale. In: **Proceedings of the 2019 International Conference on Management of Data**. [S.l.: s.n.], 2019. p. 362–375.
- BARATLOO, A. *et al.* Evidence based emergency medicine; part 1: Simple definition and calculation of accuracy, sensitivity and specificity. **Emergency**, v. 3, p. 48–49, 05 2015.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.
- BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.
- BROWN, P. F. *et al.* Class-based n-gram models of natural language. **Computational linguistics**, v. 18, n. 4, p. 467–480, 1992.
- BURGES, C. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery**, Springer, v. 2, n. 2, p. 121–167, 1998.
- CABITZA, F. *et al.* The elephant in the record: On the multiplicity of data recording work:. **Health Informatics Journal**, SAGE PublicationsSage UK: London, England, v. 25, p. 475–490, 1 2019.
- CAVNAR, W. B.; TRENKLE, J. M. *et al.* N-gram-based text categorization. In: LAS VEGAS, NV. **Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval**. [S.l.], 1994. v. 161175.

- CAÑETE, J. *et al.* Spanish pre-trained bert model and evaluation data. In: **PML4DC at ICLR 2020**. [S.l.: s.n.], 2020.
- CER, D. *et al.* Universal sentence encoder. **arXiv preprint arXiv:1803.11175**, 2018.
- CHAUDHRY, B. *et al.* Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. **Annals of internal medicine**, American College of Physicians, v. 144, n. 10, p. 742–752, 2006.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, ACM, 2016.
- CHIU, J. P.; NICHOLS, E. Named entity recognition with bidirectional lstm-cnns. **Transactions of the association for computational linguistics**, MIT Press, v. 4, p. 357–370, 2016.
- CHURCH, K. W. Word2vec. **Natural Language Engineering**, Cambridge University Press, v. 23, n. 1, p. 155–162, 2017.
- COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.
- CONNEAU, A. *et al.* Unsupervised cross-lingual representation learning at scale. **arXiv preprint arXiv:1911.02116**, 2019.
- CUI, M. *et al.* Regular expression based medical text classification using constructive heuristic approach. **IEEE Access**, v. 7, p. 147892–147904, 2019.
- CUSICK, M. *et al.* Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. **Journal of psychiatric research**, Elsevier, v. 136, p. 95–102, 2021.
- DALAL, M. K.; ZAVERI, M. A. Automatic text classification: a technical review. **International Journal of Computer Applications**, Citeseer, v. 28, n. 2, p. 37–40, 2011.
- DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.
- DONG, X. *et al.* A survey on ensemble learning. **Frontiers of Computer Science**, Springer, v. 14, n. 2, p. 241–258, 2020.
- DZISEVIČ, R.; ŠEŠOK, D. Text classification using different feature extraction approaches. In: IEEE. **2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)**. [S.l.], 2019. p. 1–4.
- FILHO, J. A. W. *et al.* The brwac corpus: a new open resource for brazilian portuguese. In: **Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)**. [S.l.: s.n.], 2018.
- FREUND, Y.; SCHAPIRE, R. E. Large margin classification using the perceptron algorithm. In: **Proceedings of the eleventh annual conference on Computational learning theory**. [S.l.: s.n.], 1998. p. 209–217.

- GARDNER, M. W.; DORLING, S. R. **Artificial Neural Networks (The Multilayer Perceptron) A Review of Applications in the Atmospheric Sciences**. [S.l.], 1998. v. 32, 627-263 p.
- GRISHMAN, R. Information extraction. **IEEE Intelligent Systems**, IEEE, v. 30, n. 5, p. 8–15, 2015.
- GRISHMAN, R.; SUNDHEIM, B. M. Message understanding conference-6: A brief history. In: **COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics**. [S.l.: s.n.], 1996.
- GUPTA, S. *et al.* Automated identification of patients with immune-related adverse events from clinical notes using word embedding and machine learning. **JCO clinical cancer informatics**, Wolters Kluwer Health, v. 5, p. 541–549, 2021.
- HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982.
- HENRY, S. *et al.* 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 27, n. 1, p. 3–12, 2020.
- HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural Networks**, Pergamon, v. 2, p. 359–366, 1 1989. ISSN 0893-6080.
- HOSNI, M. *et al.* Reviewing ensemble classification methods in breast cancer. **Computer methods and programs in biomedicine**, Elsevier, v. 177, p. 89–112, 2019.
- JOHNSON, A. E. *et al.* MIMIC-III, a freely accessible critical care database. **Scientific data**, Nature Publishing Group, v. 3, n. 1, p. 1–9, 2016.
- JOULIN, A. *et al.* FastText.zip: Compressing text classification models. **arXiv preprint arXiv:1612.03651**, 2016.
- JURAFSKY, D. **Speech & language processing**. [S.l.]: Pearson Education India, 2000.
- KAMBAR, M. E. Z. N. *et al.* Clinical text classification of alzheimer's drugs' mechanism of action. In: SPRINGER. **Proceedings of Sixth International Congress on Information and Communication Technology**. [S.l.], 2022. p. 513–521.
- KETKAR, N. Stochastic gradient descent. In: **Deep learning with Python**. [S.l.]: Springer, 2017. p. 113–132.
- KIBRIYA, A. M. *et al.* Multinomial naive bayes for text categorization revisited. In: SPRINGER. **AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17**. [S.l.], 2005. p. 488–499.
- KIM, D. *et al.* Multi-co-training for document classification using various document representations: Tf-idf, lda, and doc2vec. **Information Sciences**, Elsevier, v. 477, p. 15–29, 2019.
- KUMAR, V. *et al.* Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. **IEEE Access**, IEEE, v. 9, p. 7107–7126, 2020.
- KUNCHEVA, L. I. **Combining pattern classifiers: methods and algorithms**. [S.l.]: John Wiley & Sons, 2014.

- LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: **Proceedings of the Eighteenth International Conference on Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001. (ICML '01), p. 282–289. ISBN 1558607781.
- LAMPLE, G. *et al.* Neural architectures for named entity recognition. **arXiv preprint arXiv:1603.01360**, 2016.
- LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **biometrics**, JSTOR, p. 159–174, 1977.
- LEE, J. *et al.* Open-access mimic-ii database for intensive care research. In: IEEE. **2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society**. [S.l.], 2011. p. 8315–8318.
- LI, J. *et al.* A survey on deep learning for named entity recognition. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 34, n. 1, p. 50–70, 2020.
- LI, M. D. *et al.* Automated radiology-arthroscopy correlation of knee meniscal tears using natural language processing algorithms. **Academic radiology**, Elsevier, v. 29, n. 4, p. 479–487, 2022.
- LIU, D. C.; NOCEDAL, J. On the limited memory bfgs method for large scale optimization. **Mathematical programming**, Springer, v. 45, n. 1-3, p. 503–528, 1989.
- LOPES, F.; TEIXEIRA, C.; OLIVEIRA, H. G. Named entity recognition in portuguese neurology text using crf. In: SPRINGER. **Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part I 19**. [S.l.], 2019. p. 336–348.
- LOPES, F.; TEIXEIRA, C.; OLIVEIRA, H. G. Comparing different methods for named entity recognition in portuguese neurology text. **Journal of Medical Systems**, Springer, v. 44, p. 1–20, 2020.
- LÓPEZ-ÚBEDA, P. *et al.* Automatic medical protocol classification using machine learning approaches. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 200, p. 105939, 2021.
- MAMMONE, A.; TURCHI, M.; CRISTIANINI, N. Support vector machines. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, v. 1, n. 3, p. 283–289, 2009.
- MANASWI, N. K.; MANASWI, N. K. Understanding and working with keras. **Deep learning with applications using Python: Chatbots and face, object, and speech recognition with TensorFlow and Keras**, Springer, p. 31–43, 2018.
- MARRERO, M. *et al.* Named entity recognition: fallacies, challenges and opportunities. **Computer Standards & Interfaces**, Elsevier, v. 35, n. 5, p. 482–489, 2013.
- MENACHEMI, N.; COLLUM, T. H. Benefits and drawbacks of electronic health record systems. **Risk management and healthcare policy**, Taylor & Francis, p. 47–55, 2011.
- MIKOLOV, T. *et al.* Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

