PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

AVNER DAL BOSCO

TRATAMENTO SEMÂNTICO DE REGISTROS ELETRÔNICOS SOBRE CUIDADOS DE AVC PARA UM MODELO DE GESTÃO BASEADO EM VALOR

Porto Alegre
2021

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TRATAMENTO SEMÂNTICO DE REGISTROS ELETRÔNICOS SOBRE CUIDADOS DE AVC PARA UM MODELO DE GESTÃO BASEADO EM VALOR

AVNER DAL BOSCO

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Prof[a]. Dra. Renata Vieira

Porto Alegre
2021

Ficha Catalográfica

D136t Dal Bosco, Avner

Tratamento semântico de registros eletrônicos sobre cuidados de AVC para um modelo de gestão baseado em valor / Avner Dal Bosco. – 2021.

103 p.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Renata Vieira.

1. Ontologias. 2. Registros Eletrônicos. 3. AVC. 4. Gestão de Saúde. I. Vieira, Renata. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS com os dados fornecidos pelo(a) autor(a).

Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

AVNER DAL BOSCO

TRATAMENTO SEMÂNTICO DE REGISTROS ELETRÔNICOS SOBRE CUIDADOS DE AVC PARA UM MODELO DE GESTÃO BASEADO EM VALOR

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 22 de Março de 2021.

BANCA EXAMINADORA:

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS)

Prof. Dr. Sandro José Rigo (PGCA/UNISINOS)

Profª. Dra. Renata Vieira (PPGCC/PUCRS - Orientadora)

DEDICATÓRIA

Dedico este trabalho de pesquisa aos meus pais Deisi Silveira da Silva e Setembrino Dal Bosco. Aos meus familiares, em especial a minha avó Iara Lara da Silva e minha tia Ivanirce Dal Bosco. Aos meus recomendadores acadêmicos a D.ra Anubis Rossetto e o M.e. José Figueiredo. Aos meus amigos. As forças recebidos destes foram a mola propulsora que permitiu o meu avanço, mesmo durante os momentos mais difíceis. A todos os citados, e aos incontáveis outros apoiadores da minha jornada, agradeço com muita humildade e toda sinceridade.

"Se quisermos contribuir para uma mudança na comunidade na qual estamos inseridos, precisamos, primeiro, provocar uma transformação em nós mesmos."
(Lama Sherab Drolma)

AGRADECIMENTOS

Agradeço a minha orientadora D.ra Renata Vieira por me selecionar e me orientar ao longo de todo o trajeto. Agradeço aos meus colegas de curso e professores que contribuíram para a minha formação ao compartilharem seus conhecimento. Agradeço a PUCRS e a CAPES/CNPQ por financiarem e assim permitirem a realização dessa pequisa. Agradeço ao M.e Henrique dias por ter estabelecido uma conexão entre eu e a D.ra Ana Paula Etges, a quem também agradeço pois sem o seu projeto de pesquisa esse estudo não poderia ter sido realizado. Agradeço aos colegas colabores M.e Eduardo Cortes e Ma. Bruna Zanotto que foram insubstituíveis na condução do projeto proposto pela D.ra Ana. Por fim, agradeço a todos que, mesmo não citados, contribuíram para a criação e desenvolvimento dessa pesquisa.

TRATAMENTO SEMÂNTICO DE REGISTROS ELETRÔNICOS SOBRE CUIDADOS DE AVC PARA UM MODELO DE GESTÃO BASEADO EM VALOR

RESUMO

Modelos de gestão baseados em valor requerem a precisa análise de indicadores de saúde como eventos de risco, condições clinicas, manejo de pacientes e desfechos clínicos. Atualmente essa análise é manualmente realizada através da leitura e busca por esses indicadores nos textos presentes em registros eletrônicos de saúde.

A nossa pesquisa propõe um modelo computacional de classificação de textos livres, baseado em ontologias, que automatize essa tarefa de forma que ela possa ser realizada por um computador.

Para validar o modelo proposto nós utilizamos as evoluções clinicas de 281 pacientes sob os cuidados de AVC. Foram selecionados 30 indicadores para serem identificados nessas evoluções. Destes o modelo conseguiu processar 28, e dentre eles os resultados de classificação variam de *5,83 % de f1-score* com *mcc de 8,01 %* até *94,78 % de f1-score* com *mcc de 94,78 %*, sendo a média, considerando os 30 indicadores, de *56,8 % de f1-score* com *mcc de 57,97 %*.

Palavras-Chave: Ontologias, Registros Eletrônicos, AVC, Gestão de Saúde.

A VALUE BASED MANAGMENT SEMANTIC ANALYSES OF STROKE CASES IN ELECTRONIC HEALTH RECORDS

ABSTRACT

Value-based health managment models require a precise accounting of health indexes such as risk events monitoring, clinical conditions, pacient handling and, cases disclosures. Currently this accounting is performed by manually reading and searching through eletronic health records for this indexes.

Our research proposes a way to make this a autonomous task that can be performed by a computer using a free-text concept classifier model based on ontologies.

To validate our model we tested it with digital clinical evaluations from 281 pacients under stroke care. We've selected 30 indexes to be identified in this texts. Our model was capable of identify and classify 28 of thoses indexes varying from *'5,83 % f1-score results and mcc score of 8,01 %* to *'94,78 % f1-score results and mcc-score of 94,78 %'*. Considering all 30 indexes, our model reached, on average *'56,8 % of f1-score and a mcc-score of 57,97 %'*.

Keywords: Ontologies, Electornic health records, Stroke, Health managment.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AVC – Acidente Vascular Cerebral

GSBV – Gestão de Saúde Baseada em Valor

PEP – Prontuário Eletrônico do Paciente

IA – Inteligência Artificial

PLN – Processamento de Linguagem Natural

PLN – Inteligência Artificial

FFS – Fee For Service

RPS – Recompensa por Serviço

SUS – Sistema Único de Saúde

RES – Registros Eletrônicos de Saúde

OWL – Web Ontology Language

LISTA DE ABREVIATURAS

VP. – Verdadeiro Positivos

VN. – Verdadeiro Negativo

FP. – Falso Positivo

FN. – Falso Negativo

SUMÁRIO

1. INTRODUÇÃO

Um modelo de gestão baseado em valor (GBVS) na área da saúde permite a remuneração e reconhecimento das instituições e dos prestadores de serviço dessa área de forma proporcional a qualidade do atendimento com o paciente e dos desfechos obtidos. Dessa forma, esse sistema incentiva a busca pelo tratamento mais eficiente eficiente e pelo melhor zelo para com os pacientes. Seguindo essas diretrizes todo o sistema de saúde é beneficiado, os órgãos responsáveis pelo repasse de verbas tem a certeza de estarem fazendo investimento justos e necessários, os usuários do sistema recebem um atendimento de melhor qualidade e com melhores resultados e as instituições e profissionais de saúde são incentivados a melhorar e aprimorar as suas práticas. Para que esse modelo GBVS seja uma realidade na rotina das instituições é necessário que ocorra um avanço em direção ao uso da inteligência computacional (IA) nesse cenário. Um modelo GBVS baseia-se nos relatos dos atendimentos e serviços prestados que constam em prontuários eletrônicos. É através desses documentos digitais que são avaliados as condutas e os desfechos de um serviço prestado. Esse processo hoje, mesmo nos sistemas de prontuários eletrônicos de pacientes (PEP) mais avançados, precisa ser realizado de forma manual e por isso os avanços na IA em direção a automatização dessa tarefa é fundamental. A tarefa a ser automatizada é definida como a medição precisa de características clínicas, desfechos clínicos, eventos de risco e manejos clínicos. Para realizar tal medição é necessário a identificação de palavras chaves, indicadores e conceitos nos textos livres das evoluções clínicas. Nosso estudo busca viabilizar a automatização desse processo auxiliando a implementação de uma GSBV para a linha de cuidado de Acidente Vascular Cerebral (AVC).

Para a linha de cuidados AVC especialistas dessa área elencaram 30 indicadores que medem a qualidade dos serviços e atendimentos prestados e que se sub-dividem em: *Características clinicas*, *Manejo clínico e processos de cuidados*, *Escalas de avaliação e eventos de risco* e, *Desfechos e status do paciente*. Estes indicadores foram escolhidos pela capacidade de informar e quantificar o atendimento oferecido ao pacientes em termos de qualidade. Dentre eles podemos mencionar procedimentos como a *'Trombectomia'*, que é um indicador de *Manejo clínico e processos de cuidados*, uma condição clinica como o acontecimento de um *AVC prévio* ao atendimento, sendo este um indicador de *Características clinicas*, a *Escala Braden* em *Escalas de avaliação e Eventos de Risco* e, a capacidade de *Auto cuidado* como indicador de *Desfechos e Status do Paciente*.

Para esse mesmo desafio várias abordagens utilizando aprendizado de máquina foram implementadas e apresentaram resultados satisfatórios Zanotto et al. (2021), entretanto estas requerem grande quantidade de dados anotados e capacidade de processamento para o treinamento dos modelos. Nossa proposta é desenvolver uma ferramenta computacional de IA baseada em ontologias e Processamento de Linguagem Natural (PLN)

para automatizar a tarefa mencionada. Com essa abordagem evitamos a necessidade dos dados anotados e do tempo de treinamento dos modelos.

A proposta utiliza técnicas de Processamento de Linguagem Natural (PLN) para a identificação de termos e a ontologia para raciocinar sobre as informações encontradas e inferir a classificação dos Indicadores. A ontologia relaciona termos e palavras chaves elencados pelas especializadas com os indicadores e através dessas relações utiliza axiomas que permitem a inferência das classificações dos Indicadores. As técnicas utilizadas para a detecção das informações nos textos incluem modelos de word-embeddings para a expansão de termos e listas concorrentes para a otimização do processamento.

O modelo proposto foi utilizado para classificar 46.547 sentenças que foram retiradas das evoluções clinicas de uma amostra dos dados de 191 pacientes que estavam sob os cuidados AVC do hospital base. Todas as sentenças estavam anotadas com as classificações esperadas para os 30 Indicadores elencados. O modelo após operar por 532,43 segundos conseguiu processar todas as sentenças utilizadas obtendo, para os 30 indicadores, um valor médio de F1 score de 56,8 % com MCC score de 57,97 %, precisão de 64,89 % e revocação de 54,97 %. Alguns resultados individuais de alguns indicadores como Trombólise e Fibrilação atrial, por exemplo, superam a marca de 80 % em todas as métricas. Outros, entretanto, possuem um desempenho inferior a 20 % como nos casos de Dor e Mobilidade. Uma análise detalhada desses casos revela que esses resultados são diretamente proporcionais a abrangência dos termos e palavras chaves definidos na ontologia.

Os detalhes do desenvolvimento desse trabalho são apresentados seguindo a seguinte forma, a Secção 2 irá fornecer um resumo sobre os principais conceitos e tecnologias que permeiam a concepção deste trabalho. Na Secção 3 detalharemos o projeto do estudo proposto, incluindo as informações sobre os dados e o tratamento destes. Na Secção 4 discutiremos trabalhos relacionados. A descrição dos algoritmos, da ontologia, os métodos de PLN e, os resultados são apresentados na Secção 5. Na Secção 6, discutiremos os resultados obtidos. Por fim, encerramos na Secção 7 apresentando as nossas conclusões e considerações finais.

2. CONCEITOS BASE

2.1 Modelos de gestão baseados em valor na saúde pública

Um modelo de gestão baseado em valor é uma alternativa ao atual modelo comumente adotado chamado de fee-for-service (FFS), ou recompensa-por-serviço (RPS) (Uzuelli et al., 2019). O modelo atual recompensa financeiramente as instituições de saúde proporcionalmente ao volume de atendimentos e procedimentos realizados (Bessa, 2011; George e Engel, 1980). Isso é, quanto mais procedimentos são feitos e quanto mais pacientes são atendidos maior é a remuneração. Dentro da realidade do Sistema Único de Saúde brasileiro (SUS) isso implica que o repasse de verbas e remunerações vindas dos cofres públicos é feita de acordo com o volume dos serviços prestados pelos hospitais e instituições credenciadas (Sharecare, 2020). Nessa conta não é considerada a qualidade dos atendimentos e nem os desfechos obtidos ou, a real necessidades dos procedimentos realizados. Essas condições tornam favoráveis os atos de diminuição da qualidade dos serviços prestados, do prolongamento dos tratamentos e, da solicitação de exames desnecessários pois por esses meios é possível conquistar melhores repasse de verbas e remunerações (Uzuelli et al., 2019; Sharecare, 2020).

A proposta de uma GSBV se propõem a modificar esse cenário (da Silva Etges et al., 2020). Adicionando o fator qualidade na equação, propõe-se por esta abordagem trazer o foco para a entrega, na perspetiva do paciente, dos melhores resultados ao menor custo possível (Gonçalves, 2019). Através desse modelo busca-se valorizar as instituições proporcionalmente à excelência de seus atendimentos e dos desfechos obtidos, incentivando assim um maior zelo pelos pacientes, elevando o nível de qualidade do SUS, aumentando, a adesão do paciente ao tratamento e a satisfação da relação médico-paciente (Uzuelli et al., 2019; Putnam e Lipkin, 1995). Neste sistema, as instituições que obtiverem os melhores atendimentos e os melhores resultados através dos cuidados oferecidos aos seus pacientes, serão as que receberão os melhores repasses de verbas e bonificações.

A implementação desse modelo baseia-se nas informações e dados obtidos durante os atendimentos dos pacientes (da Silva Etges et al., 2020; Gonçalves, 2019). Uma forma de fazer a avaliação da qualidade dos atendimentos é através da analise do histórico individual de cada paciente tratado pelas instituições, desde o momento de sua admissão até o momento de sua alta. Nesses históricos, também chamados de evoluções médicas, constam as informações dos procedimentos, dos tratamentos, dos manejos e dos eventos adversos que ocorreram durante o período que esse paciente esteve sob os cuidados de uma instituição (Lee, 2010; Tsai et al., 2018). Através dessas informações é possível quantificar e qualificar os serviços prestados convertendo esse conteúdo em uma métrica que irá auxiliar a elencar as instituições com os melhores desempenhos (da Silva Etges et al.,

2020). Para que a GSBV seja possível é fundamental compreender, ter acesso e ser familiarizado aos Registros Eletrônicos de Saúde (RES) pois através deles são identificados os eventos chaves ocorridos durante o tratamento de um paciente.

Por isso, na próxima seção brevemente explicaremos o que é um RES, quais as suas aplicações, vantagens, desvantagens e, como estes estão inseridos no cenário atual da saúde,

2.2 Registros Eletrônicos de Saúde

Registros eletrônicos de saúde podem ser definidos como o último nível evolutivo de um projeto conhecido como o Prontuário Eletrônico do Paciente (Anderson, 1999). Os prontuários eletrônicos, por sua vez, tem por objetivo o uso da informática como forma de organizar e armazenar a informação dos prontuários em papel. Este projeto possuí cinco níveis evolutivos: o registro médico automatizado, o registro médico eletrônico computadorizado, o registro médico eletrônico, o registro eletrônico do paciente e o registro eletrônico de saúde. Nesse último nível apenas é que as informações são interligadas de forma interinstitucional, ou seja, os dados são mantidos através das visitas e atendimentos ocorridos em diferentes instituições com diferentes especialidades (Borges et al., 2007).

Os RES apresentam vários benefícios, dentre eles o compartilhamento dos dados, a capacidade de recuperar informação, a melhora da assistência ao paciente, a viabilidade de trabalhar com padrões universalmente aceitos e com vocabulários pré-definidos, a atualização continua em nível municipal, estadual e regional o que permite o apoio a definição de politicas públicas e a regulamentação de demandas. De forma geral o RES facilita o acesso às informações do paciente, baliza as decisões inerentes aos cuidados de saúde do paciente e são essenciais para prevenir erros médicos e acompanhar o andamento da evolução de quadros clínicos (Borges et al., 2007; Standford, 2018; Patrício et al., 2011; Galvao e Ricarte, 2011; Ltda., 2016). Porém, mesmo com essa grande lista de vantagens, um dos principais desafios dessa tecnologia é a percepção que os médicos possuem dela.

A Standford Medicine e a The Harris Poll avaliaram a percepção dos médicos de cuidados primários quanto aos Registros Eletrônicos de Saúde (Standford, 2018). Dentre as métricas obtidas, destacamos quatro delas. A primeira, indica que 71% dos pesquisados concordam que o uso de sistema de registros eletrônicos de saúde contribuem para o cansaço e aumenta o tempo trabalhado por dia. Na segunda métrica, 69% concordam que mais tempo é gasto preenchendo os campos padronizados dos sistema que com o paciente, pois eles encontram dificuldades com interfaces de pouca intuitividade. A terceira revela que um em cada dois usuários apontou que os sistema computacionais para saúde são muito desfragmentados e desconexos. A última destacada aponta que sete dentre dez médicos acreditam que solucionar a interoperabilidade semântica é uma das prioridade

para que, na próxima década, os registros eletrônicos possam se tornar mais viáveis e conquistar mais espaço entre os prestadores de serviços.

Em uma outra pesquisa, feita pela Physicians Foundation (Brown, 2018), dentre 8774 prestadores de saúde que foram indagados sobre o quanto os registros eletrônicos de saúde afetaram suas atividades em diferentes categorias foi identificado que 39% concordaram não estarem satisfeitos com o design e a interoperabilidade desses registos.

A percepção dos agentes de saúde com relação a esse nível de compartilhamento de informações pode ser indicado com um dificultador à identificação dos indicadores necessários para um modelo de GSBV. As instituições se limitam a utilizar níveis inferiores do projeto de prontuários eletrônicos que não fornecem o compartilhamento dos dados e não fornecem as funcionalidades de padronização de cadastramento desses dados. Por isso, a maior parte das informações importantes que precisam ser analisadas estão registradas em formato de texto livre, o que dificulta e retarda a extração tanto manual quanto automatizada desses dados.

Conhecer e entender esses sistemas é um passo importante para que possamos conseguir o acesso aos dados neles contidos. Fazer isso de forma que seja esteja de acordo com a percepção dos prestadores de saúde é um dos principais objetivos desse trabalho para que possamos caminhar em direção a implementação de um sistema de GSBV.

Apresentados os conceitos sobre GSBV e RES, ainda temos que apresentar um terceiro conceito para que a base de conhecimentos desse trabalho seja construída. Por isso, na próxima seção escreveremos sobre ontologias, suas definições, tipos e usos.

2.3 Ontologias

A palavra ontologia é um termo vindo da filosofia onde nesse campo significa a explicação sistemática das coisas (Gomez-Perez et al., 2006).

No campo da computação uma definição comumente acordada de uma ontologia é a especificação, explicita e formal, da conceitualização de um domínio de interesse, onde por formal entende-se que seja uma especificação compreensível por um computador que com esse entendimento seja capaz de realizar deduções sobre os conceitos (Davies et al., 2006; Barcellos e Peixoto, 2004).

Uma ontologia normalmente inclui um vocabulário de termos e as especificações dos seus significados, o que engloba as definições e as relações de como os conceitos estão inter-relacionados impondo uma estrutura sobre um domínio e limitando as possíveis interpretações dos termos (Gomez-Perez et al., 2006).

Os componentes básicos de uma ontologia são as classes, estas organizadas em uma taxonomia, as relações, que retratam o tipo de interação entre os conceitos do domínio, os axiomas, que são usados para moldar sentenças que devem ser sempre verdadeiras, e instâncias, que são utilizadas para representar os próprios dados (Barcellos e Peixoto, 2004).

A tendência que se observa é o largo uso de ontologias em áreas como engenharia de conhecimento, inteligência artificial, design e integração de base de dados, gerenciamento de informações relacionadas ao processamento de linguagem natural, integrações e recuperações de informações, e em campos emergentes como a web semântica (Gomez-Perez et al., 2006).

Ontologias podem ser classificadas de acordo com o seu nível de formalidade. Um desses níveis é conhecido como *'is-a'*, em que os conceitos são organizados de acordo com uma organização de conjuntos e subconjuntos (Lassila e McGuinness, 2001). Uma forma de descrever uma ontologia nesse nível é utilizando uma linguagem de ontologia web (OWL).

2.3.1 Ontologias OWL

A *W3C Web Ontology Language (OWL)* é uma linguagem da web semântica projetada para representar ricos e complexos conhecimentos sobre coisas, grupos de coisas, e as relações entre as coisas (Group, 2012).

Uma ontologia OWL pode formalizar um domínio, definindo classes e propriedades destas classes, definir indivíduos e afirmações sobre eles e, usando-se a semântica formal OWL, especificar como derivar consequências lógicas, isto é, fatos que não estão presentes na ontologia, mas são vinculados pela semântica (Welty et al., 2009).

Uma ontologia OWL é composta por *Indivíduos*, que representam objetos de interesse de um domínio, *Propriedades* que são relações binárias entre indivíduos e, *Classes* que são interpretadas como conjuntos que contém indivíduos (Horridge et al., 2004).

A OWL foi projetada para prover uma linguagem de ontologia que pudesse ser usada para descrever, de um modo natural, classes e relacionamentos entre elas em documentos e aplicações Web (de Lima e de Carvalho, 2005).

Observando esse conceito percebemos que podemos utilizar essa tecnologia para organizar as informações presentes nos RES, relacionado termos e dados-chaves à conceitos específicos, possibilitando que uma aplicação computacional possa usar essa organização para realizar a operação de detecção automática das informações relevantes para a implementação de um modelo de GSBV.

2.4 Word Embeddings

O último conceito que apresentamos é o de *'Word Embeddings'*. Este conceito pode ser descrito como a vetorização de palavras e compõem uma das unidades fundamentais dos algoritmos de processamento de linguagem natural e são utilizados para modelar matematicamente a representação de palavras considerando suas relações de similaridade semântica e sintática no contexto em que ocorrem (Corrêa Cordeiro et al., 2018). Com essa ferramenta, é possível inferir relações de similaridade entre duas palavras a partir do cálculo da distância entre seus vetores (Mikolov et al., 2013).

Exemplificando, suponha que nós temos um conjunto de textos com sete palavras *(bee, eagle, goose, helicopter, drone, rocket, e jet)* e três conceitos *(wings, engine, e sky)*, no espaço vetorial cada palavra é caracterizada por três coordenadas que correspondem ao número de vezes que a cada palavras é encontrada em cada contexto. Supondo que, *helicopter* não apareça no contexto *wings* e ocorra duas vezes em engine e quatro em *sky* as suas coordenadas seriam então (0,2,4), de forma similar cada palavra ocupa uma posição especifica nesse espaço vetorial, como é ilustrado na figura 2.1 (Desagulier, 2008).

