**PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ**
**ESCOLA POLITÉCNICA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE**

**ARNON BRUNO VENTRILHO DOS SANTOS**

# CODIFICAÇÃO DE NARRATIVAS CLÍNICAS POR MEIO DO DEEP LEARNING

**CURITIBA**
**2018**

**ARNON BRUNO VENTRILHO DOS SANTOS**

# CODIFICAÇÃO DE NARRATIVAS CLÍNICAS POR MEIO DO DEEP LEARNING

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologia em Saúde da Escola Politécnica da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de mestre em Tecnologia em Saúde.
Área de concentração: Informática em Saúde

**Orientadora**: Profa. Dr.ª Deborah Ribeiro Carvalho

**CURITIBA**

**2018**

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Edilene de Oliveira dos Santos CRB 9 / 1636

Santos, Arnon Bruno Ventrilho dos
S237c Codificação de narrativas clínicas por meio do DEEP Learning / Arnon
2018 Bruno Ventrilho dos Santos ; orientadora, Deborah Ribeiro Carvalho. -- 2018
96 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2018
Bibliografia: f. 81-88

1. Engenharia biomédica. 2. Aprendizado do computador. 3. Processamento de linguagem natural (Computação). 4. Codificação clínica. 5. Aparelho urinário. I. Carvalho, Deborah Ribeiro. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Tecnologia em Saúde. III. Título.

CDD 20. ed. – 610.28

Pontifícia Universidade Católica do Paraná

Escola Politécnica

Programa de Pós Graduação em Tecnologia em Saúde

**ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO**
**PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EM SAÚDE**

**DEFESA DE DISSERTAÇÃO Nº 256**

ÁREA DE CONCENTRAÇÃO: TECNOLOGIA EM SAÚDE

Aos trinta dias do mês de maio de 2018 às 14:00h, no Auditório Carlos Chagas, 2º Andar-Bloco Verde, realizou-se a sessão pública de Defesa da Dissertação: **"CODIFICAÇÃO DE NARRATIVAS CLÍNICAS POR MEIO DO DEEP LEARNING"** apresentado pelo aluno Arnon Bruno Ventrilho dos Santos sob orientação da Professora Dr.ª Deborah Ribeiro Carvalho como requisito parcial para a obtenção do título de **Mestre em Tecnologia em Saúde**, perante uma Banca Examinadora composta pelos seguintes membros:

**Prof.ª Dr.ª** Deborah Ribeiro Carvalho
PUCPR (Presidente) _____ (assinatura) Aprovado (Aprov/Reprov.)

**Prof. Dr.** Emerson Cabrera Paraiso
PUCPR (Examinador) _____ (assinatura) Aprovado (Aprov/Reprov.)

**Prof. Dr.** Myriam Regattieri Delgado
UTFPR (Examinador) _____ (assinatura) Aprovado (Aprov/Reprov.)

Início: 14:00 Término: 16:30

Conforme as normas regimentais do PPGTS e da PUCPR, o trabalho apresentado foi considerado APROVADO (aprovado/reprovado), segundo avaliação da maioria dos membros desta Banca Examinadora.
Observações: _____

O(a) aluno(a) está ciente que a homologação deste resultado está condicionada: (I) ao cumprimento integral das solicitações da Banca Examinadora, que determina um prazo de 60 dias para o cumprimento dos requisitos; (II) entrega da dissertação em conformidade com as normas especificadas no Regulamento do PPGTS/PUCPR; (III) entrega da documentação necessária para elaboração do Diploma.

ALUNO(A): Arnon Bruno Ventrilho dos Santos _____ (assinatura)

_____
Profº. Drº. Percy Nohama,
**Coordenador do PPGTS PUCPR**

## AGRADECIMENTOS

Agradeço inicialmente a minha orientadora, professora Dra. Deborah Ribeiro Carvalho por toda atenção, compreensão e ensinamentos durante o desenvolvimento desse projeto. Sem dúvidas, foi com sua doçura e rigor que mais aprendi sobre ciência.

Agradeço ao professor Dr. Renato Camargos Couto e equipe, que tão bem me receberam em sua cidade e na instituição que o Prof. Renato administra, proporcionando valioso suporte intelectual ao meu trabalho.

Agradeço também aos demais professores que, ao longo da minha trajetória acadêmica, sempre me inspiraram de alguma forma e fizeram com que perseguisse o caminho do constante aperfeiçoamento.

Agradeço a minha família, na figura da minha mãe Wilma, sempre presente me dando forças para seguir, meu irmão William com sua atenção e expertise em assuntos dos mais variados e que gentilmente cedeu recursos computacionais para o desenvolvimento desse trabalho, meu pai Adelson pela presença e minha noiva Caroline pelo apoio, carinho e compreensão durante as infindáveis horas de desenvolvimento desse projeto, além das contribuições para com essa pesquisa, mesmo não sendo especialista nos assuntos aqui tratados. Vocês sempre me deram forças para seguir em frente mesmo quando o caminho parecia incerto. Sem vocês, não seria possível! Obrigado!

Agradeço meus amigos Jean Karax, Paulo Silas, Luã Del Zotto, Adriel Smailey, Jailson Carvalho, Alisson Galbine, Clayton Poitevin Jr e Filipe Gusmão por debates enriquecedores e conversas pouco produtivas que sempre ajudaram a ver as coisas de formas diferentes.

Agradeço aos colegas do PPGTS, Marcelo, Barbara, Fernanda, João e Ana e do grupo de PLN, Yohan e Lucas, que dedicaram seu tempo ao prover valiosas contribuições para esse trabalho.

Agradeço ao Luciano Cebula, que viu em mim uma vontade e ajudou a tornar um sonho em realidade.

Por fim, agradeço a CAPES pelo apoio financeiro.

## RESUMO

**Introdução:** A codificação clínica para os códigos da Classificação Internacional de Doenças garante a uniformização dos diagnósticos e do pagamento de contas clínicas. Neste contexto, torna-se relevante a tarefa do profissional codificador, responsável por realizar a leitura das narrativas clínicas e a partir dessas definir quais códigos corresponde aos achados lá existentes. Essa tarefa, no entanto, é morosa, pois predominantemente manual e as estratégias de automação compreendem modelos tradicionais de processamento de linguagem natural, que apresentam dificuldades. A literatura sugere que uma alternativa às estratégias tradicionais é o Deep Learning para o processamento de linguagem natural a partir do uso de representações numéricas do significado semântico das palavras, os *word embeddings*. Isto aplicado aos textos clínicos possibilita que um algoritmo reconheça e aprenda conceitos existentes em narrativas clínicas, o que pode aperfeiçoar o processo de codificação ao sugerir códigos da Classificação Internacional de Doenças para os diagnósticos identificados. **Objetivo:** Conceber um modelo baseado em Deep Learning para o processamento de linguagem natural que classifique sumários de alta de acordo com os códigos da Classificação Internacional de Doenças para outros distúrbios do trato urinário (CID N39). **Método**: Pesquisa de natureza aplicada, exploratória e experimental, composta por três etapas. A primeira etapa é a aquisição das bases de dados, que compreendem os sumários de alta e uma base contendo os nomes próprios do Brasil. A segunda etapa é o pré-processamento dos dados, onde se adequou os textos. A última etapa é o desenvolvimento dos modelos, onde se utilizou os textos pré-processados dos sumários de alta para criar os *word embeddings*usando o algoritmo *GloVe*. Utilizaram-se esses *embeddings* como camada escondida dos modelos testados. Testou-se o modelo *baseline* e a partir dos resultados, verificou-se que esse modelo não conseguiria ser um classificador adequado para o problema em questão, demandando adaptações. Adaptou-se então 20 modelos diferentes a partir do *baseline*. Avaliaram-se os modelos a partir do micro-*Fscore* médio e desvio padrão amostral, utilizando uma estratégia de validação cruzada estratificada (10-*folds*), além da acurácia na classificação de exemplos da base balanceada. **Resultados:**Treinou-se o classificador seguindo o modelo *baseline*, obtendo-se micro-*Fscore* médio de 0,08 e desvio padrão amostral de 0,05, com acurácia de 24% na base de testes complementar. A partir das adaptações, o melhor modelo é aquele que tem como entrada um *embedding* de 500 dimensões, 4 camadas de convolução com *kernel* de tamanho 128 e ativação "*ReLu*", janelas de tamanho 5, 8, 10,12 e 4 camadas de *1-max-pooling*, com amostragem global que alimenta uma camada densa de tamanho 128 com *dropout* de 0.2 e camada de saída de tamanho 10 e ativação do tipo *softmax*. Essa estratégia obteve um micro-*Fscore* médio de 0.97 com desvio padrão amostral de 0.04 e 82.85% de instâncias corretamente classificadas no conjunto complementar de testes. **Conclusões:**A partir dos resultados, verifica-se que a adaptação do modelo *baseline* contribui para automação da codificação clínica. Além disso, demonstrou-se que o aprendizado de *wordembeddings* a partir da própria base de textos contribui para a classificação de textos contendo vocabulário clínico.



**Palavras-chave:** Aprendizado de Máquina. Processamento de Linguagem Natural. Codificação Clínica. Trato Urinário

## ABSTRACT

**Introduction:** The clinical coding for transposition of the International Classification of Diseases (IDC) codes is a task that guarantees the standardization of diagnoses and payment of clinical accounts. In this context, it becomes relevant the task of the coding professional, who is responsible for reading the clinical narratives and, from these, define which codes correspond to the existing findings. This task, however, is very time-consuming because it is still predominantly manual and automation strategies comprise traditional natural language processing model that present difficulties. The literature suggests that an alternative to traditional strategies is Deep Learning for the natural language processing, by using numerical representations of the semantic meaning of words, known as word embeddings. This alternative applied to clinical texts enables algorithms to recognize and learn existing concepts in clinical narratives, which can optimize the coding process by suggesting codes of the International Classification of Diseases from the clinical narratives. **Objective**: Design a Deep Learning based model for natural language processing that classifies discharge summaries according to the codes of the International Classification of Diseases for other urinary tract disorders (ICD N39).**Method**: Research of applied, exploratory and experimental nature, composed of three stages. The first step is the acquisition of the databases, which comprise the discharge summaries and a database containing people names in Brazil. The second stage is the pre-processing
of the data, where the necessary adjustments were made in the texts. The last step is the development of the models, where the pre-processed texts of the discharge summaries were used to create word embeddings using the GloVe algorithm. These embeddings were used as the first hidden layer of the tested models. It was tested the baseline model and from the obtained results, it was verified that this model could not be a suitable classifier for the problem in question, demanding adaptations. Then 20 different models were adapted from the baseline. The models were evaluated by using the averagemicro-Fscore and respective sample standard deviation, using a stratified cross-validation strategy (10-folds), in addition to the accuracy in classifying examples of the artificially balanced database, verifying that the proposed model better classifies clinical documents if compared to the baseline. **Results**: Aclassifier was trained following the baseline model with a convolution layer, obtaining an average micro-Fscore of 0.08 and sample standard deviation of 0.05, with a 24% accuracy in the balanced test base. From the adaptations, the best model obtained is the one that has as input a 500-word word embedding, 4 layers of convolution with size 128 kernel and "ReLu" activation, windows size 5,8,10,12 and 4 layers of 1-max-pooling, with global pooling that feeds a dense layer of size 128 with dropout of 0.2 and output layer of size 10 and softmax activation. From this strategy, we obtained an average micro-Fscore of 0.97 with sample standard deviation of 0.04 and 82.85% of correctly classified instances in the artificially created set.**Conclusions:** From the results, it is verified that the adaptation of the baseline model contributes to the automation of clinical coding. In addition, it has been shown that learning word embeddings from the text base itself contributes to the learning of specific patterns of clinical vocabulary.



**Keywords**: Machine Learning. Natural Language Processing. Clinical Coding. Urinary Tract.

## LISTA DE FIGURAS

## LISTA DE QUADROS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ADAM | Adaptative Moment Estimation |
| AG | Acurácia Geral |
| AHIMA | American Health Information Management Association |
| AM | Aprendizado de Máquina |
| BoW | Bag-of-words |
| CAC | Computer-Assisted Coding |
| CB | Causa Básica |
| CID | Classificação Internacional de Doenças |
| CNN | Convolutional Neural Network |
| CUDA | Compute Unified Device Architecture |
| DATASUS | Departamento De Informática Do Sistema Único De Saúde Do Brasil |
| DL | Deep Learning |
| GLOVE | Global Vectors for Word Representations |
| IA | Inteligência Artificial |
| ITU | Infecção do Trato Urinário |
| LSTM | Long Short-Term Memory |
| MLP | Multilayer Perceptron |
| NLTK | Natural Language Toolkit |
| OGCR | Official Guidelines for Coding and Reporting |
| ONU | Organização das Nações Unidas |
| PLN | Processamento de Linguagem Natural |
| PRISMA | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| PUCPR | Pontifícia Universidade Católica do Paraná |
| RI | Recuperação de Informações |
| RNA | Rede Neural Artificial |
| RNN | Recurrent Neural Network |
| *S* | Desvio Padrão Amostral |

| | |
|---|---|
| SA | Sumário de Alta |
| SIH | Sistema de Informação de Internações Hospitalares |
| SIM | Sistema de Informação em Mortalidade |
| SINAN | Sistema de Informação de Agravos de Notificação |
| SLP | Single Layer Perceptron |
| SNOMED CT | Systematized Nomenclature of MedicineClinical Terms |
| WE | Word Embedding |

**SUMÁRIO**

# 1 INTRODUÇÃO

A codificação clínica é o processo onde parte da informação clínica de um paciente, usualmente armazenada em narrativas clínicas, é transformada em códigos(PREDA; CHIRIAC; MUSAT, 2012).

Esse processo também é descrito como a tradução e agrupamento de conceitos clínicos em códigos que visam padronizar a nomenclatura de sintomas, diagnósticos e demais situações clínicas em linguagem única (AALSETH, 2006).

Dentre as possíveis codificações clínicas, a codificação tratada neste trabalho é aquela cujos códigos utilizados para transposição são os códigos da Classificação Internacional de Doenças (CID). A CID é um instrumento inicialmente proposto no “Primeiro Congresso Internacional de Estatística”, e que foi então constituída por Jacques Bertillon no final do século XIX. Atualmente este instrumento é organizado e permanece em constante revisão pela Organização Mundial de Saúde (OMS) (LAURENTI, 1993).

Os dados utilizados para a codificação são textos livres contidos em narrativas clínicas e dispostos em linguagem natural. Usualmente, essas narrativas são codificadas através de um processo manual. Esse processo envolve a revisão da documentação clínica do paciente por um agente humano, o codificador: Médico ou outro profissional clínico responsável por aplicar os códigos da CID de acordo com o que está descrito na documentação clínica, realizando o processo conhecido como codificação clínica(FENTON et al., 2010).

É importante ressaltar a relevância da codificação clínica enquanto sendo um dos principais componentes do processo de coordenação de todos os atores do sistema de saúde que estão envolvidos na prestação ou financiamento de serviços de saúde. Esta tarefa tem um grande impacto sobre as atividades financeiras dos prestadores de cuidados de saúde, no monitoramento de suas atividades e também na avaliação e estimativa da necessidade de serviços de saúde (PREDA; CHIRIAC; MUSAT, 2012).

Também é uma importante ferramenta de auxílio na investigação epidemiológica, já que através da padronização das informações clínicas que a codificação viabiliza é possível identificar como situações de saúde pública se distribuem em uma população (AALSETH, 2006).

A necessidade por qualidade no processo de codificação é importante não só para fins estatísticos ou de planejamento de recursos, mas também para a padronização da terminologia médica, que quando não padronizada pode resultar em vários termos diferentes para os mesmos diagnósticos (AALSETH, 2006).

Erros de transcrição para o meio eletrônico, falta de especificidade no diagnóstico, documentação clínica ilógica são alguns dos erros comuns que podem impactar a qualidade da codificação (AALSETH, 2006)

Há também a possibilidade de erro em decorrência da exaustiva leitura ou ainda por problemas de visão do codificador, que podem aumentar a fadiga e reduzir a desempenho de leitura ou a qualidade da compreensão do que está escrito (GRISHAM; SHEPPARD; TRAN, 1993).

Aplicações para computador que visam automatizar este processo estão disponíveis, mas atualmente não são amplamente utilizadas, provavelmente porque esses sistemas estão em constante adaptação e não foram amplamente testados em produção(FENTON et al., 2010).

O processamento de linguagem natural (PLN), subárea de Inteligência Artificial (IA) e da linguística tem como objetivo a extração de informações computáveis a partir de textos (NOHAMA; PACHECO; SCHULZ, 2013).

Os esforços para utilização do PLN têm resultado em importantes progressos na extração de informação de grandes bases de dados não estruturados da *web*, análise de sentimento em redes sociais ou análise gramatical para avaliação de ensaios (SOCHER, 2014; LECUN; ZHANG 2016)

Empresas como a 3M, Dolbey Systems, Artificial Medical Intelligence e Optum360 desenvolveram e patentearam ferramentas comerciais (conhecidas pelo acrônimo “CAC”, *computer automated coding*) que, fazendo uso de técnicas de AM como o PLN, realizam a extração de informações de textos clínicos e os codificam para códigos da CID, auxiliando o trabalho do profissional codificador. O PLN, portanto,já é utilizado para essa finalidade.

A tarefa de compreensão de textos através do PLN, no entanto, é um problema tradicionalmente difícil devido à extrema variabilidade de formação da linguagem, diferenciação entre idiomas e significados das palavras (LECUN; ZHANG 2016; NOHAMA; PACHECO;SCHULZ, 2013).

Além disso, os métodos tradicionais de PLN requerem que as características extraídas a partir dos textos sejam manualmente definidas e ajustadas de acordo com a natureza do problema (SOCHER, 2014; AHUJA et al., 2014).

Particularidades como essas tornam o PLN, de certa forma, especializado em um determinado idioma, de maneira que se o idioma for alterado, muitas características precisam ser redesenhadas manualmente (SOCHER, 2014; LECUN; ZHANG 2016).

Uma alternativa ao processo tradicional de PLN é a utilização de representações númericas das palavras, os *wordembeddings (*WE*)*, que se valem da hipótese distributiva da linguística a fim de capturar o significado semântico e o contexto no qual as palavras estão inseridas. Essas representações numéricas, usualmente extensas e não esparsas, são organizadas em vetores usualmente utilizados como entrada em classificadores em AM (SOCHER, 2014; BENGIO; COURVILLE; GOODFELLOW, 2015)

Avanços no AM e na capacidade computacional dos computadores atuais viabilizaram o surgimento de novas estratégia de extração de informações de dados não estruturados mesmo quando os dados são muito extensos, como em imagens ou em representações numéricas de textos. O principal expoente desses avanços são as técnicas de Deep Learning (DL)(CARVALHO; SANTOS, 2015).

Métodos que fazem uso do DL aprimoraram muito o estado-da-arte no reconhecimento de fala, reconhecimento de imagens, e muitos outros domínios (BENGIO; HINTON; LECUN, 2015).

DL é uma subárea de AM que faz uso de redes neurais artificiais de múltiplas camadas e que lida com o reconhecimento, processamento, interpretação e classificação de imagens, textos, fala etc. fazendo uso do aprendizado por representações (CARVALHO; SANTOS, 2015).

Os métodos de DL são métodos de aprendizagem de representação de múltiplos níveis, obtidos através da composição de módulos simples, mas não lineares, que transformam representações simples (em níveis superiores) em representações cada vez mais complexas, na medida em que os níveis de representação se aprofundam. Com a composição de tais transformações, funções muito complexas podem ser aprendidas (BENGIO; HINTON; LECUN, 2015).

O DL já teve demonstrada sua capacidade de identificação de padrões em imagens, inclusive superando agentes humanos (ESTEVA; KUPREL; THRUN, 2015; HE et al., 2016).

A utilização de DL em textos escritos em linguagem natural vem sendo bastante estudada dada sua performance em conjuntos de dados muito extensos, como por exemplo:WE. Avanços significativos em análise de sentimentos em redes sociais e mesmo na compreensão de linguagem sem conhecimento prévio sobre as características de determinado idioma foram obtidos a partir da utilização desta estratégia e sua eficácia demonstrada em comparação a outras técnicas de AM (SANTOS; GATTI, 2014; LeCUN; ZHANG, 2016; BLUNSOMet al., 2015; CARLSON et. al., 2017)

Além disso, vários trabalhos que utilizam o DL para o PLNsurgiram ao longos dos anos a fim de vencer algumas das dificuldades encontradas no processo manual de definição de características e otimização dos resultados inicialmente encontrados com PLN(LECUN; ZHANG 2016; BLUNSOMet al., 2015).

Dessa forma, o DL para oPLN fazendo uso dos WE, além de ser uma abordagem moderna para um problema tradicionalmente complexo, representa potencial solução para a automação parcial da codificação de narrativas clínicas.

A fim de testar o potencial das técnicas de DL na codificação de narrativas clínicas, que representa um problema de classificação multi-classe, utilizar-se-á uma base de dados de narrativas clínicas de transtornos do trato urinário para classificar a qual (quais) dos 7 códigos específicos da CID-10 para patologias dessa natureza a narrativa analisada está associada.

A pergunta que orienta esse estudo, portanto, é como a utilização de um modelo de DL para o PLN pode auxiliar o processo de codificação clínica?

## 1.1 OBJETIVO GERAL

Conceber um modelo baseado em Deep Learning (DL) para o PLN que classifique sumários de alta de acordo com os códigos CID para “outros transtornos do trato urinário” (N39).

## 1.2 OBJETIVOS ESPECÍFICOS

Faz parte dos objetivos específicos dessa pesquisa:

a) Produzir WEde vocabulário clínico em Português do Brasil a partir dos textos clínicos;
b) Produzir uma base de testes artificiais e alternativos, a partir da utilização desses WE;
c) Avaliar os modelos desenvolvidos com o uso desta base artificial alternativa.

## 1.3 CONTRIBUIÇÕES

Com os objetivos constituídos, essa pesquisa passa a oferecer contribuições de diferentes naturezas, que serão descritas a seguir.

### 1.3.1 Para a ciência

Essa pesquisa visa aproximar o DLdo estado-da-arte para o PLN em Português do Brasil.Adicionalmente, espera-se também que possa contribuir com a expansão de modelos de DL, que utilizam WE treinados a partir de base dados de natureza específica, como essa, integralmente composta por textos clínicos.

### 1.3.2 Para a sociedade

A codificação clínica influencia em quais serviços médicos são pagos e como são pagos, de maneira que uma codificação de má qualidade ou inexistente pode acarretar em prejuízos financeiros à instituição que presta o serviço ou para quem paga pelo serviço realizado, seja o indivíduo ou a seguradora (AALSETH, 2006). Dessa forma, a contribuição social se evidencia por meio da necessidade por qualidade e agilidade do processo de codificação clínica a partir do uso de automações, para que haja uma melhor alocação e planejamento de recursos em saúde, objetivando maior qualidade na prestação de serviços de saúde.

# 2 REFERENCIAL TEÓRICO

Este capítulo está organizado de forma a viabilizar ao leitor o entendimento do processo de codificação clínica, conhecendo algumas das características e limitações que esse processo apresenta, passando pela Classificação Internacional De Doenças (CID)e as alternativas de automação de codificação sugeridas pela literatura,culminando na proposta aqui oferecida.

É importante mencionar o critério utilizado para adoção dos trabalhos utilizados para compor o referencial teórico desta pesquisa. A partir do uso do método "*Preferred Reporting Items for Systematic Reviews and Meta-Analysis*" (PRISMA) (ALTMAN et al. 2009) foram selecionados estudos que demonstram a utilização do PLN, de DL e dos WE utilizados para o PLN em DL, bem como estudos que indiquem como se dá o processo de codificação clínica e quais as dificuldades identificadas nesse processo, objetivando identificar quais as ferramentas disponíveis e quais lacunas que não tenham sido completamente preenchidas pelas ferramentas tecnológicas disponíveis.

A subseção 2.1 traz uma breve definição e descrição da CID, sustentando o processo de codificação, trazido na subseção 2.2, onde também descreve-se o trabalho do profissional codificador, responsável por realizar a codificação a partir das narrativas clínicas.

A subseção 2.3 traz a definição de "Transtornos do Trato Urinário", sua descrição na CID (Outros Transtornos do Trato Urinário)e suas subdivisões, exemplificando sua codificação e pormenorizando cada uma das 7 subcategorias de diagnóstico.A subseção 2.4, traz a perspectiva de extração de conhecimento de narrativas clínicas, a partir da qual apresentam-se as estratégias disponíveis para obtenção de conhecimento em textos, especialmente as estratégias que fazem uso do AM, como o PLN.A subseção 2.5 aborda o conceito de AM e do PLN, abordando algumas características do AM, além de vantagens e dificuldades apontadas pela literatura em estratégias tradicionais de PLN.