- MIKOLOV, T.; LE, Q. V.; SUTSKEVER, I. Exploiting similarities among languages for machine translation. **arXiv preprint arXiv:1309.4168**, 2013.
- MOHIT, B. Named entity recognition. **Natural language processing of semitic languages**, Springer, p. 221–245, 2014.
- MOSCHITTI, A. A study on optimal parameter tuning for rocchio text classifier. In: SPRINGER. **European Conference on Information Retrieval**. [S.l.], 2003. p. 420–435.
- MUJTABA, G. *et al.* Clinical text classification research trends: Systematic literature review and open issues. **Expert systems with applications**, Elsevier, v. 116, p. 494–520, 2019.
- NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. **Lingvisticae Investigationes**, John Benjamins, v. 30, n. 1, p. 3–26, 2007.
- OLSON, D. L.; DELEN, D. **Advanced data mining techniques**. [S.l.]: Springer Science & Business Media, 2008.
- PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543.
- PETERS, A. C. *et al.* Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. **Journal of Biomedical Semantics**, BioMed Central, v. 13, n. 1, p. 1–19, 2022.
- PETERSON, L. E. K-nearest neighbor. **Scholarpedia**, v. 4, n. 2, p. 1883, 2009.
- PIRES, T.; SCHLINGER, E.; GARRETTE, D. How multilingual is multilingual bert? **arXiv preprint arXiv:1906.01502**, 2019.
- RAJPUT, A. *et al.* J48 and jrip rules for e-governance data. **International Journal of Computer Science and Security (IJCSS)**, Citeseer, v. 5, n. 2, p. 201, 2011.
- RAMOS, J. *et al.* Using tf-idf to determine word relevance in document queries. In: CITeseer. **Proceedings of the first instructional conference on machine learning**. [S.l.], 2003. v. 242, n. 1, p. 29–48.
- RATNER, A. *et al.* Snorkel: Rapid training data creation with weak supervision. **The VLDB Journal**, Springer, v. 29, n. 2, p. 709–730, 2020.
- RISH, I. *et al.* An empirical study of the naive bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46.
- RUMELHART, D.; HINTON, G.; WILLIAMS, R. **Learning representations by back-propagating errors**. [S.l.], 1986.
- SAFAVIAN, S. R.; LANDGREBE, D. A survey of decision tree classifier methodology. **IEEE transactions on systems, man, and cybernetics**, IEEE, v. 21, n. 3, p. 660–674, 1991.
- SAINI, R.; GHOSH, S. K. Ensemble classifiers in remote sensing: A review. In: IEEE. **2017 International Conference on Computing, Communication and Automation (ICCCA)**. [S.l.], 2017. p. 1148–1152.

- SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information processing & management**, Elsevier, v. 24, n. 5, p. 513–523, 1988.
- SANG, E. F. T. K.; VEENSTRA, J. Representing text chunks. In: **Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics**. [S.l.: s.n.], 1999. p. 173–179.
- SANTOS, H. D. P. d. *et al.* An initial investigation of the charlson comorbidity index regression based on clinical notes. In: IEEE. **2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)**. [S.l.], 2018. p. 6–11.
- SCHILLINGER, D. *et al.* Closing the loop: physician communication with diabetic patients who have low health literacy. **Archives of internal medicine**, American Medical Association, v. 163, n. 1, p. 83–90, 2003.
- SCHNEIDER, E. T. R. *et al.* Biobertpt-a portuguese neural language model for clinical named entity recognition. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 3rd Clinical Natural Language Processing Workshop**. [S.l.], 2020. p. 65–72.
- SCHONLAU, M.; GUENTHER, N.; SUCHOLUTSKY, I. Text mining with n-gram variables. **The Stata Journal**, SAGE Publications Sage CA: Los Angeles, CA, v. 17, n. 4, p. 866–881, 2017.
- SCHUSTER, M.; PALIWAL, K. K. Bidirectional recurrent neural networks. **IEEE transactions on Signal Processing**, Ieee, v. 45, n. 11, p. 2673–2681, 1997.
- SHI, S. *et al.* Applications of blockchain in ensuring the security and privacy of electronic health record systems: A survey. **Computers & security**, Elsevier, v. 97, p. 101966, 2020.
- SILVA, D. A. da *et al.* Predicting the occurrence of surgical site infections using text mining and machine learning. **PloS one**, Public Library of Science San Francisco, CA USA, v. 14, n. 12, p. e0226272, 2019.
- SILVA, M. G. da; OLIVEIRA, H. T. A. de. Combining word embeddings for portuguese named entity recognition. In: SPRINGER. **Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings**. [S.l.], 2022. p. 198–208.
- SOUSA, F. S. *et al.* Categorização automática de conteúdos web de saúde em português brasileiro com classificador bayesiano. **Journal of Health Informatics**, v. 4, n. 1, 2012.
- SOUSA, O. L. de *et al.* Deep learning in image analysis for covid-19 diagnosis: a survey. **IEEE Latin America Transactions**, IEEE, v. 19, n. 6, p. 925–936, 2021.
- SOUSA, O. L. V. de *et al.* Natural language processing for clinical data classification. **iSys - Brazilian Journal of Information Systems**, v. 15, n. 1, p. 13:1–13:17, 2022.
- SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. **Brazilian Conference on Intelligent Systems**. [S.l.], 2020. p. 403–417.
- SOUZA, J. V. A. de *et al.* Named entity recognition for clinical portuguese corpus with conditional random fields and semantic groups. In: SBC. **Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde**. [S.l.], 2019. p. 318–323.