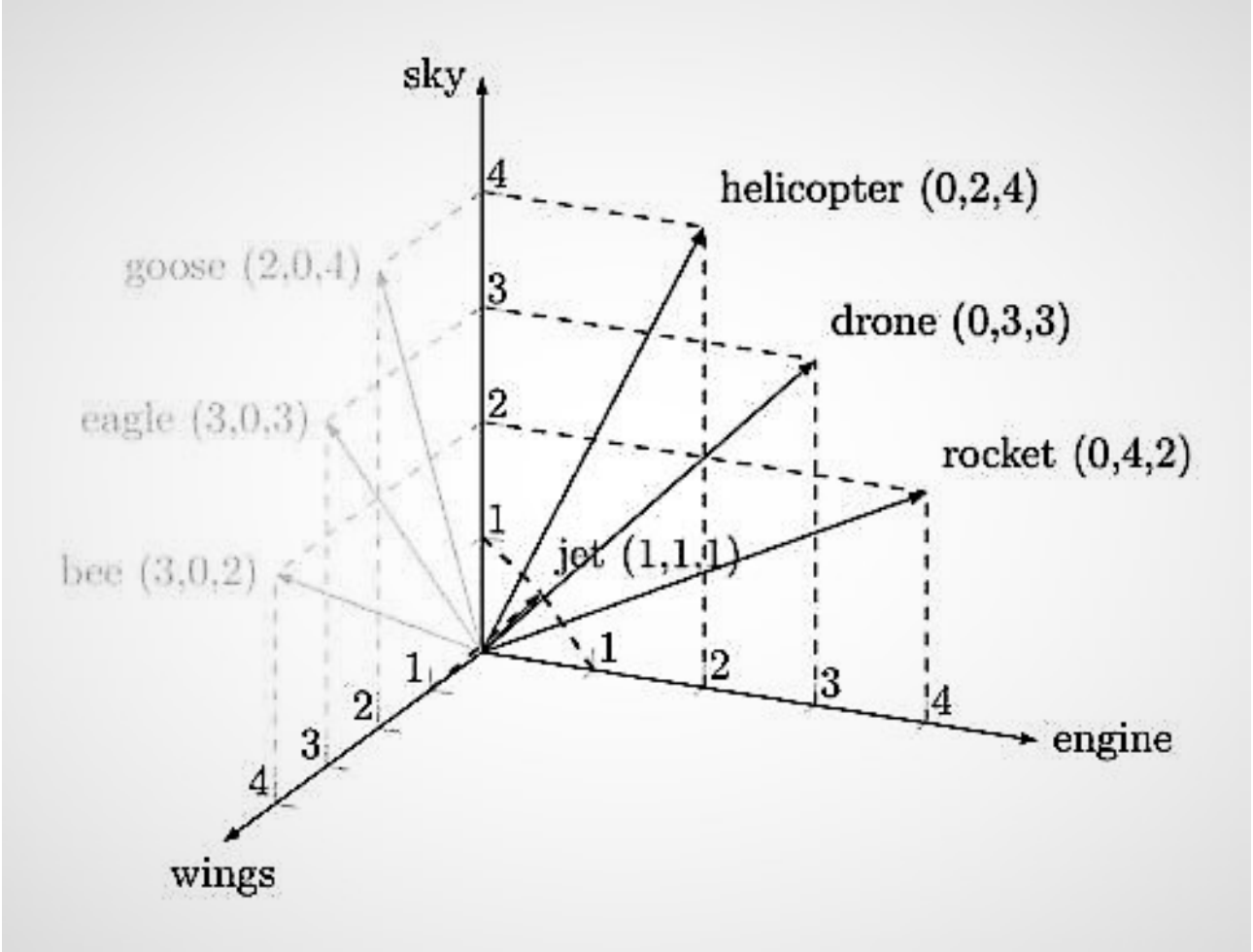

Figura 2.1 – Ilustração de um espaço vetorial com Word Embedding (Desagulier, 2008).

O processo de aprendizado de um modelo Word Embeddings é feito através de um rede neural sobre uma tarefa específica, como classificação de documentos ou é feita através de um processo não supervisionado usando uma análise estatística de documentos. Com esse treinamento é possível captar o significado das palavras e quais são usadas em contextos similares e portanto retratando a similaridade entre elas (Brownlee, 2019).

Dessa forma é possível expandir um conjunto de palavras e termos conhecidos com as palavras aprendidas como similares dentro de um contexto.

Como a premissa de uma GSBV se baseia na análise de textos livres em prontuários as ferramentas de word embeddings podem ser exploradas para realizar a expansão dos termos e conceitos que precisam ser encontrados nesses textos e assim, possivelmente, melhorar o desempenho de modelos computacionais nessa tarefa.

3. PROJETO DO ESTUDO

3.1 Questionamento proposto

O principal propósito desse estudo é verificar a aplicabilidade e o desempenho de um modelo computacional baseado em ontologias na tarefa de detectar de forma automatizada e precisa as informações em texto livre de prontuários eletrônicos.

3.2 Dados disponíveis

Para responder o nosso questionamento trabalhamos com a detecção de informações das evoluções médicas de pacientes da linha de cuidados de acidentes vasculares cerebral.

Esta decisão foi tomado devido à disponibilidade dos dados pertencentes a este domínio e, pela proximidade à uma equipe de especialistas desse domínio. Os dados disponíveis para esse estudo são os textos de evoluções clínicas de 281 pacientes hospitalizados no hospital-base para tratamento de AVC. Estes dados datam da primeira internação no dia 01/01/19 até a última alta no dia 23/07/2019.

3.3 Caracterização do problema

O desafio apresentado caracteriza-se pela detecção de informações em textos livres de prontuários eletrônicos. Essas informações foram organizadas, com o auxilio de uma equipe de especialistas de domínio, em 30 indicadores. Estes indicadores foram divididos em quatro sub-grupos: Características clínicas; Manejo clínico e processo de cuidado; Escalas de avaliação e eventos de risco; Indicadores clínicos, desfechos e status do paciente. Essa divisão é proposta para que seja possível diferenciar os indicadores que são específicos da linha de cuidados AVC, dos indicadores que são genéricos à outras linhas de cuidados. Na Tabela 3.1 apresentamos todos os indicadores e seus sub-grupo.

No Anexo A são descrito em detalhes esses indicadores, e são definido quais são as sentenças, os termos e as palavras chaves que são utilizados para identificar a presença ou confirmação de um determinado indicador em uma determinada parte da evolução. Este levantamento de termos além de auxiliar na tarefa sendo abordada também apresenta potencial valia para servidores de saúde que podem usar esta lista como referência para

Tabela 3.1 – Resumo das variáveis e seus sub-grupos

Sub-Grupo	Variaveis			
Caracteristicas Clinicas	**Doença Coronaria**	**AVC Prévio**	**Dislipidemia**	**Etilista**
	Fibrilação Atrial	**Hipertensão Arterial Sistemica**	**Câncer**	**Diabetes**
	Obesidade	**Tabagismos**		
Manejo Clínico e Processo de Cuidado	Localização	Trombectomia	Trombólise	
Escala de Avaliação e Eventos de Risco	**Hemorragia Intracraniana**	**Queda**	**Escala Brande**	**Risco de Queda**
	Indicativo de Infecção			
Desfechos e Status do Paciente	Óbito	Dor	Alimentação	Força
	Paresia	Mobilidade	Nível de Mobilidade	Comunicação
	Capacidade Cognitiva	Rankin	Auto Cuidado	NIH

unificar e alinhar as formas de se expressar e se comunicar durante a composição das evoluções.

Na Tabela 3.2 apresentamos as descrições de alguns dos indicadores, como devem ser feita a classificação destes, quais são as características generalizadas dessas classificações, quais são os exemplos de sentenças que são utilizados pelas especialistas para identifica-los e, quais dos demais indicadores compartilham as mesmas características e portanto são classificadas de modo semelhante.

Tabela 3.2 – Descrição dos indicadores e de suas classificações

Indicador	Descrição do Indicador	Classificações	Característica de Classificação	Exemplos de Sentenças	Indicadores com característica similares
Trombólise	Refere ao possível manejo clínico que o paciente AVCi recebeu.	Sem menção ao procedimento	Sem menção explicita	'Sem história pregressa de dislipidemia'	**Doença Coronária** **Fibrilação Atrial** **Diabetes** **Hipertensão Arterial Sistêmica** **Dislipidemia** **Trombectomia**
		Procedimento não realizado	Menção explicita da não ocorrência do indicador através de termos	'Não trombolisado - delta' 'Contraindicação a trombólise' 'Sem janela trombolítica'	
		Procedimento Realizado	Menção explicita da ocorrência do indicador através de termos	'Trombólise endovenosa' 'Alteplase' 'Avci trombolisado'	**Hemorragia Intracrâniana** **Óbito**
Obesidade	Obesidade	Doença ausente	Sem menções	'Trombólise endovenosa'	
		Doença presente	Menção explicita da ocorrência do indicador através de termos,acompanhados ou não de valores numéricos	'imc > 30' 'obeso'	**Paresia**
AVC prévio	Acidente vascular cerebral, não especificado como hemorrágico ou isquêmico e Infarto Cerebral	Sem menção ao evento	Sem menções	'índice de massa corporal = 30'	
		Evento não ocorrido	Menção explicita da não ocorrência do indicador através de termos	'Sem história de AVC prévio'	**Câncer**
		Evento ocorrido	Menção explicita da ocorrência do indicador através de termos e em um determinado tempo.	'AVC isquêmico prévio' 'História prévia: AVCi AVCh' 'AVC em 2008'	
Tabagismo	Registro do consumo condicionado a dependência de cigarros ou outros produtos que contenham tabaco.	Sem menção a condição	Sem menções	'AVC isquêmico prévio'	
		Tabagista	Menção explicita do indicador através de termos	'Tabagista' 'Fumante'	
		Ex-tabagista	Menção explicita do indicador através de termos e em um determinado tempo	'Ex-tabagista' 'Tabagista no passado'	**Etilismo**
		Não tabagista	Menção explicita da não ocorrência do indicador	'Nega tabagismo' 'Não tabagista'	
Localização	Indica o local de presente momento do paciente.	Sem menção a condição	Sem menções	'Tabagista no passado'	**Alimentação**
		Emergência	Menção explicita da classificação do indicador através de termos	'evolução diária - emergênca'	**Queda**
		CTI	Menção explicita da classificação do indicador através de termos	'Interna na UTI após procedimento'	**Comunicação**
		Unidade de internação	Menção explicita da classificação do indicador através de termos	'Enfermagem C1'	
Escala Braden	Escala de Braden é um recurso utilizado nas Unidades de Terapia Intensiva para medir o risco dos pacientes críticos de desenvolverem lesões por pressão.	Sem menção a condição	Sem menções	'Paciente deu entrada na UI após procedimento'	**Risco de Queda**
		Risco baixo (braden>=17)	Menção explicita da classificação do indicador através de termos, acompanhados ou não de valores numéricos	'Baixo risco: escore = 20'	**Indicativo de Infecção**
		Risco médio (16<braden<13)	Menção explicita da classificação do indicador através de termos, acompanhados ou não de valores numéricos	'escore: 15'	**Dor**
		Risco alto (braden<=12)	Menção explicita da classificação do indicador através de termos, acompanhados ou não de valores numéricos	'escore = 13'	**Capacidade Cognitiva**
Força	Indica a capacidade de superar ou opor-se a uma resistência por meio da atividade muscular, conforme escala de grau de força, composições descritivas ou nível de mobilidade associado.	Sem menção a situação atual	Sem menções	'Enfermagem C1'	**Nível de Mobilidade**
		Escore 0	Classificação do escore por termos, acompanhados ou não de valores numéricos	'Não percebe contração. Plegia.'	**Escala Rankin**
		Escore 1	Classificação do escore por termos, acompanhados ou não de valores numéricos	'Traço de contração sem produção de moviemento'	**NIH**
		Escore 2	Classificação do escore por termos, acompanhados ou não de valores numéricos	'Contração fraca, elimina gravidade'	
		Escore 3	Classificação do escore por termos, acompanhados ou não de valores numéricos	'Realiza movimento contra a gravidade, porém sem resistência adicional'	
		Escore 4	Classificação do escore por termos, acompanhados ou não de valores numéricos	'Realiza movimento contra a gravidade e resistência externa'	
		Escore 5	Classificação do escore por termos, acompanhados ou não de valores numéricos	'Supera maior quantidade de resistência'	
Paresia	Movimento limitado ou fraco, motilidade num padrão abaixo do normal. Precisão do movimento, amplitude do movimento e a resistência muscular localizada, ou seja, refere-se a um comprometimento parcial, uma perda de força.	Sem menção a situação	Sem menções	'Contração fraca, elimina gravidade'	**Mobilidade**
		Registro de paresia	Menção explicita do indicador, através de termos, ou relação com a classificação de outras classes	'Apresenta hemiparesia' Se indicador Força tiver escore < 3	

Na tabela 3.2 observamos que, por exemplo, para identificar que um paciente possui a condição clinica de obesidade é necessário que ao longo da evolução esteja presente termos como *'obesidade'* ou alguma a menção ao seu *'imc'* sendo que o valor deste deve ser *maior que 30*.

Tomemos por exemplo o seguinte trecho de uma evolução médica retirada dos dados disponibilizados:

> Cardiologia - inicio acompanhamento a pedido da Dra. *nome do médico*-
> Sra.*nome do paciente*, 78 anos
> # HAS desde os 35 anos
> # DM em tto desde os 60 anos
> # Fibrilação atrial em 2008-2009 (uso de amiodarona até fev/2017)
> recorrência de FA documentada desde março/2019 - reiniciou amiodarona e usou até abril/19
> # Obesidade
> Eco com dilatação biatrial (AE 4.9cm)
> ...
> AVCi - trombolise com delta t 4h
> NIHSS chegada: 5 / Após trombólise: 4 / Após 24h: 1
> Após trombolise, vomitos e hipotensão (60x40) - revertido com SF 250ml
> ...

Esse texto, através da presença do termo *'obesidade'*, seria classificado como indicador de que o paciente possui a condição clínica de obesidade. Além disso, percebemos a presença do termo *'trombólise'*, sem negações explicitas como nas sentenças de exemplos *'Sem janela trombolítica'* ou *'Não trombolisada'*, por isso, este texto também seria um indicador de que este paciente realizou este procedimento. Também podemos perceber a indicação da condição de Diabetes através do termo *'DM'*, da realização da Fibrilação atrial com a explicita menção de *'Fibrilação atrial'*, da condição de Hipertensão com o termo *'HAS'* e, do escore de NIHSS.

Através dessa identificação de termos em sentenças que as classificações são feitas pelos avaliadores da evolução. Por isso, conhecer esses conjuntos de termos e como eles se relacionam com os indicadores é importante para o processo de automatização do processo de classificação das evoluções.

Na Tabela 3.3 mostramos os principais termos utilizados para cada classe dos indicadores expostos na Tabela 3.2. Na Tabela 3.4 contamos o total de termos identificados para todos os indicadores exemplificados na Tabela 3.2, no anexo B listamos os termos e a contagem de termos para todos os indicadores. Todos os termos foram elencados e definidos pelos especialistas de domínio.

Do levantamento dos termos percebeu-se que alguns termos são relevantes a mais de um indicador, como é o caso do termos *'delta'* que é utilizado na identificação

Tabela 3.3 – Termos elencados para os indicadores exemplificados

Indicador	Classificações	Termos identificados
Trombólise	Procedimento não realizado	reperfusão, terapia, indicação, contraindicado, trombolítico, contraindicação, trombólise, não, trombolisada, sem, delta, janela
	Procedimento Realizado	delta, reperfusão, terapia, sem, contraindicação, alteplase, trombolisada, trombolítico, trombólise
Obesidade	Doença ausente	imc
	Doença presente	obesa, obeso, obesidade, imc, índice, massa, corporal
AVC prévio	Evento não ocorrido	sem, avc
	Evento ocorrido	acidente, vascular, cerebral, encefálico, prévia, avch, avci, prévio, avc
Tabagismo	Tabagista	fumante, tabagismo, tabagista
	Ex-tabagista	ex-fumante, ex-tabagista, tabagista, passado
	Não tabagista	nega, tabagismo, tabagista
Localização	Emergência	emergência
	CTI	fisioterapia, intensiva, terapia
	Unidade de internação	internação, unidade, a1, a2, a3, a4, b1, b2, b3, b4, c1, c2, c3, c4, d1, d2, d3, d4, e1, e2, e3, e4
Escala Braden	Risco baixo (braden>=17)	braden
	Risco médio (16<braden<13)	braden
	Risco alto (braden<=12)	braden
Força	Escore 0	paralisia, plegia, plégico, não, contração
	Escore 1	sem, contração, movimento
	Escore 2	contração, fraca, elimina, gravidade
	Escore 3	resistência, contra, gravidade, movimento
	Escore 4	resistência, contra, gravidade, movimento, força, perda, leve, sutil
	Escore 5	maior, resistência, supera, força
Paresia	Registro de paresia	força, mie, perda, hemiparesia, hemiparético, parético

dos indicadores *'Trombólise'* e *'Trombectomia'*. Percebemos que se a classificação deste indicadores dependesse apenas desse termo teríamos dificuldades em des-ambíguar a classificação. Entretanto, também foi constatado que poucos indicadores possuem apenas um termo como palavra chave para classificação e que na maioria dos indicadores as composições dos termos elencados são únicas e não geram ambiguidades.

Com os exemplos e termos definidos podemos perceber que existe uma relação de pertencimento ou existência entre os termos e os textos das evoluções. Ou seja, um conjunto de termos se relaciona com um conjunto de sentenças de forma que ao estarem presentes nesses textos eles as qualificam e indicam a presença dos indicadores ao quais queremos identificar para fazer a avaliação do atendimento prestado.

Tabela 3.4 – Termos elencados para os indicadores exemplificados

Indicador	Classificações	Total de Termos	Total de Termos Diferentes
Trombólise	Procedimento não realizado	11	14
	Procedimento Realizado	9	
Obesidade	Doença ausente	1	4
	Doença presente	4	
AVC prévio	Evento não ocorrido	2	10
	Evento ocorrido	9	
Tabagismo	Tabagista	3	7
	Ex-tabagista	4	
	Não tabagista	3	
Localização	Emergência	1	26
	CTI	3	
	Unidade de internação	22	
Escala Braden	Risco baixo (braden>=17)	1	1
	Risco médio (16<braden<13)	1	
	Risco alto (braden<=12)	1	
Força	Escore 0	5	18
	Escore 1	3	
	Escore 2	4	
	Escore 3	3	
	Escore 4	7	
	Escore 5	4	
Paresia	Registro de paresia	6	6
Total			86

3.4 Definição do método

A nossa abordagem para a detecção dos indicadores contidos nos diferentes níveis de prontuários eletrônicos é a sumarização em classes dos textos livres contidos nas evoluções. Essa solução permeia os sistemas de prontuários eletrônicos já existentes, dando a eles a capacidade de continuar a operar da forma como foram projetados e ainda compartilhar os dados requeridos. Dessa forma acreditamos que os profissionais da saúde poderão continuar utilizando o nível de prontuários eletrônicos aos quais estão acostumados e não precisaram ceder à padronizações de um sistema unificado, como sugere um sistema de RES. Estes profissionais poderão também continuar utilizando a linguagem natural em suas evoluções de forma a conservar a privacidade de seus métodos e condutas, uma vez que a informação exposta das evoluções será um resumo em conceitos-chaves dos eventos e das situações ocorridas durante o serviço prestado à um paciente.

A Figura 3.1 representa todo o processo do projeto sendo que o projeto de estudos presente se propõe a abordar somente a parte destacada, isso é, a composição do modelo de classificação.

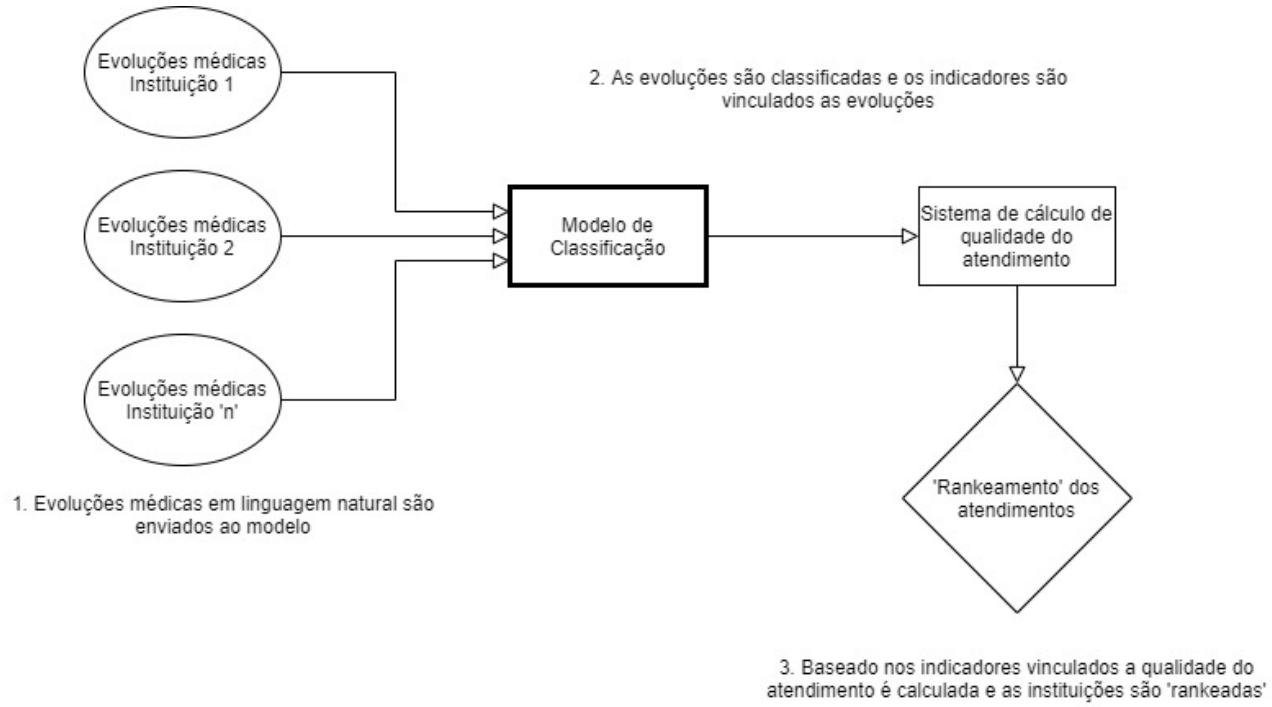

Figura 3.1 – Contextualização do modelo.

Por termos a colaboração de especialistas de domínio optamos por construir um modelo baseado em uma ontologia e um algoritmo de detecção de textos. Objetivamos aproveitar o conhecimento trazido por essa equipe organizando as informações fornecidas e levantadas por ela de uma forma que seria possível relacionar os conjuntos de termos e palavras chaves com os textos das evoluções permitindo, de forma automatizada, o processo de dedução das classificações como exemplificado anteriormente.

Para esse fim, a ontologia criada tem três principais características. A primeira é como um dicionário, dos termos e palavras chaves que precisam ser encontrados nos textos e evoluções e que são utilizados pelo algorítimo. A segunda característica é conceitualizar os indicadores a serem classificados e modelar a relação deles com os termos elencados. A terceira é, após a identificação dos termos nas sentenças, armazenar as sentenças com a associação dos termos identificados pelo algoritmo e a partir dessas associações realizar o processo de inferência para a classificação dessas sentenças nas classes que modelam os indicadores. A Figura 3.2 apresenta o fluxograma do método projetado.

3.5 Medidas de avaliação adotadas

Para a avaliação dos resultado obtidos foram utilizados as métricas *'f1-score'*, *'mcc-score'*, *'Precision'*, *'Recall'* e, *'Accuracy'*. Estas métricas são baseadas na matriz de confusão que relaciona os resultados obtidos das classificações em *'Falsos Positivos'*,

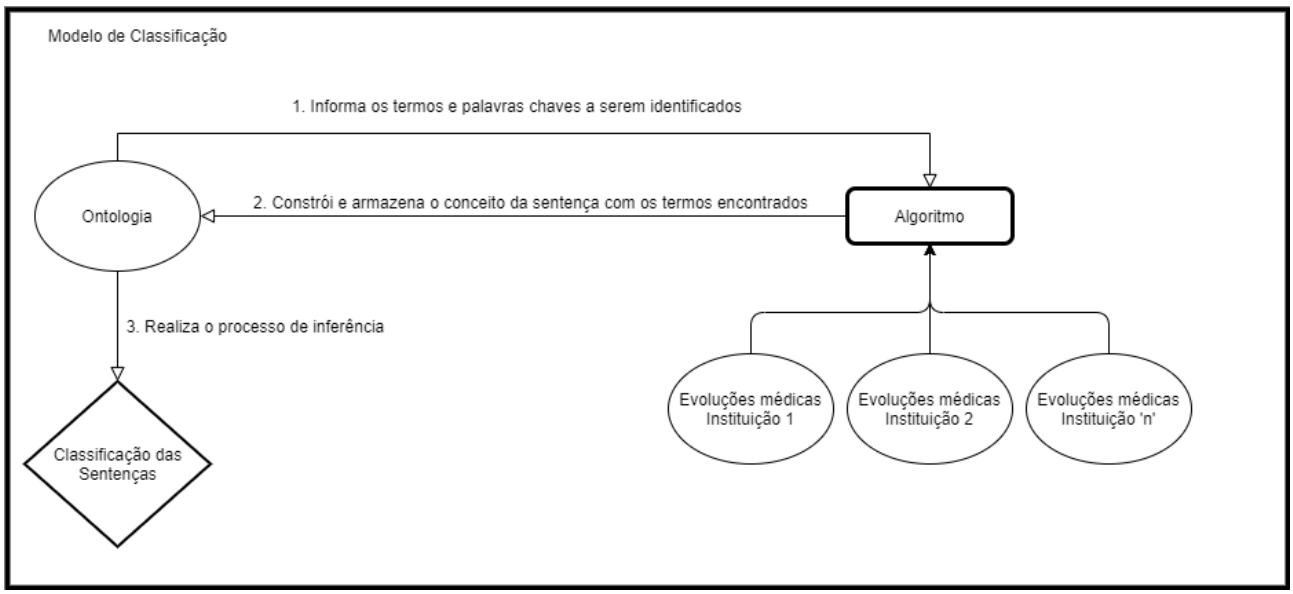

Figura 3.2 – Fluxo do método adotado

'Falso Negativos', *'Verdadeiros Positivos'* e, *'Verdadeiros Negativos'* (Joseph, 2009; Chicco e Jurman, 2020).