A subseção 2.6 aborda os WE, tratados na literatura como uma das alternativas ás dificuldades do PLN tradicional, e que são usualmente utilizados como entrada de algoritmos de aprendizado de máquinapara a classificação de textos, como os algoritmos de redes neurais artificiais. A subseção 2.7 define as redes neurais artificiais, exemplificando as características desse algoritmo e introduzindo o DL, que

é uma ramificação das redes neurais artificias e que adquiriu bastante notoriedade pelos avanços na classificação de imagens e textos, e a partir da qual objetiva-se constituir uma solução de suporte para codificação de narrativas clínicas. A subseção 2.8 traz algumas das aproximações frequentes para definição de parâmetros e construção de modelos de DL, especialmente o modelo escolhido para compor a solução proposta nesta pesquisa.

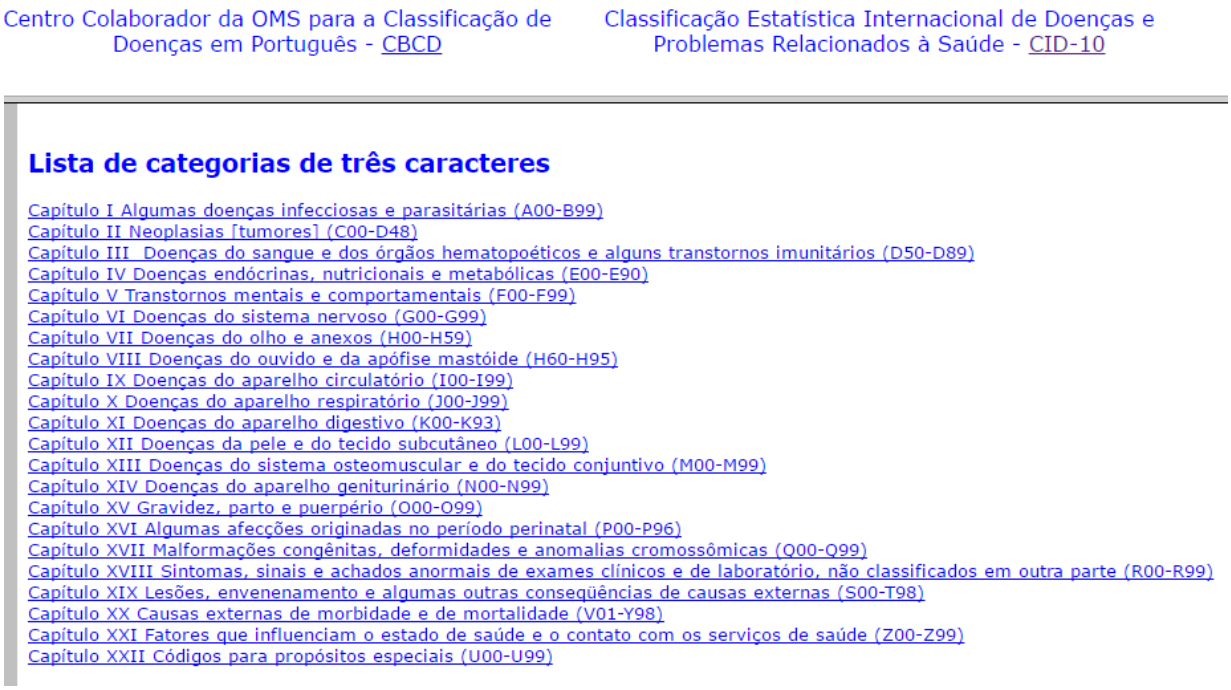
## 2.1 A CLASSIFICAÇÃO INTERNACIONAL DE DOENÇAS

A classificação internacional de doenças (CID) é instrumento constituído pela Organização das Nações Unidas (ONU) e administrada pela Organização Mundial de Saúde (OMS) cujo objetivo é promover a comparabilidade epidemiológica, classificação, processamento e apresentação de estatísticas de morbidade e mortalidade. É a classificação reconhecida internacionalmente que ajuda profissionais clínicos, formuladores de políticas e pacientes a navegar, entender e comparar sistemas e serviços de saúde. (BOWLES; SHAFRAN-TOPAZ; TOPAZ, 2013; AALSETH, 2006).

Desde seu lançamento como classificação internacional de doenças pela ONU em 1900, a CID é constantemente atualizada a fim de que reflita alterações conceituais e incorpore novas nomenclaturas de morbilidade e (ou) mortalidade (BOWLES; SHAFRAN-TOPAZ; TOPAZ, 2013; AALSETH, 2006).

Atualmente, a CID encontra-se em sua décima revisão (vigente desde 1990) que compreende 3 volumes e 22 capítulos com mais de 12 mil códigos (Figura 1) que servem de base para orientação do processo de codificação clínica (BRASIL; FUNASA, 2001).

Figura 1 - Base de dados de consulta da CID-10 do DATASUS



Centro Colaborador da OMS para a Classificação de Doenças em Português - CBCD

Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde - CID-10

**Lista de categorias de três caracteres**

Capítulo I Algumas doenças infecciosas e parasitárias (A00-B99)
Capítulo II Neoplasias [tumores] (C00-D48)
Capítulo III Doenças do sangue e dos órgãos hematopoéticos e alguns transtornos imunitários (D50-D89)
Capítulo IV Doenças endócrinas, nutricionais e metabólicas (E00-E90)
Capítulo V Transtornos mentais e comportamentais (F00-F99)
Capítulo VI Doenças do sistema nervoso (G00-G99)
Capítulo VII Doenças do olho e anexos (H00-H59)
Capítulo VIII Doenças do ouvido e da apófise mastóide (H60-H95)
Capítulo IX Doenças do aparelho circulatório (I00-I99)
Capítulo X Doenças do aparelho respiratório (J00-J99)
Capítulo XI Doenças do aparelho digestivo (K00-K93)
Capítulo XII Doenças da pele e do tecido subcutâneo (L00-L99)
Capítulo XIII Doenças do sistema osteomuscular e do tecido conjuntivo (M00-M99)
Capítulo XIV Doenças do aparelho geniturinário (N00-N99)
Capítulo XV Gravidez, parto e puerpério (O00-O99)
Capítulo XVI Algumas afecções originadas no período perinatal (P00-P96)
Capítulo XVII Malformações congênitas, deformidades e anomalias cromossômicas (Q00-Q99)
Capítulo XVIII Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte (R00-R99)
Capítulo XIX Lesões, envenenamento e algumas outras conseqüências de causas externas (S00-T98)
Capítulo XX Causas externas de morbidade e de mortalidade (V01-Y98)
Capítulo XXI Fatores que influenciam o estado de saúde e o contato com os serviços de saúde (Z00-Z99)
Capítulo XXII Códigos para propósitos especiais (U00-U99)

Fonte: DATASUS, 2017.

De acordo com o DATASUS (2017), A CID-10 é apresentada em volumes, que são descritos no quadro abaixo:

Quadro 1 - Descrição dos volumes da CID-10

| Volume | Descrição |
|---|---|
| I | Conhecido como “Lista Tabular”, onde estão dispostas as classificações. É formada por códigos de três caracteres (uma letra e dois algarismos) e subcategorias que subdividem as categorias em especificidades, quando existentes, por meio de ponto sucedido de algarismo arábico (Ex.: A00.1, Cólera devido a *Vibrio cholerae* 01, biótipo *El Tor*) |
| II | Além de apresentar o histórico e desenvolvimento da classificação, apresenta também guias e orientações para usuários da CID, em especial, possui as regras e disposições gerais para a codificação de morbidade e mortalidade. Importante ressaltar que embora possua essas orientações para codificação, não é o único instrumento que orienta os profissionais que codificam. |

| | |
|---|---|
| III | É o índice alfabético da CID, onde os usuários podem procurar por determinada causa básica (CB) de forma facilitada, também é onde pode se ter a nomenclatura adequada para a CB. Idealmente, este volume deve ser utilizado em conjunto com o VOLUME I para a realização da codificação (AALSETH, 2006) |

Fonte: Aalseth, 2006.

Há também a utilização de outras versões da CID em outros países que, eventualmente, podem discordar de seções da CID que não incorporem especificidades daquela região. É o caso dos Estados Unidos, que desde a CID-8 vem adotando versões adaptadas a realidade do território norte-americano (AALSETH, 2006).

Usualmente, países que adotam versões customizadas da CID classificam suas versões atribuindo a elas um acrônimo que indique essa customização, é o caso da CID-10-CM, utilizada nos Estados Unidos e que indica uma classificação internacional de doenças com "modificações clínicas" (do inglês "*Clinical Modification*"), sendo então necessária para a codificação de diagnósticos clínicos enquanto a CID-10-PCS ("*Procedure Coding System"* que traz os códigos para procedimentos de admissão de pacientes.Situação semelhante é encontrada na classificação utilizada na Austrália, a CID-10-AM ("*Australian modification*") e no Canadá com a CID-10-CA ("*Canadian Adaptation*"). (BOWLES; SHAFRAN-TOPAZ; TOPAZ, 2013; MCKENZIE et al., 2010; AALSETH, 2006).

O Brasil utiliza a CID-10 sem adaptações, seu uso é compulsório para compor a base de códigos de todas as informações em Mortalidade (Portaria GM/MS n° 1832/94, publicada no DOU n° 218, de 03 de novembro de 1994) e em Morbidade (Portaria 1311/GM de 12 de setembro de 1997). Os códigos da CID são utilizados no SIM (Sistema de Informação em Mortalidade), no SIH (Sistema de Informação de Internações Hospitalares) e no SINAN (Sistema de Informação de Agravos de Notificação), ou mesmo no campo destinado ao diagnóstico de malformações congênitas na Declaração de Nascidos-Vivos que está incluído no SINASC (Sistema de Informação sobre Nascidos Vivos) (LAURENTI et. al., 2013; DATASUS, 2017).

A OMS prevê a décima primeira revisão da CID (CID-11) para 2018, mas não se espera que seja utilizada antes de 2020. Para compor essa nova revisão, a OMS analisou quase toda a classificação em uma perspectiva estrutural. Essas avaliações

contribuíram para a edição da estrutura e índice da classificação. A OMS também se encontrou com as principais partes interessadas em temas críticos, incluindo dermatologia, diabetes, demência, doenças cerebrovasculares, incluindo acidente vascular cerebral e cuidados primários. A partir deste ponto e de maneira concorrente, a OMS atualizou várias seções-chave para melhorar a transparência e fornecer informações sobre o progresso do projeto(OMS, 2017).

As estruturas que foram configuradas para a fase de design da CID-11 estão agora sendo redesenhadas para atender às necessidades de finalização e manutenção contínua da CID que, embora ainda não completa, apresenta notável progresso. (OMS, 2017; LAURENTI et. al., 2013).

## 2.2 A CODIFICAÇÃO CLÍNICA

A codificação clínica, neste contexto,corresponde à transposição dos diagnósticos das CB nas narrativas clínicas para os códigos correspondentes na CID (CHIRIAC; MUSAT; PREDA, 2012; LOPES, 2009; BRASIL, 2001).

Para Camargo Jr. e Favoreto (2011), a abordagem da narrativa é trazida para a clínica como uma ferramenta que pode facilitar a percepção e a interpretação do significado do processo de adoecimento, como um modo de o profissional de saúde incorporar novos enunciados ao seu repertório interpretativo e, assim, ampliar a dimensão dialógica, hermenêutica e integral do saber e da prática clínica.

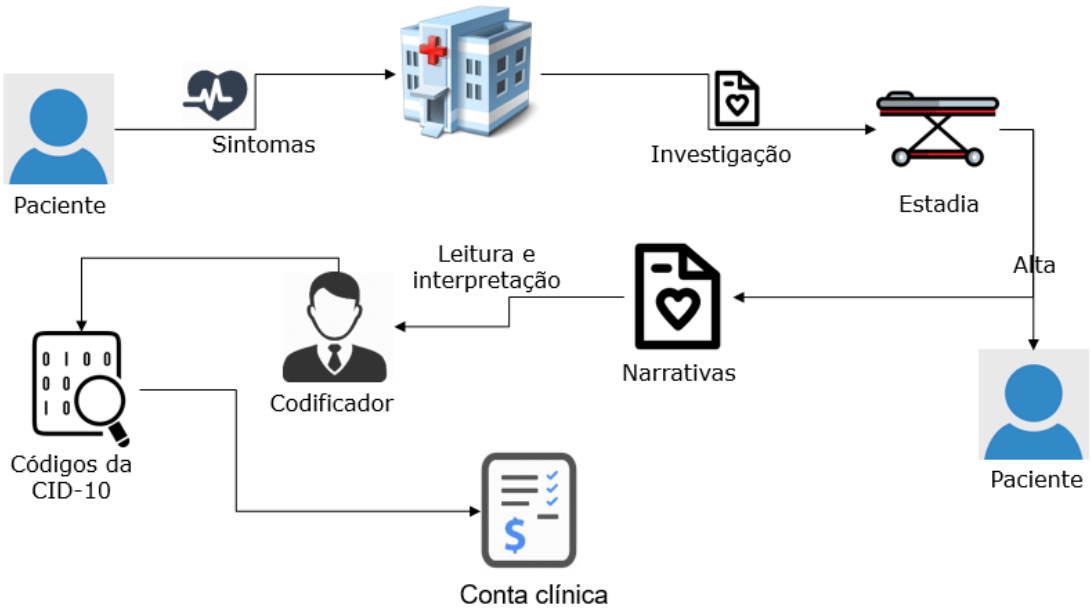
É tarefa do profissional codificador ler as narrativas clínicas, identificar conceitos, como o diagnóstico principal, diagnósticos adicionais de comorbidade ou de complicação, CB, os procedimentos cirúrgicos e a cada um desses conceitos ou termos associar um código da CID adotada naquela região (LOPES, 2009).

Os dados dos prontuários fazem parte dessas narrativas e são essenciais para dar continuidade ao tratamento do paciente, pois, além do cadastro, estes dados referem-se ao seu acompanhamento, podendo existir relatórios de comportamento preventivo que são utilizados em identificação de riscos, avaliação de novas intervenções, diagnósticos, entre outros até que se receba a alta. Estes dados fazem partedo registro de evolução ou história clínica do paciente, que, quando preenchido

de forma correta, alerta sobre variações e resultados das consultas, diagnósticos, medicamentos e comportamento do paciente (LAPELLE et al., 2006; NARYSHKIN; SCHULTZ, 2009; BULEGON, 2011). Os dados em formatos de texto constituem as narrativas clínicas (CHU, 2002).

O fluxo do processo de codificação clínica se dá quando o paciente dá entrada no hospital, onde então se inicia uma investigação dos sintomas e, a partir da alta,reúnem-se os documentos clínicos obtidos durante esse período para leitura e interpretação do profissional codificador (Figura 2).

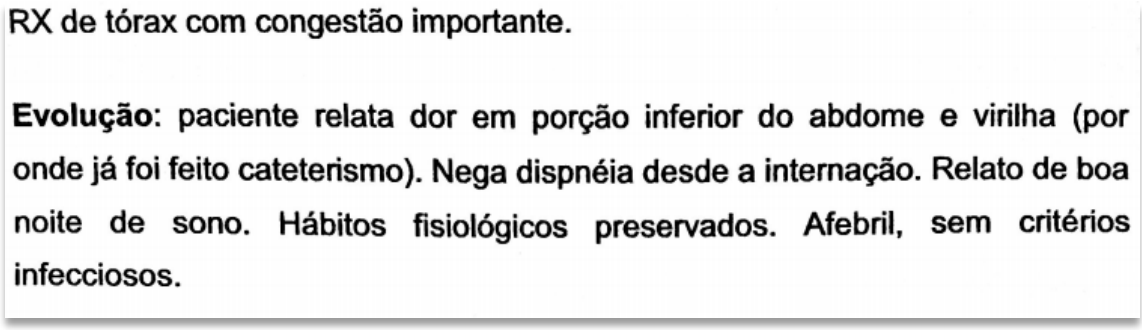
Figura 2–Fluxo do processo de codificação clínica




Fonte: O autor,2018.

É a partir das narrativas clínicas, que consistem em textos escritos em linguagem natural (Figura 2) que é realizado o processo de codificação.

Figura 3- Narrativa Clínica com a evolução do paciente



RX de tórax com congestão importante.

**Evolução**: paciente relata dor em porção inferior do abdome e virilha (por onde já foi feito cateterismo). Nega dispnéia desde a internação. Relato de boa noite de sono. Hábitos fisiológicos preservados. Afebril, sem critérios infecciosos.

Fonte: O autor,adaptado de um documento clínico real, 2018.

O profissional codificador, no entanto, pode encontrar dificuldades enquanto interpretando as informações contidas nas narrativas clínicas. Essas dificuldades ao longo do processo de codificação podem prejudicar a qualidade da codificação com base nestas informações. Se o registo clínico for incompleto ou impreciso, por exemplo, ou se o codificador não identificar os diagnósticos de complicação ou de comorbidade, o hospital pode ter um prejuízo real, uma vez que a codificação clínica também alimenta a base de dados hospitalar e nacional, a partir da qual se faz investigação epidemiológica e se calculam indicadores hospitalares, além disso, em países onde o processo de codificação e pagamento de contas clínicas é constantemente supervisionado e auditado, contas clínicas codificadas erroneamente podem ensejar investigações por fraude, ou ainda sanções financeiras e (ou) judiciais à instituição que relatou determinado custo (AALSETH, 2006; LOPES, 2009).

Há ainda a possibilidade de erro na codificação em decorrência da exaustiva leitura, cansaço ou ainda por problemas de visão do codificador, que podem aumentar a fadiga e reduzir a performance de leitura ou a qualidade da compreensão do que está escrito (GRISHAM; SHEPPARD; TRAN, 1993).

A imperícia do codificador é um aspecto qualitativo, cuja mensuração é dificilmente constatável se não pela experiência que o profissional possui com a tarefa ou pela auditoria do processo de codificação (LOPES, 2009).

No Brasil, a codificação deve ser realizada por técnicos qualificados (BRASIL; FUNASA, 2001). Não foram encontradas diretrizes da OMS que determinem qual profissional deve realizar a codificação clínica, no entanto, uma breve revisão da literatura mostra que o entendimento comum é de que a codificação deve ser realizada por profissionais com conhecimento clínico ecom treinamento em codificação. O Quadro 2exemplifica qual o cenário para os profissionais codificadores em diferentes

países do mundo, evidenciando que a codificação clínica é realizadapor profissionais treinados, o que a rigor poderia afastar a imperícia como causa de dificuldades durante a o codificação.

Quadro 2 - Perfil dos Codificadores clínicos em diferentes países

| Referência | País | Agente Codificador |
|---|---|---|
| FUNASA; BRASIL (2001) | Brasil | Médico, Técnico treinado, ou Diretoria Regional de Saúde |
| CHIRIAC; MUSAT; PREDA (2012) | Romênia | Médico ou outro profissional (treinado ou não) |
| CACE (2012) | Argentina | Médico ou Técnico treinado |
| MINISTÉRIO DE SANIDAD, (2016) | Espanha | Médico ou Técnico treinado |
| LOPES (2009) | Portugal | Técnico treinado |
| CHESTER (2016), DEAR et. al (2010) | Reino Unido | Técnico treinado |
| AALSETH (2006) | Estados Unidos | Técnico treinado |
| DEAR et. al (2010) | Canadá | Técnico treinado |
| DEAR et. al (2010) | Austrália | Técnico treinado |

Fonte: Adaptado da literatura, 2017.

Em linhas gerais, afirma-se que o processo de codificação tem início assim que há um diagnóstico documentado, de forma que possa ser convertido em um número de código de diagnóstico de acordo com a CID (AALSETH, 2006).

Para realizar a codificação, o profissional pode se guiar através do volume II da CID-10, ou ainda usar a ferramenta instituída pela "*American Health Information Management Association" (*AHIMA), o "*Official Guidelines for Coding and Reporting*" (OGCR), documento que orienta as instruções gerais para codificação e que é utilizado por profissionais codificadores em todo o mundo (CHESTER, 2016; BULEGON, 2011; LOPES, 2009; AALSETH, 2006).

Este documento proposto pela AHIMA consiste em 71 páginas com convenções e instruções de codificação além de orientações detalhadas por capítulo da CID, seleção do diagnóstico principal de entrada e de saída. (AALSETH, 2006).

## 2.3 TRANSTORNOS DO TRATO URINÁRIO

Segundo a base do DATASUS (2017), transtornos do trato urinário são caracterizados por disfunções na micção associadas a alguma patologia clínica.

A CID-10 reserva uma classe específica para problemas dessa natureza (N39 – Outros transtornos do trato urinário) e 7 subcategorias que melhor detalham o distúrbio que foi verificado a partir da avaliação clínica 39.0 fazem referência a “infecções do trato urinário de localização não especificada”. Segundo Leal et al. (2010), esse sintoma é verificado a partir da presença de bactéria na urina, que indica comprometimento da flora bacteriana do trato urinário.

O código N39. 1 representa “Proteinúria persistente não especificada” e é caracteriza por uma liberação anormal de proteínas através da urina. Segundo Miller (2010), pode indicar lesão renal ou doenças cardiovasculares de maneira geral.

De maneira semelhante, o código N39. 2 também representa proteinúria, mas é caracterizado por “Proteinúria ortostática não especificada” que indica a liberação anormal de proteínas identificada a partir da análise da primeira atividade urinária do dia.

O código N39. 3 representa o diagnóstico de “Incontinência de tensão (*stress*)”, que indica perda involuntária da urina durante esforço, prática de exercício, ao tossir ou espirrar (GOMES; RIOS, 2010).

O código N39. 4 (outras incontinências urinárias não especificadas) também pode ser utilizado para incontinência urinária, no entanto, segundo a CID-10 (DATASUS, 2017), seu diagnóstico deve estar associado a outros códigos da CID-10 que indiquem refluxo, sobre fluxo ou urgência.

Os códigos N39. 8 (Outros transtornos especificados do aparelho urinário) e, especialmente o N39.9 (Transtornos não especificados do aparelho urinário), são utilizados para condições clínicas que envolvam o trato urinário e não podem ser caracterizadas pelos demais códigos da CID-10.

## 2.4 EXTRAÇÃO DE CONHECIMENTO A PARTIR DE NARRATIVAS CLÍNICAS

De acordo com Nohama, Pacheco e Schulz (2013), a representação do conhecimento contido das narrativas clínicas é objeto de grande interesse científico com o objetivo de estabelecer novas bases conceituais e tecnológicas que auxiliem na extração dessa informação.

Oleynik et. al. (2010) indicam que os textos que constam nas narrativas clínicas, por serem escritos em linguagem natural, apresentam variabilidade em decorrência do idioma, da região ou de jargões médicos. A extração de conhecimento a partir de narrativas clínicas passa a ser, portanto, difícil do ponto de vista tecnológico, dado que seria preciso um processo de filtragem, adaptação ou padronização de palavras, acrônimos e conceitos sem clareza.

Fazendo uso do AM, é possível automatizar parcialmente esse processo ao realizar a extração das informações em textos livres a partir do PLN, ramo da linguística e do AM que se concentra na identificação e extração de conhecimento a partir de textos escritos em linguagem natural (OLEYNIK et al., 2010). No âmbito clínico, um dos documentos das narrativas clínicas que apresenta de maneira resumida as ocorrências clínicas e que pode ser utilizado para extração de informação, conhecimento e automação da codificação é os sumários de alta (SA) hospitalar (Figura 3) (NOHAMA, PACHECO; SCHULZ, 2013; BULEGON, 2011).

Figura 4 - Exemplo de sumário de alta, com dados que podem servir para o PLN

**RELATÓRIO DE ALTA HOSPITALAR**

**Paciente:** **Atend:**

**Idade:** 78 **Sexo:** F **Endereço:** **Tel.:**

**Convênio**

**Plano:**

**Data da Internação:** 12/03/2015 **Data da Alta:** 14/03/2015

**Tipo de Alta:** Alta melhorado

**Retorna em:** Não

**ORIENTAÇAO DE ALTA/EVOLUÇÃO**

* Evoluções
Data: 14/03/15 Usuário:
depressão, obesidade, sequela avc
internado por mal estar geral e itu

paciente com quadro estavel, afebril, eupneica, niveis pressoricos controlados
cd alta com ornetações, receita e prescrição, retorno em caso de piora,oriento acompanhamnto ambulatorial, oriento manter fisioterapia
orientações, gerais a acompanhante

* Diagnósticos
Data: 14/03/15 -
INFECCAO DO TRATO URINARIO DE LOCALÍZACAO NAO ESPECIFICADA

Fonte: O autor, adaptado de sumário de alta real. 2018.

Esse documento condensa informações médicas do paciente e facilita sua eventual readmissão ou consulta no hospital. Contêm sinais e sintomas do paciente, antecedentes pessoais e familiares, exame físico, laudos, medicações usadas e planas para o seguimento do caso. Estratégias para extração de informações em narrativas clínicas, como o PLN, podem, portanto, utilizar este documento clínico para extrair informações sobre diagnóstico de um paciente, embora se apresentem algumas dificuldades (OLEYNIUK et al. 2010; GUIMARAES; KLÜCK, 1999).