- SOUZA, J. V. A. de *et al.* A multilabel approach to portuguese clinical named entity recognition. **Journal of Health Informatics**, v. 12, 2020.
- STIELL, A. *et al.* Prevalence of information gaps in the emergency department and the effect on patient outcomes. **Cmaj**, Can Med Assoc, v. 169, n. 10, p. 1023–1028, 2003.
- STURMAN, D. J.; ZELTZER, D. A survey of glove-based input. **IEEE Computer graphics and Applications**, IEEE, v. 14, n. 1, p. 30–39, 1994.
- SUTTON, C.; MCCALLUM, A. *et al.* An introduction to conditional random fields. **Foundations and Trends® in Machine Learning**, Now Publishers, Inc., v. 4, n. 4, p. 267–373, 2012.
- SWAIN, P. H.; HAUSKA, H. The decision tree classifier: Design and potential. **IEEE Transactions on Geoscience Electronics**, IEEE, v. 15, n. 3, p. 142–147, 1977.
- TANG, P. C.; MCDONALD, C. J. Electronic health record systems. **Biomedical informatics: computer applications in health care and biomedicine**, Springer, p. 447–475, 2006.
- UYSAL, A. K.; GUNAL, S. The impact of preprocessing on text classification. **Information processing & management**, Elsevier, v. 50, n. 1, p. 104–112, 2014.
- VIEIRA, P. *et al.* Detecting pulmonary diseases using deep features in x-ray images. **Pattern Recognition**, Elsevier, v. 119, p. 108081, 2021.
- WANG, D.; SU, J.; YU, H. Feature extraction and analysis of natural language processing for deep learning english language. **IEEE Access**, IEEE, v. 8, p. 46335–46345, 2020.
- WANG, Y. *et al.* A clinical text classification paradigm using weak supervision and deep representation. **BMC medical informatics and decision making**, Springer, v. 19, n. 1, p. 1–13, 2019.
- WEHNERT, S. *et al.* Applying bert embeddings to predict legal textual entailment. **The Review of Socionetwork Strategies**, Springer, p. 1–23, 2022.
- WEI, Q. *et al.* A study of deep learning approaches for medication and adverse drug event extraction from clinical text. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 27, n. 1, p. 13–21, 2020.
- WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics and intelligent laboratory systems**, Elsevier, v. 2, n. 1-3, p. 37–52, 1987.
- WOLF, T. *et al.* Transformers: State-of-the-art natural language processing. In: **Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations**. [S.l.: s.n.], 2020. p. 38–45.
- WRIGHT, R. E. **Logistic regression**. [S.l.]: American Psychological Association, 1995.
- WU, S.; MANBER, U. Fast text searching: allowing errors. **Communications of the ACM**, ACM New York, NY, USA, v. 35, n. 10, p. 83–91, 1992.
- XU, S.; LI, Y.; WANG, Z. Bayesian multinomial naïve bayes classifier to text classification. In: SPRINGER. **Advanced Multimedia and Ubiquitous Engineering: MUE/FutureTech 2017 11**. [S.l.], 2017. p. 347–352.

YATSKO, V. Methods and algorithms for automatic text analysis. **Automatic Documentation and Mathematical Linguistics**, Springer, v. 45, n. 5, p. 224–231, 2011.

YUN-TAO, Z.; LING, G.; YONG-CHENG, W. An improved tf-idf approach for text classification. **Journal of Zhejiang University-SCIENCE A** 2005 6:1, Springer, v. 6, p. 49–55, 8 2005. ISSN 1862-1775.

ZAREI, S.; BOZORG-HADDAD, O.; NIKOO, M. R. The basis of artificial neural network (ann): Structures, algorithms and functions. In: **Computational Intelligence for Water and Environmental Sciences**. [S.l.]: Springer, 2022. p. 225–250.

ZHANG, Y. *et al.* Biowordvec, improving biomedical word embeddings with subword information and mesh. **Scientific data**, Nature Publishing Group, v. 6, n. 1, p. 1–9, 2019.

ZHANG, Y.; JIN, R.; ZHOU, Z.-H. Understanding bag-of-words model: a statistical framework. **International Journal of Machine Learning and Cybernetics**, Springer, v. 1, n. 1, p. 43–52, 2010.

ZHOU, P. *et al.* Attention-based bidirectional long short-term memory networks for relation classification. In: **Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)**. [S.l.: s.n.], 2016. p. 207–212.