Estes valores são obtidos conforme as equações exposta a seguir:

$$Recall = \frac{\text{VP}}{\text{VP + FN}} \tag{3.1}$$

$$Accuracy = \frac{\text{VN}}{\text{VN + FP}} \tag{3.2}$$

$$Precision = \frac{\text{VP}}{\text{VP + FP}} \tag{3.3}$$

$$F_1\ score = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}} \tag{3.4}$$

$$MCC = \frac{\text{VP * VN - FV * FN}}{\sqrt{\text{(VP + FP) * (VP + FN) * (VN + FP) * (VN + FN)}}} \tag{3.5}$$

A Tabela 3.5 explicita quais as condições, dentro do contexto do problema proposto, que caracterizam as classificações de verdadeiros positivos, verdadeiros negativo, falsos positivos e falsos negativos.

Tabela 3.5 – Descrição das condições de classificação

Verdadeiros Positivos	Verdadeiros Negativos
Valor anotado = Valor classificado E Valor Anotado $\neq (0 ou -1)$	Valor anotado = 0 E Valor Classificado = 0
	Valor anotado = -1 E Valor Classificado = -1
Falsos Positivos	Falsos Negativos
Valor Anotado = 0 E Valor Classificado $\neq (0 E -1)$	Valor anotado $\neq -1 E Valor classificado = -1$
Valor anotado $\neq Valor classificado E Valor Anotado \neq (0 ou -1)$	Valor anotado $\neq (-1 E 0) E Valor classificado = 0$
Valor anotado = -1 E Valor classificado $\neq (0 ou -1)$	

4. TRABALHOS RELACIONADOS

Nesta seção apresentamos alguns trabalhos relacionados do campo da ciência da computação, para que possamos ter um compreendimento melhor de como a utilização de ontologias tem sido aplicada em processos de classificação de textos. Fazemos uma revisão de 9 trabalhos que empregam ontologias na resolução do problema de classificação de textos livres. Os 9 trabalhos estudados foram: *'A Comparative Study for Domain Ontology Guided Feature Extraction'* (Wang et al., 2003), *'Ontology-Based Multilabel Text Classification of Construction Regulatory Documents'* (Zhou e El-Gohary, 2015), *'Ontology-based text classification into dynamically defined topics'* (Allahyari et al., 2014), *'Ontology-guided feature engineering for clinical text classification'* (Garla e Brandt, 2012), *'Using ontology-based text classification to assist Job Hazard Analysis'* (Chi et al., 2014), *'Fostering natural language question answering over knowledge bases in oncology EHR'* (Schwertner et al., 2019), *'Ontology-based information extraction for juridical events with case studies in Brazilian legal realm'* (de Araujo et al., 2017), *'Ontology based Concept Extraction and Classification of Ayurvedic Documents'* (Gayathri e Kannan, 2020), *'Ontology-based clinical information extraction from physician's free-text'* (Yehia et al., 2019).

Da análise dos trabalhos vemos que (Wang et al., 2003), (Zhou e El-Gohary, 2015) e, (Garla e Brandt, 2012) utilizam algoritmos de aprendizado de máquina em alguma parte do processo de classificação, tanto para aprendizado dos conceitos-termos ou para a o processo de classificação em si. Para evitar a necessidade de treinamento de um modelo de aprendizado de máquina, reduzir a quantidade de dados anotados, em nossa abordagem sugerimos que a ontologia deveria ser a ferramenta classificadora baseando-se no conhecimento nela modelado e sendo auxiliada por um algoritmo de detecção de textos. No trabalho (Allahyari et al., 2014) encontramos exatamente essa abordagem, onde utiliza-se a ontologia como classificador. Nesse trabalho os autores fazem a projeção e seleção de grafos para efetivar a classificação. Em nossa abordagem buscamos uma solução mais simplificada através da detecção de termos em sentenças. Em (Chi et al., 2014) a ontologia foi criada de forma semi-automática. Porém devido a baixa quantidade de dados a abrangência dos resultados não atinge bons índices. Aprendemos desse trabalho que para obter bons resultados precisamos criar a ontologia do forma manual, ou termos uma grande quantidade de dados. Vendo o trabalho (Schwertner et al., 2019) notamos que a ontologia foi criada a partir da análise textual das sentenças identificando conceitos chaves em cada uma delas. A ontologia é então utilizada para a definição da relação entre entidades e as sentenças de forma a utilizar essas relações para melhorar as buscas de respostas. A similaridade ao nosso trabalho vem nesse ponto onde também temos termos chaves identificados nas sentenças. No estudo feito em (Gayathri e Kannan, 2020) vemos um desafio similar ao nosso, onde é preciso identificar informações dentro dos textos, a divergência é que além de detectar essas informações os autores precisavam definir a importância e

relevância dessas informações. Sendo esta última parte fora do nosso escopo ressaltamos o uso de uma ontologia de domínio como suporte e base para o processo de classificação e extração de informação. No artigo (Yehia et al., 2019) após superado o desafio da conversão dos textos escritos a mão para formatos digitais, podemos observar o uso de ontologias de domínio para classificar sentenças em conceitos presentes nessas ontologias. As classificações são feitas através de um conjunto de regras baseados nas relações de pertencimento entre os dados extraídos e os conceitos definidos.

O trabalho (de Araujo et al., 2017) apresenta a ontologia como ferramenta classificadora através do processo de inferência realizado através de regras linguísticas elencadas por especialistas do domínio criadas através da análise dos termos presentes nas sentenças dos textos. Divergindo apenas no domínio este trabalho se alinha com o nosso em todos os outros aspectos, mostrando excelentes resultados para esse tipo de abordagem.

Na Tabela 4.1 compilamos os principais desafios de cada trabalho, assim como as abordagens e metodologias utilizadas para a solução dos problemas.

Todos esses trabalhos e seus resultado nos indicam que o uso de ontologias de domínio podem auxiliar e até mesmo serem os fatores principais dos processos e sistemas de classificação de textos e nos encorajam a seguir o caminho traçado e previsto na Seção 3. Assim na próxima secção detalharemos como a nossa abordagem foi implementada para resolver o nosso problema exposto.

Tabela 4.1 – Resumo dos trabalhos selecionados

Trabalho	Desafio	Solução	Dados	Resultados
[WMAB03]	Resolver as dificuldades sofridas por algoritmos de classificação de textos e, algoritmos de seleção de características.	Uma ontologia de domínio hierárquica que reduz a quantidade de características a serem extraídas dos documentos e guia o processo de extração e classificação.	Documentos de 10 journals da base de dados MEDLINE. 150 artigos foram selecionados. Destes 50 aritgos foram utilizados para testes e 100 para validação.	A ontologia proposta conseguiu reduzir a quantidade de características necessárias para os algoritmos de KNN melhorando a performance destes no processo de classificação atingindo uma acuracidade de 83%
[ZEG16]	Classificar documentos de textos em categorias pré-definidas para automatizar a verificação de regras regulamentorias de projetos de construções.	Um algoritmo de classificação de textos baseado em ontologias de domínio que utilizam as características semânticas dos textos.	Vinte e cinco documentos regulatórios de construções foram utilizados e deles foram selecionados 2400 clausulas.	Os resultados mostram que o algoritmo baseado em ontologia obteve melhores resultados quando comparados com modelos de classificação baseados em aprendizado de máquinas em 4 métricas diferentes. Os valores obtidos foram recall de 98,7% e precision 92,7%.
[AKJ14]	Classificar automaticamente documentos de textos em tópicos definidos dinâmicamente.	Uma ontologia de domínio e um conjuntos de tópicos são utilizados para medir a similaridade semântica entre o grafo temático das projeções dos documentos de textos e o sub-grafo projetado das definições dos contexto.	Corpus de noticias extraídos do Reulers RSS feed entre 10/2013 e 01/2014 totalizando 3872 documentos de texto.	A ontologia atingiu uma precisão média de 93,8% ao classificar os textos em tópicos de contextos de alto nível. Ela atingiu uma precisão média de 87,6% na classificação em tópicos de sub-categorias. E obteve uma precisão média de 89,3% na classificação em tópicos compostos.
[GB12]	Aprimorar a classificação de textos clínicos baseadas em aprendizado de máquina.	Utilização da estrutura taxonômica do Sistema Médica de Linguagem Unificada para aprimorar o ranqueamento de características e, projeção da similaridade semântica pelos textos clínicos no espaço de características.	Conjunto de dados do desafio I2B2 2008 contendo 1237 avaliações textuais de pacientes diabéticos ou com sobrepeso.	Dentre os cinco experimentos realizados os resultados de classificação alcançaram um micro-f1 de 63,55% e um macro-f1 de 95,94% sendo estes valores maiores que 8 dos melhores classificados no desafio I2B2 2008.
[CLH14]	Aprimorar a identificação de cenários de riscos em industrias de construção e suas possíveis medidas de segurança através da análise semi-automática de textos de guias de segurança.	Uma ontologia de domínio construída semi-automaticamente que mapeia as situações de risco, os cenários de uma da construção e, verifica a aplicabilidade das medidas de segurança em cada cenários através de um algoritmo de classificação de textos baseado na ontologia.	Três conjuntos de fontes foram utilizados para criar a ontologia de domínio, sendo elas, a base de dados CPWR, os relatórios da NIOSH FACE e, os padrões OSHA. No total foram 1166 documentos de textos selecionados.	Dentre os experimentos executados as classificações obtiveram alta precisão chegando até 100% porém apresentaram baixo recall, em moda 20%. Os autores concluem que haveria a necessidade de uma maior quantidade de dados para a composição da ontologia e dos testes.
[SRA+19]	Compor um sistema de perguntas e respostas que informe através da linguagem natural o conteúdo presente nos textos livre de evoluções médicas.	A integração de três recursos: Um processo de extração de informações, uma base de conhecimento e um sistema de perguntas e respostas.	Foram utilizados 3.309 documentos com dados clínicos compilados de Abril de 2017 até Novembro de 2018, que geraram 212.829 palavras e 28.178 linhas.	A precisão do sistema obteve uma recuperação de respostas corretas de 87,5%.
[dARB17]	Extrair de textos livres informações e eventos jurídicos.	Uma ontologia de domínio com regras linguísticas integrado com um mecanismo de inferência utilizada para a extração das informações.	Um corpus de 200 documentos compostos de 39.895 sentenças foi utilizado.	Os melhores resultados obtidos atingem a marca de 100% de precision e 100% de recall
[GK20]	Encontrar os conceitos mais relevantes dentro de textos e escrituras sobre a medicina tradicional indiana.	Um sistema de extração e classificação de conceitos baseado em uma ontologia de domínio.	Aproximadamente 3.600 documentos foram coletados. Desses 2850 documentos possuíam conteúdos sobre a medicina tradicional e 750 não possuíam conteúdos pertinentes ao domínio.	O sistema baseado na ontologia obteve como resultado um recall superior a 80% e uma precision superior a 60%.
[YBA+19]	Extrair conceitos clínicos de notas escritas à mão por médicos e converte-las em dados estruturados para que possam ser acessados através de EHRs	Um sistema de extração de informações baseado em ontologia.	Foram utilizados 150 notas clinicas de três hospitais diferentes.	O desempenho geral do sistema atingiu um recall de 94%, precision de 99% e f-score 96%

5. METODOLOGIA E DESENVOLVIMENTO

5.1 Pré-processamento e seleção dos dados

Para obtermos um conjunto de dados balanceado foi selecionado um sub-conjunto dos dados contendo as evoluções de 191 pacientes.

Na primeira etapa de pré-processamento, realizou-se o processo de anonimizar os nomes dos pacientes e do corpo clínico contidos nas textos. Todas as evoluções foram processadas por um algoritmo que substituía o nome dos pacientes e dos médicos por caracteres genéricos. Os nomes dos médicos e pacientes são informações estruturadas no dado conjunto de dados, assim utilizamos as informações registras nos referentes campos que continham essa informação para detectar e substituir elas no corpo dos textos das evoluções.

A segunda etapa incluiu o processo de separação das evoluções em sentenças. Foi teorizado que com a segmentação dos textos em porções menores, os modelos desenvolvidos teriam melhor desempenho na avaliação por sentença do que por evolução completa (Schütze et al., 2008). Para esse procedimento um algoritmo realiza a separação em sentenças através da detecção nos textos de caracteres de quebra de linha. Na imagem 5.1 mostramos o algoritmo utilizado para essa tarefa.

```
def pre_processing(text):
    new_text = ""
    text = text.replace("\r", "").replace("_x000D_", "")
            .replace("x000D_", "").replace("_x000D", "")
            .replace("x000D","").strip()

    for line in text.split("\n"):
        if line.replace(" ", "").replace("\t", "").strip() == "":
            new_text += ".\n"
        else:
            new_text += " * " + line
    return new_text.strip()
```

Figura 5.1 – Algoritmo de separação em sentenças

Para compreender esse processo, retomemos como exemplo o seguinte trecho de evolução:

Cardiologia - inicio acompanhamento a pedido da Dra. *nome do médico*
Sra.*nome do paciente*, 78 anos
HAS desde os 35 anos
DM em tto desde os 60 anos
Fibrilação atrial em 2008-2009 (uso de amiodarona até fev/2017)
recorrência de FA documentada desde março/2019 - reiniciou amiodarona e usou até abril/19
Obesidade
Eco com dilatação biatrial (AE 4.9cm)
...

Após o processo de separação em sentenças o conjunto de sentenças obtidos é apresentado na Tabela 5.1:

Tabela 5.1 – Sentenças obtidas a partir do texto da evolução

#	Sentença
1	Cardiologia - inicio acompanhamento a pedido da Dra.nome do médico - Sra.nome do paciente, 78 anos
2	# HAS desde os 35 anos
3	# DM em tto desde os 60 anos
4	# Fibrilacão atrial em 2008-2009 (uso de amiodarona até fev/2017) recorrência de FA documentada desde março/2019 - reiniciou amiodarona e usou até abril/19
5	# Obesidade
6	Eco com dilatação biatrial (AE 4.9cm)

Como resultado desse processo os textos das evoluções resultaram em um conjunto de 46.547 sentenças.

5.2 Anotação dos dados

A equipe de especialistas fez a anotação em pares para registrar a ocorrência de cada variável em cada sentença. Para essa tarefa duas especialistas foram selecionadas, ambas com graduação no domínio da saúde. Cada especialistas leu cada uma das 46.547 sentenças e atribui a elas a classificação do indicador encontrado conforme definido no Anexo A. As sentenças foram randomizadas para garantir que não haveria bias sequencial de anotação, isso é, as informações textuais presentes em uma sentença passada afetasse a anotação de uma sentença presente ou futura. Os resultados obtidos pelas anotações individuais foram automaticamente comparados para que fosse identificado divergências entre as anotação. Após a divergências serem identificadas, ambas anotadoras trataram em conjunto cada caso encerrando assim o processo de anotação.

A Tabela 5.2 apresenta o total de sentenças classificadas nos indicadores selecionados como exemplos. No anexo C listamos a contagem para todos os indicadores.

Tabela 5.2 – Total de sentenças identificadas para os indicadores exemplificados

Indicador	Classificação Atribuída	Total de Sentenças
Trombólise	Sem menção	46048
	Não realizou ou sem janela para terapia	107
	Trombólise	392
Obesidade	Sem menção	46460
	Doença Ausente	58
	Doença presente	28
AVC prévio	Sem menção	46309
	Não	29
	Sim	209
Tabagismo	Sem menção	46264
	Não Tabagista	32
	Tabagista	74
	Ex-Tabagista	177
Localização	Sem menção	45035
	Emergência	109
	Unidade de Internação	1073
	UTI/CTI	330
Braden	Sem menção	46287
	Baixo Risco	61
	Risco Médio	59
	Risco Alto	140
Grau de Força	Sem menção	45857
	escore 0	143
	escore 1	19
	escore 2	32
	escore 3	47
	escore 4	252
	escore 5	197
Paresia	Sem menção	46037
	Registro de paresia	510

5.3 Modelo de classificação

5.3.1 Ontologia

O primeiro elemento do nosso modelo de classificação é a ontologia. Para a implementação desse passo a ferramenta de criação e edição de ontologias *Protégé* foi utilizada para criar a ontologia OWL que permite a classificação das sentenças. O *Protégé*, é

uma ferramenta gráfica que nos permite pensar sobre os modelos de domínios a um nível conceitual permitindo que nos concentremos nos conceitos, em suas relações dentro do domínio e nos fatos que queremos expressar (Grosso et al., 1999). A versão mais recente dessa ferramenta foi construída para operar em diferentes plataformas, oferece suporte de customização e vem sendo utilizada por mais de 300 individuais e grupos de pesquisa, majoritariamente interessados na área da medicina informatizada (Noy et al., 2001).

Na criação da ontologia três conjuntos de conceitos foram criados. O primeiro conjunto é o dos *'Termos'*, onde todos os sub-conceitos são as palavras chaves a serem identificadas nas sentenças.

O segundo conjunto são as das *'Sentença'*, onde todos os sub-conceitos são as sentenças analisadas pelo algorítimo e as quais devem ser classificadas pela ontologia.

O terceiro conjunto de conceitos são os *'Indicadores'* as quais as Sentenças devem ser classificadas. Quando necessário, para cada indicador foi criado um subconjunto para cada classificação possível do Indicador.

Uma relação de propriedade de objeto foi criada para expressar o relacionamento entre as Sentenças e os Termos. A relação de objetos criada é chamada *'contain'*, e é utilizada sempre que um dos termos descritos é encontrado em uma sentença pelo algoritmo formado a tripla *'(concept, relation, concept)'*, como por exemplo se uma sentença identificada por *'Sentença_026_45'* possuir o texto *'AVCi - trombolise com delta t 4h'* a seguinte tripla é gerada *'(Sentença_026_45, contain, trombólise)'*

Ao conjuntos dos Termos foram adicionados alguns sub-conjuntos. O primeiro deles foi chamado de *'Valorados'* que possui a finalidade de conter todos os termos que precisam ser acompanhados de um valor numérico para qualificarem uma classificação. Para cada termos descrito nesse sub-conjunto uma relação de propriedade de dados é criada para compor a tripla *'(concept, relation, value)'*. Por exemplo, a classe *Obesidade* requer para a sua classificação que o valor do *'imc'* seja superior ou igual a 30. Assim a propriedade de dados *'imc'* também é adicionada na ontologia, assim o termo é utilizada para sinalizar ao algoritmo que é preciso identificar, entre as próximas palavras, um valor numérico e a propriedade diz a ontologia o valor numérico encontrado. Por exemplo, se uma sentença identificada por *'Sentença_026_45'* possuir o texto *'Paciente obeseo, imc = 33'* as seguintes triplas são geradas *'(Sentença_026_45, contain, obeso)'* e *'(Sentença_026_45, imc, 33)'*.

O segundo sub-conjunto foi adicionado para podermos detectar negações explicitas dos Indicadores, este foi chamado de *'Negações'* e engloba todos os termos que indicam uma negação dentro das frases. De forma similar ao executado com os termos *'Valorados'*, para cada termo sob esse sub-conjunto um propriedade de objetos foi adicionada de forma a compor uma tripla que explicitamente indica quais termos estão sendo negados. Por exemplo se uma sentença identificada por *'Sentença_026_45'* possuir o texto *'Paciente não possui diabetes'* a seguinte tripla é gerada *'(Sentença_026_45, não, diabetes)'*.

Um terceiro sub-conjunto foi adicionado ao conjunto dos *'Termos'* chamado *'TempoPassado'*. Sob ele estão todos os termos que podem indicar que um outro termo ocorreu em um tempo no passado. Para cada termo deste conjunto também foi criado uma relação de propriedade para que fique indicado quais termos estão ocorrendo no passo. Mais uma vez, tomemos por exemplo uma sentença identificada por *'Sentença_026_45'*, caso esta possua o texto *'Ex-tabagista'* ou *'história prévia: AVC'* as seguintes triplas são geradas respectivamente *'(Sentença_026_45, ex, tabagista)'* e *'(Sentença_026_45, prévia, avc)'*.

Ainda para contemplar mais um caso de discurso no tempo passo um quarto subgrupo foi adicionado, esta chamado de *'TempoPassadoRetroativo'* que contempla todos os termos que indicam o tempo passado, porém normalmente estão sintaticamente à frente na sentença e por isso eles mudam o tempo de um situação previamente citada na sentença. Para exemplificar se uma sentença identificada por *'Sentença_026_45'*, possua o texto *'tabagista no passado'* ou *'AVC prévio'* as seguintes triplas são geradas respectivamente *'(Sentença_026_45, passado, tabagista)'* e *'(Sentença_026_45, prévio, avc)'*. Na figura 5.2 ilustramos a estruturação da ontologia conforme foi descrita.

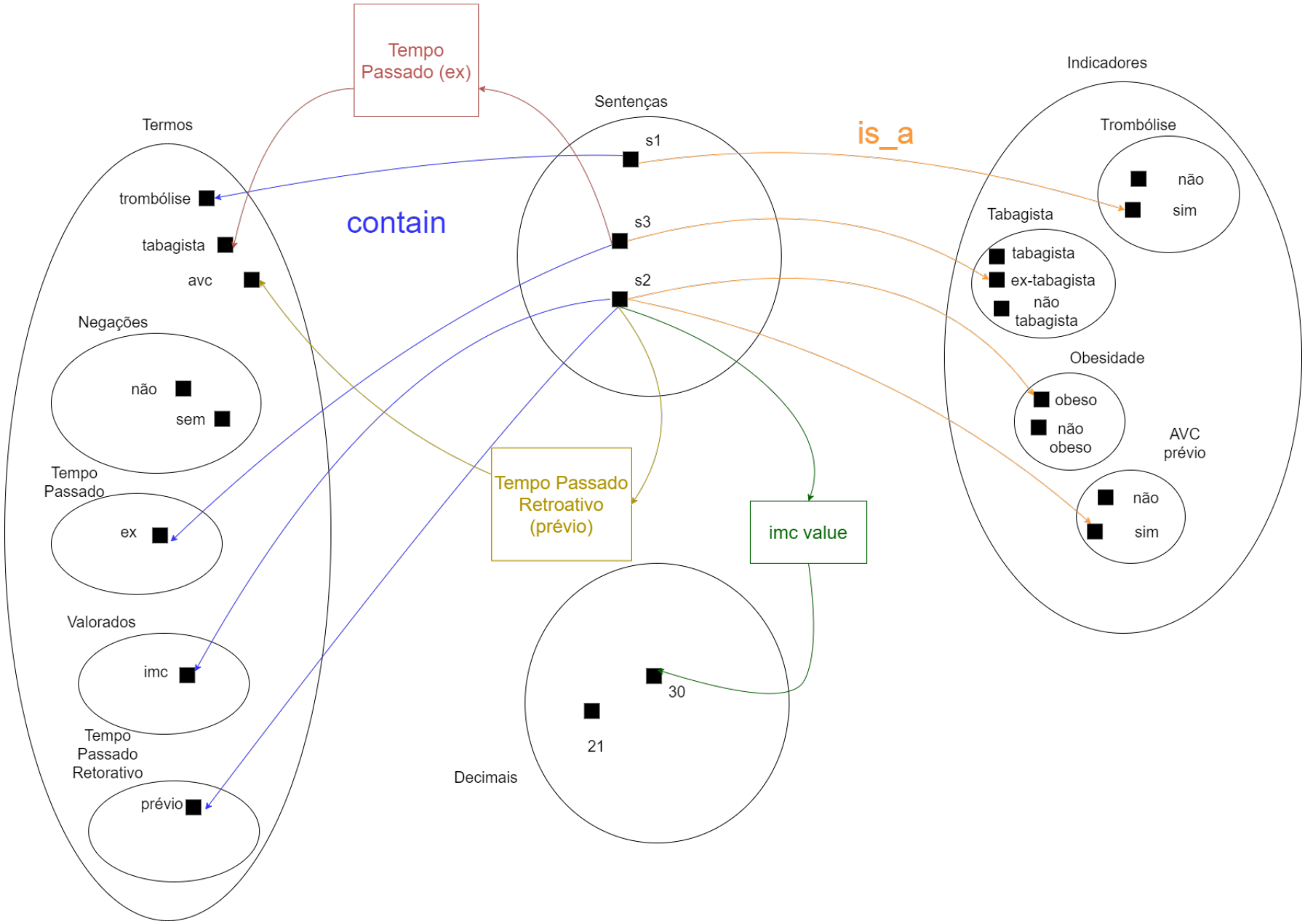

Figura 5.2 – Estrutura dos conjuntos da ontologia e sua relações

Para o processo de classificação e inferência axiomas gerais foram compostos. Para a classe *'Trombólise'*, por exemplo, o axioma *'(contain some trombólise) SubClassOf trombólise1'*, foi criado, assim sentenças que possuem o termo *'trombólise'* na frase serão classificadas como membros da classe *'Trombólise'*, sendo que o procedimento foi explicitamente realizado. Para a classe *'Obesidade'* o axioma *'imc some xsd:decimal[>= 30] SubClassOf obesidade1'*, por exemplo, foi criado, assim sentenças que possuem o termo imc seguido de um valor maior que 30 serão consideradas como um caso da afirmativo de *'Obesidade'*. As Figuras 5.3 e 5.4 mostram como as características descritas são vistas na ontologia através do software *Protègé*.