Uma dificuldade inerente à interpretação das narrativas clínicas é a utilização de acrônimos, inclusive para os diagnósticos. Apesar de alguns acrônimos não serem oficiais, são frequentemente utilizados. De forma que há a necessidade de expandir acrônimos em linguagem única e padronizada. Há iniciativas correntes com esse objetivo em países como os Estados Unidos, onde já se usa o instrumento conhecido como *Sistematize Nomenclatura of. Medicine - Clínica Tiros* (SNOMED CT), vocabulário médico padronizado que define o significado dos termos médicos, visando melhorar a interoperabilidade entre sistemas (NOHAMA, PACHECO; SCHULTZ, 2013; AALSETH, 2006).

Alguns autores também citam outras dificuldades em obter informações importantes a partir do sumário de alta que, por ser documento composto por profissional clínico, pode ter sua qualidade comprometida por falta de preenchimento

em decorrência do grande volume de atendimentos em alguns hospitais, ou ainda pela falha na legibilidade dos demais documentos clínicos que ajudam a compor este importante documento (SOUZA, 2012).

Há, no entanto, diversos trabalhos que fazem uso desse documento e que demonstraram que apesar das dificuldades supracitadas, o sumário de alta pode ser utilizado para o PLN. É o caso dos trabalhos de Souza (2012), Bulegon (2011), Melton E Hripcsak (2005).

## 2.5 APRENDIZADO DE MÁQUINA E O PLN

A definição formal de aprendizado de máquina (AM), conforme originalmente proposto, é do “campo de estudo que dá ao computador a habilidade de aprender sem ser explicitamente programado” (SAMUEL, 1959).

Há, no entanto, definições que melhor caracterizam essa área de IA. É o caso da definição de Bengio, Courville e Goodfellow (2015) que definem o termo aprendizagem de máquina como sendo a capacidade do computador na “detecção automatizada de padrões significativos em dados”.

O campo de estudo doAM se ramifica em subáreas e vários subcampos que lidam com diferentes tipos de tarefas de aprendizado, como o aprendizado supervisionado, não-supervisionado, por reforço, indutivo etc. No entanto, a título de contextualização, os tipos a serem definidos nesta pesquisa são o aprendizado supervisionado e o aprendizado não supervisionado:

O aprendizado supervisionado é o processo do aprendizado de máquina a partir do qual a máquina aprende pela experiência, ou seja, aprende utilizando exemplos bem definidos, rotulados e com informações sobre a sua natureza. Usualmente estes exemplos são bem subdivididos em conjuntos que servem como base de treinamento e também como base de testes do que foi aprendido. O aprendizado supervisionado foi originalmente descrito como “o aprendizado onde se aprende com um professor (exemplos)” (FELLOW et al. 1976; NORVIG; RUSSEL, 2009).

Já no aprendizado não-supervisionado, o aprendizado é realizado sem exemplos rotulados. O algoritmo de aprendizado não-supervisionado processa dados

de entrada com o objetivo de apresentar algum resumo, versão compactada desses dado, ou mesmo um agrupamento de dados em subconjuntos de objetos semelhantes. Formalmente, esse tipo de aprendizado foi definido como aquele onde a máquina não conta com um “professor” e precisa reconhecer padrões a partir da identificação de características semelhantes em um determinado conjunto de dados (FELLOW et al. 1976; BEN-DAVID, 2014).

Enquanto estratégias supervisionadas assumem que há exemplos rotulados para o aprendizado, estratégias não-supervisionadas de aprendizado assumem que o aprendizado deve ser realizado pela organização de características comuns entre os exemplos. Em linhas gerais, abordagens supervisionadas são importantes para problemas de classificação onde se têm exemplos rotulados cujas características podem ser aprendidas por um algoritmo de aprendizado supervisionado, e então generalizadas para exemplos ainda não vistos, objetivando sua classificação de acordo com o que foi aprendido. No aprendizado não-supervisionado, a análise se dá a partir da identificação das características que os dados compartilham, para que então se consiga determinar a qual grupo estes dados pertencem. Por isso, é importante haver um conjunto grande de dados para que o algoritmo de aprendizado não-supervisionado consiga identificar a maior quantidade de características e distinções possíveis nos dados, assim viabilizando a organização dos dados em conjuntos que compartilham de características semelhantes (BEN-DAVID, 2014).

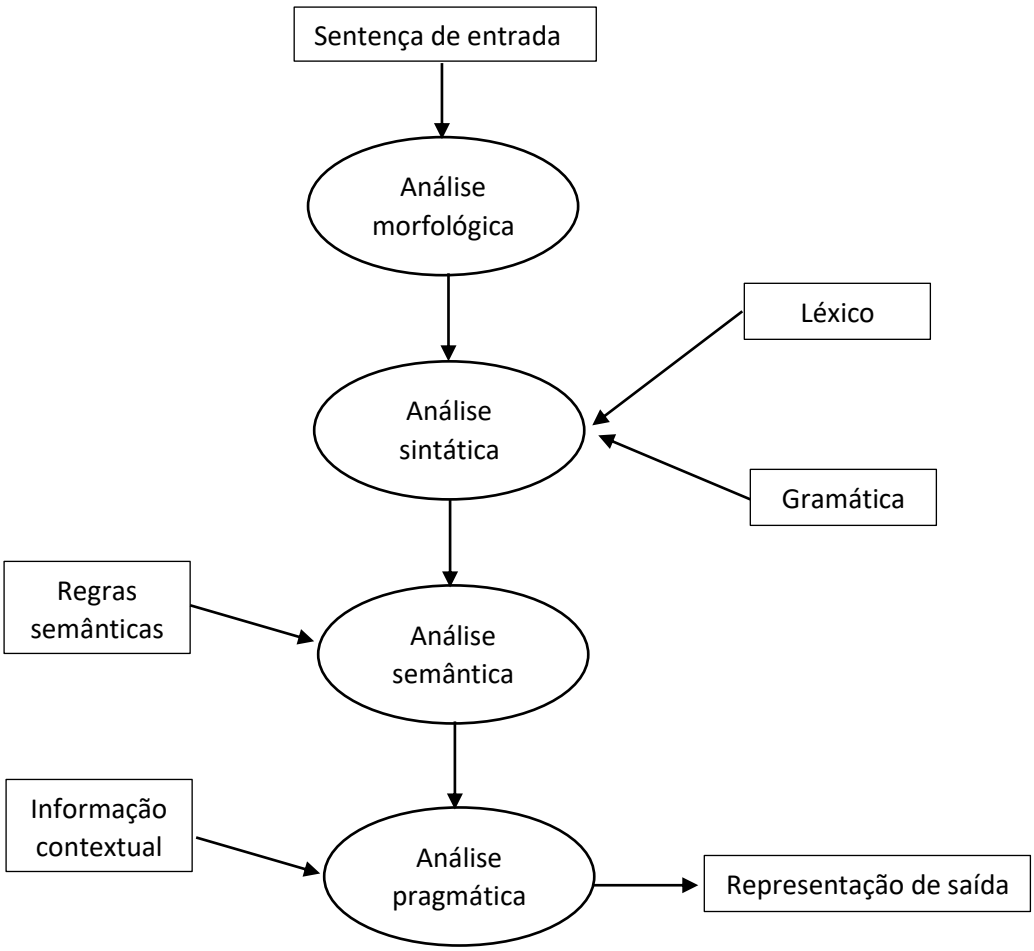
Estratégias de aprendizado não-supervisionado são interessantes para o PLN, pois as palavras contidas em textos, usualmente sem rótulo, podem ser organizadas de acordo com as características que compartilham (BEN-DAVID, 2014; SOCHER, 2014)

O PLN é uma subárea de AM e da linguística que se concentra no processo de extração e interpretação de informações em textos (BATES, 1993). Nos recentes anos, o PLN avançou em tarefas como extração de informações de grandes conjuntos de dados não estruturados da *web*, análise de sentimentos em redes sociais ou análise gramatical para classificação de ensaios. Um dos objetivos do PLN é o desenvolvimento de algoritmos gerais e escaláveis que possam resolver conjuntamente tarefas de PLN e aprender as representações intermediárias necessárias das unidades linguísticas envolvidas (SOCHER, 2014; HUGHES; KOTOULAS; SUZUMURA, 2017). É tido como uma técnica para extrair informações

de textos, muito útil para recuperação da informação (RI), tradução automática, sumarização de texto (Da SILVA E SOUZA, 2014).

O PLN está tradicionalmente dividido em quatro etapas: análise morfológica, análise sintática, análise semântica e análise pragmática, conforme figura abaixo (Figura 4) (CHOPRA et al. 2013):

Figura 5- Representação gráfica do fluxo do PLN




Fonte: O autor, 2018.Adaptado de CHOPRA et al., 2013.

Na etapa de análise morfológica, o vocabulário, palavras e expressão de um determinado idioma são organizados, descritos, analisados e eventualmente separados em palavras, parágrafos ou sentenças, sob a perspectiva de identificação e descrição da estrutura das palavras (CHOPRA et al. 2013).

A análise sintática envolve a análise das palavras em determinada frase a partir do seu sentido léxico para descrever a estrutura gramatical da sentença. As palavras são transformadas em estruturas que mostram como as palavras estão relacionadas umas às outras. Segundo Chopra et al. (2013), a frase "o menino o vai para a escola" é uma frase que definitivamente seria rejeitada por um analisador sintático do idioma português por não fazer sentido sintático.

Na análise semântica, há a abstração do significado das palavras em relação a um dicionário ou contexto (regras semânticas). As estruturas são criadas pelo analisador sintático com significado atribuído. Há um mapeamento entre as estruturas sintáticas e os objetos contidos na sentença, de maneira que sentenças sem sentido lógico jamais seriam aceitas por um analisador semântico (CHOPRA et al. 2013).

Por fim, a análise pragmática interpreta os resultados da análise semântica a partir da perspectiva de um contexto específico (contexto do diálogo ou estado do mundo etc.), ou seja, o que foi dito ao longo da sentença é reinterpretado sobre o que realmente significa (CHOPRA et al. 2013).

Existem trabalhos com resultados muito interessantes utilizando o PLN para extração de informação a partir de textos clínicos, como demonstrado por Bulegon (2011), Nohama, Pacheco e Schulz (2013) e Oleynik et al. (2011).

No entanto, nos últimos anos surgiram alguns autores apontaram estratégias alternativas ao modelo tradicional de PLN e que objetivam otimizar diversas das suas características e definições manuais. Autores como Socher (2014) e Chen et al. (2013) criticam a efetividade das estratégias tradicionais de PLN, alegando que estas possuem duas deficiências comuns para o processamento de linguagem natural em qualquer domínio, sendo a primeira delas a simplificação de suposições de idioma. Segundo este autor, no PLN algoritmos são desenvolvidos e, em seguida, os dados são “forçados” a um formato compatível com esse algoritmo. Por exemplo, um primeiro passo comum na classificação ou agrupamento de texto é ignorar a ordem das palavras e a estrutura gramatical e representar textos em termos de listas de palavras não ordenadas, processo conhecido como *word-bagging*. Ainda segundo Socher (2014), essa estratégia leva a problemas no entendimento do sentimento geral expresso nas frases analisadas. Socher (2014) então propõe que um modelo ideal aprenda que enquanto as palavras compõem um determinado sentimento inicial, o sentimento geral precisa ser o absoluto, ou seja, mesmo que em uma frase o significado associado seja, por exemplo, positivo, a frase precisa ser análise sob seu

contexto todo, eventualmente levando a um sentimento ou significado negativo sob contexto geral. Abordagens de *word-bagging* em sua maioria não conseguem realizar essa distinção, pois realizam a contagem de palavras em um texto.

Outra deficiência apontada por Socher (2014) é que enquanto muito tempo é dedicado ao desenvolvimento do modelo de PLN, o desempenho da maioria dos sistemas de aprendizagem depende fundamentalmente das representações de recursos da entrada. Por exemplo, os sistemas que fazem uso do PLN tradicional usam marcações de parte da fala, anotações especiais para cada local, pessoa ou organização (assim chamadas entidades nomeadas), recursos de árvore de análise ou a relação de palavras em uma grande taxonomia. Cada um desses recursos leva muito tempo para serem desenvolvidos e integrados para cada nova tarefa, retardando tanto o desenvolvimento quanto o tempo de execução do algoritmo final (SOCHER, 2014).

Ferramentas que fazem uso do PLN tradicional para extração de conhecimento em narrativas clínicas, portanto, também possuem as mesmas deficiências, dificultando seu desenvolvimento e adaptação. Algumas das ferramentas conhecidas para automatizar a codificação clínica utilizam o PLN. Essas ferramentas são conhecidas como CAC (*computer-assisted coding*), e estão listadas no Quadro 3. Lamentavelmente, essas ferramentas não possuem suporte documental que habilite uma análise mais profunda da metodologia utilizada, fato que pode ser em vinculado ao fato de se tratarem de tecnologias patenteadas pelas empresas que as desenvolveram.

Quadro 3 - Ferramentas atualmente utilizadas para extração de informação de textos clínicos

| Referência | Documento-base | Aplicação | Características funcionais |
|---|---|---|---|
| 3M (2017) | Narrativas Clínicas | Codificação Clínica | Sistema comercial que alegadamente extrai códigos de narrativas clínicas já codificadas verifica se estes são adequados e provê códigos adicionais quando necessário. |
| Dolbey Systems (2017) | Narrativas Clínicas | Codificação Clínica | Sistema comercial (Fusion CAC) que, a partir de narrativas clínicas, extrai códigos da CID-10 e os indica ao profissional codificador para que esta escolha qual o código melhor se adéqua ao diagnóstico em questão. |
| EMScribe CAC (2017) | Narrativas Clínicas | Codificação Clínica | Sistema comercial que extrai conceitos clínicos das narrativas e converte em códigos da CID-10 e sugere códigos da CID-10 para que o profissional codificador escolha o código adequado. |
| Optum360 (2017) | Narrativas Clínicas | Codificação Clínica | Sistema comercial que extrai códigos da CID-10 a partir dos registros clínicos do paciente. |

Fonte: O autor, adaptado da literatura, 2018.

Há também ferramentas de codificação computadorizadas menos sofisticadas do que opções que utilizam o PLN. Essas ferramentas são conhecidas como "*encoders*", que também visam facilitar o processo de codificação. Esses programas variam de programas simples que são apenas replicações dos livros de codificação em um formato computadorizado, para softwares interativos que fazem todas as perguntas necessárias para chegar ao código correto da categoria de diagnóstico (AALSETH, 2006).

## 2.6 WORD EMBEDDINGS

*Word* Embeddings (WE) são representações numéricas do significado semântico de palavras criadas a partir do uso de algoritmos de AM. Segundo Socher (2014) e Chen et al. (2013), modelos que fazem uso de WE representam uma importante alternativa às estratégias tradicionais de PLN, por não serem dependentes do idioma e de um determinado vocabulário, além de serem capazes capturar o significado semântico das palavras dentro de um espaço vetorial sem a necessidade da definição manual de características como no PLN tradicional. Isso se torna possível, pois o conceito de WE se sustenta na hipótese distribucional, originalmente descrita por Harris (1954).

De acordo com esta hipótese, se observarmos duas palavras que constantemente ocorrem dentro dos mesmos contextos, é possível assumir que significam coisas semelhantes. Note-se que a hipótese não exige que as palavras ocorram em conjunto (como em *word-bagging*), mas que as palavras ocorram com o mesmo conjunto de outras palavras. Citemos como exemplos as palavras “nadar” e “natação”. Seguindo a proposta de Harris (1954), essas duas palavras devem carregar significado semelhante, pois frequentemente ocorrem com as mesmas palavras vizinhas. A importância da hipótese distribucional foi demonstrada em inúmeras experiências (GOODENOUGH; RUBENSTEIN, 1965).

A ideia geral por trás dos modelos de WE é motivada por esta hipótese distribucional, produzindo espaços vetoriais com várias dimensões, nas quais as palavras são representadas por vetores de *N* dimensões de contexto cujas orientações relativas em um plano são assumidas como indicadores de semelhança

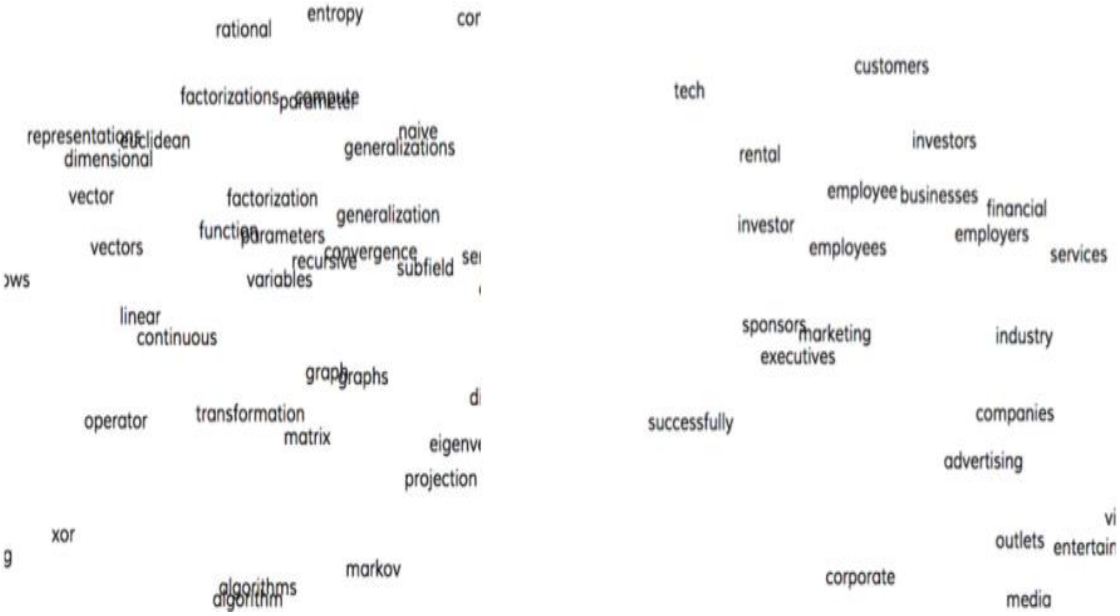
semântica (Figura 6). Uma orientação comum é de que vetores com um tamanho *N* próximo a 0 tendem a representar de maneira limitada o significado das palavras, enquanto vetores com *N* maiores que 1000 tendem a representar significados de maneira exageradamente abstrata (HEUER, 2016; SOCHER, 2014).

O objetivo da hipótese distribucional aplicada ao PLN é encontrar uma representação e um vetor que se aproxima do significado de uma determinada palavra, evitando dessa forma o processo tradicional de PLN, que demanda muita experimentação e ajustes manuais para um funcionamento adequado (SOCHER, 2014; CHEN et al 2013).

Os *word embeddings* são usualmente representados por “dimensões” (*N*). A palavra “dimensões” no contexto dos *word embeddings* representa quantos números compõem a representação numérica de uma determinada palavra. Na Figura 7, por exemplo, a palavra “linguistics” é representada por um vetor de 8 dimensões (*N* = 8). Palavras cujo significado semântico é expresso em muitas dimensões (por exemplo, N>100), usualmente possuem uma melhor distinção das demais palavras (SOCHER, 2014; CHEN et al., 2013).

A Figura 16 descreve como WE podem ser representados em um plano, indicando que palavras de significado semelhante tendem a permanecer próximas umas das outras.

Figura 6- Representação de um vetor de palavras em um plano




Fonte: Heuer, 2016.

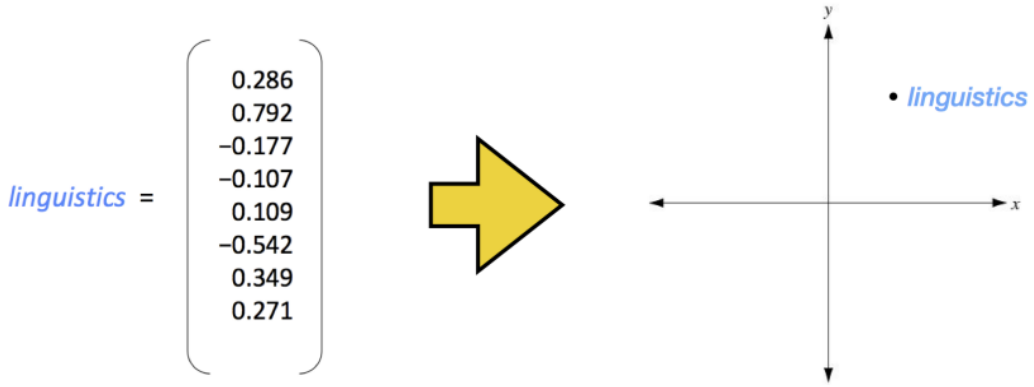
A utilização dos WE está relacionada a um número grande de palavras para criação desses vetores, dado que o treinamento destes vetores é feito de maneira não-supervisionada, o que torna sua utilização limitada para situações onde o conjunto de palavras está associada à existência de textos onde essas palavras se encontram (SOCHER, 2014; CHEN et al. 2013).

Socher (2014), por exemplo, utilizou 50 bilhões de palavras para criar um WE com seu algoritmo conhecido como "*GloVe*" (acrônimo para *Global Word Vectors*). Esse algoritmo cria a relação semântica entre as palavras com base na contagem de co-ocorrência entre as palavras dado um determinado contexto.

Chen et al. (2013) utilizaram cerca de 6 bilhões de palavras para criar um WE a partir do algoritmo "*word2vec*". Esse algoritmo busca identificar o significado semântico e as relações entre as palavras a partir da previsão do contexto de uma palavra alvo (estratégia conhecida como *Skip-gram*), ou prever as palavras alvo a partir do contexto (estratégia conhecida como *continous bag-of-words)*. Socher (2014), no entanto, comparou o algoritmo GloVe ao *Word2Vec* utilizando as estratégias *Skip-gram* e *continous bag-of-words*, concluindo que o GloVe é capaz de produzir os melhores WE, retendo maior parte do significado semântico das palavras e mais rapidamente. Os mesmos resultados foram verificados por e Botvinick (2016), Rodrigues (2016) e Cohen et al. (2017), enquanto e Berardi et al. (2015) e Chandrasekan et al. (2016) apontam para uma melhor acurácia utilizando o *Word2Vec* com *skip-gram*.

Modelos de PLN que se valem de WE também possuem uma característica importante, e que os tornam particularmente interessantes para tarefas de classificação utilizando AM. O fato das palavras serem representadas por vetores de números torna os WE uma entrada qualificada para modelos de AM que recebem como entrada dados numéricos, como por exemplo, as redes-neurais artificiais (SOCHER, 2014; HEUER, 2016). A Figura 6 representa como uma palavra pode ser representada numericamente e, dado o devido ajuste de dimensões, sua representação em um plano cartesiano.

Figura 7 - WE contendo o vetor de 8 dimensões que representa a palavra "linguistics"




Fonte: Heuer, 2016.

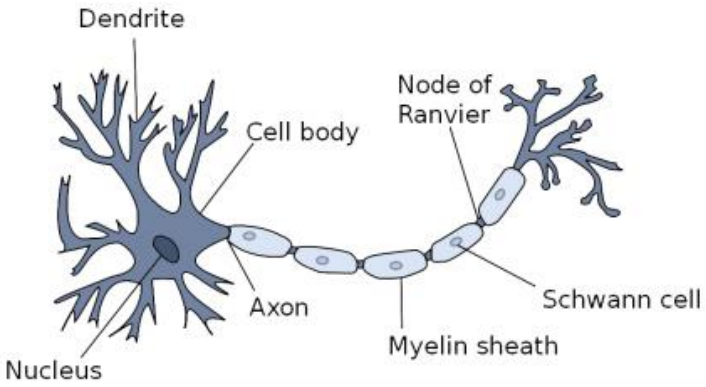
## 2.7 REDES NEURAIS ARTIFICIASE O DEEP LEARNING

O termo “redes neurais artificiais” (RNA) representa uma série de modelos computacionais que, simulando a atividade dos neurônios biológicos, objetivam identificar determinados padrões a partir de sinais disparados por esses neurônios artificiais (MCCULLOCH; PITTS, 1943).

Os neurônios biológicos (Figura 7) estão ligados uns aos outros de maneira que quando estimulados transmitem um sinal elétrico, ou sinapse, através do axônio. Do axônio, esses sinais não são diretamente transferidos para os neurônios seguintes, mas primeiro eles devem atravessar a fenda sináptica, onde o sinal é alterado novamente por processos químicos variáveis (KRIESEL, 2006).

No neurônio receptor, as várias entradas que foram pós-processadas na fenda sináptica são somadas ou acumuladas em um único pulso. Dependendo de como o neurônio é estimulado pela entrada, o próprio neurônio emite um pulso, assim, a saída não é linear e não proporcional à entrada original e o sinal é propagado (KRIESEL, 2006).

Figura 8- Ilustração de um neurônio biológico




Fonte: Kriesel, 2006.