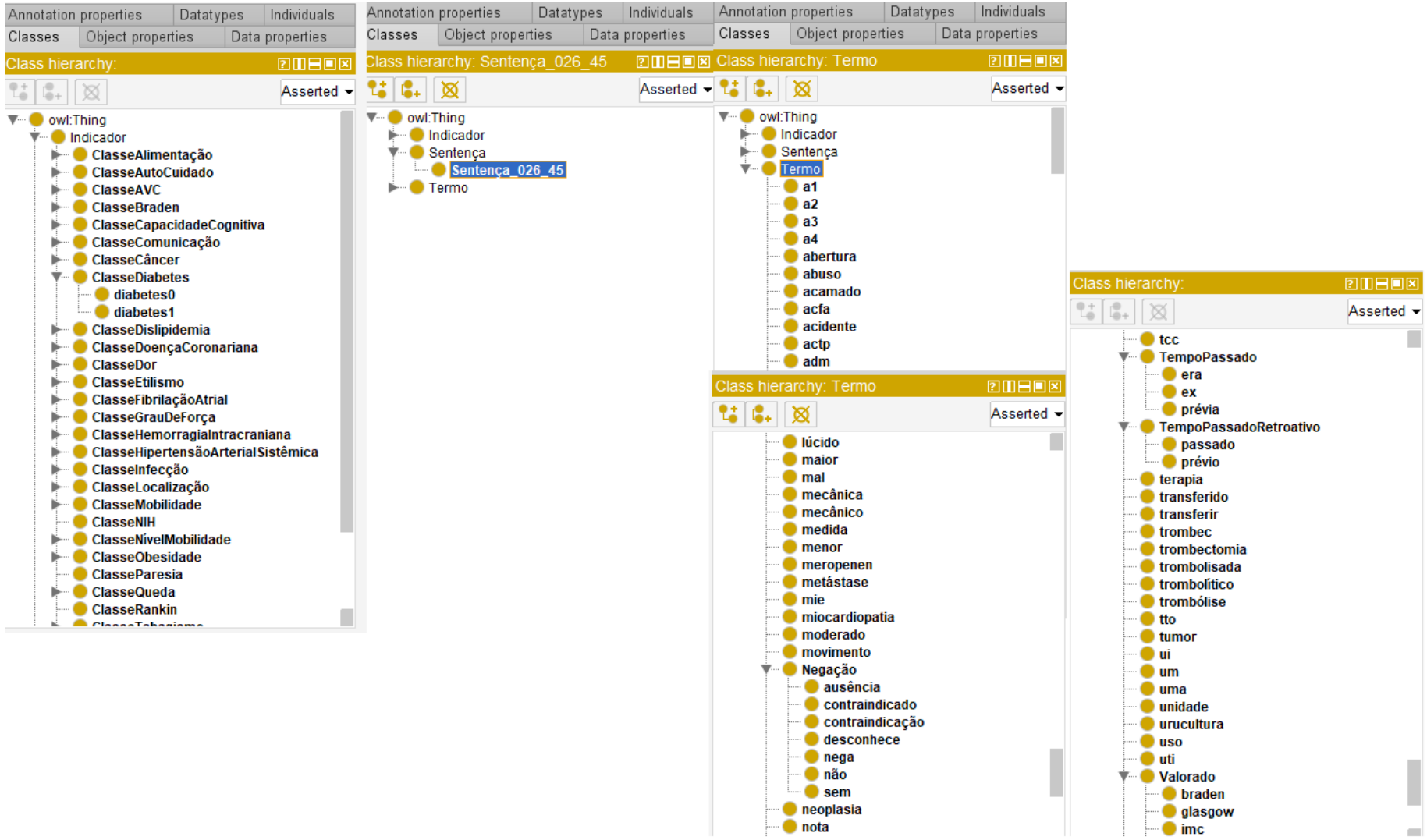

Figura 5.3 – Visão da ontologia no software *Protègé*: classes e conceitos

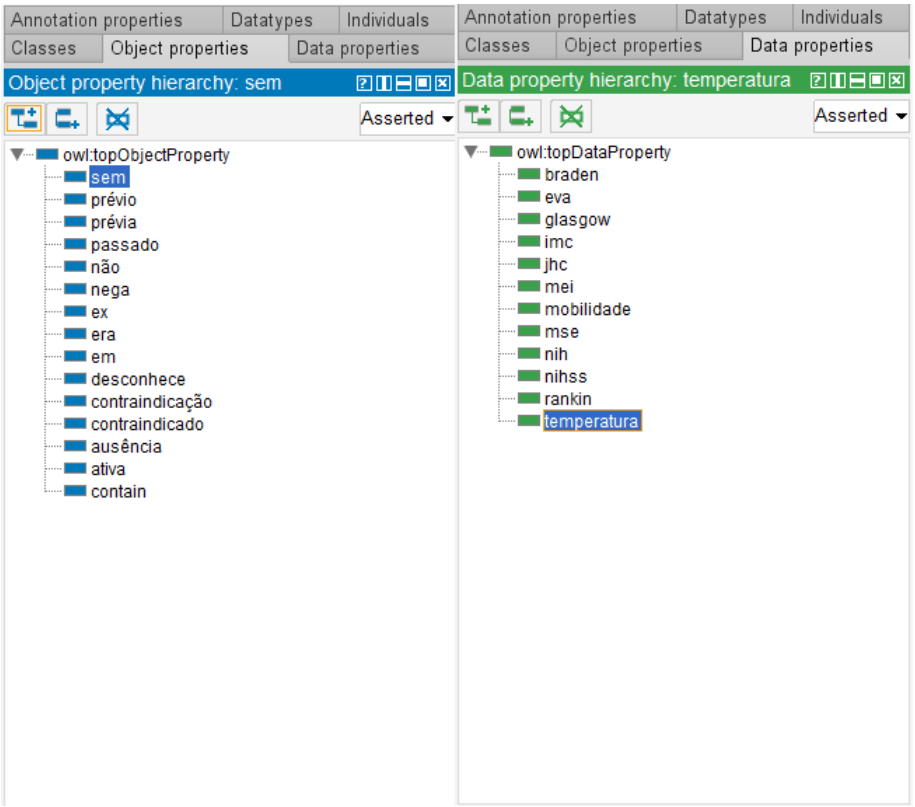

Figura 5.4 – Visão da ontologia no software *Protègé*, propriedades e relações

Na Figura 5.5 mostramos como uma sentença das evoluções fica configurada, na ontologia, como uma subclasse da classe *'Sentenças'* após ter sido processada pelo algoritmo. Na Figura 5.6 mostramos um exemplo dos axiomas e na Figura 5.7 como, após o processo de inferência ser realizado, a hierarquia de classes fica construída.

Na Figura 5.7 vemos que a *'Senteça_026_45'* foi organizada como um sub-conceito da classe *'tabagismo1'* por conter o termo *'tabagista'* conforme implica o axioma, apresentado na Figura 5.6, *'(contain some fumante) or (contain some tabagismo) or (contain some tabagista) SubClassOf tabagismo1'*.

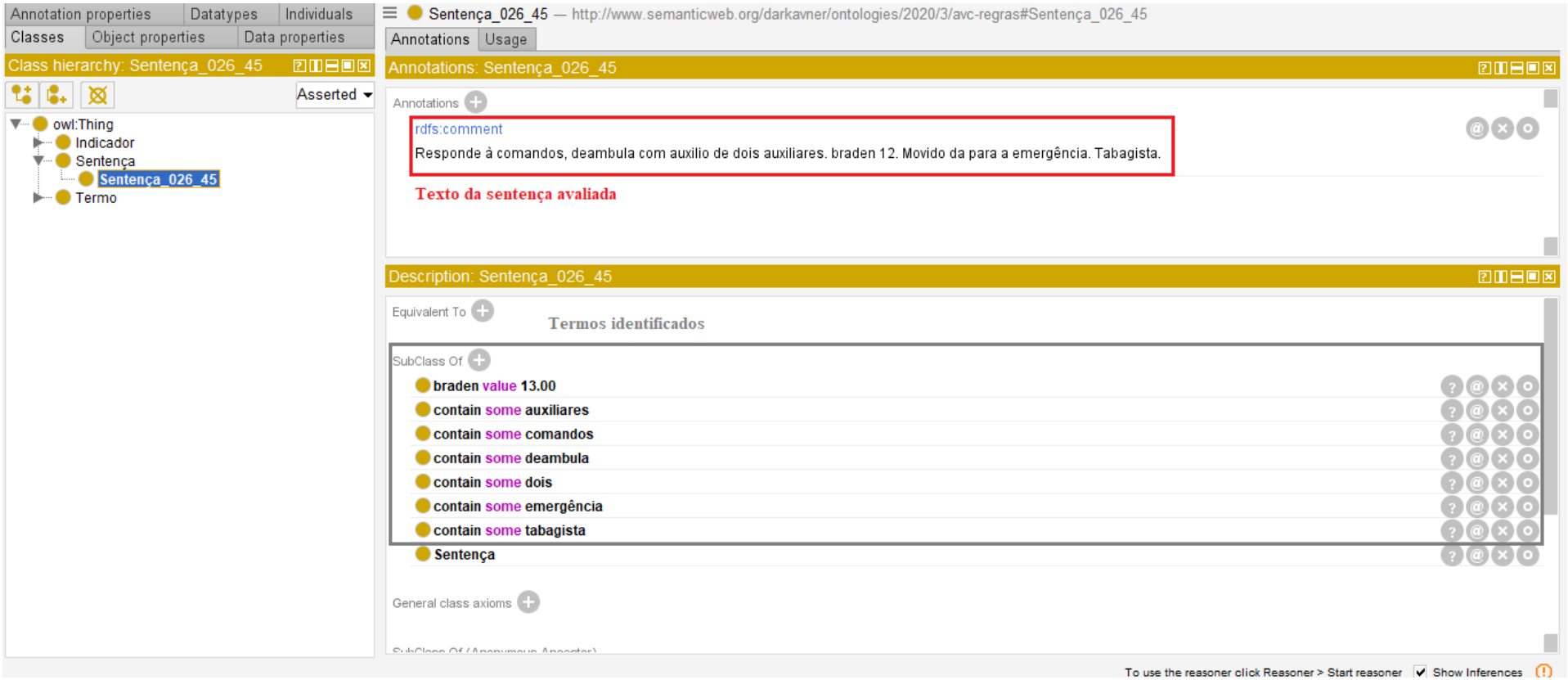

Figura 5.5 – Visão da ontologia no software *Protègé*, exemplo de sentença

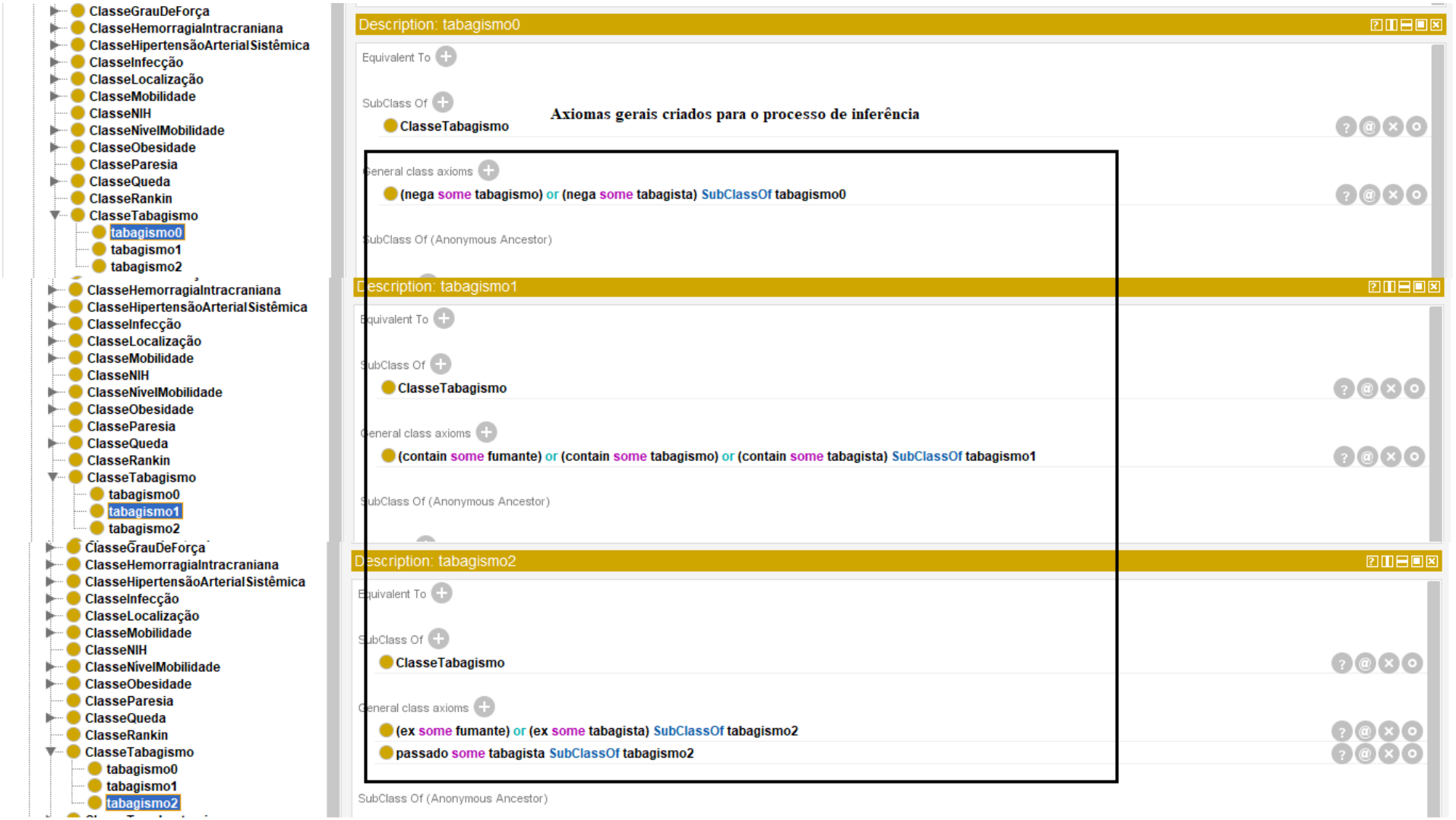

Figura 5.6 – Axiomas na ontologia

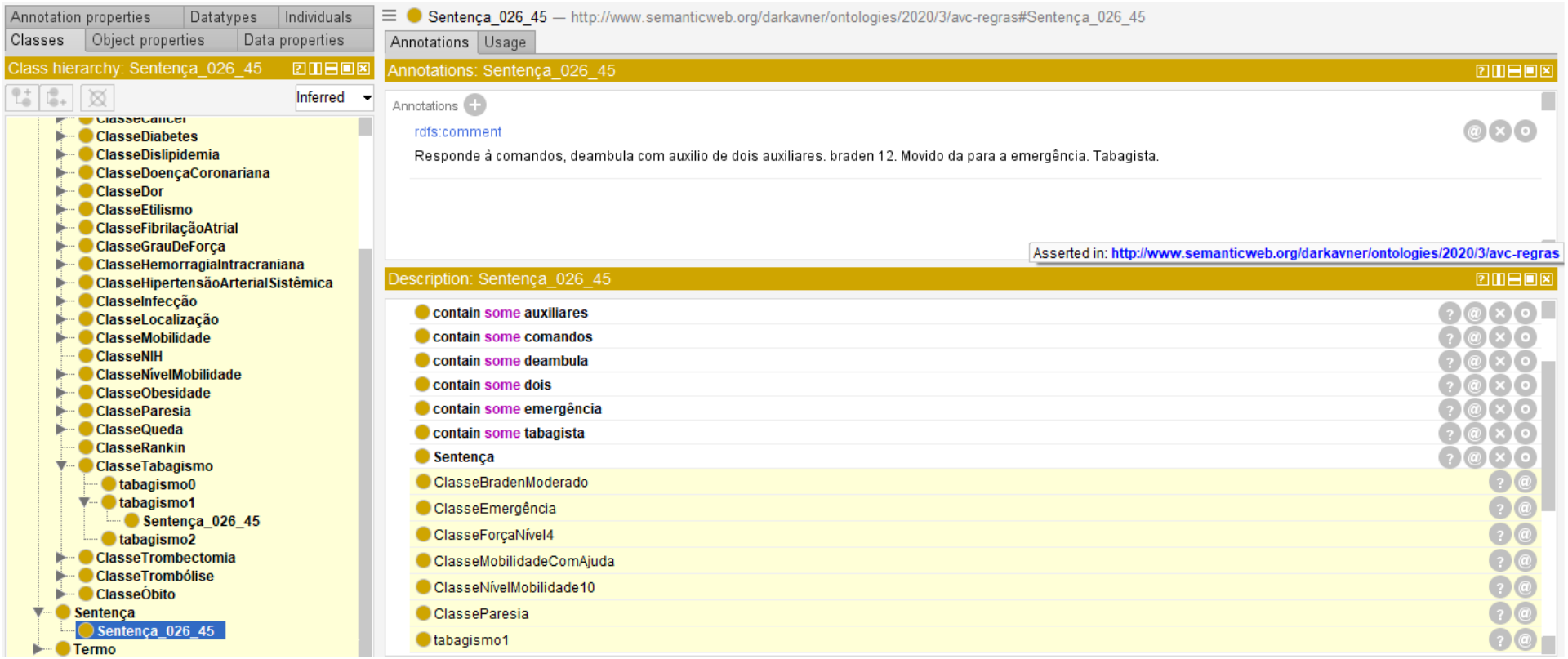

Figura 5.7 – Hierarquia de classes após processo de inferência

5.3.2 Algoritmo

Com a ontologia descrita, construímos o algoritmo que faz a detecção dos termos nas evoluções e população das sentenças na ontologia. Em uma visão geral o algoritmo tem o trabalho de detectar, nas sentenças os termos descritos na ontologia e caso acha correspondências construir as triplas de relações seguindo o padrão explicitado na secção 5.3.1.

Um ponto que deve ser explicitado é que os termos descritos na ontologia estão gramaticalmente corretos e em capitalização minuscula. Assim para que as palavras nas sentenças que aparecem, por exemplo, como *'Trombolise'*, *'EMERGENCIA'* e, *'CTI'* possam ser consideradas correspondentes, pelo algoritmo, aos termos descritos como *'trombólise'*, *'emergência'* e, *'cti'* na ontologia, um processo de tratamento e expansão de termos foi adotado. Para cobrir esses erros gramaticais e ainda expandir a quantidade de termos sem precisar elencar todas as possibilidades de variações de cada termo na ontologia foi decidido utilizar um modelo de *'word embeddings'*. O modelo adotado foi o *'WORD EMBEDDINGS PARA SAÚDE'*[1](dos Santos et al., 2018), um recurso que corresponde à três modelos pré-treinados de palavras retiradas de textos médicos utilizando 21 milhões de sentenças para criar os modelos das palavras resultando em 63 mil palavras com relação semântica e sintática. Os algoritmos de treinamento utilizadas foram Word2Vec e FastText com as estratégias CBOW e Skip-Gram. As 4 combinações de algoritmo e estratégia foram avaliadas para determinar qual par retornaria os melhores resultados de classificação. Para essa avaliação o mesmo conjunto de dados foi classificado utilizando cada par e utilizando os 10 termos mais similares por similaridade de cosseno. Ao final foi observado os resultados das métricas definidas na seção 3. Os melhores resultados, em média, foram obtidos pela combinação FastText CBOW. Os valores obtidos nessa, e nas demais, combinações são demonstrados na seção 6. Por isso, na nossa metodologia, cada palavra das sentenças são expandidas utilizando os 10 termos mais similares por similaridade de cosseno do modelo *'WORD EMBEDDINGS PARA SAÚDE'* utilizando a combinação de algoritmo e treinamento *FastText CBOW*, com essa implementação ao invés do algoritmo do modelo proposto detectar na lista de termos da ontologia apenas a palavra dada no texto, tenta-se detectar, além da palavra dada, todas as 10 palavras mais similares retornadas pelo *word embeddings*.

A implementação de um modelo word embbeddings apesar de melhorar os resultados de classificação, como veremos na seção 6, trouxe um novo desafio que foi o aumento no tempo de processamento dos textos uma vez que todas as palavras de todas as sentenças deveriam ser expandidas, algumas vezes repetidamente. Para abordar esse problema duas listas concorrentes foram implementadas. A primeira encarrega-se de ar-

[1] https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/word-embeddings-para-saude/

mazenar todos os termos que mesmo expandidos não possuem correspondências com a lista de termos descrito na ontologia, assim caso esse termo venha a ser apresentado novamente para o algoritmo este saberá que não precisa avalia-lo. A segunda lista concorrente é a lista de sinônimos. Caso um termo tenha sido expandido e algum dos termos obtidos possuem correspondências com a lista de termos da ontologia essa lista irá armazenar o termo original e os termos expandidos correspondentes. Assim na próxima vez que esse termo for apresentado ao algoritmo este não precisará expandi-lo novamente. Dessa forma conseguimos melhorar o tempo de execução do algoritmo como veremos nas comparações que serão apresentadas na seção 6.

Para compreendermos mais o funcionamento do algoritmo a seguinte lista enumerada descreve os pontos principais da sequência lógica construída para o algoritmo:

1. O algoritmo recupera da ontologia a lista de termos, a lista de negações, a lista de valorados, a lista de de termos que indicam tempo passado e, a lista de termos que indicam a tempo passado de um um termo retroativo;
2. As sentenças são lidas das evoluções pelo algoritmo;
3. As sentenças, uma por vez, são tratadas para remover caracteres especiais e separadas em palavras;
4. As palavras de cada sentença, uma por vez, são tratadas para ficarem com todas as letras em minúsculo.
5. Verifica-se a necessidade do termo ser um valor numérico:
   (a) Caso negativo:
      i. Verifica-se a presença do termo na lista de termos sem correspondentes:
         A. Caso negativo:
            - Verifica-se a presença no termo na lista de sinônimos, ou seja, os termos já avaliados e expandidos:
              – Caso negativo:
                * Utiliza-se o modelo de word embeddings para obter os 10 termos mais similares
                * Os termos e os similares que possuem correspondência com as listas da ontologia são salvos em uma lista.
              – Caso positivo:
                * Recupera-se os termos já expandidos que estão salvos na lista de sinônimos
                * Salva-se o termos e seus similares na lista de termos detectados para a sentença sendo avaliada

B. Caso positivo:
   - O termo atual não é avaliado e o algoritmo passa a avaliar o próximo termo da sentença

(b) Caso positivo:

i. Verifica-se se o termo é um valor numérico:

A. Caso negativo:
   - Repete-se o teste para os próximos dois termos
   - Após a terceira tentativa desativa-se a busca por um valor numérico

B. Caso positivo:
   - Adiciona-se a lista de termos e relações o termo valorado encontrado e o valor referente a esse termo
   - Desativa-se a busca por um valor numérico

6. Verifica-se a presença do termo e seus similares na lista de negações da ontologia:

(a) Caso positivo:

- Sinaliza que o termo atual possui correspondências
- Adiciona-se a lista de sinônimos os termos comuns entres as listas
- Sinaliza que a relação com o próximo termo é uma negação e o tipo da relação é o termo atual

7. Verifica-se a presença do termo e seus similares na lista de termos de tempo passado da ontologia:

(a) Caso positivo:

- Sinaliza que o termo atual possui correspondências
- Adiciona-se a lista de sinônimos os termos comuns entres as listas
- Sinaliza que a relação com o próximo termo é em um tempo passado e o tipo da relação é o termo atual

8. Verifica-se a presença do termo e seus similares na lista de termos valorados da ontologia:

(a) Caso positivo:

- Sinaliza que o termo atual possui correspondências
- Adiciona-se a lista de sinônimos os termos comuns entres as listas

- Sinaliza que a relação com o próximo termo é valorado, o tipo da relação é o termo atual e sinaliza a necessidade de encontrar um valor numérico

9. Verifica-se a presença do termo e seus similares na lista de termos da ontologia:

   (a) Caso positivo:

   - Sinaliza que o termo atual possui correspondências
   - Adiciona-se a lista de sinônimos os termos comuns entres as listas
   - Caso a relação não esteja definida como negação ou tempo passado, defini o uso da relação 'contain'
   - Adiciona à lista de termos e relações a relação corrente definida e como valor o termo atual

10. Verifica-se a sinalização de correspondência do termo atual:

    (a) Caso negativo:

    - Adiciona-se o termo a lista de sem correspondentes

11. Após avaliar todas as palavras da sentença adiciona a lista de termos e relações na ontologia

Neste fluxo do algoritmo o processo de inferência e classificação pode ser executado em dois momentos, o primeiro é após a análise de cada sentença, logo em seguida ao item 11 da lista acima, ou em um segundo momento após todas as sentenças terem sido analisadas pelo algoritmo. Neste trabalho o segundo momento foi escolhido para avaliar as métricas do modelo proposto. Neste caso, após toda a hierarquia de classes ser gerada na ontologia com todas as sentenças avaliadas, os conceitos aos quais cada sentença foram classificados são recuperados e comparados com as anotações feitas pelas especialistas.

5.3.3 Exemplo do processamento

Nessa secção iremos amostrar algumas sentenças do conjunto de dados e demonstrar o passo a passo do processo de classificação, detalhando a entrada e a saída obtida em cada etapa.

A primeira sentença que iremos amostrar é *'# EX-TABAGISTA (IT 60 MAÇOS/ANO) # DPOC?'*. No anexo D mostramos todas as saídas do processamento dessa sentença. No que segue iremos focar nossa atenção apenas nos eventos mais relevantes desse processamento.

Nesta sentença o algoritmo ao detectar o termo *'ex'*, verifica na lista recebida da ontologia que este é um termo que indica uma condição passada. Assim ele busca por um próximo termo conhecido e válido e ao encontrar o termo *'tabagista'* o algoritmo então conclui que nesta sentença deve-se construir a relação *ex some tabagista*. Assim, pela regra de inferência *ex some tabagista* essa sentença é classificada como *tabagista2* que é o alias na ontologia para o indicador *Ex-tabagista*

A segunda sentença que vamos usar como exemplo é a *'D # AVC isquêmico previo - sem sequelas aparentes -mRankin previo: 3 # Demência de Alzheimer - Tem vida de relação, corversa, caminha, alimenta-se.'*. No anexo E veremos toda a saída do processamento. Novamente no que segue iremos avaliar apenas os eventos mais relevantes para a compreensão do processamento.

Nesse exemplo o algoritmo detecta o termo *'rankin'* e verifica que na lista de termos da ontologia este é um termo valorado, ou seja, ele acusa um indicador mas para isso precisa estar acompanhado de um valor numérico. Assim o algoritmo busca nos próximos termos por tal valor e ao encontrar o termo *'3'* este conclui que a relação *'rankin some 3'* deve ser vinculada a esta sentença e dessa forma a ontologia consegue armazenar classificar essa sentença como um indicador de Rankin do paciente assim como o valor atribuído pelo descritor da evolução. Outro ponto que destacamos nessa sentença é a atuação da lista recorrente de termos sem correspondências que após avaliar o termo *de* pela primeira e não encontrar um correspondente com as listas na ontologia passa a impedir que o termo seja expandido pelo modelo *wordembeddings* novamente. Um terceiro caso a ser ressaltado é com relação a passagem da sentença *AVC isquêmico previo*, que deveria ter indicado a ocorrência de um AVC em um momento no passado devido a presença do termo *'previo'* estando posicionando posteriormente ao termo *'AVC'*. Porém, o termo registrado na ontologia que indicaria o tempo passado retroativo é *prévio*, e a expansão do modelo *wordembeddings* não conseguiu corrigir a palavra da sentença para a grafia correta impedindo assim a classificação adequada. Esse é um dos principais geradores de erros do modelo construído e discutiremos melhor e com mais exemplos essas situações na secção 6.

5.4 Ambiente de desenvolvimento e processamento do algoritmo

Para a construção do algoritmo descrito foi utilizado a linguagem de programação *'Python'*[2]. Para a recuperação das informações da ontologia e o processo de inferência em tempo de execução de código foi utilizado a biblioteca *Owlready* (Lamy, 2017)[3]. A análise

[2]https://www.python.org
[3]https://owlready2.readthedocs.io/en/latest/index.html

e manipulação dos dados foi feita utilizando as bibliotecas *Numpy*[4] e *Pandas*[5]. O ambiente de desenvolvimento adotado foi o *Google Colab*[6].

[4]https://numpy.org
[5]https://pandas.pydata.org
[6]https://colab.research.google.com

6. RESULTADOS

Em todos os testes as 46.547 sentenças anotadas foram utilizadas como entrada no algoritmo. Na saída além das métricas *'precision'*, *'recall'* ,*'accuaracy'*, *'mccscore'* e, *'f1-score'* também era gerado uma tabela, para cada indicador, com todos os falsos positivos e falsos negativos que foram identificados e o tempo de execução da classificação também foi registrado.

6.1 Resultados das combinações do modelo Word embeddings

Como mencionado no capítulo 5, inicialmente foi avaliado o desempenho das 4 combinações possíveis de treinamento e algoritmo do modelo *word embeddings*. As tabelas 6.2, 6.1, 6.3 e, 6.4 mostram os desempenho de cada combinação. Com esses resultados montamos a tabela 6.5 onde podemos comparar a média obtida por cada combinação e assim, juntamente com uma análise dos resultados das tabelas anteriores, decidir pela utilização da combinação *FastText + CBOW*.

Tabela 6.1 – Resultado da combinação Word2Vec + SKIP

Métricas com word embedding estratégia Word2Vec + SKIP				
Tempo de Classificação	**1532.72 segundos**			
Classes	f1 (%)	mcc (%)	precision (%)	recall (%)
Trombólise	**64,87**	**67,49**	**50,2**	**91,67**
Óbito	0,71	0,16	4,43	0,39
Localizacao	**62,46**	**63,95**	**84,84**	**49,42**
Doenca Coronariana	18,91	29,9	10,56	90,42
Fibrilacao Atrial	**80,37**	**80,82**	**71,78**	**91,29**
Diabetes	51,12	54,84	37,04	82,48
Avc Prévio	**3,23**	**2,79**	**3,57**	**2,94**
Hipertensão	76,7	77,37	66,2	91,15
Obesidade	**42,74**	**48,89**	**28,74**	**83,33**
Dislipidemia	31,93	41,53	19,4	90,21
Câncer	**17,5**	**18,29**	**12,96**	**26,92**
Tabagismo	35,82	44,35	22,41	89,29
Etilismo	**24,71**	**35,72**	**14,32**	**89,83**
Trombectomia	0	0	0	0
Hemorragia Intracraniana	**4,39**	**4,78**	**8,33**	**2,98**
Queda	14,02	25,59	7,63	86,36
Braden	**91,13**	**91,08**	**91,87**	**90,4**
Risco de queda	47,22	47,43	55,56	41,06
Indicativo de Infecção	**27,26**	**26,45**	**24,43**	**30,82**
Dor	12,75	13,17	9,17	20,9
Nível de Mobilidade	**51,87**	**55,26**	**81,66**	**38**
Mobilidade	20,11	27,77	66,67	11,84
Grau de Força	**21,52**	**20,9**	**17,9**	**26,97**
Paresia	29,29	33,68	18,98	64,12
Comunicação	**33,43**	**33,91**	**24,43**	**52,89**
Capacidade Cognitiva	13,4	18,11	7,77	48,48
Rankin	**17,73**	**25,62**	**64,29**	**10,29**
Auto Cuidado	0	-0,71	0	0
Alimentacao	**57,63**	**59,23**	**44,17**	**82,9**
NIH	59,07	60,13	48,84	74,73
Média	**33,73**	**36,95**	**33,27**	**52,07**

Tabela 6.2 – Resultado da combinação Word2Vec + CBOW

Métricas com word embedding estratégia Word2Vec + CBOW				
Tempo de Classificação	**1635.62 segundos**			
Classes	f1 (%)	mcc (%)	precision (%)	recall (%)
Trombólise	**67,09**	**69,06**	**53,44**	**90,1**
Óbito	0,71	0,88	8,05	0,37
Localizacao	**52,34**	**50,94**	**55,12**	**49,83**
Doenca Coronariana	70,66	71,85	59,06	87,94
Fibrilacao Atrial	**78,4**	**79,17**	**68,12**	**92,33**
Diabetes	72,85	73,6	86,31	63,03
Avc Prévio	**16,12**	**23,98**	**62,86**	**9,24**
Hipertensão	89,28	89,18	86,86	91,84
Obesidade	**63,29**	**65,18**	**51,02**	**83,33**
Dislipidemia	65,3	67,59	51,63	88,81
Câncer	**16,7**	**19,82**	**10,68**	**38,28**
Tabagismo	92,49	92,51	95,8	89,41
Etilismo	**75,86**	**76,39**	**67,54**	**86,52**
Trombectomia	0	0	0	0
Hemorragia Intracraniana	**1,96**	**2,54**	**6,25**	**1,16**
Queda	32,76	41,74	20,21	86,36
Braden	**91,13**	**91,08**	**91,87**	**90,4**
Risco de queda	68,31	69,31	57,56	83,99
Indicativo de Infecção	**28,95**	**28,1**	**27,06**	**31,13**
Dor	13,96	13,96	10,84	19,6
Nível de Mobilidade	**50,28**	**50,71**	**62,96**	**41,86**
Mobilidade	18,97	20,87	35,35	12,96
Grau de Força	**21,05**	**20,62**	**16,88**	**27,95**
Paresia	47,32	51,89	32,92	84,12
Comunicação	**43,34**	**42,93**	**35,31**	**56,1**
Capacidade Cognitiva	32,76	35,73	22,4	61
Rankin	**87,61**	**87,69**	**92,36**	**83,33**
Auto Cuidado	0	-0,41	0	0
Alimentacao	**67,79**	**67,9**	**81,56**	**58,01**
NIH	76,11	75,96	75,74	76,49
Média	**48,11**	**49,36**	**47,53**	**56,18**

Tabela 6.3 – Resultado da combinação FastText + CBOW

Métricas com word embedding estratégia FastText + CBOW				
Tempo de Classificação	**7504.59 segundos**			
Classes	f1 (%)	mcc (%)	precision (%)	recall (%)
Trombólise	**87,75**	**87,66**	**89,35**	**86,21**
Óbito	0,2	-0,24	2,86	0,11
Localizacao	**89,04**	**88,7**	**90,6**	**87,53**
Doenca Coronariana	78,5	78,37	77,06	80
Fibrilacao Atrial	**83,78**	**84,06**	**76,44**	**92,68**
Diabetes	89,89	89,86	92,63	87,31
Avc Prévio	**23,02**	**28,35**	**56,14**	**14,48**
Hipertensão	91,07	90,95	90,99	91,15
Obesidade	**81,36**	**81,36**	**82,76**	**80**
Dislipidemia	94,07	94,22	100	88,81
Câncer	**17,23**	**17,06**	**15,14**	**20**
Tabagismo	93,33	93,37	97,06	89,88
Etilismo	**82,61**	**82,59**	**80,85**	**84,44**
Trombectomia	64,97	68,82	97,46	48,73
Hemorragia Intracraniana	**5,83**	**8,01**	**19,35**	**3,43**
Queda	36,89	44,97	23,46	86,36
Braden	**94,78**	**94,78**	**97,12**	**92,55**
Risco de queda	61,77	63,23	80,16	50,25
Indicativo de Infecção	**25,58**	**30,2**	**56,9**	**16,5**
Dor	13,72	13,76	10,53	19,69
Nível de Mobilidade	**51,8**	**54,22**	**76,42**	**39,18**
Mobilidade	18,87	23,84	50,76	11,59
Grau de Força	**27,48**	**27,63**	**36,31**	**22,11**
Paresia	64,33	64,43	72,77	57,65
Comunicação	**61,5**	**62,72**	**80,97**	**49,58**
Capacidade Cognitiva	33,7	33,15	28,59	41,03
Rankin	**87,61**	**87,69**	**92,36**	**83,33**
Auto Cuidado	0	-0,74	0	0
Alimentacao	**61,83**	**64,54**	**91,17**	**46,78**
NIH	74,74	74,59	73,72	75,79
Média	**56,58**	**57,74**	**64,66**	**54,91**

Tabela 6.4 – Resultado da combinação FastText + SKIP

Métricas com word embedding estratégia FastText + SKIP				
Tempo de Classificação	**7751.66 segundos**			
Classes	f1 (%)	mcc (%)	precision (%)	recall (%)
Trombólise	**79,53**	**79,82**	**88,92**	**71,93**
Óbito	0,21	-0,18	3,08	0,11
Localizacao	**74,11**	**74,7**	**90,41**	**62,79**
Doenca Coronariana	21,66	30,29	12,59	77,46
Fibrilacao Atrial	**73,8**	**73,66**	**75,55**	**72,13**
Diabetes	58,06	60,74	44,29	84,29
Avc Prévio	**18,73**	**19,62**	**28,18**	**14,03**
Hipertensão	84,19	84,06	80,83	87,85
Obesidade	**80**	**79,99**	**80**	**80**
Dislipidemia	93,28	93,48	100	87,41
Câncer	**17,53**	**17,27**	**16,16**	**19,16**
Tabagismo	86,25	86,23	83,15	89,58
Etilismo	**35,23**	**42,66**	**22,46**	**81,58**
Trombectomia	36,3	46,63	98	22,27
Hemorragia Intracraniana	**3,69**	**4,19**	**8**	**2,4**
Queda	22,64	32,72	13,14	81,82
Braden	**95,41**	**95,4**	**97,15**	**93,73**
Risco de queda	19,03	25,84	60,26	11,3
Indicativo de Infecção	**21,49**	**26,9**	**56,16**	**13,29**
Dor	10,36	9,94	8,75	12,69
Nível de Mobilidade	**51,04**	**52,93**	**72,59**	**39,36**
Mobilidade	19,77	25,01	53,12	12,14
Grau de Força	**27,15**	**29,03**	**45,1**	**19,43**
Paresia	54,61	55,63	70,15	44,71
Comunicação	**51,91**	**54,52**	**78,73**	**38,72**
Capacidade Cognitiva	31,59	31,96	24,25	45,29
Rankin	**87,61**	**87,69**	**92,36**	**83,33**
Auto Cuidado	0	-0,73	0	0
Alimentacao	**68,03**	**68,15**	**81,77**	**58,25**
NIH	78,97	78,88	81,79	76,33
Média	**46,74**	**48,90**	**55,56**	**49,45**

Tabela 6.5 – Comparação dos tempos e médias das estratégias

Comparação média das estratégias				
	Combinações			
	Word2Vec + CBOW	Word2Vec + SKIP	**FastText + CBOW**	FastText + SKIP
Tempo (segundos)	1635,62	1532,72	7504,59	7751,66
Média F1 (%)	48,11	33,73	56,58	46,74
Média MCC (%)	49,36	36,95	57,74	48,90
Média Precision (%)	47,53	33,27	64,66	55,56
Média Recall (%)	56,18	52,07	54,91	49,45

6.2 Resultados do modelo desenvolvido

Com a configuração do *word embeddings* definida, foram feitas três avaliações do modelo de classificação visando avaliar os benefícios e ganhos trazidos pelas ferramentas e técnicas implementadas ao modelo. A primeira execução foi realizada sem a implementação do word embedding e sem as listas concorrentes de sinônimos e termos incomparáveis. Na segunda avaliação foi adicionado o word embeddings. Por fim na terceira avaliação as listas foram implementadas.

Nas tabelas 6.6, 6.7 e, 6.8, apresentamos os resultados de *'f1score'*, *'mccscore'*, *'precision'* e, *'recall'* obtidos e, o tempo de execução obtidos em cada execução do modelo. Na tabela 6.9 apresentamos a comparações das médias entre as três execuções do modelo final.

Tabela 6.6 – Resultado das classificações do modelo sem o uso de modelos word embedding e sem o uso de listas concorrentes

Métricas sem word embedding e sem listas concorrentes				
Tempo de Classificação	**72,39 segundos**			
Classes	f1 (%)	mcc (%)	precision(%)	recall(%)
Trombólise	**63,27**	**65,28**	**85,43**	**50,23**
Óbito	0,1	-0,51	1,45	0,05
Localizacao	**75,72**	**75,82**	**87,94**	**66,49**
Doenca Coronariana	84,28	84,3	89,05	80
Fibrilacao Atrial	**89,41**	**89,4**	**86,36**	**92,68**
Diabetes	89,89	89,86	92,63	87,31
Avc Prévio	**23,1**	**28,61**	**57,14**	**14,48**
Hipertensão	91,24	91,13	91,16	91,32
Obesidade	**81,36**	**81,36**	**82,76**	**80**
Dislipidemia	94,07	94,22	100	88,81
Câncer	**17,89**	**17,73**	**15,74**	**20,73**
Tabagismo	93,15	93,2	97,47	89,19
Etilismo	**82,61**	**82,59**	**80,85**	**84,44**
Trombectomia	41,08	50,44	98,39	25,96
Hemorragia Intracraniana	**1**	**1,34**	**3,85**	**0,57**
Queda	36,89	44,97	23,46	86,36
Braden	**94,14**	**94,18**	**97,9**	**90,66**
Risco de queda	60,06	61,36	77,29	49,11
Indicativo de Infecção	**25,55**	**30,24**	**57,23**	**16,45**
Dor	13,37	13,46	10,12	19,69
Nível de Mobilidade	**52,55**	**56,15**	**83,53**	**38,33**
Mobilidade	19,42	27,03	65,69	11,39
Grau de Força	**27,7**	**27,85**	**36,59**	**22,28**
Paresia	64,14	64,41	74,23	56,47
Comunicação	**56,56**	**59,13**	**83,57**	**42,74**
Capacidade Cognitiva	33,68	33,01	29,17	39,84
Rankin	**79,87**	**81,18**	**97,69**	**67,55**
Auto Cuidado	0	-0,73	0	0
Alimentacao	**61,42**	**64,13**	**90,75**	**46,42**
NIH	80,81	80,79	85	77,02
Média	**54,48**	**56,06**	**66,08**	**51,22**

Tabela 6.7 – Resultado das classificações do modelo com o uso de modelos word embedding e sem o uso de listas concorrentes

Métricas com word embedding e sem listas concorrentes				
Tempo de Classificação	**7504,59 segundos**			
Classes	f1 (%)	mcc (%)	precision (%)	recall (%)
Trombólise	**87,75**	**87,66**	**89,35**	**86,21**
Óbito	0,2	-0,24	2,86	0,11
Localizacao	**89,04**	**88,7**	**90,6**	**87,53**
Doenca Coronariana	78,5	78,37	77,06	80
Fibrilacao Atrial	**83,78**	**84,06**	**76,44**	**92,68**
Diabetes	89,89	89,86	92,63	87,31
Avc Prévio	**23,02**	**28,35**	**56,14**	**14,48**
Hipertensão	91,07	90,95	90,99	91,15
Obesidade	**81,36**	**81,36**	**82,76**	**80**
Dislipidemia	94,07	94,22	100	88,81
Câncer	**17,23**	**17,06**	**15,14**	**20**
Tabagismo	93,33	93,37	97,06	89,88
Etilismo	**82,61**	**82,59**	**80,85**	**84,44**
Trombectomia	64,97	68,82	97,46	48,73
Hemorragia Intracraniana	**5,83**	**8,01**	**19,35**	**3,43**
Queda	36,89	44,97	23,46	86,36
Braden	**94,78**	**94,78**	**97,12**	**92,55**
Risco de queda	61,77	63,23	80,16	50,25
Indicativo de Infecção	**25,58**	**30,2**	**56,9**	**16,5**
Dor	13,72	13,76	10,53	19,69
Nível de Mobilidade	**51,8**	**54,22**	**76,42**	**39,18**
Mobilidade	19,15	24,44	52,71	11,7
Grau de Força	**27,48**	**27,63**	**36,31**	**22,11**
Paresia	64,33	64,43	72,77	57,65
Comunicação	**61,54**	**62,77**	**81,09**	**49,58**
Capacidade Cognitiva	33,8	33,28	28,63	41,25
Rankin	**87,61**	**87,69**	**92,36**	**83,33**
Auto Cuidado	0	-0,74	0	0
Alimentacao	**61,83**	**64,54**	**91,17**	**46,78**
NIH	74,74	74,59	73,72	75,79
Média	**56,59**	**57,76**	**64,73**	**54,92**

Tabela 6.8 – Resultado das classificações do modelo com o uso de modelos word embedding e o uso de listas concorrentes

Métricas com word embedding e com listas concorrentes				
Tempo de Classificação	**532,43 segundos**			
Classes	f1 (%)	mcc (%)	precision	recall
Trombólise	**87,75**	**87,66**	**88,94**	**86,45**
Óbito	0,2	-0,24	2,86	0,11
Localizacao	**89,01**	**88,67**	**92,29**	**87,74**
Doenca Coronariana	78,5	78,37	77,06	80
Fibrilacao Atrial	**83,65**	**83,84**	**76,44**	**92,68**
Diabetes	89,89	89,86	92,63	87,31
Avc Prévio	**26,76**	**32,05**	**56,14**	**14,48**
Hipertensão	91,26	91,15	90,99	91,15
Obesidade	**81,36**	**81,36**	**82,76**	**80**
Dislipidemia	94,07	94,22	100	88,81
Câncer	**17,23**	**17,06**	**15,14**	**20**
Tabagismo	92,24	92,24	94,96	89,68
Etilismo	**82,61**	**82,59**	**80,85**	**84,44**
Trombectomia	64,97	68,82	97,46	48,73
Hemorragia Intracraniana	**5,83**	**8,01**	**19,35**	**3,43**
Queda	36,89	44,97	23,46	86,36
Braden	**94,78**	**94,78**	**97,12**	**92,55**
Risco de queda	61,77	63,23	80,16	50,25
Indicativo de Infecção	**25,58**	**30,2**	**56,9**	**16,5**
Dor	13,37	13,46	10,12	19,69
Nível de Mobilidade	**50,75**	**53,01**	**75,91**	**39,01**
Mobilidade	19,64	24,87	51,15	11,59
Grau de Força	**27,48**	**27,63**	**37,15**	**22,5**
Paresia	64,33	64,43	72,77	57,65
Comunicação	**61,54**	**62,77**	**81,06**	**49,49**
Capacidade Cognitiva	33,8	33,28	28,59	41,03
Rankin	**87,61**	**87,69**	**92,36**	**83,33**
Auto Cuidado	0	-0,74	0	0
Alimentacao	**61,83**	**64,54**	**90,8**	**46,68**
NIH	79,33	79,22	81,23	77,52
Média	**56,80**	**57,97**	**64,89**	**54,97**

Tabela 6.9 – Comparação das médias e tempos das três execuções

Comparação de médias do modelo			
	Execuções		
	S/ Word Embeddings e S/ Listas	C/ Word Embeddings e S/ Listas	**C/ Word Embeddings e C/ Listas**
Tempo (segundos)	72,39	7504,59	532,43
Média F1 (%)	54,48	56,59	56,80
Média MCC (%)	56,06	57,76	57,97
Média Precision (%)	66,08	64,73	64,89
Média Recall (%)	51,22	54,92	54,97

Das tabelas podemos observar que as métricas, em média entre as classes, mostram pouca variação entre o uso de word embeddings ou não. Porém ao avaliarmos algumas classes especificas vemos um aumento considerável nas métricas de classificação como no caso da classe *'Trombólise'* que passa de um f1 de 63,27 e mcc de 65,28 para 87,75 e 87,66 respectivamente. Outro casos notáveis de melhora são *'Trombectomia'* e *'Trombectomia'*. Podemos observar também a melhora de performance com a implementação das listas concorrentes, onde obtivemos um ganho de tempo de aproximadamente *'6972,16 segundos'* entre as execuções com os modelos word embeddings.