De maneira análoga, uma RNA consiste em unidades de processamento simples que possuem conexões direcionadas e com peso de conexão entre esses neurônios (Figura 8). Nesse caso, o peso de conexão entre dois neurônios *i* e *j* é definido como (KRIESEL, 2006):

$$w_{ij} \tag{2.1}$$

Os dados são transferidos entre os neurônios através de conexões com este peso de conexão (seja excitador ou inibitório). Essa conexão entre neurônios é identificada como *"função de propagação"*, e pode ser definida da seguinte forma para um neurônio *j* em relação a um neurônio *i* (KRIESEL, 2006):

$$O_{ij} \tag{2.2}$$

Dessa forma, pode-se definir a soma ponderada de pesos entre neurônios da seguinte maneira (KRIESEL, 2006):

$$rede_j = \sum_{i \epsilon I} (O_{ij} \, . \, w_{ij}) \tag{2.3}$$

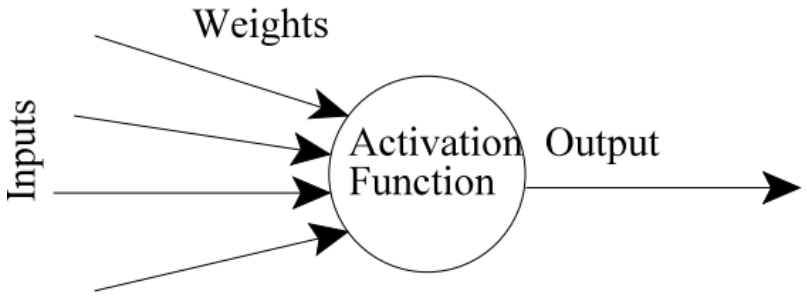
Em uma RNA (Figura 8), um neurônio é ativado com base numa função de ativação, que indica quando o neurônio deve ou não ativar, e que pode ser definida da seguinte forma (KRIESEL, 2006):

(2.4)

$$a_j(t) = f_{ativação}(rede_j(t), a_j(t-1), \Theta_j)$$

De acordo com Kriesel (2006), essa função transforma a rede de entrada $rede_j$, bem como o estado anterior de ativação $a_j(t-1)$ em um novo estado de ativação $a_j(t)$, com valor limite $\Theta$, que é apenas assinalado a *j* e marca a posição do valor máximo da função de ativação.

Figura 9 - Representação de um neurônio artificial



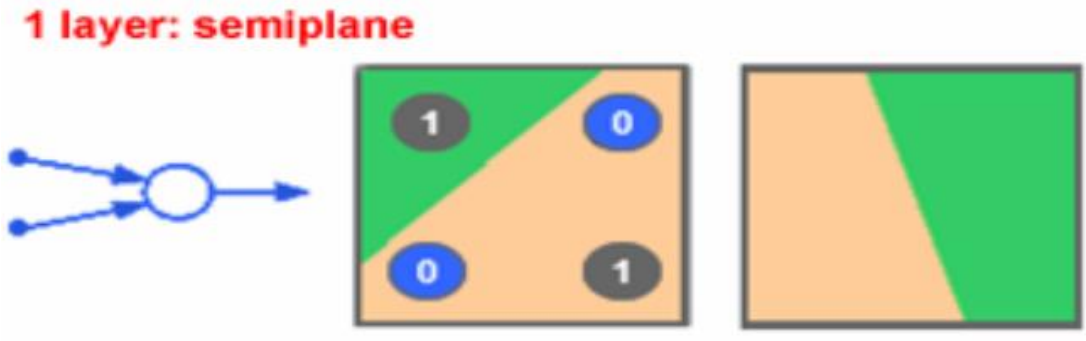

Fonte: Gershenson, 2005.

De maneira geral, as RNAs podem se apresentar em dois tipos: *Single-layer perceptrons* (SLP) e *Multi-layer perceptrons* (MLP) (KRIESEL, 2006).

RNA do tipo SLP é um *perceptron* tendo apenas uma camada de pesos variáveis e uma camada de neurônios de saída, sendo uma ferramenta para solução de problemas linearmente separáveis, dado que um *perceptron* com uma única camada gera regiões de decisão sob a forma de semiplano, como representado na Figura 9(BALAS et. al. 2009; KRIESEL, 2006).

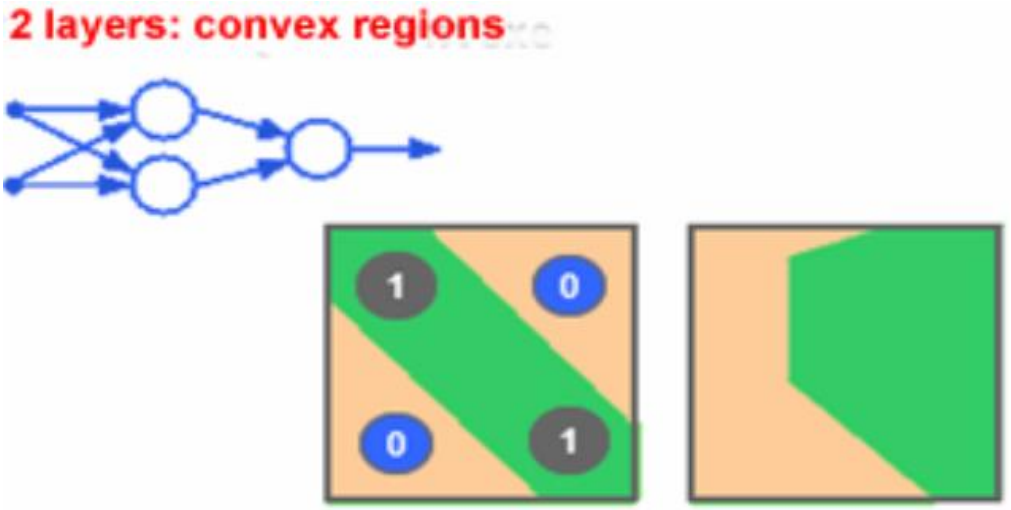
Nas RNA do tipo MLP, cada neurônio atua como um *perceptron* padrão para as saídas dos neurônios na camada anterior, assim a saída da rede pode estimar regiões de decisão convexa (Figura 10), resultante da interseção dos semiplanos gerado pelos neurônios e até mesmo produzir regiões de decisão arbitrárias (Figura 11) (BALAS et. al. 2009; KRIESEL, 2006)

Figura 10- Representação do espaço de decisão de uma RNA do tipo SLP



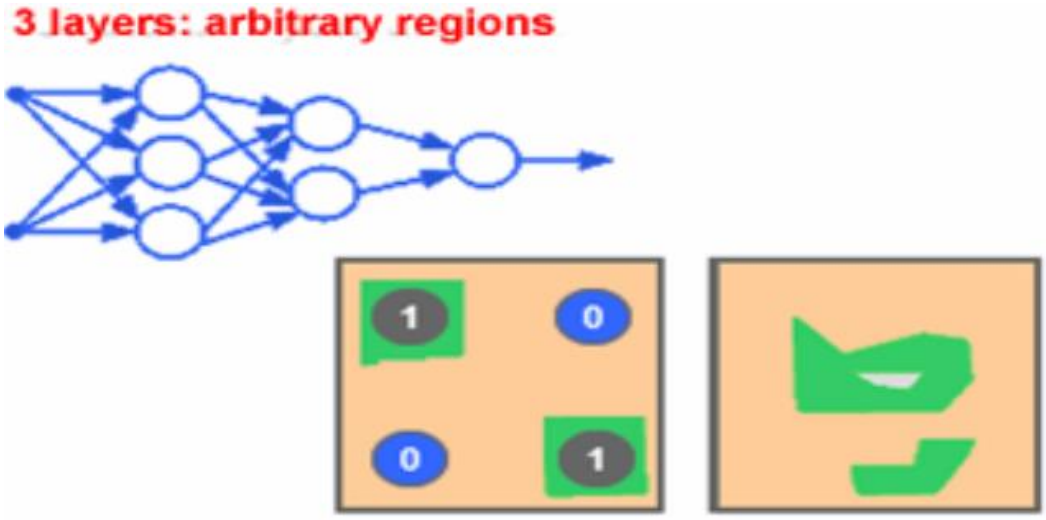

Fonte: Balas et al., 2009.

Figura 11 - Representação do espaço de decisão de uma RNA do tipo MLP




Fonte: Balas et al., 2009.

Figura 12 - Representação do espaço de decisão arbitrário de uma RNA do tipo MLP




Fonte: Balas et al., 2009.

A maneira como os *perceptrons* são conectados para formar uma rede é também denominada “Topologia” da rede (usualmente essa relação também é chamada de “arquitetura” ou “estrutura” da rede), e é de fundamental importância na funcionalidade e performance da RNA (FIESLER, 1996).

De acordo com Norvig e Russell (2009) as conexões entre neurônios podem variar entre densas (usualmente descrito na literatura como “*fully-connected”*), recorrente, em malha (*mesh*) etc. Nas RNAs densas, que são as mais comuns e também as mais simples, todas as possíveis conexões “intercamada” (ou seja, camada inferior com camada superior) estão presentes. Essas redes são utilizadas para problemas comuns de classificação (FIESLER, 1996).

Nas redes recorrentes as saídas dos neurônios são usadas como entradas de *feedback* (resposta) para outros neurônios. O recurso de *feedback* qualifica essas redes para o processamento de informações dinâmicas, o que significa que elas podem ser empregadas em sistemas previsão de séries temporais, identificação e otimização de sistemas, controle de processo etc. (NORVIG; RUSSEL, 2009)

Um tipo importante de MLP que pode assumir as topologias mencionadas e que adquiriu notoriedade recentemente são as redes que realizam o que conhecemos como Dep. Learning (DL). DL é também reconhecida como uma subárea de aprendizado de máquina que, a partir do uso de algoritmos de RNA do tipo MLP, permite que modelos computacionais compostos por múltiplas camadas de processamento aprendam representações de dados com múltiplos níveis de abstração (CARVALHO; SANTOS, 2015; BENGIO; HINTON; LECUN, 2015).

Sua origem remete às primeiras RNA do tipo MLP propostas por Ivakhnenko cujo propósito era a solução de problemas que cresciam em complexidade e em nível de abstração, consequentemente aumentando a quantidade de camadas escondidas da *perceptrons*. Dessa forma, pode-se afirmar que RNAs usadas em DL são MLP( CARVALHO; SANTOS, 2015).

A principal característica de DL e que torna essa área tão relevante para o AM, é sua capacidade de aprender por meio de representações. O aprendizado de representações permite que um algoritmo receba como entrada dados brutos e descubra automaticamente as representações necessárias para a detecção ou classificação de padrões na medida em que as representações são transferidas entre camadas (BENGIO; HINTON; LECUN, 2015).

As RNAs deDL são algoritmos de aprendizagem de representação com vários níveis de representação, obtidos através da composição de módulos simples, mas não-lineares, que transformam a representação de um nível inferior (e consequentemente mais simples) em representações cada vez mais complexas na medida em que o número de camadas de neurônios vai aumentando, tornando a topologia da RNA mais “profunda” (daí o termo “*Deep”*). Com a composição de tais transformações, funções muito complexas podem ser aprendidas. Para tarefas de classificação, as camadas mais próximas da entrada ampliam os aspectos que são importantes para a discriminação e eliminam variações irrelevantes (BENGIO; HINTON; LECUN, 2015).

Dessa forma, o DL é, dentre outras finalidades, utilizado para o PLN ao extrair padrões complexos a partir de textos convertidos em vetores numéricos, como nos WE (SOCHER, 2014; BENGIO; COURVILLE; GOODFELLOW, 2015).

Ayyar e Bear (2017), Hughes, Kotoulas e Suzumura (2017), por exemplo, fizeram uso do DL para PLN com WE treinado a partir do algoritmo *GloVe*. Essa estratégia permite que a partir da utilização de vetores de palavras pré-treinados (como o *word2vec* ou *GloVe*) seja possível alimentar um modelo baseado em DL para classificar textos, dessa forma o modelo de DL não precisa aprender a estrutura do idioma ou as características da linguagem, pois essas características já foram absorvidas pelo WE (AYYAR; BEAR, 2017; HUGHES; KOTOULAS; SUZUMURA, 2017; SOCHER, 2014).

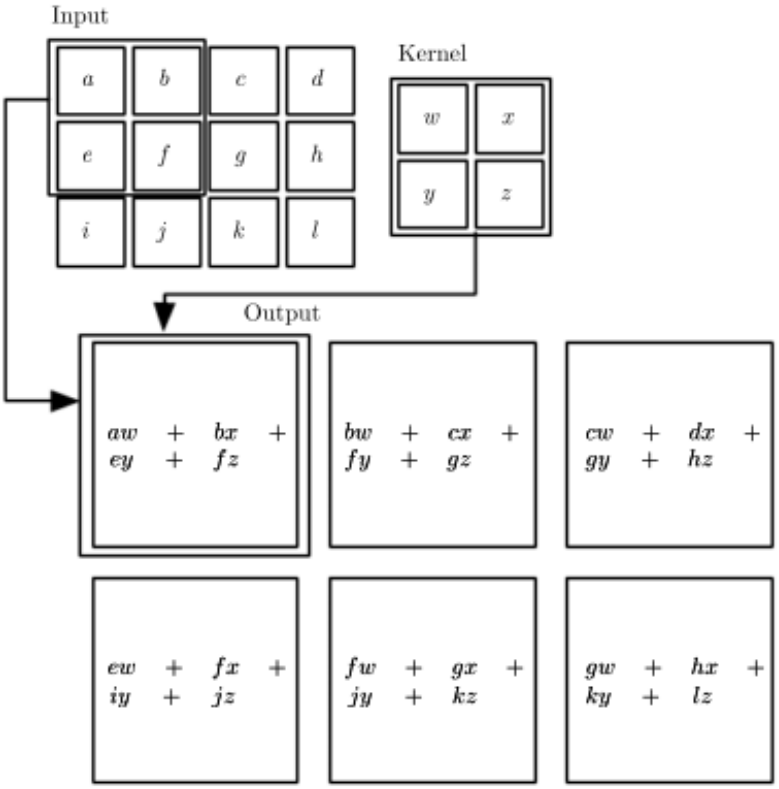
A utilização dos WE para o DL é ainda sustentada pelo fato de que muitos modelos tradicionais de PLN são baseados em contagens de palavras onde a classificação de determinado texto é realizada de acordo com a contagem de ocorrências das palavras em um contexto pré-definido(SOCHER, 2014; BENGIO; COURVILLE; GOODFELLOW, 2015). Isso pode prejudicar o desempenho de generalização do modelo ou mesmo levar ao problema conhecido como "maldição da dimensionalidade", onde um vetor de palavras em um grande vocabulário é muito esparso, de maneira que o tamanho do vetor de palavras pode facilmente superar o tamanho dos dados de treinamento. As soluções clássicas para este problema envolvem tanto a engenharia de características manuais mencionadas anteriormente, quanto o uso de funções de alvo muito simples como em modelos lineares (SOCHER, 2014; BENGIO; COURVILLE; GOODFELLOW, 2015; CHEN et al., 2013).

Dentre os modelos de DL para o PLN existentes na literatura, há predominância na utilização dos modelos de redes neurais de convolução (CNN) e redes neurais recorrentes (RNN). Ambos modelos também são utilizados para problemas de classificação de imagens e textos. Como verificado nos trabalhos de Esteva et al (2015), Socher (2014), Kavuluru e Rios (2015) e Hughes, Kotoulas e Suzumura (2017), CNN são voltadas à extração de características gerais em textos e imagens. Já as RNN como as do tipo *Long Short-Term Memory* (LSTM) são mais comumente utilizadas na classificação de sentimentos em textos, onde uma característica identificada como importante precisa ser temporariamente armazenada e propagada para a classificação final (AYYAR; BEAR, 2017; BENGIO; COURVILLE; GOODFELLOW, 2015). Essa característica torna as CNNs uma ferramenta importante para problemas de classificação multi-classe no PLN.

As CNN são um tipo especializado de RNA de DL para processamento de dados que possuem uma topologia de matriz. A palavra convolução em seu nome sugere que esse modelo implementa a operação matemática de convolução, que representa uma operação entre dois sinais, um de entrada e um de interação, chamado *kernel*, tendo um terceiro sinal como saída que relaciona os dois sinais anteriores (Figura 12). O *kernel* desempenha papel fundamental em modelos de convolução. Também conhecido como “mapas de características”, o *kernel* é a ferramenta que efetivamente analisa os padrões durante a convolução. Cada *kernel* possui um tamanho, que indica o tamanho das características que serão analisadas (janelas) (BENGIO; COURVILLE; GOODFELLOW, 2015).

A implementação de convolução em DL se dá a partir da assunção de que *pixels* em imagens ou WE representam sinais que, quando interagindo com a camada de convolução podem ou não ativar um sinal que, em modelos de CNN, é utilizado como padrão para entrada de uma próxima camada de identificação de características (BENGIO; COURVILLE; GOODFELLOW, 2015).

Figura 13 - Representação da operação de convolução




Fonte: Bengio; Courville; Goodfellow, 2015.

Essa próxima camada é denominada camada de *pooling*, que substitui a saída da rede em um determinado local por uma estatística resumida das saídas próximas. Em outras palavras, a operação de *pooling* fornece como saída os sinais que foram mais fortemente acionados nas interações anteriores, isso viabiliza que os padrões mais importantes em dados não-estruturados sejam identificados pelo modelo e propagados para um classificador (BENGIO; COURVILLE; GOODFELLOW, 2015).

A operação de convolução permite que uma rede compartilhe parâmetros ao longo do tempo, assemelhando-a a forma como as Redes Neurais Recorrentes (RNN) trabalham, embora as CNNs façam esse compartilhamento por característica padrão da operação de convolução. Dessa forma, a operação de convolução quando extraindo parâmetros de WE também considera a forma como as palavras estão posicionadas dentro de um contexto (SOCHER, 2014; BENGIO; COURVILLE; GOODFELLOW, 2015).

A saída da convolução é uma sequência em que cada membro da saída é uma função de um pequeno número de membros vizinhos da entrada. A ideia do compartilhamento de parâmetros se manifesta na aplicação do mesmo *kernel* de convolução a cada etapa de tempo (BENGIO; COURVILLE; GOODFELLOW, 2015).

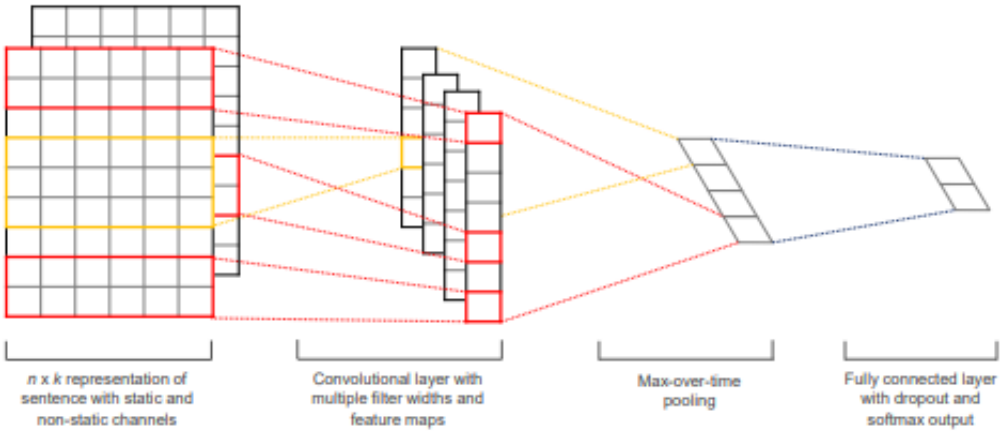
A partir de todas essas características das CNN, Kim (2014) verificou a partir do seu modelo *baselin* que mesmo modelos simples de CNN têm bom desempenho na classificação de textos e sentenças, se comparado a modelos mais complexos, como as RNN.

## 2.8 SELEÇÃO DE PARÂMETROS E TREINAMENTO DECNN

De acordo com De Andrade (2013), determinar o *design* apropriado de arquitetura de um modelo de AM é um desafio porque essa configuração difere para cada conjunto de dados e, portanto, requer ajustes para cada um destes. Para CNN, essa afirmação mantém-se adequada. Para tanto, Wallace e Zhang (2015) propuseram um trabalho exploratório a fim de estabelecer parâmetros e arquiteturas ótimas para CNNs aplicadas ao PLN. Os autores se baseiam na estratégia proposta por Kim (2014) (Figura 13), tida como o *baseline* para o PLN em classificação de textos e sentenças. Kim (2014) propôs uma CNN simples que alcançou resultados comparáveis aos modelos mais complexos para a classificação de textos.

Wallace e Zhang (2015) indicam que esses resultados sugerem que a arquitetura de CNN adotada por Kim (2014) pode servir de *baseline* para novos modelos. Outros trabalhos, como o de Hughes, Kotoulas e Suzumura (2017) também tratam o modelo de Kim (2014) como os modelos do qual novos modelos podem derivar (KIM, 2014; HUGHES; KOTOULAS; SUZUMURA, 2017).

Figura 14 - Modelo *baseline* de CNN para classificação de textos




Fonte: Kim, 2014.

O modelo de Kim (2014) consiste em uma camada de entrada composta por sentenças transformadas em WE, usando vários filtros (com tamanhos variáveis de janela) para obter várias características distintas dos dados de entrada. Os sinais ativados a partir da convolução são propagados e identificados nas camadas de *pooling*. Esses recursos então formam a penúltima camada e são passados para uma camada densa com ativação *softmax*, cuja saída é a distribuição de probabilidade sobre os rótulos das sentenças ou textos (KIM, 2014).

Wallace e Zhang (2015) propõem que a quantidade de camadas de convolução da CNN deve ser utilizada com moderação, prezando pela simplicidade, uma vez que modelos muito complexos tendem ao sobre-ajuste (*overfitting*), ou seja, se adaptam aos dados utilizados para treinamento, enquanto modelos demasiadamente simples sofrem com sob-ajuste (*underfitting*), que representa uma adaptação muito simples aos dados. Dessa forma, é importante utilizar uma quantidade de camadas de convolução que seja suficiente para garantir que o modelo seja simples, mas com boa capacidade de generalização avaliada a partir do desempenho de classificação.

Quanto aos filtros de convolução (*kernel*), Wallace e Zhang (2015) recomendam executar uma procura pelo melhor número entre 100 e 600, levando em consideração o fato de que o tempo de execução aumenta exponencialmente para filtros cada vez mais próximos desse limite superior. Wallace e Zhang (2015) indicam que, assim que um número ideal de filtros é encontrado, esse número deve ser replicado para as eventuais próximas camadas de convolução, variando então o tamanho da janela da região dos filtros entre 1 e 10, podendo ser maior para textos longos. Os autores justificam essa combinação de diferentes tamanhos de janelas

para garantir que padrões diferentes são propagados pela CNN em cada camada (WALLACE; ZHANG, 2015).

Essa orientação quanto ao tamanho de janela não é acompanhada por Hughes, Kotoulas e Suzumura (2017) que optaram por utilizar janelas de mesmo tamanho em todas as camadas. Wallace e Zhang (2015), no entanto, realizaram testes em diferentes conjuntos de dados e identificaram que essa uniformidade no tamanho das janelas de procura não contribui para melhoria preditiva do modelo.

Wallace e Zhang (2015) em sua exploração também apontam que a camada de *pooling* deve seguir a estratégia de *1-max-pooling*, dado seu desempenho consistentemente melhor do que estratégias alternativas para a tarefa de classificação de textos. Isso pode ocorre porque certos *n*-gramas na sentença pode ser mais preditivos por si mesmos do que toda a sentença considerada em conjunto. Tanto Kavuluru e Rios (2015), Huang e Li (2016) quanto Hughes, Kotoulas e Suzumura (2017), fazem uso dessa estratégia.

Em todos os modelos de CNN verificados para classificação, a camada final é a camada *softmax,* que fornece uma distribuição de probabilidade por classe (totalizando 1, indicando a força com a qual os sinais de saída foram ativados), indicando a probabilidade individual de cada uma das classes previstas.

Além dessas escolhas, também é necessário realizar a exploração e a escolha dos hiperparâmetros que serão considerados pelo algoritmo para treinamento do modelo. Segundo Duffy et al. (2017), o sucesso de um modelo de AM também depende da seleção dos melhores hiperparâmetros, que são parâmetros que influenciam na capacidade de aprendizado e generalização do modelo.

A taxa de aprendizagem, por exemplo, controla quanto os pesos da RNA ajustados em relação ao gradiente. O *dropout* fornece um método computacionalmente barato, mas poderoso de regularizar uma ampla família de modelos. Pode ser visto como um método que transforma um modelo complexo em diversos modelos mais simples a partir da desativação aleatória de perceptrons em uma RNA. Esse processo contribui para evitar que o modelo de rede neural se adapte ao problema em questão – reduzindo o *overfitting* (BENGIO; COURVILLE; GOODFELLOW, 2015).

O *momentum* controla o quão rápido deve ocorrer o aprendizado, e *decay* indica o quanto os pesos dos outros hiperparâmetros precisam ser ajustados para maximizar o resultado preditivo (BENGIO; COURVILLE; GOODFELLOW, 2015).