6.3 Análise dos erros

Na tabela 6.10 listamos alguns exemplos de sentenças que foram erroneamente classificadas pelo modelo. Através desses exemplos buscamos compreender melhor os pontos falhos do modelo e como poderíamos aprimora-lo.

Tabela 6.10 – Tabela de exemplos dos erros mais comuns na classificação

Indicador	#	Matriz Confusão	Anotado	Classificado	Sentença
Óbito	**1**	**Falso Negativo**	**0.0**	**-1**	**Condição Ventilatória: ar ambiente, eupneico.**
	2	Falso Negativo	0.0	-1	Ambiente - Na poltrona, estável, colaborativo, sem queixas, acompanhado da filha.
AVC Prévio	**3**	**Falso Negativo**	**1.0**	**-1**	**D # AVC isquêmico previo - sem sequelas aparentes -mRankin previo: 3 # Demência de Alzheimer - Tem vida de relação, corversa, caminha, alimenta-se.**
	4	Falso Negativo	1.0	-1	Paciente com história de AVCisquêmico em out/18 e dezembro de 2018.
	5	**Falso Negativo**	**1.0**	**-1**	**#Atual: AVCI #Prévio: AVC / DM2 / HAS / DPOC.**
	6	Falso Positivo	0.0	1.0	# Nega AVC ou Infarto prévio
Câncer	**7**	**Falso Positivo**	**0.0**	**1.0**	**# CA bexiga em 2012 #**
	8	Falso Positivo	0.0	1.0	# 2016 - Ca de células claras rim direito - Nefrectomia parcial Dir # Descolamento de retina há 1 ano - Olho Esq # Adenocarcinoma com células em anel de sinete do esôfago distal - QT até junho/19 com progressão da doença + prótese esofágica # medicações em uso:
Hemorragia Intracrâniana	**9**	**Falso Negativo**	**1.0**	**-1**	**> transformação hemorrágica..**
	10	Falso Positivo	0.0	1.0	Não há evidência de lesão expansiva, hemorragia intracraniana ou desvios da linha média .
Dor	**11**	**Falso Positivo**	**0.0**	**1.0**	**Sem queixas de dor ou desconforto.**
Localização	12	Falso Negativo	1.0	-1	Emergencia/Enfermagem
	13	**Falso Negativo**	**1.0**	**-1**	**> EMG HMV->**

Analisando a tabela de exemplos de erros, observou-se que mesmo com a implementação do modelo de word embeddings alguns termos não eram detectados devido a sua acentuação ou formas de abreviação utilizadas pelo funcionário. Nos exemplos *3, 12 e, 13* da tabela 6.10, vemos alguns desses casos onde para o modelo *word embedding* foi observado que as palavras *'emergencia'* e *'emergência'* possuíam uma similaridade menor que 2, assim como, *'previo'* e *'prévio'*, sendo outras palavras melhores elencadas como similares e dessa forma quando a evolução apresentava os termos sem a correta acentuação, o seu correto correspondente não estava entre os 10 mais similares. Alguns casos de abreviaturas para a palavra *'emergência'* como *'EMG'* também não eram identificados.

Um segundo erro comum foi com relação a abrangência dos termos e exemplos elencados no manual de anotação. Nos exemplos *1, 2 e, 9* da tabela temos exemplos de sentenças e termos que identificavam o indicador porém não haviam sido descritas e portando nenhuma regra com relação a esses casos foi implementada na ontologia.

Classes que indicam acontecimentos no passado também contribuíram nas classificações errôneas. Nos exemplos *4, 7, 8* o tempo passado é indicado através das datas, o qual nem a ontologia ou o algoritmo estavam preparados para processar. No caso do exemplo *5* observamos o uso do termo *'prévio'* para indicar um acontecimento passado de eventos que são mencionados posteriormente a sua aparição no texto, porém na ontologia o termo *'prévio'* foi organizado como um *Termo passo retroativo* que condiciona eventos passados mencionados anteriormente a sua aparição no texto, como seria o caso de uma sentença como *'AVC isquêmico prévio'*.

Com os exemplos *6, 10 e 11* vemos um termo que nega vários eventos em sequência, as regras na ontologia foram construída para que um termo negue apenas uma palavra sequente a ela na sentença como seria o caso de uma sentença como *'Sem dor'*.

Essa análise realizada indica que ainda existe um conjunto de melhorias que podem ser feitas tanto em questões de abrangência dos termos, da ontologia, da abordagem lógica e semântica dos textos. Algumas melhorias seriam: a adição de novos termos e axiomas à ontologia e; o aperfeiçoamento dos axiomas já descritos. Com essas alterações seria possível melhorar a generalização das regras e assim as classificações serem mais adequadas.

6.4 Comparação com modelos de Aprendizado de Máquina

Em (Zanotto et al., 2021) o mesmo desafio foi abordado utilizando soluções baseadas em aprendizado de máquinas. na figura 6.1 evidenciamos o modelo de aprendizado de máquina W+C+SVM que usou as técnicas Word-TFIDF and Character-TFIDF e obteve os melhores resultados de classificação e comparamos com os resultados obtidos pelo nosso

modelo. No trabalho (Zanotto et al., 2021) apenas 28 Indicadores foram classificados pelos modelos de Aprendizado de Máquina.

Classes	W+C+SVM	ONTOLOGIA
Alimentação	89,5	61,83
Localização	88,9	89,01
Risco de queda	89,6	61,77
Diabetes	89	89,89
Paresia	88,7	64,33
Trombólise	85,8	87,75
Óbito	89,5	0,2
Hipertensão	86	91,26
Obesidade	81,7	81,36
Mobilidade	75,7	19,64
Indicativo de infecção	79,9	25,58
Hemorragia intracraniana	66,4	5,83

Classes	W+C+SVM	ONTOLOGIA
Comunicação	74,4	61,54
Dislipidemia	83,2	94,07
Tabagismo	82,1	92,24
Trombectomia	72,6	64,97
Fibrilação atrial	71,3	83,65
AVC prévio	67,1	26,76
Doença coronária	61,2	78,5
Dor	52	13,37
Etilismo	38,6	82,61
Nível de mobilidade	40,5	50,75
Rankin (mRS)	26,9	87,61
NIHSS	12,4	79,33

Figura 6.1 – Comparação das métricas entre modelos

Através dos números evidenciados mostramos que a abordagem de classificação por ontologias possui resultados similares aos obtidos com técnicas de aprendizado de máquina. Outro ponto a ser ressaltado é que em indicadores que os modelos de aprendizado de máquina se mostraram ineficientes o nosso modelo obteve altos resultado. Observamos que o contrário também é verdadeiro. Essa observação aponta que a resolução completa do desafio proposto pode estar contida em um ambiente que combine os métodos de forma que este cubram os pontos fracos um do outro.

7. CONCLUSÃO

7.1 Conclusões sobre o modelo

Neste trabalho nós propomos uma abordagem de classificação de textos através do uso de ontologias. A análise dos resultado mostram que a abordagem adotada para a classificação desses textos é promissora para pelo menos 18 dos 30 indicadores e também nos indica quais desses precisam ser melhores tratados a fim de obter melhores resultados. Assim com esse estudo fornecemos não só uma abordagem para o processo de classificação de textos médicos, mas também um indicador de quais classes podem se beneficiar do modelo proposto e quais precisam ser detectadas com um modelo mais elaborado.

7.2 Limitações

Reconhecemos como limitação desse estudos que este foi elaborado sob a análise de apenas um conjunto de dados proveniente de um hospital parceiro. Devido a contratempos não tivemos acesso aos dados de outros hospitais parceiros com o qual entramos em contato e assim não conseguimos nesse momento validar a abrangência do modelo proposto. Como trabalho futuro devemos aplicar o modelo gerado em uma nova amostra de dados para validar o desempenho e em cima desses novos resultado decidir qual devem ser os próximos passos para o aperfeiçoamento das técnicas utilizadas. Outro caminho a ser trilhado como trabalho futuro é a implementação, ou criação, de modelos *word embeddings* especializados ao domínio dos pacientes de AVC para que a abrangência e precisão do modelo sejam ampliadas. Ainda, para esse trabalho havia sido planejado o uso de alinhamentos de ontologias como forma de realizar esse processo de expansão do conhecimentos que melhoraria abrangência e precisão dos termos, não tendo tido recursos hábeis de explorar tal solução também apontamos esse caminho como um estudo futuro a ser realizado. Outro ponto a ser ressaltado é com relação a escalabilidade da ontologia e a sua performance. A medida que outras áreas de cuidado forem abrangidas a manutenção da ontologia pode se tornar onerosa. Também é possível que área de cuidados possuam uma quantidade muito volumosa de termos e axiomas que podem comprometer a construção e o desempenho da ontologia. Para esses desafios pensamos que podemos adotar uma estratégia de divisão do problema em que diferentes ontologias podem ser construídas para diferentes áreas de cuidado, e mesmo dentro de uma área de cuidado ainda realizar outra sub-divisão caso esta possua uma definição de conhecimento muito extensa. Com o alinhamento das ontologias seria possível reunir todos esses esforços distribuídos construindo assim um modelo unificado com melhoras em escalabilidade e manutenção.

7.3 Conclusões sobre o projeto

O trabalho elaborado não foi um projeto individual e isolado. O modelo criado está inserido em um contexto com um escopo multidisciplinar que visa solucionar um desafio de saúde pública. Por isso, para o desenvolvimento deste não foi suficiente utilizarmos dos conhecimentos dos campos da computação e também não conseguimos trabalhar sozinhos. Para que este estudo tivemos a oportunidade de trabalhar com pesquisadores das área da saúde, engenharia de produção e de outras áreas da ciência da computação. A equipe construída precisou atuar em sinergia e trocar conhecimentos constantemente para que a comunicação fosse constantemente clara, objetiva e, unificada. A principal vantagem de um trabalho com essa característica é sem dúvidas a integração dos conhecimentos dessas áreas diversas assim como a absorbição de conceitos e informações que operando apenas em nosso campo de atuação normalmente não teríamos acesso. O maior desafio sem dúvidas é manter todos os integrantes sempre alinhados com relação a comunicação e a adequação do projeto às limitações técnicas de cada especialidade. Um segundo desafio com relação ao projeto é o tamanho do escopo do mesmo. Em sua totalidade é almejado a re-invenção dos processos e modelos de gestão de toda a saúde pública do país. Dividir esse problema em partes menores, mas sem perder a visão do todo com certeza foi um desafio que é acentuado com a integração dos diversos campos de conhecimento envolvidos

7.4 Repercussões do projeto

Os desafios entretanto não vem sozinhos, com a integração de uma equipe versátil abordando um problema de grande escala alguns resultados variados são produzidos. Nos três principais campos de atuação envolvidos conhecimentos foram criados e assimilados que irão auxiliar na resolução do problema exposto. Essa produção de conhecimento está sendo compilada em formato de um artigo ao qual será submetido a revistas da comunidade cientifica para divulgação dos frutos desse trabalho dentro de todas as áreas envolvidas. Especificamente pontuando na área da computação tivemos a oportunidade de abordar o problema com diferentes frentes. O modelo proposto por esse trabalho foi apenas uma das soluções. Outros modelos baseados em aprendizado de máquina foram também implementados e avaliados. No artigo sendo desenvolvidos serão apresentados os comparativos entre as frentes escolhidas e as conclusões e aprendizados tirados.

7.5 Considerações finais

Observando em retrospectiva todo o conteúdo exposto acreditamos que podemos afirmar que com a realização desse projeto saímos dele com muito mais conhecimento do que começamos, não só no campo de atuação ao qual havíamos nos proposto a operar, mas com conhecimentos sobre os desafios e conceitos de outras área que sem a realização de um trabalho como este não teríamos tido contato tão facilmente. Também acreditamos que com esse estudo damos um passo para frente na viabilização de um modelo de gestão baseado em valor na saúde pública e esperamos que com esse passo possamos nos aproximar de uma sociedade mais justa e que zele de forma ímpar pela saúde de seus integrantes.

REFERÊNCIAS BIBLIOGRÁFICAS

Allahyari, M., Kochut, K. J. e Janik, M. (2014). Ontology-based text classification into dynamically defined topics. In: *Proceedings of the 8th IEEE International Conference on Semantic Computing*, pp. 273–278, Newport Beach, Estados Unidos da America. IEEE.

Anderson, J. (Jan, 1999). Increasing the acceptance of clinical information systems. *M.D. computing: computers in medical practice*, vol. 16, pp. 62—65.

Barcellos, A. M. e Peixoto, B. M. (Fev, 2004). Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. *Ciência da Informação*, vol. 32, pp. 7–20.

Bessa, R. d. O. (2011). *Análise dos modelos de remuneração médica no setor de saúde suplementar brasileiro*. Dissertação de mestrado, FGV, São Paulo, Brasil. 107p.

Borges, M., Torres, A., Araújo, E. e Figueiredo, L. (Dez, 2007). Prontuário eletrônico do paciente — a funcionalidade do registro informatizado. *Revista de Enfermagem UFPE on line*, vol. 1, pp. 254–261.

Brown, D. (2018). Study: Physicians are unhappy with ehr design, interoperability. Recuperado de "https://www.aiin.healthcare/topics/connected-care/physicians-unhappy-ehr-design-interoperability". Outubro 2019.

Brownlee, J. (2019). What are word embeddings for text? Recuperado de "https://machinelearningmastery.com/what-are-word-embeddings/". Agosto 2019.

Chi, N.-W., Lin, K.-Y. e Hsieh, S.-H. (Out, 2014). Using ontology-based text classification to assist job hazard analysis. *Advanced Engineering Informatics*, vol. 28, pp. 381–394.

Chicco, D. e Jurman, G. (Jan, 2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, vol. 21, pp. 1–13.

Corrêa Cordeiro, F., Evsukoff, A. e Gomes, D. (2018). Word embeddings in portuguese for the specific domain of oil and gas. In: *Proceedings of the 19th Rio Oil & Gas*, pp. 10, Rio de Janeiro, Brasil. IBP.

da Silva Etges, A. P. B., Ruschel, K. B., Polanczyk, C. A. e Urman, R. D. (Mai, 2020). Advances in value-based healthcare by the application of time-driven activity-based costing for inpatient management: A systematic review. *Value in Health*, vol. 23, pp. 812–823.

Davies, J., Studer, R. e Warren, P. (2006). *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. Wiley, Chichester, Inglaterra.

de Araujo, D. A., Rigo, S. J. e Barbosa, J. L. V. (Jul, 2017). Ontology-based information extraction for juridical events with case studies in brazilian legal realm. *Artificial Intelligence and Law*, vol. 25, pp. 379–396.

de Lima, J. C. e de Carvalho, C. L. (2005). Ontologias-owl (web ontology language). Relatório técnico, Universidade Federal de Goiás.

Desagulier, G. (2008). Word embeddings: the (very) basics. Recuperado de "https://corpling.hypotheses.org/495". Abril 2020.

dos Santos, H. D. P., Ulbrich, A. H. D., Woloszyn, V. e Vieira, R. (2018). An initial investigation of the charlson comorbidity index regression based on clinical notes. In: *Proceedings of the 31st IEEE International Symposium on Computer-Based Medical Systems*, pp. 6–11, Karlstad, Suécia. IEEE.

Galvao, M. C. B. e Ricarte, I. L. M. (Dez, 2011). O prontuário eletrônico do paciente no século xxi: contribuições necessárias da ciência da informação. *InCID: Revista de Ciência da Informação e Documentação*, vol. 2, pp. 77–100.

Garla, V. N. e Brandt, C. (Out, 2012). Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics*, vol. 45, pp. 992–998.

Gayathri, M. e Kannan, R. (Jan, 2020). Ontology based concept extraction and classification of ayurvedic documents. *Procedia Computer Science*, vol. 172, pp. 511–516.

George, E. e Engel, L. (Mai, 1980). The clinical application of the biopsychosocial model. *American journal of Psychiatry*, vol. 137, pp. 535–544.

Gomez-Perez, A., Fernández-López, M. e Corcho, O. (2006). *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer Science & Business Media, Madri, Espanha.

Gonçalves, F. (Fev, 2019). Optimizing patients' pathways in international cooperation, by doing value based healthcare (vbhc). *Acta Médica Portuguesa*, vol. 32, pp. 167–168.

Grosso, W., Eriksson, H., Fergerson, R., Gennari, J., Tu, S. e Musen, M. (1999). Knowledge modeling at the millennium : The design and evolution of protégé-2000. In: *Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management*, pp. 16–21, Juan-les-pin, França. EKAW.

Group, O. W. (2012). Owl. Recuperado de https://www.w3.org/OWL/. Dezembro 2020.

Horridge, M., Knublauch, H., Rector, A., Stevens, R. e Wroe, C. (2004). *A Practical Guide To Building OWL Ontologies Using The Prot'eg'e-OWL Plugin and CO-ODE Tools*. University of Manchester, Manchester, Inglaterra.

Joseph, J. (2009). Best metric to measure accuracy of classification models. Recuperado de https: //clevertap.com/blog/the-best-metric-to-measure-accuracy-of-classification-models/. Fevereiro 2020.

Lamy, J.-B. (Jul, 2017). Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine*, vol. 80, pp. 11–28.

Lassila, O. e McGuinness, D. L. (2001). The role of frame-based representation on the semantic web. Relatório técnico, Stanford University, Palo Alto, Estados Unidos da América.

Lee, T. H. (Dez, 2010). Putting the value framework to work. *New England Journal of Medicine*, vol. 363, pp. 2481–2483.

Ltda., M. I. N. (2016). A importância dos dados dos clientes para hospitais e clínicas. Recuperado de "http://www.mv.com.br/pt/blog/a-importancia-dos-dados-dos-clientes-para-hospitais-e-clinicas". Outubro 2019.

Mikolov, T., Chen, K., Corrado, G. S. e Dean, J. (2013). Efficient estimation of word representations in vector space. Recuperado de "http://arxiv.org/abs/1301.3781". Agosto 2020.

Noy, N., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. e Musen, M. (Abr, 2001). Creating semantic web contents with protege-2000. *Intelligent Systems, IEEE*, vol. 16, pp. 60 – 71.

Patrício, C., Maia, M., Machiavelli, J., Novaes, M. e Navaes, A. (Jan, 2011). O prontuário eletrônico do paciente no sistema de saúde brasileiro: uma realidade para os médicos? *Scientia Medica*, vol. 21, pp. 121–131.

Putnam, S. M. e Lipkin, M. (1995). *The Patient-Centered Interview: Research Support*, cap. 47, pp. 530–538. Springer, Nova York, Estados Unidos da América.

Schütze, H., Manning, C. D. e Raghavan, P. (2008). *Introduction to information retrieval*, vol. 39. Cambridge University Press Cambridge, Cambridge, Inglatera.

Schwertner, M. A., Rigo, S. J., Araújo, D. A., Silva, A. B. e Eskofier, B. (2019). Fostering natural language question answering over knowledge bases in oncology ehr. In: *Proceedings of the 32nd International Symposium on Computer-Based Medical Systems*, pp. 501–506, Cordoba, Espanha. IEEE.

Sharecare (2020). Fee for service na saúde: veja os desafios desse modelo de pagamento. Recuperado de "https://sharecare.com.br/noticias/fee-for-service-na-saude/". Janeiro 2020.

Standford (2018). How doctors feel about eletronic health records. Recuperado de "https://med.stanford.edu/content/dam/sm/ehr/documents/EHR-Poll-Presentation.pdf". Outubro 2019.

Tsai, M., Porter, J. e Adams, D. (Abr, 2018). The denominator in value-based health care: Porter's hidden costs. *Anesthesia And Analgesia*, vol. 127, pp. 1.

Uzuelli, F. H. d. P., Costa, A. C. D. d., Guedes, B., Sabiá, C. F. e Batista, S. R. R. (Jun, 2019). Reforma da atenção hospitalar para modelo de saúde baseada em valor e especialidades multifocais. *Ciência & Saúde Coletiva*, vol. 24, pp. 2147 – 2154.

Wang, B. B., Mckay, R. I. B., Abbass, H. A. e Barlow, M. (2003). A comparative study for domain ontology guided feature extraction. In: *Proceedings of the 26th Australasian computer science conference*, pp. 69–78, Adelaide, Austrália. Australian Computer Society, Inc.

Welty, C., McGuinness, D. L. e Smith, M. K. (Nov, 2009). Owl web ontology language guide. *W3C recommendation, W3C*, vol. 1, pp. 21.

Yehia, E., Boshnak, H., Abdelgaber, S., Abdo, A. e Elzanfaly, D. (Ago, 2019). Ontology-based clinical information extraction from physician's free-text notes. *Journal of Biomedical Informatics*, vol. 98, pp. 103–117.

Zanotto, B., da Silva Etges, A. P. B., Bosco, A. D., Cortes, E. G., Ruschel, R., Martins, S. O., Souza, A. C., Valiense, C., Viegas, F., Canuto, S., Luiz, W., Vieira, R., Gonçalves, M. e Polanczyk, C. A. (2021). Automatic classification of electronic health records for a value-based program through machine learning. In: *Value in Health*, pp. à aparecer, Virtual edition. The International Society for Pharmacoeconomics and Outcomes Research. Aceito para publicação.

Zhou, P. e El-Gohary, N. (Set, 2015). Ontology-based multilabel text classification of construction regulatory documents. *Journal of Computing in Civil Engineering*, vol. 30, pp. 40–54.

ANEXO A – Manual de anotação

1. Características Clínicas

Variável	Definição	Opções de Resposta	Orientações para resposta / Sentenças
Doença coronariana	CID10 I20 – Angina Pectoris	0- Doença ausente	# Sem histórico de CI # Não possui história pregressa de IAM # Paciente sem histórico de DAC
	CID10 I21 – Infarto Agudo do Miocárdio		
	CID10 I22 – Infarto do Miocárdio Recorrente		
	CID10 I24 – Outras Doenças Isquêmicas Agudas do Coração	1- Doença presente	# Angina estável # Angina instável # ACTP # IAM # IAMSSST # IAMCSST # CRM # Cardiopatia isquêmica # CI # DAC # SCA # DCC # Aterosclerose coronariana/coronária # Estenose coronariana/coronária
	CID10 I25 – Doença Isquêmica Crônica do Coração		

Variável	Definição	Opções de Resposta	Orientações para resposta / Sentenças
Fibrilação Atrial	CID10 I48 - Flutter e fibrilação atrial	0- Doença ausente	# Sem história pregressa de FA # Não apresenta fibrilação atrial
		1- Doença presente	# FA # Fibrilação atrial # Fibrilação atrial paroxística # ACFA
Diabetes	CID10-E10 – Diabetes Mellitus Insulino-dependente	0- Doença ausente	# Nega DM # Não possui diabetes # Sem história pregressa de DM
	CID10-E11 – Diabetes Mellitus Não-insulino-dependente		
	CID10-E12 – Diabetes Mellitus Relacionado Com a Desnutrição	1- Doença presente	# DM # Diabetes mellitus # Diabete mellitus # Diabetes # DM2
	CID10-E13 – Outros Tipos Especificados de Diabetes Mellitus		
	CID10-E14 – Diabetes Mellitus Não Especificado		
AVC prévio	CID10-I64 Acidente vascular cerebral, não especificado como hemorrágico ou isquêmico	0- Não	# Sem história prévia de AVC
	CID10-I63 – Infarto Cerebral	1- Sim	# AVC isquêmico prévio # AVC ISQ em 2018 # História prévia: AVCi AVCh # Acidente vascular cerebral # Acidente vascular encefálico
Hipertensão arterial sistêmica	CID10-I10 – Hipertensão Essencial (primária)	0- Doença ausente	# Nega HAS. # Não possui história pregressa de hipertensão # Sem histórico de HAS
	CID10- I11 – Doença Cardíaca Hipertensiva	1- Doença presente	# HAS # Hipertensão arterial Hipertensa # História prévia de Has.

Variável	Definição	Opções de Resposta	Orientações para resposta / Sentenças
Obesidade	CID10- E66 – Obesidade	0- Doença ausente	-
		1- Doença presente	# Obesidade \# Obeso \# IMC > 30 Índice de massa corporal > 30
Dislipidemia	CID10-E78 - Distúrbios do metabolismo de lipoproteínas e outras lipidemias	0- Doença ausente	# Nega DLP \# Sem história pregressa de dislipidemia
		1- Doença presente	# Dislipidemia \# Dislipidêmico \# DLP
Câncer	Registro de câncer em atividade, indicativo de realização de tratamentos para a doença (radioterapia, quimioterapia) ou indicativo de câncer recente. Quando indicar que o paciente já teve câncer curado ou no passado não deverá ser considerado.	0- Doença ausente	#2005 câncer de pele #Ca no passado
		1- Doença presente	# Ca em atividade \# Câncer em tratamento #em tto quimioterápico \# Neoplasia de próstata ativa
Tabagismo	Registro do consumo condicionado a dependência de cigarros ou outros produtos que contenham tabaco (possível relato do paciente). O relato de negação ou ex-fumante também deve ser considerado	0- Não tabagista	# Nega tabagismo \# Não tabagista
		1- Tabagista	# Tabagismo \# Tabagista \# Fumante
		2- Ex-tabagista	# Ex-tabagista \# Tabagista no passado

Variável	Definição	Opções de Resposta	Orientações para resposta / Sentenças
Etilista	Registro do consumo condicionado a dependência de álcool (possível relato do paciente). O relato de negação ou ex-alcoolista também deve ser considerado	0- Não etilista	# Nega etilismo # Não etilista
		1- Etilista	# Etilismo # Alcoolismo # Abuso de álcool # Alcoolista
		2- Ex etilista	# História prévia de alcoolismo # Ex alcoólatra #Ex elitista

2. Manejo Clínico e Processo de cuidado

Variável	Definição	Opções de Resposta	Orientações para resposta / Sentenças
Localização	Indica o local de presente momento do paciente. ***Sentenças que indicam deslocamento do paciente sem informativo do destino são desconsideradas	1- Emergência	# Evolução diária- emergência- enfermagem
		3- CTI	## CTIA ## # Paciente proveniente da emergência por volta das 15:50h pós trombólise por AVCi. # Interna na UTI após procedimento. # Fisioterapia terapia intensiva
		4-Unidade Internação	# Enfermagem C1. # Paciente deu entrada na UI
Trombectomia	Refere ao possível manejo clínico que o paciente AVCi recebeu. ***A contraindicação, sem delta ou sem janela também deve ser considerada como não realização do procedimento**	0- Não realizou ou sem janela para terapia	# Sem indicação de trombólise ou trombectomia. # Sem janela para trombectomia
		1- Trombectomia	# Trombectomia mecânica – # Trombectomia mecânica as 12 horas
Trombólise	Refere ao possível manejo clínico que o paciente AVCi recebeu. ***A contraindicação, sem delta ou sem janela também deve ser considerada como não realização do procedimento.**	0- Não realizou ou sem janela para terapia	# Não trombolisada - delta # Contraindicação a trombólise (hipodensidade >1/3). # Sem janela trombolítica
		1- Trombólise	# Trombolise endovenosa # ALTEPLASE # AVCi trombolisado

3. Escalas de Avaliação e Eventos de Risco

Variável	Definição	Opções de Resposta	Orientações para resposta/ Sentenças
Hemorragia intracraniana	CID10 - I61 Hemorragia intracerebral.	0- Hemorragia ausente	# TC controle ontem sem sangramento # TCC sem alterações # Sem indícios de sangramento
	CID10 - I62 Outras hemorragias intracranianas não-traumáticas.	1- Hemorragia presente	# Hemorragia cerebelar # Hemorragia cerebral com transformação hemorrágica # Hemorragia em fossa posterior próximo ao IV ventrículo.
Queda	Registro de queda do paciente durante período de internação.	1- Sim, queda	# Paciente relata insegurança de usar andador após queda # Paciente perdeu o equilíbrio e apresentou queda da própria altura. # Caiu no quarto a noite
Escala Braden	Escala de Braden é um recurso utilizado nas Unidades de Terapia Intensiva para medir o risco dos pacientes críticos de desenvolverem lesões por pressão.	1- Risco baixo	Baixo risco: escore >17
		2-Risco moderado	Risco moderado: escore 16-13
		3-Risco alto	Risco alto: escore =<12
Risco de queda	Variável com objetivo de identificar o risco de queda do paciente de acordo com a escala John Hopkins. Atenta-se também a descritivos da escala que podem ser ajustados de acordo com o risco definido por consenso dos profissionais.	1- Baixo risco	Baixo risco: escore de 0-5 pontos ou atentar para medidas preventivas de queda. Ex.: # Risco para Quedas (Escala John Hopkins 02); # Manter medidas de prevenção de queda;
		2- Moderado risco	Risco moderado: escore de 6-13 pontos. Ex.: #Reforço orientações com a familiar sobre o risco de quedas. # risco de queda. *Sentenças que apresentam apenas "risco de queda", sem score ou contexto, podem ser consideradas como risco moderado.
		3-Alto risco	Alto risco: escore > 13 pontos. Implementação do protocolo de risco de queda. Ex.: i) Alto risco de queda (14). ii) Protocolo de alto risco para queda;

Variável	Definição	Opções de Resposta	Orientações para resposta/ Sentenças
Indicativo de Infecção	Sentenças que indiquem possível infecção viral, bacteriana ou fúngica no paciente. Observar: sinais flogísticos (Calor (febre), rubor ou hiperemia (vermelhidão), edema (inchaço), hiperestesia (dor ao toque), perda de função), uso de antibióticos, exames de investigação como hemocultura, início de secreção por infecção **desconsiderar resultados de exames sem indicativo explicito de infecção ***** em anexo alguns antimicrobianos comuns na prática clínica**	0- Indicativo negativo de infecção	#Sem sinais flogísticos #Afebril, sinais vitais estáveis #Temp máx 37,2ºC # Culturais HMCs 09/08: negativas ATQ 09/08: negativo HMCs 12/08: negativas ATQ
		1- Indicativo alerta de infecção	# Pneumonia base direita-< BACTRIN # EQU 13/09: urocultura em andamento. #NOVA ITU - MEROPENEN -> E.COLI

4. Indicadores Clínicos Desfechos e Status do Paciente

Variável	Definição	Opções de Resposta	Orientações para resposta/ Sentenças
Óbito	Sentenças que indiquem nota de óbito/falecimento. Indicativos de não-óbito podem ser marcados e são considerados: sinais vitais estáveis, registros de paciente presente no leito, bem, respirando espontaneamente, previsão de alta do paciente.	0- Indicativos de presença e sinais vitais	# No leito, tranquilo, acompanhado. # Condição Ventilatória: ar ambiente, eupneica # Eupneica, ventilando espontâneo em ar ambiente, sem esforço ventilatório; # Hemodinamicamente estável # SV estáveis
		1- Nota de óbito	# Nota de falecimento # Paciente faleceu hoje pela manhã
Dor	Indica se o paciente apresenta sinais clínicos de dor. **observar: escala EVA de dor, escala PAINAD, escala BPS e composições de palavras que justifiquem marcação.	0- Sem dor	# Paciente sem queixas # Paciente bem, sem dor #Paciente não refere dor # Sem queixas álgicas # Nega dor.
		1- Dor leve/moderada	Escores (EVA/PAINAD/BPS) 1 a 6 Ex.: #Dor indefinida #Dor leve #Queixa de dor.
		2- Dor intensa	Escores (EVA/PAINAD/BPS) 7 a 10 Ex.: #Dor difusa #Dor grave #Dor intensa
Alimentação	Indica se o paciente precisa de sonda ou gastrostomia para se alimentar.	1- Alimentação por via oral	#Paciente alimentando-se por via oral. #Dieta para semi pastosa #Boa aceitação VO.
		2- Uso de sondas ou gastrostomia	#Paciente alimentando-se por via alternativa # Cuidados com SNE; #Dieta em BI; #Cuidados com gastrostomia #GTT #Dieta polimérica

Variável	Definição	Opções de Resposta	Orientações para resposta/ Sentenças
Força	Indica a capacidade de superar ou opor-se a uma resistência por meio da atividade muscular, conforme escala de grau de força, composições descritivas ou nível de mobilidade associado. Levar em conta apenas membros superiores e inferiores (facial não), no caso de uma sentença apresentar vários graus de força, priorizar força de Membros Inferiores e proximais e menor grau apresentado. *****Verificar relação com escala NIH	Escore de 0 a 5	Apresentação da escala: 0 – Não percebe contração. Plegia/paciente plégico é marcado com força 0. 1 – Traço de contração sem produção de movimento. *Passível de correlação com escala NIH domínio Membros Superiores (5) e Membros Inferiores (6) pontuação 4. 2 – Contração fraca, elimina gravidade. *Passível de correlação com escala NIH domínio Membros Superiores (5) e Membros Inferiores (6) pontuação 3. 3 – Realiza movimento contra a gravidade, porém sem resistência adicional. *Passível de correlação com escala NIH domínio Membros Superiores (5) e Membros Inferiores (6) pontuação 2 ou 1. 4 – Realiza movimento contra a gravidade e resistência externa. *Passível de correlação com escala NIH domínio Membros Superiores (5) e Membros Inferiores (6) pontuação 0. 5 - Supera maior quantidade de resistência. Sem alteração em força.
Paresia	Movimento limitado ou fraco, motilidade num padrão abaixo do normal. No que se refere à força muscular, precisão do movimento, amplitude do movimento e a resistência muscular localizada, ou seja, refere-se a um comprometimento parcial, uma perda de força. ***Levar em conta apenas membros superiores e inferiores (paresia facial não)**	1- Registro de paresia	# Apresenta hemiparesia # Paciente apresenta perda de força em MIE # Hemiparesia à E de predomínio braquial ***** Em casos de força grau 1 a 3 indicar paresia.**

Variável	Definição	Opções de Resposta	Orientações para resposta/ Sentenças
Mobilidade	Variável indica se paciente é capaz de andar com autonomia. Associada com escala de nível de mobilidade utilizada pela instituição *(Callen, BL., et al. Medsurg nursing, 2004)* ou a descrição dos seus níveis, podendo ser adaptada conforme necessário. *****Usualmente associada com grau de força e escala de nível de mobilidade da instituição (se aplicável) conforme orientações de resposta.**	0 - Incapaz de andar	Incapacidade de andar podem ser relacionados aos níveis de 1 a 8 na escala de mobilidade, ou descrições como: i) paciente dependente: equipe promove trocas de decúbito, posicionamento e ADM; ii) paciente participa com a equipe das trocas de decúbito, posicionamento e ADM; iii) paciente é independente no leito; iv)transferido para a cadeira; v) auxilio mecânico ou 3 pessoas para a cadeira /cadeira de rodas; vi) transferência para a cadeira/de rodas com auxilio de duas pessoas; vii)auxilio de duas pessoas , ortostase e pivô para a cadeira; viii) auxílio de uma pessoa, ortostase e pivô para a cadeira;
		1 - Capaz de andar **sem ajuda** de outra pessoa ou dispositivo	Capaz de andar sem ajuda de outra pessoa ou dispositivo pode ser relacionados aos níveis de mobilidade de 12 a 15 na escala de mobilidade, ou descrições como: i) deambula com um auxiliar de prontidão; ii) deambula independente somente no quarto; iii) deambula fora do quarto, distância menor que um corredor; iv) deambula fora do quarto, distância maior que um corredor. ***** Quando mobilidade 1, marcar força 5.**
		2 - Capaz de andar **com ajuda** de outra pessoa ou dispositivo	Capaz de andar com a ajuda de outra pessoa ou dispositivo podem ser relacionados aos níveis de 9 a 11 na escala de mobilidade, ou descrições como: i) uma pessoa ao lado de prontidão para transferir para a cadeira; ii) deambula com dois auxiliares; iii) deambula com um auxiliar. ***** Quando mobilidade 2, marcar força 4.**

Variável	Definição	Opções de Resposta	Orientações para resposta/ Sentenças
Nível de mobilidade	Facultativo: variável dependente do uso, pela instituição, de escalas de rastreamento para níveis de mobilidade do paciente. Caso não haja uma escala padrão proposta, podem ser feitas de acordo com as descrições indicadas. Escala de nível de mobilidade utilizada pela instituição *(Callen, BL., et al. Medsurg nursing, 2004)*	Escore de 1 a 15	1 - Paciente dependente 2- Paciente participa com a equipe das trocas de decúbito, posicionamento e ADM 3- Paciente é independente no leito 4- Transferido para a cadeira 5-auxílio mecânico ou 3 pessoas para a cadeira /cadeira de rodas 6- Transferência para a cadeira/de rodas com auxílio de duas pessoas 7-Auxílio de duas pessoas, ortostase e pivô para a cadeira 8-Auxílio de uma pessoa, ortostase e pivô para a cadeira 9-Uma pessoa ao lado de prontidão para transferir para a cadeira 10-Deambula com dois auxiliares 11-Deambula com um auxiliar 12-Deambula com um auxiliar de prontidão 13-Deambula independente somente no quarto 14-Demabula fora do quarto, distância menor que um corredor 15-Deambula fora do quarto, distância maior que um corredor

Variável	Definição	Opções de Resposta	Orientações para resposta / Sentenças
Comunicação	Variável indica se paciente tem problemas de comunicação verbal. *Verificar relação com escala NIH.	0- Sem comunicação verbal	Impossibilidade de efetuar comunicação verbal. Afasia grave, toda a comunicação é feita através de expressões fragmentadas; necessidade de interferência, a quantidade de informação que pode ser trocada é limitada. Ex.: # Paciente comunica por gestos. *Passível de correlação com escala NIH domínio Linguagem (9) e Disartria (10) pontuação 2 ou 3.
		1- Comunica bem	Comunicação sem dificuldades/assintomáticas. Ex.: # Paciente dá bom dia; # Fala ok; # Recuperado dos sintomas de fala; # Sem afasia, normal; # Comunicativo *Passível de correlação com escala NIH domínio Linguagem (9) e Disartria (10) pontuação 0.
		2- Comunica pouco ou mal	Comunicação verbal com dificuldade/ problemas. Afasia leve a moderada; perda óbvia de alguma fluência, sem limitação significativa das ideias expressas ou formas de expressão. Ex.: # Paciente com afasia de expressão e agora disartria. # Fala enrolada *Passível de correlação com escala NIH domínio Linguagem (9) e Disartria (10) pontuação 1.

Variável	Definição	Opções de Resposta	Orientações para resposta/ Sentenças
Capacidade cognitiva	Variável sobre o entendimento, situação de lucidez e orientação do paciente. Deve-se sempre avaliar o contexto da sentença, pois palavras como "colaborativo", "bem", "interagindo", "atendem comandos" quando sozinhas não indicam entendimento pleno, porém quando agrupadas em contexto podem traduzir algum discernimento do paciente. *Verificar relação com escala NIH.	0- Não estado de clareza	Paciente sem clareza ou em estado de sedação. Ex.: # Paciente sedado, não responsivo, # Paciente no leito, abertura ocular espontânea, não interage, sem atender comandos. *Passível de correlação com escala NIH domínio Nível de Consciência (1,1a,1b) pontuação 2.
		1- Lúcido, orientado, coerente	Situação de lucidez e orientação do paciente. Ex.: # LOC # Escala glasgow 15 # Paciente lúcido, responsivo, tranquilo, obedece a comandos *Passível de correlação com escala NIH domínio Nível de Consciência (1,1a,1b) pontuação 0.
		2- Confuso, compreende mal	Paciente confuso, compreendendo mal, delirium. Ex.: # Paciente relata confusão em pensamentos, desorientado em espaço e tempo. *Passível de correlação com escala NIH domínio Nível de Consciência (1,1a,1b) pontuação 1.

Variável	Definição	Opções de Resposta	Orientações para resposta / Sentenças
Escala Rankin	Escala de avaliação funcional pós-AVC. Preenche a variável conforme indicação do score na evolução do paciente (de 0 a 6). ***** Quando rankin 4 ou 5, marcar auto-cuidado = 0**	Escore de 0 a 6	Apresentação da escala: 0 - Sem sintomas 1 - Nenhuma deficiência significativa, a despeito de sintomas 2 - Leve deficiência - Incapaz conduzir todas as atividades de antes, mas é capaz de cuidar dos próprios interesses sem assistência 3 - Deficiência moderada -Requer alguma ajuda, mas é capaz de caminhar sem assistência (pode usar bengala ou andador) *4 - Deficiência moderadamente grave - Incapaz de caminhar sem assistência e incapaz de atender às próprias necessidades fisiológicas sem assistência *5 - Deficiência grave - Confinado à cama, incontinente, requerendo cuidados e atenção constante de enfermagem 6 – Óbito
Auto cuidado	Capacidade de autonomia em atividades básicas como vestir-se, ir sozinho ao banheiro. *A não autonomia também é indicada, inclusive pela relação com rankin 4 e 5. **Considera-se também relações descritivas de dependência, como quando restrito ao leito ou níveis de mobilidade correspondente a 1 ou 2.	0- Dependente para atividades básicas/de higiene	# Eliminações espontâneas, em fraldas # Rankin 4 # Restrito ao leito no momento # Precisa de auxílio para as atividades da vida diária.
		1- Independente para atividades básicas/de higiene	# Tem vida de relação, conversa, caminha, alimenta-se. # Paciente faz uso banheiro/WC
NIH	A National Institute of Health Stroke Scale (NIHSS) é uma escala padrão, validada, segura, quantitativa da severidade e magnitude do déficit neurológico após o AVC. Seu escore varia de 1 a 42, porém não é incomum esta escala aparecer de forma a indicar o nível de cada domínio considerado, quando isso acontece os valores podem ser traduzidas a outras variáveis que temos presente no manual, conforme instruções acima.	Escore de 1 a 42	# NIH 14 (sonolento 1, errou 1 pergunta, paresia face 2, desvio do olhar 1, paresia MSE 3, MIE3, disartria 1, Heminegligencia 2) # NIH: face: 1, 2, negligência 2, MSE 3, MIE 3 - Total: 11. # NIHSS 5

	***** em anexo a escala NIH**		

A variável **"status_chegada"** é uma variável-suporte nos casos em que a sentença diz respeito a uma informação claramente no passado do paciente (por exemplo, situação do paciente na chegada ao hospital), e **não refere seu quadro atual**; isso não desonera o fato da marcação da variável correspondente.

ANEXO B – Total de Termos dos Indicadores

Indicador	Classe	Termos	Termos por Classe	Termos por Indicador
Alimentação	Via Oral	pastosa', 'via', 'alimentando', 'oral', 'dieta'	5	11
	Sondas Gastronomicas	bi', 'poliméria', 'via', 'gastronomia', 'alimentando', 'gtt', 'dieta', 'sne', 'alternativa'	9	
Auto-cuidado	Dependente para atividades básicas/de higiene	auxilio', 'eliminações', 'restrito', 'alternada', 'funcionalidade', 'leito', 'fraldas', 'atividades', 'ajuda', 'rankin', 'diárias'	11	16
	Independente para atividades básicas/de higiene	conversa', 'paciente', 'uso', 'caminha', 'banheiro'	5	
AVC prévio	Não	avc', 'sem'	2	10
	Sim	encefálico', 'prévio', 'vascular', 'avci', 'acidente', 'cerebral', 'avc', 'avch', 'prévia'	9	
Braden	Baixo Risco	braden	1	1
	Risco Médio	braden	1	
	Risco Alto	braden	1	
Capacidade Cognitiva	Não estado de clareza	ocular', 'interage', 'comandos', 'sem', 'não', 'sedado', 'abertura', 'atender', 'espontânea'	9	20
	Lúcido, orientado, coerente	colaborativo', 'responsivo', 'loc', 'orientado', 'glasgow', 'obedece', 'alerta', 'compreende', 'comandos', 'lúcido'	10	
	Confuso, compreende mal	desorientado', 'confuso'	2	
Comunicação	Sem comunicação verbal	'incompreensível', 'gestos', 'comunica', 'comunicação'}	4	18
	Comunica bem	'preservado', 'fala', 'paciente', 'ok', 'comunicativo', 'recuperado', 'compreensível', 'sintomas', 'colóquio', 'verbaliza', 'dia', 'bom'}	12	
	Comunica pouco ou mal	'desorientado', 'confuso'}	2	
Câncer	Doença ausente	câncer', 'tumor', 'prévio', 'ca', 'neoplasia', 'passado'	6	10
	Doença presente	câncer', 'tumor', 'cx', 'ca', 'tto', 'metástase', 'neoplasia', 'quimioterápico'	8	
Diabetes	Doença ausente	dm', 'diabetes', 'nega', 'desconhece', 'não', 'sem', 'diabete'	7	8
	Doença presente	diabete', 'diabetes', 'dm2', 'dm'	4	
Dislipidemia	Doença ausente	sem', 'dlp', 'nega', 'dislipidemia'	4	5
	Doença presente	dlp', 'dislipidêmico', 'dislipidemia'	3	
Doença Coronária	Doença ausente	dac', 'iam', 'sem', 'não', 'ci'	5	19
	Doença presente	iamcsst', 'prévio', 'sca', 'iamssst', 'crm', 'angina', 'estenose', 'coronaria', 'coronariana', 'dac', 'actp', 'isquêmica', 'ci', 'iam', 'dcc', 'altersosclerose', 'cardiopatia'	17	
Dor	Sem Dor	dor', 'paciente', 'nega', 'queixas', 'sem', 'não'	6	10
	Dor leve/moderada	eva', 'dor'	2	
	Dor intensa	grave', 'difusa', 'dor', 'eva', 'intensa'	5	
Etilismo	Não Etilista	não', 'etilismo', 'nega', 'etilista'	4	14
	Etilista	alcoólatra', 'alcoolista', 'alcoolismo', 'abuso', 'etilismo', 'etilista', 'álcool'	7	
	Ex Etilista	prévio', 'ex-alcoolismo', ' ex-alcoólatra', 'alcoolismo', 'ex-etilista', 'ex-acoolista'	6	
Fibrilação Atrial	Doença ausente	atrial', 'fa', 'desconhece', 'fibrilação', 'não', 'sem'	6	7
	Doença presente	fa', 'fibrilação', 'acfa', 'atrial'	4	
Grau de Força	escore 0	plégia', 'plégico', 'contração', 'paralisia', 'não'	5	18
	escore 1	contração', 'sem', 'movimento'	3	
	escore 2	contração', 'gravidade', 'fraca', 'elimina'	4	
	escore 3	gravidade', 'contra', 'movimento'	3	
	escore 4	contra', 'sutil', 'perda', 'leve', 'movimento', 'gravidade', 'resistência'	7	
	escore 5	maior', 'força', 'supera', 'resistência'	4	
Hemorragia Intracraniana	Hemorragia ausente	hemorragica', 'transformação', 'alterações', 'sem', 'sgto', 'sangramento', 'tcc', 'tc'	8	12
	Hemorragia presente	cerebelar', 'hemorragica', 'transformação', 'intracraniana', 'cerebral', 'hemorragia'	6	
Hipertensão Arterial Sistêmica	Doença ausente	não', 'has', 'desconhece', 'sem', 'nega', 'hipertensão', 'hipertensa'	7	11
	Doença presente	arterial', 'prévia', 'intracraniana', 'has', 'hipertensão', 'hemorragia', 'hipertensa'	7	
Indicativo de Infecção	Indicativo negativo de infecção	sem', 'flogísticos', 'sinais', 'afebril'	4	17
	Indicativo alerta de infecção	temperatura', 'e.coli', 'urucultura', 'atb', 'meropenen', 'pneumonia', 'febre', 'febril', 'leucocitose', 'bactrin', 'antibiocoterapia', 'sepse', 'hemocultura'	13	
Localização	Emergência	emergência'	1	26
	Unidade de Internação	a3', 'e1', 'e3', 'b2', 'c4', 'c3', 'c2', 'e4', 'd4', 'e2', 'unidade', 'd2', 'd3', 'b1', 'a4', 'internação', 'b3', 'a2', 'd1', 'b4', 'c1', 'a1'	22	
	UTI/CTI	fisioterapia', 'intensiva', 'terapia'	3	