É difícil encontrar a combinação de hiperparâmetros apropriada para um dado conjunto de dados porque não é bem entendido como esses hiperparâmetros interagem uns com os outros para influenciar a precisão do modelo resultante (DE SALVO et al. 2016)

Além disso, não existe uma formulação matemática para calcular os hiperparâmetros apropriados para um determinado conjunto de dados, de modo que a seleção depende de tentativa e erro. Dessa forma, a experiência ou o conhecimento prévio acerca dos dados dão ao pesquisador maiores chances de encontrar ajustes ótimos para o problema que se deseja tratar (ALBELWI E MAHMOOD, 2017).

Há, no entanto, parâmetros que podem ser ajustados seguindo algoritmos conhecidos como otimizadores, estes iniciam valores randômicos para os hiperparâmetros de um algoritmo de AM e, a partir de formulações matemáticas, tentam encontrar a melhor combinação de hiperparâmetros para minimizar o valor da função custo de um modelo, aumentando a capacidade preditiva do modelo. Dentre os otimizadores mais utilizados em DL, está o “Adam” (LEI BA; KINGMA, 2014), “Adagrad” (KIM, 2014), “Adadelta” (KIM, 2014; WALLACE; ZHANG, 2015), “RMS Prop” (BENGIO et al., 2015). Há evidências, a partir de Lei Ba e Kingma (2014), que o algoritmo “Adam” converge mais rapidamente na direção de parâmetros ótimos se comparado a outros otimizadores, isto é, encontra os melhores hiperparâmetros mais rapidamente, diminuindo o tempo de treinamento total.

Outra questão que influencia a capacidade preditiva de modelos de AM, especialmente os modelos de DL é a quantidade de dados disponíveis para treinamento e a distribuição desses dados em relação às classes que representam. Entretanto, não é bem entendido o tamanho “mínimo” do conjunto de dados para experimentos DL bem-sucedidos. Kavuluru e Rios (2015) usam aproximadamente 90.000 exemplos de textos para um modelo baseado em CNN, Hughes, Kotoulas e Suzumura (2017) usaram aproximadamente 15.000 exemplos para uma abordagem similar e Li et al. (2016) até mesmo utilizaram conjuntos de dados com aproximadamente 4400 exemplos para PLN com CNN para treinar *word embeddings*. Todas essas experiências não demonstram um padrão de *overfitting*. Algumas estratégias visam combater eventuais dificuldades com a quantidade de dados disponíveis ou o desbalanceamento de classes. Uma das estratégias comumente adotada é a de sobre-amostragem (*oversample*) dos dados, onde se criam exemplos artificiais a fim de aumentar o tamanho do conjunto de dados, ou torná-lo balanceado,

criando exemplos artificiais nas classes menos representativas (minoritárias). Essa estratégia, no entanto, torna a base de treinamento e a classe minoritária pobre na variabilidade de exemplos (BENGIO; COURVILLE; GOODFELLOW, 2015).

Há ainda a estratégia de sob-amostragem (*undersample*), cujo objetivo é eliminar exemplos da classe mais representativa (majoritária) até que ela possua distribuição similar às minoritárias. A grande dificuldade encontrada nessa estratégia é o fato de que, na medida em que exemplos são eliminados, perdem-se também características que podem ser importantes para o treinamento do modelo, além de se diminuir ainda mais o conjunto de dados (BENGIO; COURVILLE; GOODFELLOW, 2015).

Outra estratégia comumente utilizada para criação de exemplos artificiais tanto para treinamento quanto para teste é a denominada “aumento de dados” *(data augmentation)*. Essa estratégia consiste em replicar alguns dos exemplos da amostra de dados inserindo ruídos de forma a torná-los ligeiramente diferentes. Essa medida tem precedentes na natureza dos dados. Muitos exemplos em imagens são muito semelhantes, variando apenas a maneira como a imagem é representada, iluminação, sombreamento etc. O mesmo vale para textos, pois palavras que comumente aparecem dentro de um mesmo contexto podem possuir o mesmo significado (como também postulado pela hipótese distribucional de Harris). Dessa forma, um conjunto de dados artificiais em textos pode fazer uso de dicionários de sinônimos, como proposto por LeCun e Zhang (2015), ou experimentalmente, utilizando os resultados mais similares dentro de um WE para criação de exemplos artificiais (WANG, 2017; BENGIO; COURVILLE; GOODFELLOW, 2015).

Alternativamente, mesmo que haja uma quantidade significativa de dados para treinamento, mas com algum desbalanceamento entre classes, pode se realizar o treinamento sensível a custos (*cost-sensitive learning*). Essa estratégia visa aplicar pesos diferentes para treinamento das classes majoritária e minoritária, onde as classes menos representativas recebem um peso, ou importância, no treinamento superior as classes majoritárias. De forma que o otimizador ajusta o modelo para que os erros de classificação na classe minoritária sejam menos frequentes, pois erros nas classes majoritárias representam um peso menor em busca da menor função custo do modelo. Em outras palavras, o modelo é forçado através dos pesos de cada classe a prestar maior atenção nos exemplos minoritários (HUANG et al. 2016).

No entanto, a utilização dessa heurística requer cautela, uma vez que a representatividade das classes deve ser mantida durante a divisão do conjunto de dados em teste e treino numa rotina de validação cruzada, de forma que os pesos continuem representando o desbalanceamento em cada um dos conjuntos de treinamento e teste (BENGIO; COURVILLE; GOODFELLOW, 2015; HUANG et al. 2017).

# 3 ENCAMINHAMENTOS METODOLÓGICOS

A partir da apresentação do referencial teórico que norteia essa pesquisa, apresentam-se as etapas que compõem essa pesquisa, bem como quais foram as escolhas metodológicas adotadas.

Esse projeto conta com bases de dados secundárias compreendendo sumários de alta de pacientes de dois hospitais brasileiros, um do Paraná e outro de Minas Gerais.

A fim de seguir os critérios éticos e metodológicos da pesquisa científica envolvendo seres humanos, fez-se necessária a devida aprovação dos envolvidos e garantida da instituição que mantém os dados da base secundária, de maneira que seja a confidencialidade dos dados, bem como o consentimento dos indivíduos que se envolvem na pesquisa.

Para tal, o pesquisador solicitou a participação e aprovação dos profissionais codificadores que se envolverão na pesquisa durante seu desenvolvimento (ANEXO 1, ANEXO 2, ANEXO 3).

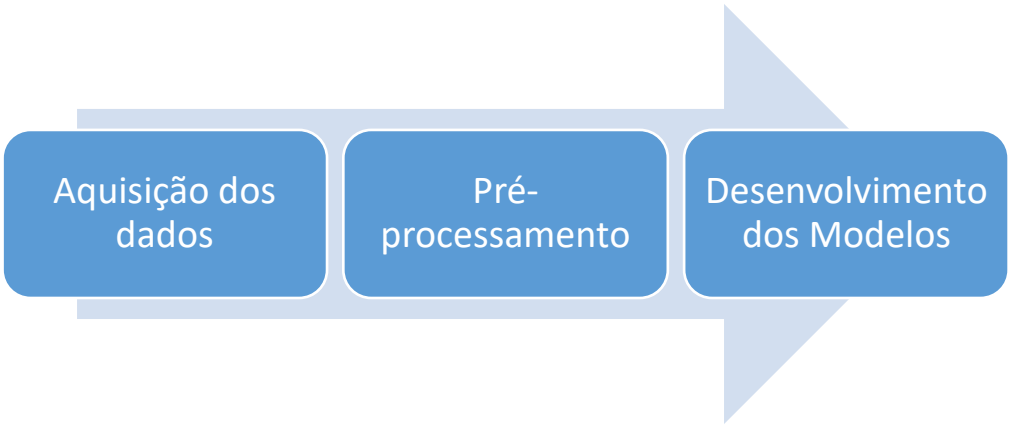
Também é requerida a disponibilização da base de dados secundária (ANEXO 5), bem como foi indicado pelo pesquisador o compromisso com a confidencialidade dos dados (ANEXO 4).

Essa pequisa foi aprovada pelo Comitê de Ética em Pesquisa da Pontifícia Universidade Católica do Paraná (PUCPR), sob o Parecer nº 2.251.282 no dia 31 de Agosto de 2017. A partir da aprovação pelo Comitê de Ética em Pesquisa, o pesquisador envolvido comprometeu-se em manter a confidencialidade dos dados coletados, bem como a privacidade do conteúdo, conforme preconizam os documentos internacionais e a Resolução MS nº 466/2012.

## 3.1 ETAPAS DE PESQUISA

Essa é uma pesquisa de natureza aplicada, exploratória e experimental, composta por 3 etapas (Figura 14) (BONITA;BEAGLEHOLE; KJELLSTRÖM, 2010).

Figura 15 - Etapas da Pesquisa




Fonte: O autor, 2018.

Todas as etapas dessa pesquisa são pormenorizadas nas subseções a seguir, de forma a evidenciar o método proposto. A primeira etapa, apresentada na subseção 3.2, é a aquisição de dados, onde obtiveram-se os dados dos sumários de alta de dois hospitais do Brasil, além de uma base compreendendo os nomes próprios do Brasil. Essas bases foram então utilizadas na próxima etapa, apresentadas na subseção 3.3, que é o pré-processamento de dados, onde os textos contidos nos sumários de alta foram descaracterizados a partir da base de nomes e adaptados para utilização na última etapa apresentada na subseção 3.4, que é o desenvolvimento do modelo, onde os dados pré-processados foram utilizados para elaboração de um modelo de DL para o PLN utilizando WE.

## 3.2 AQUISIÇÃO DOS DADOS

A primeira etapa é a aquisição das bases de dados, que compreendem os 4030 SA e uma base contendo os nomes próprios do Brasil, que ocorrem pelo menos 20 vezes no país, de acordo com levantamento realizado pelo CENSO em 2010. Destes 4030, SA, 2030 estão codificados com algum código referente ao diagnóstico de “outros transtornos do trato urinário (CID N39)”, disponíveis em arquivo de formato

*.pdf* e provenientes de um hospital de Minas Gerais, enquanto os outros 2000 SAforam disponibilizados em arquivo de formato *.txt*, não se relacionam a transtornos do trato urinário e são provenientes de um hospital do Paraná. Essa medida de combinação de bases se faz necessária para garantir que a base de dados apresente exemplos diversificados e com características de escrita diferentes. Para a base de dados do Paraná, rotulou-se todos os exemplos com o código “N39.5”. Esse código não existe na CID e, portanto, foi o código atribuído para especificar um código de diagnóstico “geral” e “não relacionado a outros transtornos do trato urinário”.

Devido ao fato dos textos destas serem organizadas por indivíduos e instituições diferentes, é sabido que essa diferença pode influenciar no modelo. A avaliação aqui é de que essa influência é positiva, dado que essa mescla traz variabilidade semântica aos textos.

A Tabela abaixo mostra a representatividade de cada código dos SA, compreendendo as duas bases de dados.

Tabela 1 - Códigos da CID-10 para transtornos do trato urinário e sua ocorrência na base

| **Código da CID-10** | **Descrição** | **Ocorrência** |
|---|---|---|
| N39.0 | Infecção do trato urinário de localização não especificada | 1762 |
| N39.1 | Proteinúria persistente não especificada | 11 |
| N39.2 | Proteinúria ortostática não especificada | 0 |
| N39.3 | "Incontinência de tensão (""*stress*"")" | 47 |
| N39.4 | Outras incontinências urinárias especificadas | 51 |
| N39.5 | Diversos, Exceto outras transtornos do traro urinário | 2000 |
| N39.8 | Outros transtornos especificados do aparelho urinário | 12 |
| N39.9 | Transtornos não especificados do aparelho urinário | 147 |
| **Total** | | **4030** |

Fonte: O autor, 2018.

Além destas bases, também foi utilizada para essa pesquisa a base contendo todos os nomes do Brasil de acordo com CENSO de 2010, obtidos através de interação com a Controladoria Geral da União (CGU) através da requisição 03950.000489/2018-14. Essa base faz-se necessária para garantir a

descaracterização dos dados contidos nos sumários de alta durante a etapa de pré-processamento.

## 3.3 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento dos dados disponíveis se faz necessário para remoção de ruídos, coletar dados suficientes para modelar os ruídos esperados em uma base e viabilizar estratégias para trabalhar com dados vazios ou inexistentes (FAYYAD; PIATETSKY-SHAPO; SMYTH, 1996).

Desta forma, foi realizado o pré-processamento dos textos contidos nos SA para melhorar os resultados esperados com o modelo proposto, para tal realizaram-se todas as etapas utilizando a linguagem de programação Python em sua versão 3.5.2, a mais estável até a data de desenvolvimento dessa pesquisa.

Os SA utilizados nessa pesquisa compreendem textos que foram escritos de acordo com o entendimento que o profissional de saúde teve da situação clínica do paciente em questão. Como visto anteriormente, esse entendimento pode ter sido constituído com base em exames, evolução clínica do paciente, conduta etc.

Para o pré-processamento dos dados em *.pdf*, fez-se uso da biblioteca PyPDF2 do Python, que faz a leitura de arquivos em formato PDF e os carrega em memória para posterior processamento.Também foi realizado o carregamento dos dados em *.txt*, concatenando ambas as bases a fim de formar apenas um conjunto de dados com seus respectivos rótulos, que são os códigos já atribuídos por codificadores, representados na base pelo último número do código de subcategoria da CID (por exemplo, um SA rotulado como “N39.4” recebe o rótulo “4”)

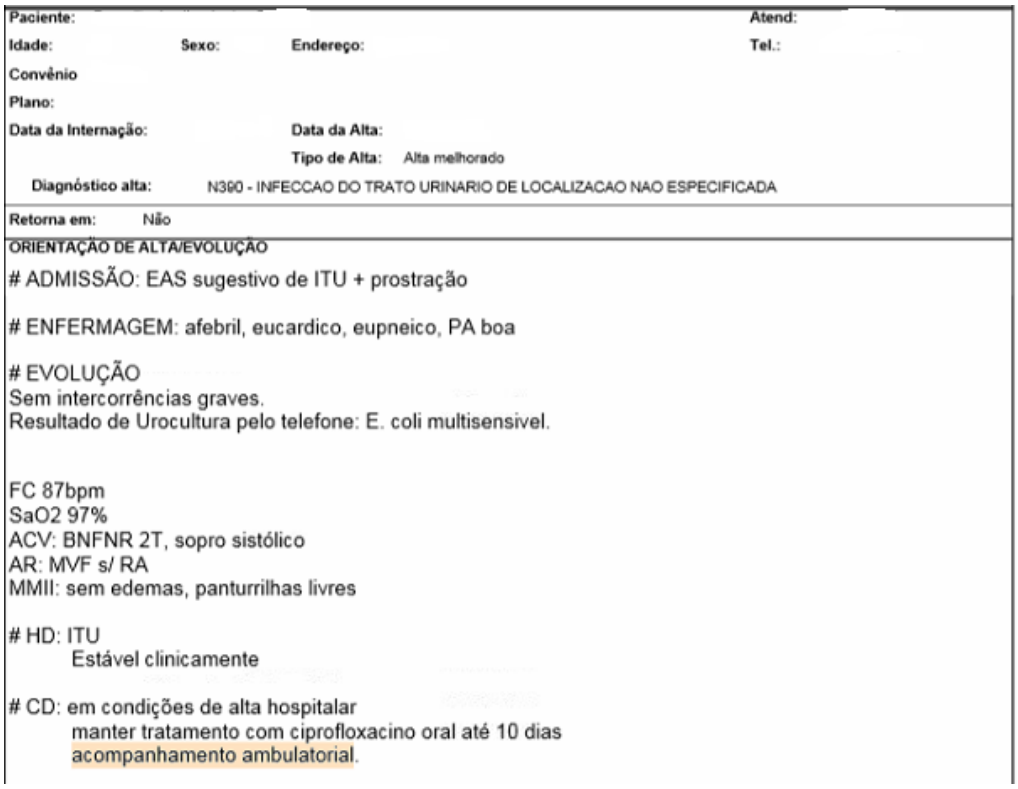
Com todos os dados carregados, utilizando o Python, efetuou-se a *tokenização* dos textos, que representa a separação de cada palavra ou sinal individualmente em *tokens*. Após isso, efetuou-se a normalização dos textos, removendo acentos, pontuações e caracteres inválidos, além de tornar todas as letras dos textos minúsculas e removerdas *stop-words* (palavras de interrupção que pouco contribuem para o significado geral de um texto) contidas na biblioteca NLTK (*Natural Language Toolkit*). O processo de remoção de stop-words, embora seja necessário para remover

palavras que pouco contribuem para obtenção de padrões, pode representar alguma perda de palavras que, eventualmente, possam ajudar a exemplificar algum conceito. Essa característica do processo de remoção de *stop-words* representa uma troca importante.

Com os textos já separados por palavras e normalizados, efetuou-se a descaracterização dos textos a partir do uso de expressões regulares que removem datas, números de documentos e telefones. Além disso, também efetuou-se a descaracterização dos SAs fazendo uso da base de nomes do Brasil, obtida durante a etapa de aquisição de bases. Para a descaracterização, desenvolvou-se um processo iterativo em Python que compara cada um dos nomes contidos na base de nomes aos *tokens* dos SA, removendo os nomes do SA quando encontrados. Com esse processo, removeu-se do texto todos os nomes contidos na base de nomes. Não houve indicativo de palavras de negação para o desenvolvimento do modelo, de forma que não foi possivel verificar sua influência.

A partir deste pré-processamento, foi possível transformar os dados contidos na Figura 15 para os dados na Figura 16.

Figura 16 - SA antes do pré-processamento, em formato .pdf



**Paciente:** **Atend:**
**Idade:** **Sexo:** **Endereço:** **Tel.:**
**Convênio**
**Plano:**
**Data da Internação:** **Data da Alta:**
**Tipo de Alta:** Alta melhorado
**Diagnóstico alta:** N390 - INFECCAO DO TRATO URINARIO DE LOCALIZACAO NAO ESPECIFICADA
**Retorna em:** Não

**ORIENTAÇÃO DE ALTA/EVOLUÇÃO**

# ADMISSÃO: EAS sugestivo de ITU + prostração

# ENFERMAGEM: afebril, eucardico, eupneico, PA boa

# EVOLUÇÃO
Sem intercorrências graves.
Resultado de Urocultura pelo telefone: E. coli multisensivel.

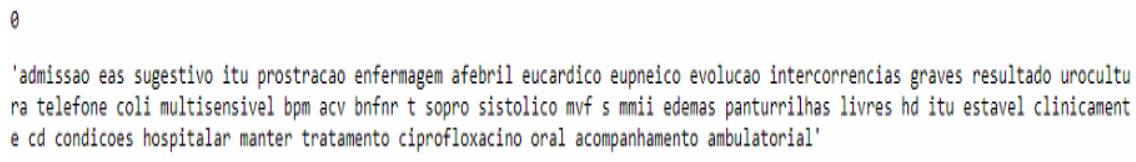
FC 87bpm
SaO2 97%
ACV: BNFNR 2T, sopro sistólico
AR: MVF s/ RA
MMII: sem edemas, panturrilhas livres

# HD: ITU
Estável clinicamente

# CD: em condições de alta hospitalar
manter tratamento com ciprofloxacino oral até 10 dias
acompanhamento ambulatorial.

Fonte: O autor, 2018.

Figura 17 - SA depois do pré-processamento




Fonte: O autor, 2018.

Nota-se no entanto, que nem todas as palavras contidas no sumário de alta antes do pré-processamento estão presentes nos sumários processados. Isso se deve à remoção das *stop-words* e dos nomes contidos na lista de nomes do Brasil, que fez com que palavras como “boa”, que também é um nome próprio contido na listagem de nomes, tenham sido removidos.

Verificou-se ainda que na lista de nomes encontram-se palavras que representam características muito importantes para essa pesquisa, é o caso de palavras (e nomes próprios) como “clara” e “Itu”, a primeira representando uma caracerística de cor da urina e a segunda é frequentemente utilizada como acrônimo para “Infecção do Trato Urinário” no vernáculo clínico. Antevendo possíveis interferências no modelo a partir da remoção destas, optou-se por preservá-las, de forma que não foram removidas dos SAs como as demais palavras.

## 3.4 DESENVOLVIMENTO DOS MODELOS

Os modelos treinados nesse trabalho foram desenvolvidos utilizando a linguagem de programação Python, com auxílio da biblioteca “*Keras*”, que implementa DL a partir da biblioteca “*Tensorflow*” e da biblioteca “Scikit Learn”, que implementa outros modelos de ML. Esses modelos implementam DL a partir de CNN para classificar SAs de acordo com os códigos específicos da CID. Para tal, utilizou-se como entrada dos modelos textos clínicos convertidos em WE e seus respectivos rótulos, para classificação supervisionada, que necessita de exemplos rotulados para classificar os textos de acordo com os códigos da CID.

Esses WE foram treinados e obtidos de maneira não-supervisionada a partir da própria base de dados, fazendo uso da biblioteca “*glove-python*” que implementa o algoritmo GloVe para criação de vetores de palavras baseados em co-ocorrência de palavras dentro de um determinado contexto. A escolha pelo algoritmo GloVe se deu devido ao trabalho de Socher (2014) que, comparando o GloVe a outros algoritmos como o *word2vec*, verificou que o GloVe capturade maneira mais eficiente o significado semântico das palavras, além de precisar de um tempo menor para treinamento.

Uma alternativa comum ao desenvolvimento do WE, a partir dos dados é utilizar WE disponíveis na internet e pré-treinados partindo de textos jornalísticos e outras fontes de dados abertos. No entanto, a opção aqui foi pelo desenvolvimento do WE a partir da própria base de dados devido ao fato do vocabulário clínico possuir especificidades (como acrônimos) que não são possíveis obter utilizando WE pré-treinados a partir de outros textos, o que potencialmente pode interferir na qualidade dos modelos testados. Um exemplo disso, é a palavra “Itu”, que no vocabulário clínico representa um acrônimo que significa “Infecção do Trato Urinário”, enquanto também representa o nome de uma cidade do Brasil.

Dessa forma, o WE desenvolvido possui as características semânticas das palavras do vocabulário clínico em português do Brasil, removendo a necessidade em definir essas características manualmente, como no processo de PLN tradicional, fartamente abordado no referencial teórico.

A fim de testar a influência de vetores de palavras de diferentes tamanhos (dimensões) no classificador final, desenvolveu-se com essa pesquisa WE com 50, 100, 300 e 500 dimensões. Essa variabilidade no tamanho dos WE é também verificada no trabalhoLenc e Král (2017) eno de Socher (2014). Todos os WE foram treinados com os mesmos textos, oriundos dos SAs, e com parâmetros definidos de acordo com Socher (2014), que prevê treinamento dos WE com taxa de aprendizagem de 0.5, janelas de tamanho 10, durante 50 épocas para WE com menos de 300 dimensões e 100 épocas para modelos com mais de 300 dimensões.

Para o desenvolvimento dos modelos de DL, adotou-se o modelo CNN, pois é o modelo predominante na literatura para extração de características gerais em textos e para o qual se possui diversos trabalhos que mostram sua utilização no PLN a partir do uso de WE, incluindo o modelo *baseline*de Kim (2014).

Realizou-se ainda a aproximação sugerida por Wallace e Zhang (2015), onde a partir da implementação do modelo *baseline*, procura-se por arquitetura e parâmetros ótimos até se obter a melhor performance do modelo. Para tanto, implementou-se o modelo *baseline* de Kim (2014) em sua versão com 1 camada de convolução e, a partir dos resultados obtidos com esse modelo, treinou-se outros 20 modelos a fim de comparar os resultados frente ao *baseline*e encontrar o melhor modelo para o problema proposto.

Para treinamento dos modelos, utilizou-se a estratégia k-*fold* estratificado com aprendizado sensível a custos (*cost-sensitive learning*), com conjunto de validação de 10% do total de registros e 10% para teste. Nessa estratégia de treinamento, utilizou-se*k*=10, pois é heurística comum na literatura, além do fato de que esse valor de *k* garante que haja pelo menos um exemplo para treinamento e teste da classe minoritária (conforme distribuição de classes verificada na Tabela 1). Já essa versão estratificada do algoritmo k-*fold* permite que as 10 divisões da base possuam a mesma representatividade de classes que a base original e que durante cada ciclo de treinamento, 1/10 da base seja utilizada para teste. Desta forma, ao fim dos 10 *folds* toda a base foi utilizada para treino e para teste. Essa medida, associada ao aprendizado sensível a custos (*cost-sensitive learning*) visa minimizar os efeitos do severo desbalanceamento de base, conforme sugerido por Huang et al. (2016).

A partir das aproximações iniciais e da exploração sugerida por Wallace e Zhang (2015), treinou-se as CNN durante 20 épocas por *fold,* com *batch* de tamanho 50 (como originalmente proposto por Kim (2014)), e hiperparâmetros otimizados usando o algoritmo *Adam*. Embora Wallace e Zhang (2015) sugiram a utilização do otimizador “AdaDelta”, optou-se pelo otimizador “Adam”,pois Lei Ba e Kingma (2015) concluíram que esse otimizador supera o “AdaDelta” ao encontrar resultados ótimos de hiperparâmetros em um tempo reduzido.