Mobilidade	Incapaz de andar	Depende do nível de mobilidade	0	0
	Capaz de andar sem ajuda de outra pessoa ou dispositivo	Depende do nível de mobilidade	0	
	Capaz de andar com ajuda de outra pessoa ou dispositivo	Depende do nível de mobilidade	0	
NIH	Escore de 1 a 42	NIH', 'MEI', 'NIHSS', 'MSE'	4	4
Nível de Mobilidade	Escore de 1	acamado', 'paciente', 'restrito', 'dependente', 'leito', 'mobilidade'	6	35
	Escore de 2	trocas', 'paciente', 'posicionamento', 'decúbito', 'mobilidade', 'adm'	6	
	Escore de 3	paciente', 'leito', 'mobilidade', 'independente'	4	
	Escore de 4	cadeira', 'transferido', 'mobilidade'	3	
	Escore de 5	mecânico', 'mobilidade', 'auxilio'	3	
	Escore de 6	pessoas', 'duas', 'transferência', 'mobilidade', 'cadeira', 'auxilio'	6	
	Escore de 7	pessoas', 'duas', 'transferência', 'mobilidade', 'cadeira', 'auxilio'	6	
	Escore de 8	pessoas', 'duas', 'pivô', 'ortostase', 'mobilidade', 'cadeira', 'auxilio'	7	
	Escore de 9	pivo', 'ortostase', 'mobilidade', 'uma', 'cadeira', 'pessoa', 'auxilio'	7	
	Escore de 10	prontidão', 'mobilidade', 'uma', 'cadeira', 'transferir', 'pessoa'	6	
	Escore de 11	dois', 'auxiliares', 'mobilidade', 'deambula'	4	
	Escore de 12	um', 'deambula', 'mobilidade', 'auxiliar'	4	
	Escore de 13	quarto', 'deambula', 'mobilidade', 'independente'	4	
	Escore de 14	quarto', 'fora', 'corredor', 'deambula', 'mobilidade', 'menor'	6	
	Escore de 15	quarto', 'fora', 'corredor', 'deambula', 'mobilidade', 'maior'	6	
Obesidade	Doença Ausente		0	4
	Doença presente	obesidade', 'imc', 'obesa', 'obeso'	4	
Paresia	Registro de paresia	hemiparesia', 'mei', 'perda', 'parético', 'força', 'hemiparético'	6	6
Risco de Queda	Baixo Risco	risco', 'baixo', 'caiu', 'queda', 'jhc'	5	6
	Moderado Risco	risco', 'baixo', 'caiu', 'queda', 'jhc'	5	
	Alto Risco	risco', 'alto' 'caiu', 'queda', 'jhc'	5	
Queda	Sim, queda	caiu', 'queda'	2	2
Rankin	Escore de 0 a 6	rankin'	1	1
Tabagismo	Não Tabagista	nega', 'tabagista', 'tabagismo'	3	7
	Tabagista	tabagista', 'fumante', 'tabagismo'	3	
	Ex-Tabagista	passado', 'ex-tabagista', 'tabagista', 'ex-fumante'	4	
Trombectomia	Não realizou ou sem janela para terapia	sem', 'trombectomia'	2	4
	Trombectomia	trombectomia', 'mecânica', 'trombec'	3	
Trombólise	Não realizou ou sem janela para terapia	delta', 'janela', 'não', 'trombolisada', 'reperfusão', 'contraindicação', 'sem', 'indicação', 'trombólise', 'trombolítico', 'contraindicado'	11	14
	Trombólise	delta', 'trombolisada', 'contraindicação', 'sem', 'alteplase', 'trombolítico', 'reprefusão', 'trombólise', 'terapia'	9	
Óbito	Indicativos de presença e sinais vitais	paciente', 'estável', 'estáveis', 'leito', 'sv', 'hemódicamente'	6	15
	Nota de óbito	ausência', 'faleceu', 'paciente', 'falecimento', 'sem', 'sinal', 'vital', 'nota', 'óbito', 'constato'	10	
Total Termos				331

ANEXO C – Total de Sentenças Anotadas para Cada Classificação dos Indicadores

Indicador	Classe	Total de Sentenças
Alimentação	Sem menção	44971
	Via Oral	695
	Sondas Gastronomicas	881
Auto-cuidado	Sem menção	46065
	Dependente para atividades básicas/de higiene	410
	Independente para atividades básicas/de higiene	72
AVC prévio	Sem menção	46309
	Não	29
	Sim	209
Braden	Sem menção	46287
	Baixo Risco	61
	Risco Médio	59
	Risco Alto	140
Capacidade Cognitiva	Sem menção	45788
	Não estado de clareza	124
	Lúcido, orientado, coerente	385
	Confuso, compreende mal	250
Comunicação	Sem menção	45413
	Sem comunicação verbal	175
	Comunica bem	269
	Comunica pouco ou mal	690
Câncer	Sem menção	46300
	Doença ausente	189
	Doença presente	58
Diabetes	Sem menção	46193
	Doença ausente	46
	Doença presente	308
Dislipidemia	Sem menção	46404
	Doença ausente	8
	Doença presente	135
Doença Coronária	Sem menção	46231
	Doença ausente	24
	Doença presente	292
Dor	Sem menção	45910
	Sem Dor	555
	Dor leve/moderada	58
	Dor intensa	23
Etilismo	Sem menção	46438
	Não Etilista	22
	Etilista	63
	Ex Etilista	24
Fibrilação Atrial	Sem menção	46255
	Doença ausente	14
	Doença presente	278

Grau de Força	Sem menção	45857
	escore 0	143
	escore 1	19
	escore 2	32
	escore 3	47
	escore 4	252
	escore 5	197
Hemorragia Intracraniana	Sem menção	46331
	Hemorragia ausente	87
	Hemorragia presente	129
Hipertensão Arterial Sistêmica	Sem menção	45958
	Doença ausente	32
	Doença presente	557
Indicativo de Infecção	Sem menção	45573
	Indicativo negativo de infecção	611
	Indicativo alerta de infecção	363
Localização	Sem menção	45035
	Emergência	109
	Unidade de Internação	1073
	UTI/CTI	330
Mobilidade	Sem menção	45702
	Incapaz de andar	503
	Capaz de andar sem ajuda de outra pessoa ou dispositivo	110
	Capaz de andar com ajuda de outra pessoa ou dispositivo	232
NIH	Sem menção	46227
	Escore de 1 a 42	320
Nível de Mobilidade	Sem menção	45783
	Escore de 1	136
	Escore de 2	24
	Escore de 3	5
	Escore de 4	157
	Escore de 5	0
	Escore de 6	0
	Escore de 7	31
	Escore de 8	71
	Escore de 9	3
	Escore de 10	61
	Escore de 11	167
	Escore de 12	59
	Escore de 13	16
	Escore de 14	12
	Escore de 15	22
Obesidade	Sem menção	46460
	Doença Ausente	58
	Doença presente	28
Paresia	Sem menção	46037
	Registro de paresia	510
Risco de Queda	Sem menção	46100
	Baixo Risco	116
	Moderado Risco	135
	Alto Risco	196
Queda	Sem menção	46525
	Sim, queda	22
Rankin	Sem menção	46358
	Escore de 0 a 6	189

Tabagismo	Sem menção	46264
	Não Tabagista	32
	Tabagista	74
	Ex-Tabagista	177
Trombectomia	Sem menção	46311
	Não realizou ou sem janela para terapia	25
	Trombectomia	211
Trombólise	Sem menção	46048
	Não realizou ou sem janela para terapia	107
	Trombólise	392
Óbito	Sem menção	44212
	Indicativos de presença e sinais vitais	2326
	Nota de óbito	9

ANEXO D – Exemplo 1 do processamento de sentenças pelo algoritmo

```
Sentença de entrada: # EX-TABAGISTA (IT ~ 60 MAÇOS/ANO)  # DPOC?


Sentença após primero tratamento, removendo "-" e "=": # EX TABAGISTA (IT ~ 60 MAÇOS/ANO)
# DPOC?


Vetor de tokens da sentença:
['#', 'EX', 'TABAGISTA', '(IT', '~', '60', 'MAÇOS/ANO)', '', '#', 'DPOC?']


Token sendo avaliado: #


O token só é avaliado se possuir pelo menos 2 caracteres.


Token sendo avaliado: EX


Token após tratamento, removendo espaços em branco e caracteres especiais:
ex


Vetor de palavras similares ao token, obtitas com o modelo word-embeddings:
[('øex', 0.9564199447631836), ('wpwex', 0.9549883604049683), ('iex', 0.9411382079124451),
('ápex', 0.9131456613540649), ('eex', 0.9014943838119507), ('aex', 0.9000544548034668),
('bvex', 0.899472713470459), ('0ex', 0.8882730603218079), ('bppex', 0.8795141577720642),
('fscex', 0.8780627846717834)]


Termos em comum entre a lista de termos de tempo passado da ontologia, a lista de termos
similares e, o token sendo avaliado:
{'ex'}


Token sendo avaliado: TABAGISTA


Token após tratamento, removendo espaços em branco e caracteres especiais:
tabagista


Vetor de palavras similares ao token, obtitas com o modelo word-embeddings:
[('tababagista', 0.993459939956665), ('øtabagista', 0.9909840226173401), ('dtabagista',
0.9899802207946777), ('qdtabagista', 0.9896560907363892), ('tabagagista',
0.9874061942100525), ('ftabagista', 0.9869678020477295), ('tabagista½',
0.9842574000358582), ('tabagista_', 0.9842092990875244), ('tabagistaø',
0.9839856624603271), ('babátabagista', 0.9837406873703003)]


Termos em comum entre a lista de termos da ontologia, a lista de termos similares e, o
token sendo avaliado:
```

```
{'tabagista'}
Token sendo avaliado: (IT


Token após tratamento, removendo espaços em branco e caracteres especiais:
it


Vetor de palavras similares ao token, obtitas com o modelo word-embeddings:
[('ità', 0.8529742956161499), ('itú', 0.852589488029 48), ('ityu', 0.8506131172180176),
('itf', 0.8431515097618103), ('aiit', 0.8383986353874207), ('itd', 0.8228018879890442),
('wait', 0.8186017274856567), ('itun', 0.8115001916885376), ('0nit', 0.8069087266921997),
('itbe', 0.8037105202674866)]


Token sendo avaliado: ~


O token só é avaliado se possuir pelo menos 2 caracteres.


Token sendo avaliado: 60


Token sendo avaliado: MAÇOS/ANO)


Token após tratamento, removendo espaços em branco e caracteres especiais:
maçosano


Vetor de palavras similares ao token, obtitas com o modelo word-embeddings:
[('maçosx', 0.643078625202179), ('maçosxano', 0.6425497531890869), ('maços',
0.6346989274024963), ('00maçosxano', 0.6308329701423645), ('maçoø', 0.6303017735481262),
('maçosdia', 0.6297149062156677), ('maçoc', 0.627691924571991), ('maçotb',
0.6275813579559326), ('maçoøex', 0.6260275840759277), ('0maços', 0.6249396800994873)]


Token sendo avaliado:


O token só é avaliado se possuir pelo menos 2 caracteres.


Token sendo avaliado: #


O token só é avaliado se possuir pelo menos 2 caracteres.


Token sendo avaliado: DPOC?
```

```
Token após tratamento, removendo espaços em branco e caracteres especiais:
dpoc


Vetor de palavras similares ao token, obtitas com o modelo word-embeddings:
[('dpocdpoc', 0.9967538118362427), ('dpocø', 0.9872171878814697), ('ødpoc',
0.9853812456130981), ('0dpoc', 0.9778203964233398), ('dpocgold', 0.9759135842323303),
('dpocs', 0.9751116037368774), ('dpocmi', 0.9742422699928284), ('dpocc', 0.9719513654708862),
('eapxdpoc', 0.9708786010742188), ('dpoch', 0.9687447547912598)]


Lista de sinônimos:
{'ex': 'ex', 'tabagista': 'tabagista'}


Lista de termos sem correspondentes:
['#', 'it', '~', '60', 'maçosano', '', 'dpoc']


Lista das relações a serem adicionadas na ontologia:
[{'relation': 'ex', 'value': 'tabagista', 'valorado': 0, 'position': 3}]



Classificações e relações após processo de inferência:

[file_sentence_sample_1611768001.6380897.Sentença,
 file_sentence_sample_1611768001.6380897.tabagismo2,
 Regras 7.ex_p.some(Regras 7.tabagista)]
```

ANEXO E – Exemplo 2 do processamento de sentenças pelo algoritmo

```
Sentença de entrada: D  # AVC isquêmico previo - sem sequelas aparentes  -mRankin previo: 3
# Demência de Alzheimer - Tem vida de relação, corversa, caminha, alimenta-se.

Sentença após primero tratamento, removendo "-" e "=": D  # AVC isquêmico previo   sem
sequelas aparentes   mRankin previo: 3  # Demência de Alzheimer   Tem vida de relação,
corversa, caminha, alimenta se.

Vetor de tokens da sentença:
['D', '', '#', 'AVC', 'isquêmico', 'previo', '', '', 'sem', 'sequelas', 'aparentes', '', '',
'mRankin', 'previo:', '3', '', '#', 'Demência', 'de', 'Alzheimer', '', '', 'Tem', 'vida',
'de', 'relação,', 'corversa,', 'caminha,', 'alimenta', 'se.']

Token sendo avaliado: AVC

Token após tratamento, removendo espaços em branco e caracteres especiais:
avc

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('avchø', 0.9796327352523804), ('øavc', 0.9783550500869751), ('avcvjd', 0.9780991077423096),
('rbeavc', 0.9763976335525513), ('avceny', 0.975973904132843), ('avcsd', 0.9729700088500977),
('uavc', 0.9720901846885681), ('avcc', 0.9717864990234375), ('iavc', 0.9698855876922607),
('avchp', 0.9698476195335388)]

Termos em comum entre a lista de termos da ontologia, a lista de termos similares e, o token
sendo avaliado:

{'avc'}
Token sendo avaliado: isquêmico

Token após tratamento, removendo espaços em branco e caracteres especiais:
isquêmico

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('isquêmicoø', 0.9794929623603821), ('isquêmicooe', 0.9748890995979309), ('isquêmicoi',
0.9733677506446838), ('isquêmicop', 0.9684271216392517), ('isquêmicoc', 0.9657162427902222),
('isquêmicoem', 0.9656476974487305), ('squêmico', 0.9651217460632324), ('isquêmicoeeg',
0.9550307393074036), ('isquêmicofos', 0.9543636441230774), ('isquêmicoatbs',
0.9519003033638)]

Token sendo avaliado: previo

Token após tratamento, removendo espaços em branco e caracteres especiais:
previo

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('previoo', 0.9644258618354797), ('0previo', 0.9581046104431152), ('previoant', 0.9531316757202148), ('previop',
0.9522475600242615), ('previoem', 0.9495493769645691), ('previocaf', 0.9479995369911194),
```

```
('previos0', 0.943787693977356), ('previoc', 0.9415261745452881), ('previofaz',
0.9354027509689331), ('previodi', 0.9321310520172119)]

Token sendo avaliado: sem

Token após tratamento, removendo espaços em branco e caracteres especiais:
sem

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('semsem', 0.9402598142623901), ('_sem', 0.9209915399551392), ('uhsem', 0.9201294183731079),
('xsem', 0.9195883870124817), ('gsem', 0.9193295240402222), ('ksem', 0.9192111492156982),
('xcsem', 0.9182100892066956), ('csem', 0.9171376824378967), ('tzbsem', 0.9171339869499207),
('0qsem', 0.9169884324073792)]

Termos em comum entre a lista de termos de negações da ontologia, a lista de token similares
e, o token sendo avaliado:
{'sem'}

Token sendo avaliado: sequelas

Token após tratamento, removendo espaços em branco e caracteres especiais:
sequelas

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('rsequelas', 0.9825713634490967), ('hsdcsequelas', 0.9758013486862183), ('sequelasø',
0.974330484867096), ('asequelas', 0.9679824709892273), ('sequelasp', 0.9663152694702148),
('sequelasmi', 0.9605541825294495), ('sequelasc', 0.9604564905166626), ('sequelash',
0.9604523181915283), ('sequelasoa', 0.9550919532775879), ('sequelashá', 0.9547930359840393)]

Token sendo avaliado: aparentes

Token após tratamento, removendo espaços em branco e caracteres especiais:
aparentes

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('aparentesav', 0.9489852786064148), ('aparenteshunt', 0.9469876289367676), ('aparentesrcp',
0.9463257193565369), ('aparentesa', 0.9451962113380432), ('aparentesdisp',
0.9402044415473938), ('aparentesi', 0.9369718432426453), ('aparentesfaces',
0.9319012761116028), ('aparentescr', 0.9279047250747681), ('aparentestv', 0.9189890027046204),
('aparentescvc', 0.9174056649208069)]

Token sendo avaliado: mRankin

Token após tratamento, removendo espaços em branco e caracteres especiais:
mrankin

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
```

```
[('akin', 0.8826376795768738), ('mrankin0', 0.8731066584587097), ('skin', 0.8689863681793213),
('hodgnkin', 0.8491312265396118), ('hodjkin', 0.8473354578018188), ('hogdkin',
0.8414438366889954), ('rankin', 0.8373583555221558), ('hosggkin', 0.821370542049408),
('hodgin', 0.820633053779602), ('hodkin', 0.8202835321426392)]

Termos em comum entre a lista de termos valorados da ontologia, a lista de termos similares e,
o token sendo avaliado:

{'rankin'}
Token sendo avaliado: previo:

Buscando por um valor numérico, token sendo avaliado: previo:

Token sendo avaliado: 3

Buscando por um valor numérico, token sendo avaliado: 3

Valor numérico detectado: 3.00

Token sendo avaliado: Demência

Token após tratamento, removendo espaços em branco e caracteres especiais:
demência

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('ødemência', 0.9893346428871155), ('thbdemência', 0.9826850891113281), ('demênciaolga',
0.9709343910217285), ('demênciaadão', 0.9664850831031799), ('demênciaem', 0.9660176634788513),
('isqdemência', 0.9611543416976929), ('00demência', 0.9593740105628967), ('dmhasdemência',
0.9540921449661255), ('demênciahiv', 0.9534928202629089), ('demênciaataxia',
0.9530728459358215)]

Token sendo avaliado: de

Token após tratamento, removendo espaços em branco e caracteres especiais:
de

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('dee', 0.7262868285179138), ('dew', 0.7176570296287537), ('deh', 0.7134923338890076),
('deø', 0.7095134854316711), ('de0', 0.7051140666007996), ('deom', 0.7026863098144531),
('dews', 0.6960441470146179), ('dersde', 0.6705716848373413), ('deem', 0.6678256988525391),
('dejavu', 0.6601331830024719)]

Token sendo avaliado: Alzheimer

Token após tratamento, removendo espaços em branco e caracteres especiais:
alzheimer

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('alzhheimer', 0.986768364906311), ('alzheimerø', 0.9817374348640442), ('alhzheimer', 0.977109968662262), ('alzeheimer',
0.9728602170944214), ('alzheimers', 0.9682837128639221),
```

```
0.9524657130241394), ('alzheimerem', 0.9436348080635071), ('alzheimeruso',
0.9414306282997131), ('alzheirmer', 0.9379398822784424)]

Token sendo avaliado: Tem

Token após tratamento, removendo espaços em branco e caracteres especiais:
tem

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('temem', 0.9201308488845825), ('vtem', 0.9177613854408264), ('dtem', 0.9083892107009888),
('item', 0.9083139300346375), ('vdtem', 0.9043123126029968), ('ihtem', 0.9004751443862915),
('temb', 0.90020352602005), ('0dtem', 0.8977198004722595), ('omtem', 0.8886024951934814),
('vdptem', 0.8884862661361694)]

Token sendo avaliado: vida

Token após tratamento, removendo espaços em branco e caracteres especiais:
vida

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('vidapn', 0.9396309852600098), ('vidarn', 0.9343447089195251), ('àvida',
0.9229457974433899), ('dívida', 0.9196599125862122), ('ávida', 0.9116578698158264),
('vidamãe', 0.9041496515274048), ('vidad', 0.8952152729034424), ('vidal', 0.8886168003082275),
('benvida', 0.8782338500022888), ('duvida', 0.872809886932373)]

Token sendo avaliado: de

Token após tratamento, removendo espaços em branco e caracteres especiais:
de

O token de já foi avaliado e não foi encontrado nenhuma expansão ou correspondência e não
precisa ser expandido novamente

Token sendo avaliado: relação,

Token após tratamento, removendo espaços em branco e caracteres especiais:
relação

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('relaçãom', 0.9800077676773071), ('relaçãop', 0.976003885269165), ('0relação',
0.974817156791687), ('relaçãp', 0.9743456840515137), ('relaçã', 0.9737211465835571),
('relaçãoa', 0.9730480313301086), ('relaçãomed', 0.972754716873169), ('erelação',
0.9706792831420898), ('emrelação', 0.969520092010498), ('relaçlão', 0.9606861472129822)]

Token sendo avaliado: corversa,

Token após tratamento, removendo espaços em branco e caracteres especiais:
corversa
```

```
Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('versa', 0.5933812260627747), ('anversa', 0.5929152369499207), ('disversa',
0.5883746147155762), ('coversarei', 0.5797482132911682), ('tranversa', 0.5761317610740662),
('trasversa', 0.5751370787620544), ('reversa', 0.5739848613739014), ('corveloeq',
0.5732448101043701), ('cordioversao', 0.5730204582214355), ('adversa', 0.5727380514144897)]

Token sendo avaliado: caminha,

Token após tratamento, removendo espaços em branco e caracteres especiais:
caminha

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('dcaminha', 0.9635627865791321), ('caminhaa', 0.9633505940437317), ('caminhava',
0.9534345269203186), ('caminharo', 0.9349479079246521), ('caminhao', 0.9324288964271545),
('caminhaenmg', 0.9252079725265503), ('caminhão', 0.9248256087303162), ('caminhandoo',
0.908381998538971), ('caminhas', 0.9042920470237732), ('camisinha', 0.9034737348556519)]

Termos em comum entre a lista de termos da ontologia, a lista de termos similares e, o token
sendo avaliado:

{'caminha'}
Token sendo avaliado: alimenta

Token após tratamento, removendo espaços em branco e caracteres especiais:
alimenta

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('alimentamndo', 0.9646124243736267), ('alimentanndo', 0.9616190195083618), ('alimentá',
0.9612510204315186), ('alimentava', 0.960787832736969), ('alimentarp', 0.9590476155281067),
('alimentars', 0.9556906819343567), ('alimentandi', 0.9505292773246765), ('alimentara',
0.948761522769928), ('alimentari', 0.9485595226287842), ('alimentaro', 0.9484350681304932)]

Token sendo avaliado: se.

Token após tratamento, removendo espaços em branco e caracteres especiais:
se

Vetor de palavras similares ao token, obtidas com o modelo word-embeddings:
[('vfse', 0.7693838477134705), ('mhse', 0.7664098143577576), ('jóse', 0.7565931081771851),
('fse', 0.7538988590240479), ('vse', 0.750080943107605), ('dse', 0.7478148937225342), ('seá',
0.7458064556121826), ('zuse', 0.742520809173584), ('bdse', 0.7418352365493774), ('nse',
0.739757776260376)]

Lista de sinônimos:
{'avc': 'avc', 'sem': 'sem', 'mrankin': 'rankin', 'caminha': 'caminha'}
```

```
Lista de termos sem correspondentes:
['D', '', '#', 'isquêmico', 'previo', 'sequelas', 'aparentes', '3', 'demência', 'de',
'alzheimer', 'tem', 'vida', 'relação', 'corversa', 'alimenta', 'se']

Lista das relações a serem adicionadas na ontologia:
[{'relation': 'contain', 'value': 'avc', 'valorado': 0, 'position': 4}, {'relation': 'rankin',
'value': '3.00', 'valorado': 1, 'position': 16}, {'relation': 'sem_p', 'value': 'caminha',
'valorado': 0, 'position': 29}]

Classificações e relações após processo de inferência:

 [file_sentence_sample_1611783632.5272691.Sentença,
 file_sentence_sample_1611783632.5272691.ClasseRankin,
 Regras 7.contain.some(Regras 7.avc),
 Regras 7.rankin.value(3.0),
 Regras 7.sem_p.some(Regras 7.caminha)]
```

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br