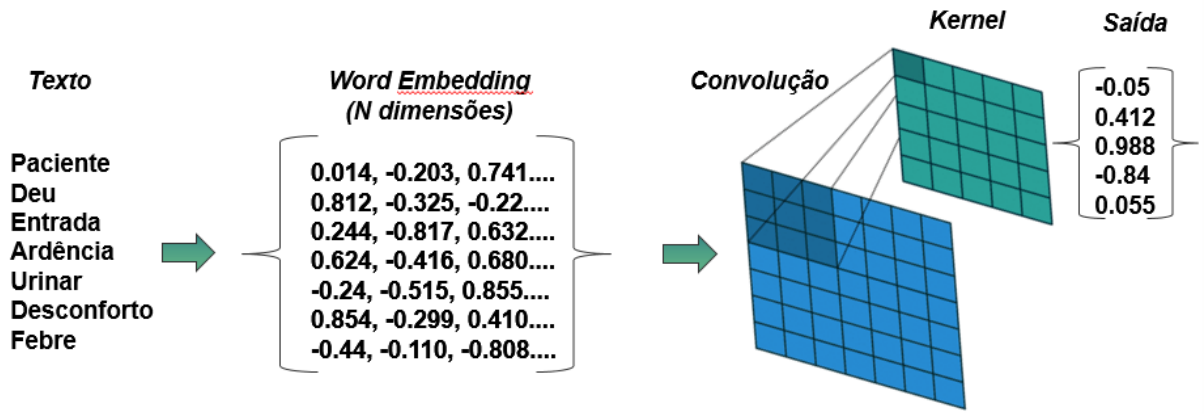
O otimizador “Adam” é projetado para combinar as vantagens de dois otimizadores populares, o “AdaGrad” e “RMSProp”. Algumas das vantagens do “Adam” frente aos demais são que a atualização de parâmetros desconsidera qualquer reescalonamento do gradiente, e o passo das atualizações são limitadas pelo hiperparâmetro, de forma que ele trabalha com gradientes esparsos e, naturalmente, executa uma forma de adaptação de tamanho de passo (LEI BA; KINGMA, 2014).

Seguindo ainda a proposta de aproximação sugerida por Wallace e Zhang (2015), optou-se que as CNN treinadas possuam 128 filtros (*kernel*), e janelas de diferentes tamanhos em cada camada, pois é desejável obter diferentes padrões a partir de cada convolução. Essas camadas de convolução possuem função de ativação do tipo “ReLu”, pois essa ativação evita problemas de fuga de gradiente (*vanishing gradient*), que reduzem a qualidade do modelo final. Depois de cada camada de convolução existe uma camada de *1-max-pooling*, que visa identificar os sinais ativados com maior força durante a convolução. Após isso, uma nova camada de *pooling* global faz a verificação dos sinais com maior ativação a partir das outras camadas e alimenta um classificador baseado em RNA do tipo MLP denso de tamanho 128, mesmo tamanho do *kernel*, com *dropout* 0.2 como medida de regularização para suavizar qualquer tipo de sobre ajuste, conforme sugerido pelo por Wallace e Zhang (2015). A ativação da camada densa é do tipo *softmax*, pois essa provê uma distribuição de probabilidades entre as 7 classes possíveis (6 códigos da CID de agora com Tabela 1 e 1 código que representa um diagnóstico geral).

Além do *dropout* e objetivando uma maior capacidade de regularização adotou-se também a estratégia de regularização conhecida por “interrupção precoce” (*early-stopping*) durante o treinamento. Nessa estratégia, instancia-se uma função que observa a evolução da função custo durante o treinamento para o conjunto de teste em cada *fold* para garantir que, na eventualidade do valor da função custo não sofrer redução durante 5 épocas, o treinamento naquele *fold* é interrompido, os pesos atualizados e parte-se para o treinamento no próximo *fold,* como sugerido por Prechetl (2012). Essa estratégia visa garantir que o treinamento é interrompido o mais próximo possível do valor mínimo global da função custo.

Os diferentes modelos treinados e testados são então exemplificados no Quadro 4. A partir das definições acima, adotou-se um protocolo de testes para comparar o desempenho do modelo *baseline*de Kim (2014) frente a outros 20 modelos de CNN diferentes com WE de tamanhos variados, com o objetivo final é encontrar o modelo que melhor codifique os sumários de alta de acordo com os códigos da CID-10 (Figura 18).

Figura 18–Modelo de Kim, adaptado para essa pesquisa




Fonte: O autor, 2018.

Adicionalmente, e além do modelo de DL utilizado nesta pesquisa, implementou-se com a mesma base de dados outro modelo que não utiliza DL, tampouco utiliza WE. Esse modelo utiliza o algoritmo de *Support Vector Machine* (SVM), também considerado um algoritmo “raso” (ou não profundo), para classificar documentos com base na ocorrência (contagem) de palavras de um conjunto de dados.

Para implementação desse modelo, utilizou-se o processo de *grid-search* (busca dinâmica) para encontrar os melhores hiperparâmetros, e os textos utilizados para treinamento e teste são os mesmos que os utilizados para os modelos de DL, com treinamento sensível a custos. Dessa forma, tem-se que o modelo de SVM aqui representado é o modelo que apresenta as melhores condições de comparação com o modelo de DL

Quadro 4 - Modelos treinados e testados nesta pesquisa

| # Modelo | Tamanho WE | # Camadas CNN | Tamanho *kernel* | Tamanho Janela | Ativação | *Pooling* | Tamanho do *batch* | Taxa de *Dropout* |
|---|---|---|---|---|---|---|---|---|
| *Baseline* | 300 | 1 | 100 | 3 | Relu | *1-max* | 50 | 0.5 |
| 1 | 50 | 1 | 128 | 5 | ReLu | 1-*max* | 50 | 0.2 |
| 2 | 50 | 2 | 128 | 5, 8 | ReLu | 1-*max* | 50 | 0.2 |
| 3 | 50 | 3 | 128 | 5, 8, 10 | ReLu | 1-*max* | 50 | 0.2 |
| 4 | 50 | 4 | 128 | 5, 8, 10, 12 | ReLu | 1-*max* | 50 | 0.2 |
| 5 | 50 | 5 | 128 | 5, 8, 10, 12, 15 | ReLu | 1-*max* | 50 | 0.2 |
| 6 | 100 | 1 | 128 | 5 | ReLu | 1-*max* | 50 | 0.2 |
| 7 | 100 | 2 | 128 | 5, 8 | ReLu | 1-*max* | 50 | 0.2 |
| 8 | 100 | 3 | 128 | 5, 8, 10 | ReLu | 1-*max* | 50 | 0.2 |
| 9 | 100 | 4 | 128 | 5, 8, 10, 12 | ReLu | 1-*max* | 50 | 0.2 |
| 10 | 100 | 5 | 128 | 5, 8, 10, 12, 15 | ReLu | 1-*max* | 50 | 0.2 |
| 11 | 300 | 1 | 128 | 5 | ReLu | 1-*max* | 50 | 0.2 |
| 12 | 300 | 2 | 128 | 5, 8 | ReLu | 1-*max* | 50 | 0.2 |
| 13 | 300 | 3 | 128 | 5, 8, 10 | ReLu | 1-*max* | 50 | 0.2 |
| 14 | 300 | 4 | 128 | 5, 8, 10, 12 | ReLu | 1-*max* | 50 | 0.2 |
| 15 | 300 | 5 | 128 | 5, 8, 10, 12, 15 | ReLu | 1-*max* | 50 | 0.2 |
| 16 | 500 | 1 | 128 | 5 | ReLu | 1-*max* | 50 | 0.2 |
| 17 | 500 | 2 | 128 | 5, 8 | ReLu | 1-*max* | 50 | 0.2 |
| 18 | 500 | 3 | 128 | 5, 8, 10 | ReLu | 1-*max* | 50 | 0.2 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 19 | 500 | 4 | 128 | 5, 8, 10, 12 | ReLu | 1-*max* | 50 | 0.2 |
| 20 | 500 | 5 | 128 | 5, 8, 10, 12, 15 | ReLu | 1-*max* | 50 | 0.2 |

Fonte: O autor, 2018.

### 3.4.1 Protocolo de testes

Com o objetivo de testar os modelos treinados, adotou-se rotinas de testes diferentes. A primeira, utilizada apenas para avaliação dos modelos de DL, é o teste a partir da rotina de validação cruzada aplicada ao conjunto de treinamento (base desbalanceada). Na segunda estratégia, utilizada para avaliação de ambos, utilizou-se a técnica de aumento de dados (*data augmentation*) para criar exemplos artificiais que compõem uma base de testes balanceada.

A primeira estratégia está diretamente associada a forma como os modelos foram desenvolvidos. Considerando que a classe minoritária conta com 11 exemplos (Tabela 1), uma estratégia alternativa de teste (como *holdout*) diminuiria significativamente a quantidade de exemplos utilizados para treinamento, de forma que a validação cruzada por meio do algoritmo *k-fold* estratificado permite que toda a base seja utilizada como treino e teste, conforme enunciado anteriormente. Com essa estratégia, tem-se o resultado avaliado para os testes em cada *fold*, de forma a obter como avaliação uma média de desempenho associada a um desvio padrão ao final do treinamento.

Para a segunda estratégia, optou-se pela criação de exemplos artificiais seguindo a heurística proposta por Wang (2017). Dado que os WE desenvolvidos a partir dessa base são vetores numéricos que representam palavras, utilizou-se os próprios WE para identificar quais vetores alternativos mais se assemelham aos vetores das palavras contidas nos SAs utilizados para treinamento. Para tanto, utilizou-se a função "*most_similar()*" da biblioteca "glove_python". Essa função implementa o cálculo de similaridade de cossenos entre dois vetores numéricos, de forma a estabelecer qual vetor tem a maior similaridade ao vetor alvo. Em outras palavras, essa procura implica na identificação de palavras que possam ser sinônimos umas das outras dentro dos WE. O objetivo dessa estratégia é criar uma base balanceada que com esses sinônimos, que represente SAs alternativos, como se fossem escritos de forma diferente, artificialmente representando a forma como outro indivíduo escreveria o mesmo SA, preservando seu rótulo original.

Dessa forma, optou-se por realizar a seleção randômica de 10 SAs de cada uma das classes indicadas na Tabela 1 e para cada um desses criar um SA artificial

seguindo a heurística mencionada acima, substituindo cada uma das palavras pelas palavras mais semelhantes do WE, totalizando 70 SAs artificiais. Optou-se pela seleção de 10 exemplos, pois assim seria possível garantir a não repetição de nenhum SA. A Figura 17 representa um SA artificial criado a partir do SA representado na Figura 16.

Figura 19 - SA artificial

'evidenciada leucocituria rinossinusopatia repeticao inapetencia supervisao quiexas eupneico eucardico comqaudro roux reorienta mos estetico andamento entraremos multissensivel coli irpm rci bnnf margeando mie+ sopro eupneica ggeral exaltados panturrilhas livres panturrilhas dialiticas repeticao hemodinamicamente estavel cerumim clinicas relatorio fisioterapi parameningeo nzilpeni cilina vestibulo gasparin geriatrico'

Fonte: O autor, 2018

### 3.4.2 Avaliação dos modelos

.

Para avaliação dos modelos treinados, mensurou-se sua qualidade a partir das métricas comumente propostas na literatura para classificação de textos em bases balanceadas e desbalanceadas.

Para a primeira rotina de testes foi utilizada abordagem similar a aquela utilizada por Carlson et al. (2017), Kavuluru e Rios (2015), e Ayyar e Bear (2017), que usaram a medida de *Fscore* e suas variantes para avaliar a qualidade do modelo desenvolvido para classificação de textos. Dado o desbalanceamento de classes, optou-se pela avaliação a partir da variante do Fscore denominada *micro-F1* score ($F1_{\mu}$), que segundo Elkan, Lipton e Naryanaswamy (2014), é menos afetado pelo desempenho de classes minoritárias se comparado a outras métricas. O $F1_{\mu}$ relaciona a micro-precisão ($\text{precisão}_{\mu}$) (Equação 3.1) e micro-*recall* ($\text{recall}_{\mu}$) (Equação 3.2) (especificade e sensibilidade) para produzir uma medida que consiga melhor descrever a qualidade da classificação do texto em base desbalanceada, a partir da relação entre exemplos classificados como verdadeiros positivo (*tp*), falsos positivos (*fp*), e falsos negativos (*fn*). Essa medida é a métrica de avaliação sugerida por Lapalme e Sokolova (2009) para classificadores multi-classe e descrita conforme Equação 3.3 (LAPALME;  SOKOLOVA, 2009).

$$Precisão_{\mu} = \sum_{i=1}^{l} \frac{tp_i}{(tp_i + fp_i)} \tag{3.1}$$

$$Recall_{\mu} = \sum_{i=1}^{l} \frac{tp_i}{(tp_i + fn_i)} \tag{3.2}$$

$$F1_{\mu} = \frac{(\beta^2 + 1)Precisão \cdot Recall}{\beta^2 \cdot Precisão + Recall} \tag{3.3}$$

Como o treinamento dos modelos é feito em 10 conjuntos (*folds*), a qualidade dos modelos na primeira rotina de testes é feita com base no $F1_{\mu}$ médio, associado ao seu desvio padrão amostral (*s*).

Para avaliação dos modelos na segunda rotina de testes, em base balanceada, aplicou-se a avaliação a partir da acurácia geral ($AG$), ou seja, quantos exemplos foram corretamente classificados do total de exemplos do dataset. Essa medida é vista por Lapalme e Sokolova (2009) como "uma avaliação confiável para classificadores". No entanto, objetivando avaliar a classificação para cada uma das classes, expandiu-se a avaliação nessa rotina de testes para avaliar também a $\text{precisão}_{\mu}$, $\text{recall}_{\mu}$ e $F1_{\mu}$ por classe.

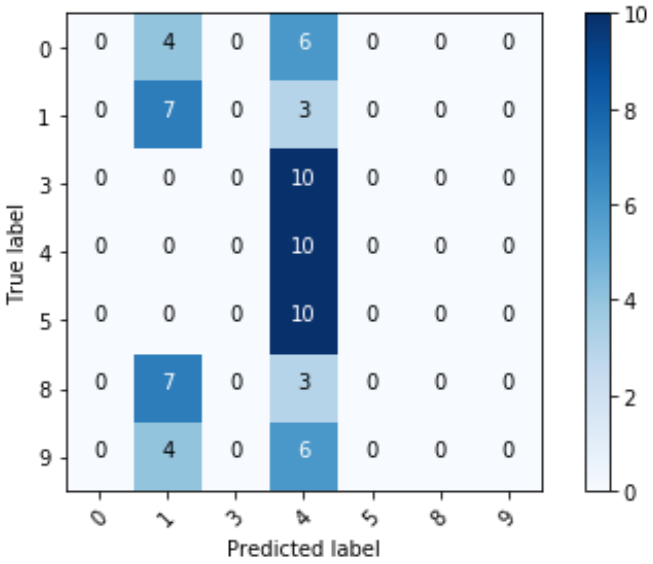
A fim de avaliar a qualidade do treinamento, analisou-se se há sobre-ajuste (*overfitting*) nos modelos treinados a partir da diferença gráfica entre função custo (*loss*) ao longo do treino e do teste. Quando esses valores são muito diferentes, particularmente quando o valor da função custo no treino é muito menor que a função custo no teste, é indício de que o modelo está sobre-ajustado, ou seja, aprendeu muito bem os dados de treinamento mas não consegue generalizar para o conjunto de teste.

Definiu-se como os melhores modelos aqueles que apresentam maior $F1_{\mu}$ médio, com menor desvio ($S$) e maior acurácia geral ($AG$). Os resultados obtidos a partir do desenvolvimento dos modelos de CNN estão apresentados por tamanho de WE. Apresentam-se de forma tabular os resultados de todos os modelos. Optou-se por representar graficamente apenas os resultados dos melhores modelos por tamanho de WE, apresentando sua matriz de confusão e expandindo seus resultados a nível de classe, conforme previamente enunciado. Os melhores modelos estão identificados com um "*" em frente ao seu número nas tabelas de resultados.

# 4 RESULTADOS

Obteve-se após o pré-processamento dos 4030 SAs um total de 327.529 *tokens* e 20.322 palavras distintas oriundas destes textos clínicos. Com esses dados, treinou-se os WE e os modelos de DL, iniciando pelo do treinamento do modelo *baseline* (KIM, 2014), que obteve $F1_{\mu}$ médio de 0.08, com $S$ de 0.05 e $AG$ 24.28% (Figura 18)

Figura 20 - Matriz de Confusão da segunda rotina de testes para o modelo *baseline*




Fonte: O autor, 2018.

Como mencionado anteriormente, também treinou-se um modelo SVM sem fazer uso dos WE. Este modelo obteve $F1_{\mu}$ médio de 0.33 e e $AG$ 32.85%.

Figura 21 - Matriz de Confusão da segunda rotina de testes para o modelo baseline

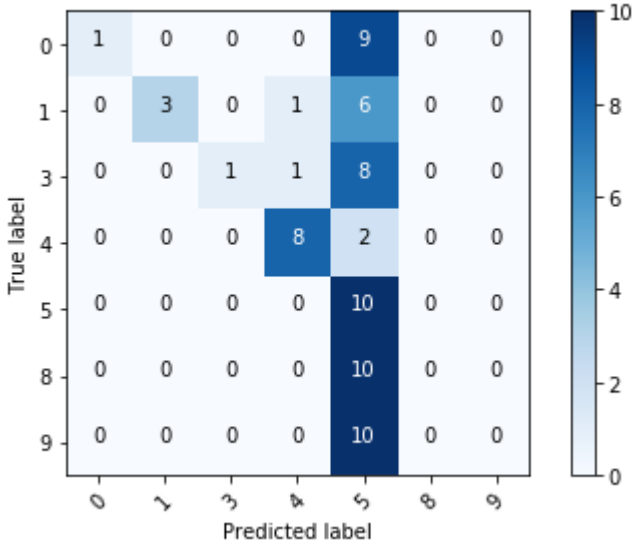



Tabela 2 - Resultados da segunda rotina de testes para o modelo SVM

| Classe | precisão $_{\mu}$ | recall$_{\mu}$ | $F1_{\mu}$ |
|---|---|---|---|
| 0 | 1.00 | 0.10 | 0.18 |
| 1 | 1.00 | 0.30 | 0.46 |
| 3 | 1.00 | 0.10 | 0.18 |
| 4 | 0.80 | 0.80 | 0.80 |
| 5 | 0.18 | 0.31 | 0.31 |
| 8 | 0.00 | 0.00 | 0.00 |
| 9 | 0.00 | 0.00 | 0.00 |

Para os modelos adaptados do *baseline*, desenvolvidos com WE de 50 dimensões, verificou-se os melhores resultados gerais no Modelo 3 (Figura 19), cuja descrição e configuração está representada no Quadro 4.

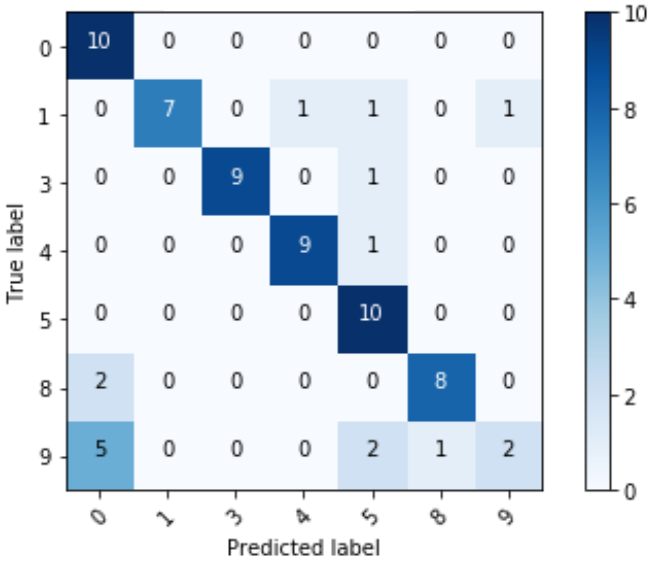
Tabela 3 - Resultados da primeira rotina de testes para os WE com 50 dimensões

| # Modelo | $F1_{\mu}$ Médio | *S* | *AG (%)* |
|---|---|---|---|
| 1 | 0.86 | 0.15 | 72.85 |
| 2 | 0.91 | 0.13 | 61.42 |
| 3* | 0.95 | 0.04 | 78.57 |

| | | | |
|---|---|---|---|
| 4 | 0.93 | 0.07 | 58.57 |
| 5 | 0.95 | 0.05 | 74.28 |

Fonte: O autor, 2018.

Figura 22 - Matriz de Confusão da segunda rotina de testes do melhor modelo de CNN com WE de 50 dimensões (Modelo#3)




Fonte: O autor, 2018.

Tabela 4 - Resultados por classe da segunda rotina de testes do melhor modelo de CNN com WE de 50 dimensões

| Classe | precisão $_{\mu}$ | recall$_{\mu}$ | $F1_{\mu}$ |
|---|---|---|---|
| 0 | 0.45 | 1.00 | 0.62 |
| 1 | 0.90 | 0.90 | 0.90 |
| 3 | 1.00 | 0.90 | 0.95 |
| 4 | 1.00 | 0.90 | 0.95 |
| 5 | 0.64 | 0.90 | 0.75 |
| 8 | 1.00 | 0.20 | 0.33 |
| 9 | 1.00 | 0.40 | 0.57 |

Fonte: O autor, 2018.

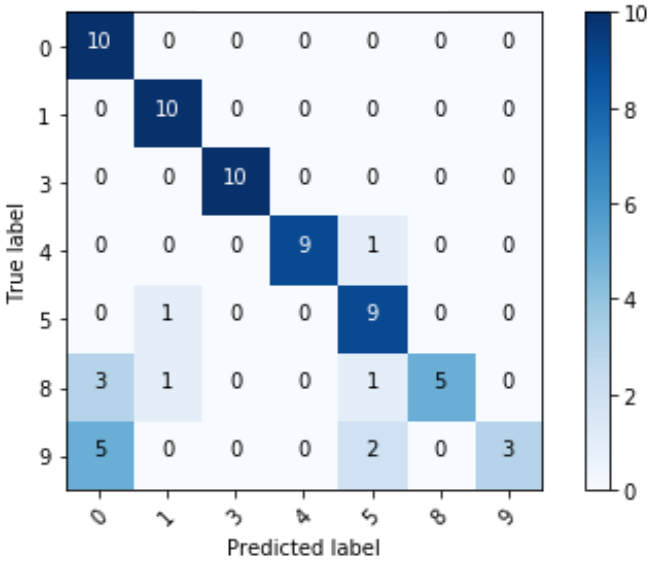
Para os resultados com WE constituídos de 100 dimensões, o modelo com quatro camadas de convolução apresentou o melhor desempenho geral, avaliado à partir do seu $F1_{\mu}$ médio e da quantidade de acertos na base artificial.

Tabela 5 - Resultados da primeira rotina de testes para os WE com 100 dimensões

| # Modelo | $F1_{\mu}$ Médio | *S* | *AG (%)* |
|---|---|---|---|
| 6 | 0.88 | 0.15 | 70.00 |
| 7 | 0.91 | 0.13 | 71.41 |
| 8 | 0.92 | 0.09 | 64.28 |
| 9* | 0.94 | 0.09 | 80.00 |
| 10 | 0.89 | 0.10 | 74.28 |

Fonte: O autor, 2018.

Figura 23 - Matriz de Confusão da segunda rotina de testes do melhor modelo de CNN com WE de 100 dimensões (Modelo#9)




Fonte: O autor, 2018.

Tabela 6 - Resultados por classe da segunda rotina de testes do melhor modelo de CNN com WE de 100 dimensões (Modelo#9)

| Classe | precisão $_{\mu}$ | recall$_{\mu}$ | $F1_{\mu}$ |
|---|---|---|---|
| 0 | 0.55 | 1.00 | 0.71 |
| 1 | 0.83 | 1.00 | 0.90 |
| 3 | 1.00 | 1.00 | 1.00 |

| | | | |
|---|---|---|---|
| 4 | 1.00 | 0.90 | 0.95 |
| 5 | 0.69 | 0.90 | 0.78 |
| 8 | 1.00 | 0.50 | 0.66 |
| 9 | 1.00 | 0.30 | 0.46 |

Fonte: O autor, 2018.

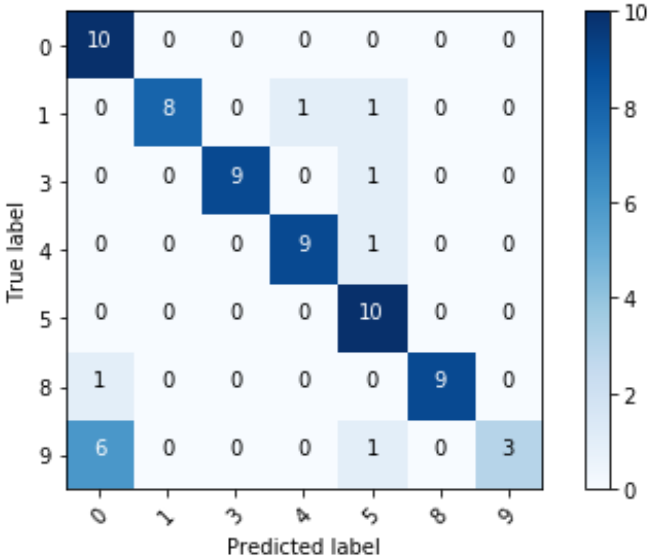
Para os modelos constituídos a partir dos WE de 300 dimensões, o modelo com duas camadas de convolução apresentou os melhores resultados, sem indícios de sobreajuste.

Tabela 7 - Resultados da primeira rotina de testes para os WE com 300 dimensões

| # Modelo | $F1_{\mu}$ Médio | *S* | *AG (%)* |
|---|---|---|---|
| 11 | 0.92 | 0.08 | 64.28 |
| 12* | 0.95 | 0.06 | 82.85 |
| 13 | 0.96 | 0.07 | 77.14 |
| 14 | 0.94 | 0.09 | 74.28 |
| 15 | 0.94 | 0.08 | 77.14 |

Fonte: O autor, 2018.

Figura 24 - Matriz de Confusão da segunda rotina de testes do melhor modelo de CNN com WE de 300 dimensões (Modelo#12)




Fonte: O autor, 2018.

Tabela 8 - Resultados por classe da segunda rotina de testes do melhor modelo de CNN com WE de 300 dimensões (Modelo#12)

| Classe | precisão $_{\mu}$ | recall$_{\mu}$ | $F1_{\mu}$ |
|---|---|---|---|
| 0 | 0.59 | 1.00 | 0.74 |
| 1 | 1.00 | 0.80 | 0.89 |
| 3 | 1.00 | 0.90 | 0.95 |
| 4 | 0.90 | 0.90 | 0.90 |
| 5 | 0.71 | 1.00 | 0.83 |
| 8 | 1.00 | 0.90 | 0.95 |
| 9 | 1.00 | 0.30 | 0.46 |

Fonte: O autor, 2018.

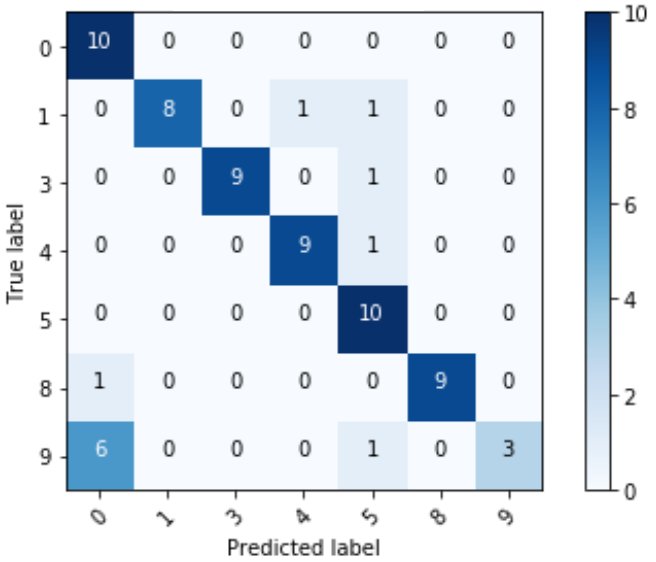
Para os modelos constituídos a partir dos WE de 500 dimensões, verifica-se que o modelo com 4 camadas de convolução apresenta os melhores resultados gerais frente a todos os demais modelos treinados, mesmo aqueles com WE de diferentes tamanhos.

Tabela 9 - Resultados da primeira rotina de testes para os WE com 500 dimensões

| # Modelo | $F1_{\mu}$ Médio | *S* | *AG (%)* |
|---|---|---|---|
| 16 | 0.91 | 0.13 | 71.42 |
| 17 | 0.90 | 0.10 | 72.85 |
| 18 | 0.95 | 0.09 | 74.28 |
| 19* | 0.97 | 0.04 | 82.85 |
| 20 | 0.96 | 0.06 | 78.57 |

Fonte: O autor, 2018.

Figura 25 - Matriz de Confusão da segunda rotina de testes do melhor modelo de CNN com WE de 500 dimensões (Modelo#19)




Fonte: O autor, 2018.

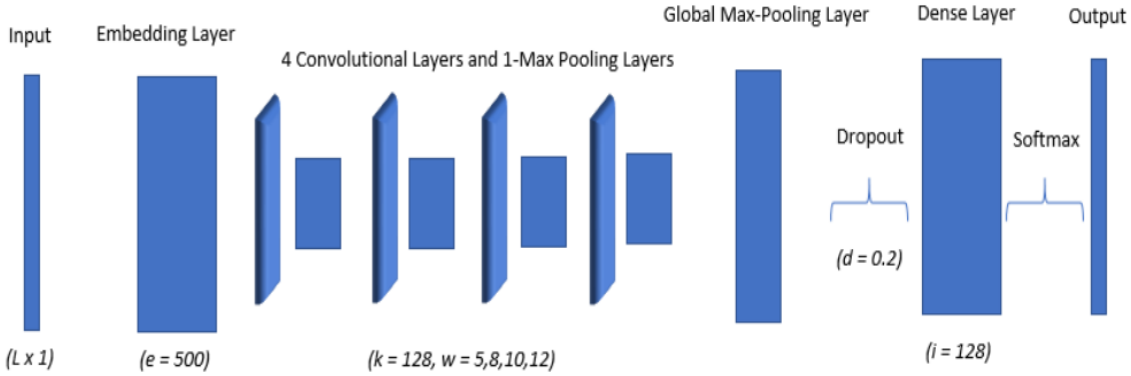
Tabela 10 - Resultados por classe da segunda rotina de testes do melhor modelo de CNN com WE de 500 dimensões (Modelo#19)

| Classe | precisão $_{\mu}$ | recall$_{\mu}$ | $F1_{\mu}$ |
|---|---|---|---|
| 0 | 0.59 | 1.00 | 0.74 |
| 1 | 1.00 | 0.80 | 0.89 |
| 3 | 1.00 | 0.90 | 0.95 |
| 4 | 0.90 | 0.90 | 0.90 |
| 5 | 0.71 | 1.00 | 0.83 |
| 8 | 1.00 | 0.90 | 0.95 |
| 9 | 1.00 | 0.30 | 0.46 |

Fonte: O autor, 2018.

Com base nestes resultados, apresenta-se graficamente o fluxo que representa o Modelo#19, demonstrando sua estrutura conforme Quadro 4.

Figura 26 - Representação gráfica do fluxo do modelo proposto (Modelo#19)




Fonte: O autor, 2018.

## 5 CONCLUSÕES

Os resultados permitem verificar que modelos com um $F1_{\mu}$ médio elevado e com um $S$ pequeno tendem a possuir melhor capacidade de generalização, avaliada à partir do $F1_{\mu}$ durante a primeira rotina de teste em cada *fold* e da $AG$ na segunda rotina de testes. Também foi possível verificar que modelos de DL que fazem uso de WE classificam melhor documentos clínicos se comparados a modelos rasos, aqui representados por modelos SVM.

Dessa forma, a partir dos resultados, identificou-se que o modelo#19 é o melhor modelo classificador obtido dentro deste percurso experimental, sendo este o modelo proposto para codificação de narrativas clínicas a partir do uso de DL. Resultados de $AG$ similares foram obtidos com o modelo#12, sugerindo que WE maiores, como no modelo#19, não são os únicos a obter bons resultados de classificação.

De igual forma, foi possível verificar excelentes resultados em todos os tamanhos de WE, com as mais variadas configurações de modelo. No entanto, é possível verificar que modelos com menos camadas de convolução tendem a possuir $F1_{\mu}$ menores e maiores erros na classificação das instâncias artificiais, fato constatável analisando os resultados para o modelo#1, modelo#6 e modelo#11. Esse achado parece ser minimizado com WE de maior tamanho, como no modelo#16.

Também verifica-se que a medida de regularização adotada (*dropout*) somada a interrupção precoce do treinamento (*early-stopping*) rapidamente diminui os efeitos do sobre ajuste verificado nas primeiras interações (*folds*), de forma que a perda (*loss*) no treino e no teste são equivalentes ao final do treinamento de todos os modelos.

Outro achado interessante dessa pesquisa, é que mesmo os melhores modelos classificam com baixa assertividade a classe 9, que representa o código N39.9 da CID-10. Avalia-se que esse resultado se deve ao fato da classe N39.9 representar uma classe muito geral (transtornos não especificados do aparelho urinário), de forma que é difícil para o classificador encontrar características comuns para esse tipo de diagnóstico. Outra possibilidade que se evidencia com base nos resultados é que o Modelo 19 classifica exemplos originalmente rotulados como N39.9 como sendo N39.0, indicando que esses diagnósticos podem compartilhar de características semelhantes, ou mesmo que os SAs originalmente rotulados como N39.9 deveriam ser codificados como N39.0.

No entanto, é importante lembrar que os resultados de $AG$ foram obtidos a partir da segunda rotina de testes, que compreende base de testes artificial. A estratégia utilizada para criação dessa base pode ter inserido ruídos indesejáveis nos textos, de forma que outras medidas de avaliação com bases balanceadas podem melhorar o entendimento e avaliação do modelo proposto.

De igual modo é possível verificar que, embora haja um desbalanceamento severo da base aqui utilizada, o classificador tratou de maneira equivalente as classes disponíveis na base artificial utilizada para validação do modelo. Isso indica que o treinamento sensível a custos e que a base estratificada foram efetivos para tratar o problema de desbalanceamento.

Em linhas gerais, verifica-se que o modelo proposto nesta pesquisa aponta para a possibilidade da automação da codificação de narrativas clínicas com o uso de DL, utilizando WE como entrada do modelo.

# 6 TRABALHOS FUTUROS

Como próximos passos dessa pesquisa, ressalta-se a importância de possuir uma base maior de exemplos de SAs para treinamento e testes, de forma que sejam possíveis novos testes com rotinas de *holdout* e abordando outras métricas de classificação de textos em bases desbalanceadas.

Essa pesquisa também demonstra a importância de expansão dos WE com vocabulário clínico para muito além daqueles exemplos utilizados para treinamento deste modelo, de forma que os mesmos resultados possam ser obtidos em outras especialidades, e esse é um dos trabalhos futuros possíveis a partir desta pesquisa.

# REFERENCIAS


3M, 3M™ 360 Encompass™ System. Disponível em: <http://www.3m.com/3M/en_US/360-encompass-system-us/computer-assisted-coding/>. Acesso em 5 de julho de 2017

AALSETH, Patricia. **Medical Coding - What it is and how it works**, v.1, Albuquerque, Novo México, 2006, 231 p.

ADAMS, Ryan P.; LAROCHELLE, Hugo; SNOEK, Jasper.Practical Bayesian Optimization Of Machine Learning Algorithms. 2012. p.1-12.

AHUJA, Arun; DOWNEY, Doug; HUANG, Fei; YANG, Yi; YATES, Yuhong G. A. Learning Representations for Weakly Supervised Natural Language Processing Tasks. **Association for Computational Linguistics**, 40(1), 2014, p.85–120.

ALBELWI, Saleh; MAHMOOD, Ausif. A FrameWork for Designing Deep Convolutional Neural Networks. **MDPI Journal**, I(19). V.242, 2017, p.2-20.

ALTMAN, Douglas G.; LIBERATI, Alessandro; MOHER, David; TETZLAFF, Jennifer. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. **BMJ -PLoS Medicine**, 339: b2535- 10.1136/bmj.b2535, 2009.

Artificial Medical Intelligence, Artificial Medical Intelligence – Artificial Intelligence and NLP Technologies. Disponível em: <http://www.artificialmed.com/computer-assisted>. Acesso em 5 de julho de 2017

AYYAR, Sandeep; BEAR, Oliver. Tagging Patient Notes With ICD-9 Codes. **29th Conference on Neural Information Processing Systems (NIPS 2016)**. Stanford University, Stanford, Estados Unidos, 2016, p-1-8.

BALAS, Valentina E.; MASTORAKIS, Nikos; POPESCU, MC; PERESCU-POPESCU, Liliana. Multilayer Perceptron and Neural Networks. WSEAS transactions on circuits and systems. V.8. I(7), 2009, p.579-588

BALDASSARE, Franca P.; FERNANDEZ, Rafael S.; OSMO, André A.; de SÁ, Márcia M.; da SILVA, Penélope R. Implantação do Sistema de Grupos Relacionados pelo Diagnóstico (Diagnosis Related Group), **XV Congresso Brasileiro de Informática em Saúde, Sociedade Beneficente de Senhoras Hospital Sírio-Libanês (HSL)**, São Paulo, SP, Brasil, 2016, p.132-134.

BATES, Elizabeth; CARNEVALE, George F. New Directions in Research on Language Development. **Developmental Review**. University of San Diego, California. 1993, p.436-470

BEN-DAVID, Shai; SHALEV-SHWARTZ, Shai. Understanding Machine Learning from Theory to Algorithms. Cambridge University Press. Cambridge, USA, 2014, p. 442

BENGIO, Yoshua.; HINTON, Geoffrey.; LECUN, Yann. Deep Learning. **Nature**, V.521, 13(1), 2015, p.436–444. https://doi.org/10.1038/nmeth.3707

BENGIO, Yoshua; BERGSTRA, James. Random Search for Hyper Parameter Optimization. **Journal of Machine Learning Research**. I(13), 2012, p.1-25.

BENGIO, Yoshua; CHUNG, Junyoung; DAUPHIN, Yann N.; DE VRIES, Harm. **RMSPropandequilibratedadaptivelearningratesfornon-convex optimization**. Dept. IRO, Universit de Montral. 2015, p.1-10

BERARDI, Giacomo; ESULI, Andrea; MARCHEGGIANI, Diego. Word embeddings go to Italy: A comparison of models and training datasets. **CEUR Workshop Proceedings**. V.1404, 2015, p.1-8

BLUNSOM, Phil; ESPELHOLT, Lasse; GREFENSTETTE, Edward; HERMANN, Karl M.; KAY,Will; KOCISKÝ, Tomás; SULEYMAN, Mustafa; **Teaching Machines to Read and Comprehend**. Google Research, 2015, p.1-13.

BONITA, Ruth; BEAGLEHOLE, Robert; KJELLSTRÖM, Tord. Tipos de estudo. In: BONITA, Ruth; BEAGLEHOLE, Robert; KJELLSTRÖM, Tord. **Epidemiologia Básica**. 2 ed. São Paulo: Editora Santos, 2010. Cap. 3, p-39-49.

BOTTOU, León; COLLOBERT, Ronan; KARLEN, Michael; KAVUKCUOGLU, Koray; KUKSA, Pavel; WESTON, Jason. Natural Language Processing (Almost) from Scratch. **Journal of Machine Learning Research**, USA, 2011, p.2493-2537.

BOTVINICK, Matthew; GERSHMAN, Samuel; PEREIRA, Francisco; RITTER. Samuel. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. **Cognitive Neuropsychology**. V. 33. I.3-4, 2016, p.175-190

BOWLES, Kathryn H.; SHAFRAN-TOPAZ, Leah; TOPAZ, Maxim. ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. **Perspectives in Health Information Management / AHIMA, American Health Information Management Association**, v.10, 1d, 2013, p.1-5.

BRASIL, Ministério da Saúde, FUNASA, Fundação Nacional da Saúde. (2001). Manual de Procedimento do Sistema de Informações sobre Mortalidade, **Vigilância Epidemiológica**, v.1, e.1, Brasília, 2001, p.1-36.

BULEGON, Hugo. **Identificação de diagnósticos contidos em narrativas clínicas e mapeamento para a classificação internacional de doenças**. Curitiba, 2011, 99 p. Dissertação (Mestrado em Tecnologia em Saúde) - Pontifícia Universidade Católica do Paraná, 2011.

CACE, Centro Argentino de Clasification de Enfermedades, Conclusiones y recomendaciones de congresos y comites de expertos, **21 boletin del Centro e Clasificacion de Enfermedades, Primero Congreso Argentino de E spectrometría de Masa**, 2012, 11 p.

CARVALHO, Deborah R. SANTOS, Arnon B. V, Deep learning for healthcare management and diagnosis, **Ibero American Journal of Applied Computing**, v.5(2), 2015, p.15–25.

CARLSON, Eric T.; CELI, Leo A.; DERNONCOURT, Franck; GEHRMANN, Sebastian; GRANT, David W.; JR. John F.; LI, Yeran; MOSELEY, E.; TYLER, Patrick D.; WELT, J.; WU, Joy T. **A Comparison of Rule-Based and Deep Learning Models for Patient Phenotyping**. USA, 2017, 18 p.

CHANDRASEKAN, Muthu K.; CHEN, Tao; KAN, Min-Yen; KANG, Hong J. A Comparison of Word Embeddings for English and Cross-Lingual Chinese Word Sense Disambiguation.School of Computing, National University of Singapore, NUS Interactive and Digital Media Institute. **Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications.** 2016, p.30-39

CHESTER, Emma. Clinical Coding Policy and Procedures**, Data Quality Suite of Policies**, NHS, Southern Health, 2016, p.1-14.

CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey; MIKOLOV, Tomas; SUTSKEVER, Ilya. **Distributed Representations of Words and Phrases and their Compositionality**. Google Research Labs. 2013, p.1-9

CHU S. Information retrieval and health/clinical management. **Yearbook of Medical Informatics**; The University of Auckland, New Zealand, 2002, p.271-275.

CHIRIAC, Nona D.; MUSAT, Simona S.; PREDA, Alin L., Aspects of Clinical Coding, **Management in Health**, Bucareste, Romenia, 2012, p.19-21

CHOPRA, Abhimanyu; PRASHAR, Abhinav; SAIN; Chandresh. Natural Language Processing. **International Journal of Technology Enhancements and Emerging Engineering Research**. 2013, p.131-134.

COLOBERT, Ronan; WESTON, Jason. **A Unified Architecture for Natural Language Processing : Deep Neural Networks with Multitask Learning,** 2008, p.1-8.

COHEN, William W.; DHINGRA, Bhuwan; LIU, Hanxiao; Salakhutdinov, Ruslan. **A Comparative Study of Word Embeddings for Reading Comprehension**. School of Computer Science Carnegie Mellon University, Pittsburgh, USA, 2017, p.1-6

COUTO, Renato C.; FILHO, José Carlos, Avaliação Da Produtividade De Hospitais Brasileiros Pela Metodologia Do Diagnosis Related Group (DRG), **XV Congresso Brasileiro de Informática em Saúde**, Faculdade de Medicina da Universidade Federal de Minas Gerais, Belo Horizonte, Brasil, 2016, p.19-28.

DATASUS. **Departamento de Informática do SUS. Cadastros Nacionais - CID-10**. [acesso em 15 maio. 2017]. Disponível em: <http://datasus.saude.gov.br/sistemas-e-aplicativos/cadastros-nacionais/cid-10>

DATASUS. **Banco de dados do Sistema Único de Saúde. CID-10 - Classificação Internacional de Doenças**. [acesso em 17 maio. 2017]. Disponível em: http://www.datasus.gov.br/cid10/v2008/cid10.htm

Da SILVA, Edson M.; SOUZA, Renato R. Fundamentos em processamento de linguagem natural : uma proposta para extração de bigramas. **Revista Eletrônica de**

**Biblioteconomia e Ciência Da Informação**, v.19(n.40), 2014, p.1–32. https://doi.org/10.5007/1518-2924.2014v19n40p1

DARIO, Giuse A.; DENNY, Joshua C; HUA, Xu; MILLER, Randolph A.; Rosenbloom, Trent. WU, Yonghui.A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. **AMIA, Annual Symposium Proceedings**, 2012, p. 997–1003.

DE ANDRADE, Anderson. Best Practices for Convolutional Neural Network Applied to Object Recognition in Images. Department of Computer Science. University of Toronto, 2013, p.1-10.

DEAR, Gareth; DIXON-LEE, Claire; MCKENZIE, Kirsten; MORAN-FUKE, Judy; WALKER, Sue. Clinical Coding Internationally: A Comparison of the Coding Workforce in Australia, America, Canada, and England. **American Health Information Management Association**. Nova Iorque. 2010, p.1-13.

DESALVO, Giulia; JAMIESON, Kevin; LI, Lisha; ROSTAMIZADEH, Afshin; TALWAKAR, Ameet. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. 2016, p.1-48.

Dolbey Systems, Dolbey Systems – Since 1914. Disponível em: <http://www.dolbey.com/solutions/coding/fusion-cac/>. Acesso em 5 de julho de 2017

DOLINSKI, Joao P. Os surtos de febre amarela na cidade de Paranaguá (1852-1878). **Revista de História Regional**, 18(2), p. 410–437, 2013. https://doi.org/10.5212/Rev.Hist.Reg.v.18i2.0007

DUFFY, Nigel; Francon.; FINK, Dan; HODJAT, Babak; LIANG, Jason; MEYERSON, Elliot; MIIKKULAINEN, Risto; NAVRUZYAN, Arshak; RAJU, Bala; RAWAL, Adyta.; Evolving Deep Neural Networks. The University of Texas, 2017, p-1-8.

ELKAN, Charles; LIPTON, Zachary C.; NARYANASWAMY, Balakrishnan**. Optimal Thresholding of Classifiers to Maximize F1 Measure**. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). I(2), 2014, p.225-239

ESTEVA, Andre; KUPREL, Brett; Thrun, Sebastian. **Deep Networks for Early Stage Skin Disease and Skin Cancer Classification**, Stanford University, Stanford, California, 2015. 8 p.

EMScribe CAC, Artificial Medical Inteligence. Disponível em: <http://www.artificialmed.com//>. Acesso em 5 de julho de 2017

FAYYAD, Usama., PIATETSKY-SHAPIRO, Gregory, Smyth, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, 17(3), 37 p, 1996. https://doi.org/10.1609/aimag.v17i3.1230

FAVORETO, Cesar A. O.; CAMARGO JR, K. R. A narrativa como ferramenta para o desenvolvimento da prática clínica, **Interface - Comunicação, Saúde, Educação,** v.15, n.37, p. 473–484, 2011.

FELLOW, King-Sun F.; ROSENFELD, Azriel. Pattern Recognition and Image Processing. **Ieee Transactions on Computers**. V.25, I(12), 1976, p.1336-1345

FENTON, Susan H.; HERSH, William R.; JENDERS, Robert A.; STANFILL, Mary H.; WILLIAMS, Margaret. A systematic literature review of automated clinical coding and classification systems.**Journal of the American Medical Informatics Association**, 17(6), p.646–651, 2010. https://doi.org/10.1136/jamia.2009.001024

FIESLER, E. Neural network topologies. **The Handbook of Neural Computation**. 1996, p.1-17

GATTI, Maira; Dos SANTOS, Cícero N. **Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts**, IBM Research, 2014, p. 69–78.

GRISHAM, David J.; SHEPPARD, Melissa; TRAN, Wendy U., Visual Symptons and Reading Performance. Symposium: **Dyslexia And Learning Disabilities: And Interdisciplinary Perspective, American Academy of Optometry**, 1993, p.18-23.

GOODENOUGH, John B.; RUBENSTEIN, Herbert. Context Correlates of Synonymy. **Communications of the ACM**. V.8, I.10, 1965. p.627-633

GOMES, Herval P.; RIOS, Luis A. S. Incontinência urinária e esforço. **Urologia Fundamental**. Cap. 26. Brasil, 2010.

GUIMARÃES JR.; KLÜCK M, Sumário eletrônico de alta: garantindo a continuidade da assistência ao paciente através da informação. **Revista de Informática Pública**. 1999 dez.;1(2):123-137.

HARRIS, Zellig S. Distributional Structure. **WORD**. V.10, I:2-3; p.146-162

HE, Kaiming; REN, Shaoqing; SUN, Jian; ZHANG, Xiangyu. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, **Proceedings of the IEEE International Conference on Computer Vision, Microsoft Research**, USA, 2015, p.1026-1034.

HEUER, Hendrik. Text comparison using word vector representations and dimensionality reduction. **Proceedings to the 8th edition of the European Conference on Python in Science**, Cambridge, 2016, p.13-16

HUANG, Chen; LI, Yining; LOY, Chen; TANG, Xiaoou. Learning Deep Representation for Imbalanced Classification. **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. 2016, p.5375-5384.

HUANG, Heng; LI, Peng. **Clinical Information Extraction via Convolutional Neural Network**. University of Texas at Arlington. 2016, p.1-5

HUGHES, Mark; KOTOULAS, Spyros; SUZUMURA, Toyotaro. Medical Text Classification using Neural Networks. **IBM Research Labs**, 2017, p.246-250.

HUTTER, Frank; LOSHCHILOV, Ilya. CMA-ES For Hyperparameter Optimization Of Deep Neural Networks. **International Conference of Robot Learning**. 2016, p.1-8.

KAVULURU, Ramakanth; RIOS, Anthony. **Convolutional Neural Networks for Biomedical Text Classification: Application in Indexing Biomedical Articles**. 2015, p.258-267

KIM, Yoon. Convolutional Neural Networks for Sentence Classification. **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)**. USA, 2014, p.1746-1751

KINGMA, Diederik; LEI BA, Jimmy. Adam: A Method for Stochastic Optimization. **2017 IEEE International Conference on Consumer Electronics**, 2015; p.434-449

KRÁL, Pavel; LENC, Ladislav. Deep Neural Networks for Czech Multi-Label Document Classification. Faculty of Applied Sciences. University of West Bohemia, Czech Republic, 2017, 12 p.

KRIESEL, David, A Brief Introduction to Neural Networks. University of Bonn, 2005, p.244

LAKATOS, Eva M.; MARCONI, Marina A. **Fundamentos de metodologia científica**. Editora Atlas S. A, 2003, 310 p. https://doi.org/10.1590/S1517-97022003000100005

LAPELLE Nancy R., LUCKMANN R, SIMPSOM EH , et al. Identifying strategies to improve access to credible and relevant information for public health professionals: a qualitative study. **BMC Public Health**. 2006; p.6:89.

LARA, Natalia, NARDI, Elene; REIS, Amanda. O financiamento da saúde no Brasil e a valorização da saúde suplementar. **Instituto de Estudos em Saúde Suplementar**, 17 p.

LAPALME, Guy; SOKOLOVA, Marina. A systematic analysis of performance measure for classification tasks, **Information Processsing and Management**. v. 45, I(4), 2009, Canada, p.427-437.

LAURENTI, Ruy. Análise da informação em saúde: 1893-1993, cem anos da Classificação Internacional de Doenças, **Revista Saúde Pública de São Paulo,**

LAURENTI, Ruy et al. A Classificação Internacional de Doencas, a Familia de Classificações Internacionais, a CID-11 e a Sindrome Pos-Poliomielite. **Arq. Neuro-Psiquiatr.**, São Paulo , v. 71, n. 9A, 2013, p. 3-10.

LEAL, Christiane; MOTA; Leticia M.; PISI, Paula C. B., RORIZ-FILHO, Jarbas S.; VILAR, Fernando C. Infecção do Trato Urinário, **Revista Brasileira de Medicina**, V. 72, I(9). 2015, p.383-387

LECUN, Yann; ZHANG, Xiang. **Text Understanding from Scratch**, Computer Science Department, Courant Institute of Mathematical Sciences, New York University, 2016, p.1-10.

LI, Ge; JIN, Zhi; MENG, Zhao; MOU, Lili; YAN, Rui. How Transferable are Neural Networks in NLP Applications?. **Conference on Empirical Methods on Natural Language Processing,** Texas,EUA, 2016, p.1-11.

LOPES, Fernando. A importância da codificação clínica como ferramenta de apoio à gestão hospitalar e os desafios que se colocam à sua evolução no contexto do sns, **Gestão Hospitalar**, Associação Portuguesa de Administradores Hospitalares (APAH), 2015, p.8-15.

LOPES, Fernando. Actividade dos codificadores deve ser reconhecida pela ordem.**Revista Norte Médico**, 2009, p.10-12.

MCCULLOCH, Warren S.; PITTS, Walter H. A Logical Calculus Of the Ideas Immanent in Nervous Activity. **Bulletin of Mathematical Biophysics.** V.(5), 1943, p.115-133

MELTON, Genevieve; HRIPCSAK, George. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. **Journal of the American Medical Informatics Association**. V.12, I(4), 2005, p.448-458

MILLER, W. G; **IFCC Working Group on Standardization of Albumin in Urine**. Questões atuais relativas à dosagem e à descrição da excreção urinária de albumina. JBPML, v. 46, n. 3, p. 187-206, 2010.

MINISTERIO DE SANIDAD, Manual de codificación CIE-10-ES Diagnósticos. **Clasificación Internacional de Enfermedades**, v.1, 303 p.

NARYSHKIN S, SCHULTZ BL. Assessment of quality of data provided on Pap test requisitions: implications for quality of care and patient safety. **Cytojournal**. 2009, p.6-11.

NOHAMA, Percy; PACHECO, Edson;SCHULZ, Stefan, Codificação de narrativas clínicas para uma ontologia de domínio. **Rev. Bras. Pesq. Saúde,** v.15(2), 2013, p.94–103.

NORVIG P.; RUSSELL S. Artificial intelligence: a modern approach. 3rd ed. Local: Prentice Hall; 2009.

OLEYNIK, Michel; NOHAMA, Percy; PACHECO, Edson; SCHULZ, Stefan. Elaboração de um Corpus Médico baseado em Narrativas Clínicas contidas em Sumários de Alta, **XII Congresso Brasileiro de Informática**, 2010, p.1-4. https://doi.org/10.13140/RG.2.1.4412.7441

OMS, Organização Mundial de Saúde, ICD-11 Update, World Health Organization - **Health Data Standards and Informatics**. 2017, 4 p.

Optum360, Optum 360. Disponível em: <https://www.optum360.com/landing/ro/enterprise-cac.html>. Acesso em 5 de julho de 2017

PEREZ-CRUZ, Fernando; READ, Jese. **Deep Learning for Multi-Label Classification**. 2014, 8 p.

PRECHELT, Lutz. Early Stopping - But When?. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2012, p.53-67

REDE Interagencial de Informação para a Saúde, Indicadores básicos para a saúde no Brasil: conceitos e aplicações.**Organização Pan-Americana Da Saúde,** 349, 2011. https://doi.org/978-85-87943-65-1

RODRIGUES, J.; ANTONIO, B.; STEVEN, N; SILVA, J. Distributional Semantics Models for Portuguese. In Computational Processing of the Portuguese Language: 12th International Conference (P**ROPOR-2016**). Springer International Publishing, 2016

SAMUEL, L. A. Some studies in machine learning using the game of checkers. **IBM Journal.** 1959, 17 p.

SOCHER, Richard,**Recursive Deep Learning For Natural Language Processing And Computer Vision,** 2014, 189 p.

SOCHER, Richard; Manning, Christopher D.; Pennington, Jeffrey. GloVe: Global Vectors for Word Representations. **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.** 2014, 12 p.

SOUZA, Andreia C. **Identificação Do Conteúdo Padronizado Do Sumário De Alta**. Curitiba, 2012, 88 p. Dissertação (Mestrado em Tecnologia em Saúde) - Pontifícia Universidade Católica do Paraná, 2012.

TURING, Alan M., Computer Machinery and Intelligence.**Mind**, New Series, Oxford University, v. 49, 1950, p.450

WALLACE, Byron C.; ZHANG, Ye. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classificatiom. **Neural and Evolutionary Computing**. 2015. p.1-18

WANG, William Y. "Liar, Liar Pants on Fire":A New Benchmark Dataset for Fake News Detection. University of California, Santa Barbara, USA. 2017. p.1-5

# ANEXO 1 – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO



## TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Você está sendo convidado (a) como voluntário(a) a participar do estudo CODIFICAÇÃO DE NARRATIVAS CLÍNICAS COM O USO DE DEEP LEARNING e que tem como objetivo conceber um modelo computacional que auxilie a codificação clínica para os padrões da CID-10, a partir da classificação de narrativas clínicas de distúrbios do trato urinário em um ou mais códigos da CID-10 e avaliar o modelo proposto frente as métricas comumente utilizadas para o processamento de linguagem natural. Acreditamos que esta pesquisa seja importante porque visa contribuir para automação do processo de codificação clínica no Brasil através de uma moderna técnica de compreensão de padrões em dados.

PARTICIPAÇÃO NO ESTUDO

A sua participação no referido estudo será de auxílio na compreensão dos acrônimos (siglas) médicos contidos nas narrativas clínicas, traduzindo-os para o seu significado literal quando necessário e disponibilizar suas conclusões para que substituam os acrônimos contidos nos textos clínicos que compõem a base de dados deste estudo. Sua participação também será importante para eventual validação dos códigos atribuídos pelo modelo computacional em relação ao que foi atribuído anteriormente na narrativa já codificada.

**RISCOS E BENEFÍCIOS**

Através deste Termo de Consentimento Livre e Esclarecido você está sendo alertado de que, da pesquisa a se realizar, pode esperar alguns benefícios, tais como: um maior contato com a pesquisa científica e participar de um processo que contribui para a automação de uma tarefa predominantemente manual. Bem como, também que é possível que aconteçam os seguintes desconfortos ou riscos em sua participação, tais como a exposição do seu nome ou a impressão de que essa pesquisa visa substituir o trabalho realizado pelo ser humano. Para minimizar tais riscos, nós pesquisadores tomaremos as seguintes medidas: Todos os dados serão descaracterizados (nenhum nome será utilizado nessa base de dados) e demonstraremos ao longo da pesquisa que a solução proposta visa otimizar e auxiliar a codificação clínica, trabalho que já é realizado pelos profissionais codificadores e para o qual deseja-se através desta pesquisa oferecer ferramentas de apoio.

**SIGILO E PRIVACIDADE**

Nós pesquisadores garantiremos a você que sua privacidade será respeitada, ou seja, seu nome ou qualquer outro dado ou elemento que possa, de qualquer forma, lhe identificar, será mantido em sigilo. Nós pesquisadores nos responsabilizaremos pela guarda e confidencialidade dos dados, bem como a não exposição dos dados de pesquisa.

**AUTONOMIA**

Nós lhe asseguramos a assistência durante toda pesquisa, bem como garantiremos seu livre acesso a todas as informações e esclarecimentos adicionais sobre o estudo e suas consequências, enfim, tudo o que você queira saber antes, durante e depois de sua participação. Também informamos que você pode se recusar a participar do estudo, ou retirar seu consentimento a qualquer momento, sem precisar justificar, e de, por desejar sair da pesquisa, não sofrerá qualquer prejuízo à assistência que vem recebendo.

**RESSARCIMENTO E INDENIZAÇÃO**

Caso tenha qualquer despesa decorrente da participação nesta pesquisa, tais como transporte, alimentação entre outros, bem como a seu acompanhante (se for o caso), haverá ressarcimento

dos valores gastos na forma seguinte: depósito em conta corrente do participante dessa pesquisa.
De igual maneira, caso ocorra algum dano decorrente de sua participação no estudo, você será devidamente indenizado, conforme determina a lei.

**CONTATO**

O pesquisador envolvido com o referido projeto é ARNON BRUNO VENTRILHO DOS SANTOS, discente da Pontifícia Universidade Católica do Paraná (PUC-PR), e com ele você poderá manter contato pelo telefone 41 9 88301388

O Comitê de Ética em Pesquisa em Seres Humanos (CEP) é composto por um grupo de pessoas que estão trabalhando para garantir que seus direitos como participante de pesquisa sejam respeitados. Ele tem a obrigação de avaliar se a pesquisa foi planejada e se está sendo executada de forma ética. Se você achar que a pesquisa não está sendo realizada da forma como você imaginou ou que está sendo prejudicado de alguma forma, você pode entrar em contato com o Comitê de Ética em Pesquisa da PUCPR (CEP) pelo telefone (41) 3271-2292 entre segunda e sexta-feira das 08h00 às 17h30 ou pelo e-mail nep@pucpr.br.

**DECLARAÇÃO**

Declaro que li e entendi todas as informações presentes neste Termo de Consentimento Livre e Esclarecido e tive a oportunidade de discutir as informações deste termo. Todas as minhas perguntas foram respondidas e eu estou satisfeito com as respostas. Entendo que receberei uma via assinada e datada deste documento e que outra via assinada e datada será arquivada nos pelo pesquisador responsável do estudo.

Enfim, tendo sido orientado quanto ao teor de todo o aqui mencionado e compreendido a natureza e o objetivo do já referido estudo, manifesto meu livre consentimento em participar, estando totalmente ciente de que não há nenhum valor econômico, a receber ou a pagar, por minha participação.

| Dados do participante da pesquisa | |
|---|---|
| Nome: | [illegible] |
| Telefone: | - |
| e-mail: | [illegible] |

Local, 10 de AGOSTO de 2017

Assinatura do participante da pesquisa

Assinatura do Pesquisador

# ANEXO 2 – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO



## TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Você está sendo convidado (a) como voluntário(a) a participar do estudo CODIFICAÇÃO DE NARRATIVAS CLÍNICAS COM O USO DE DEEP LEARNING e que tem como objetivo conceber um modelo computacional que auxilie a codificação clínica para os padrões da CID-10, a partir da classificação de narrativas clínicas de distúrbios do trato urinário em um ou mais códigos da CID-10 e avaliar o modelo proposto frente as métricas comumente utilizadas para o processamento de linguagem natural. Acreditamos que esta pesquisa seja importante porque visa contribuir para automação do processo de codificação clínica no Brasil através de uma moderna técnica de compreensão de padrões em dados.

### PARTICIPAÇÃO NO ESTUDO

A sua participação no referido estudo será de auxílio na compreensão dos acrônimos (siglas) médicos contidos nas narrativas clínicas, traduzindo-os para o seu significado literal quando necessário e disponibilizar suas conclusões para que substituam os acrônimos contidos nos textos clínicos que compõem a base de dados deste estudo. Sua participação também será importante para eventual validação dos códigos atribuídos pelo modelo computacional em relação ao que foi atribuído anteriormente na narrativa já codificada.

### RISCOS E BENEFÍCIOS

Através deste Termo de Consentimento Livre e Esclarecido você está sendo alertado de que, da pesquisa a se realizar, pode esperar alguns benefícios, tais como: um maior contato com a pesquisa científica e participar de um processo que contribui para a automação de uma tarefa predominantemente manual. Bem como, também que é possível que aconteçam os seguintes desconfortos ou riscos em sua participação, tais como a exposição do seu nome ou a impressão de que essa pesquisa visa substituir o trabalho realizado pelo ser humano. Para minimizar tais riscos, nós pesquisadores tomaremos as seguintes medidas: Todos os dados serão descaracterizados (nenhum nome será utilizado nessa base de dados) e demonstraremos ao longo da pesquisa que a solução proposta visa otimizar e auxiliar a codificação clínica, trabalho que já é realizado pelos profissionais codificadores e para o qual deseja-se através desta pesquisa oferecer ferramentas de apoio.

### SIGILO E PRIVACIDADE

Nós pesquisadores garantiremos a você que sua privacidade será respeitada, ou seja, seu nome ou qualquer outro dado ou elemento que possa, de qualquer forma, lhe identificar, será mantido em sigilo. Nós pesquisadores nos responsabilizaremos pela guarda e confidencialidade dos dados, bem como a não exposição dos dados de pesquisa.

### AUTONOMIA

Nós lhe asseguramos a assistência durante toda pesquisa, bem como garantiremos seu livre acesso a todas as informações e esclarecimentos adicionais sobre o estudo e suas consequências, enfim, tudo o que você queira saber antes, durante e depois de sua participação. Também informamos que você pode se recusar a participar do estudo, ou retirar seu consentimento a qualquer momento, sem precisar justificar, e de, por desejar sair da pesquisa, não sofrerá qualquer prejuízo à assistência que vem recebendo.

### RESSARCIMENTO E INDENIZAÇÃO

Caso tenha qualquer despesa decorrente da participação nesta pesquisa, tais como transporte, alimentação entre outros, bem como a seu acompanhante (se for o caso), haverá ressarcimento

dos valores gastos na forma seguinte: depósito em conta corrente do participante dessa pesquisa.
De igual maneira, caso ocorra algum dano decorrente de sua participação no estudo, você será devidamente indenizado, conforme determina a lei.

**CONTATO**

O pesquisador envolvido com o referido projeto é ARNON BRUNO VENTRILHO DOS SANTOS, discente da Pontifícia Universidade Católica do Paraná (PUC-PR), e com ele você poderá manter contato pelo telefone 41 9 88301388

O Comitê de Ética em Pesquisa em Seres Humanos (CEP) é composto por um grupo de pessoas que estão trabalhando para garantir que seus direitos como participante de pesquisa sejam respeitados. Ele tem a obrigação de avaliar se a pesquisa foi planejada e se está sendo executada de forma ética. Se você achar que a pesquisa não está sendo realizada da forma como você imaginou ou que está sendo prejudicado de alguma forma, você pode entrar em contato com o Comitê de Ética em Pesquisa da PUCPR (CEP) pelo telefone (41) 3271-2292 entre segunda e sexta-feira das 08h00 às 17h30 ou pelo e-mail nep@pucpr.br

**DECLARAÇÃO**

Declaro que li e entendi todas as informações presentes neste Termo de Consentimento Livre e Esclarecido e tive a oportunidade de discutir as informações deste termo. Todas as minhas perguntas foram respondidas e eu estou satisfeito com as respostas. Entendo que receberei uma via assinada e datada deste documento e que outra via assinada e datada será arquivada nos pelo pesquisador responsável do estudo.

Enfim, tendo sido orientado quanto ao teor de todo o aqui mencionado e compreendido a natureza e o objetivo do já referido estudo, manifesto meu livre consentimento em participar, estando totalmente ciente de que não há nenhum valor econômico, a receber ou a pagar, por minha participação.

| Dados do participante da pesquisa | |
|---|---|
| Nome: | |
| Telefone: | |
| e-mail: | |

Local, ___ de _____ de ____.

Assinatura do participante da pesquisa

Assinatura do Pesquisador

## ANEXO 3 – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO



**TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO**

Você está sendo convidado (a) como voluntário(a) a participar do estudo CODIFICAÇÃO DE NARRATIVAS CLÍNICAS COM O USO DE DEEP LEARNING e que tem como objetivo conceber um modelo computacional que auxilie a codificação clínica para os padrões da CID-10, a partir da classificação de narrativas clínicas de distúrbios do trato urinário em um ou mais códigos da CID-10 e avaliar o modelo proposto frente as métricas comumente utilizadas para o processamento de linguagem natural. Acreditamos que esta pesquisa seja importante porque visa contribuir para automação do processo de codificação clínica no Brasil através de uma moderna técnica de compreensão de padrões em dados.

PARTICIPAÇÃO NO ESTUDO

A sua participação no referido estudo será de auxílio na compreensão dos acrônimos (siglas) médicos contidos nas narrativas clínicas, traduzindo-os para o seu significado literal quando necessário e disponibilizar suas conclusões para que substituam os acrônimos contidos nos textos clínicos que compõem a base de dados deste estudo. Sua participação também será importante para eventual validação dos códigos atribuídos pelo modelo computacional em relação ao que foi atribuído anteriormente na narrativa já codificada.

**RISCOS E BENEFÍCIOS**

Através deste Termo de Consentimento Livre e Esclarecido você está sendo alertado de que, da pesquisa a se realizar, pode esperar alguns benefícios, tais como: um maior contato com a pesquisa científica e participar de um processo que contribui para a automação de uma tarefa predominantemente manual. Bem como, também que é possível que aconteçam os seguintes desconfortos ou riscos em sua participação, tais como a exposição do seu nome ou a impressão de que essa pesquisa visa substituir o trabalho realizado pelo ser humano. Para minimizar tais riscos, nós pesquisadores tomaremos as seguintes medidas: Todos os dados serão descaracterizados (nenhum nome será utilizado nessa base de dados) e demonstraremos ao longo da pesquisa que a solução proposta visa otimizar e auxiliar a codificação clínica, trabalho que já é realizado pelos profissionais codificadores e para o qual deseja-se através desta pesquisa oferecer ferramentas de apoio.

**SIGILO E PRIVACIDADE**

Nós pesquisadores garantiremos a você que sua privacidade será respeitada, ou seja, seu nome ou qualquer outro dado ou elemento que possa, de qualquer forma, lhe identificar, será mantido em sigilo. Nós pesquisadores nos responsabilizaremos pela guarda e confidencialidade dos dados, bem como a não exposição dos dados de pesquisa.

**AUTONOMIA**

Nós lhe asseguramos a assistência durante toda pesquisa, bem como garantiremos seu livre acesso a todas as informações e esclarecimentos adicionais sobre o estudo e suas consequências, enfim, tudo o que você queira saber antes, durante e depois de sua participação. Também informamos que você pode se recusar a participar do estudo, ou retirar seu consentimento a qualquer momento, sem precisar justificar, e de, por desejar sair da pesquisa, não sofrerá qualquer prejuízo à assistência que vem recebendo.

**RESSARCIMENTO E INDENIZAÇÃO**

Caso tenha qualquer despesa decorrente da participação nesta pesquisa, tais como transporte, alimentação entre outros, bem como a seu acompanhante (se for o caso), haverá ressarcimento

dos valores gastos na forma seguinte: depósito em conta corrente do participante dessa pesquisa.
De igual maneira, caso ocorra algum dano decorrente de sua participação no estudo, você será devidamente indenizado, conforme determina a lei.

**CONTATO**

O pesquisador envolvido com o referido projeto é ARNON BRUNO VENTRILHO DOS SANTOS, discente da Pontifícia Universidade Católica do Paraná (PUC-PR), e com ele você poderá manter contato pelo telefone 41 9 88301388.

O Comitê de Ética em Pesquisa em Seres Humanos (CEP) é composto por um grupo de pessoas que estão trabalhando para garantir que seus direitos como participante de pesquisa sejam respeitados. Ele tem a obrigação de avaliar se a pesquisa foi planejada e se está sendo executada de forma ética. Se você achar que a pesquisa não está sendo realizada da forma como você imaginou ou que está sendo prejudicado de alguma forma, você pode entrar em contato com o Comitê de Ética em Pesquisa da PUCPR (CEP) pelo telefone (41) 3271-2292 entre segunda e sexta-feira das 08h00 às 17h30 ou pelo e-mail nep@pucpr.br.

**DECLARAÇÃO**

Declaro que li e entendi todas as informações presentes neste Termo de Consentimento Livre e Esclarecido e tive a oportunidade de discutir as informações deste termo. Todas as minhas perguntas foram respondidas e eu estou satisfeito com as respostas. Entendo que receberei uma via assinada e datada deste documento e que outra via assinada e datada será arquivada nos pelo pesquisador responsável do estudo.

Enfim, tendo sido orientado quanto ao teor de todo o aqui mencionado e compreendido a natureza e o objetivo do já referido estudo, manifesto meu livre consentimento em participar, estando totalmente ciente de que não há nenhum valor econômico, a receber ou a pagar, por minha participação.

| **Dados do participante da pesquisa** | |
|---|---|
| Nome: | [illegible] |
| Telefone: | [illegible] |
| e-mail: | [illegible] |

Local, 10 de agosto de 2017

Assinatura do participante da pesquisa

Assinatura do Pesquisador

## ANEXO 4 – TERMO DE COMPROMISSO DE UTILIZAÇÃO DE DADOS

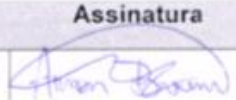
# Termo de Compromisso de Utilização de Dados (TCUD)

**Eu ARNON BRUNO VENTRILHO DOS SANTOS**, abaixo assinado(s), pesquisador envolvido no projeto de título **CODIFICAÇÃO DE NARRATIVAS CLÍNICAS COM O USO DE DEEP LEARNING**, me comprometo a manter a confidencialidade sobre os dados coletados nos arquivos do **INSTITUTO DE ACREDITAÇÃO E GESTÃO EM SAÚDE (IAG-SAÚDE)**, bem como a privacidade de seus conteúdos, como preconizam os Documentos Internacionais e as Resoluções 466/12 e 510/16, do Conselho Nacional de Saúde.

Informo que os dados a serem coletados dizem respeito a **NARRATIVAS CLÍNICAS DE DISTÚRBIOS DO TRATO URINÁRIO** ocorridos entre as datas de: **janeiro de 2005 e julho de 2017**.

Curitiba, 06, agosto e 2017.

Envolvidos na manipulação e coleta dos dados:

| Nome completo | CPF | Assinatura |
|---|---|---|
| ARNON BRUNO VENTRILHO DOS SANTOS | 07116857974 |  |
| | | |
| | | |
| | | |
| | | |
| | | |

# ANEXO 5 – AUTORIZAÇÃO DA INSTITUIÇÃO

**AUTORIZAÇÃO**

Eu **RENATO CAMARGOS COUTO**, abaixo assinado, responsável pelo INSTITUTO DE ACREDITAÇÃO E GESTÃO EM SAÚDE (IAG-Saúde), autorizo a realização do estudo **CODIFICAÇÃO DE NARRATIVAS CLÍNICAS COM O USO DE DEEP LEARNING**, a ser conduzido pelos pesquisadores abaixo relacionados. Fui informado pelo responsável do estudo sobre as características e objetivos da pesquisa, bem como das atividades que serão realizadas na instituição a qual represento.

Declaro ainda ter lido e concordar com o parecer ético emitido pelo CEP da instituição proponente, conhecer e cumprir as Resoluções Éticas Brasileiras, em especial a Resolução CNS 466/12 e/ou CNS 510/16. Esta instituição está ciente de suas corresponsabilidades como instituição coparticipante do presente projeto de pesquisa e de seu compromisso no resguardo da segurança e bem-estar dos sujeitos de pesquisa nela recrutados, dispondo de infraestrutura necessária para a garantia de tal segurança e bem-estar.

Belo Horizonte, 10 de agosto de 2017.

Assinatura e carimbo do responsável institucional

INSTITUTO DE ACREDITAÇÃO E GESTÃO EM SAUDE LTDA.

Renato C. Coutc
CRM - 18.234

LISTA NOMINAL DE PESQUISADORES:

_Arnion Bruno Ventrilho dos Santos___