CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

DETECÇÃO DE SINAIS DE EVENTOS ADVERSOS DE MEDICAMENTOS EM TEXTOS INFORMAIS

Alexandre Martins da Cunha

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador(a): Gustavo Paiva Guedes da Silva

Rio de Janeiro,

Dezembro 2019

DETECÇÃO DE SINAIS DE EVENTOS ADVERSOS DE MEDICAMENTOS EM TEXTOS INFORMAIS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Alexandre Martins da Cunha

Banca Examinadora:

Presidente, Prof. Gustavo Paiva Guedes e Silva, D.Sc. (orientador)

Profa. Kele Teixeira Belloze, D.Sc. (CEFET/RJ)

Prof. Eduardo Soares Ogasawara, D.Sc. (CEFET/RJ)

Prof. Fellipe Duarte, D.Sc.

Universidade Federal Rural do Rio de Janeiro - UFRRJ

Rio de Janeiro,

Dezembro 2019

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

C972 Cunha, Alexandre Martins da
Detecção de sinais de eventos adversos de medicamentos em textos informais / Alexandre Martins da Cunha.—2019.
133f., il., color., tabs.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2019.
Bibliografia : f. 110-133
Orientador: Gustavo Paiva Guedes da Silva

1. Computação. 2. Processamento de linguagem natural (Computação). 3. Computação semântica. I. Silva, Gustavo Paiva Guedes da (Orient.). II. Título.

CDD 004

Elaborada pela bibliotecária Teresa Cristina Gaio Mattos – CRB/7 nº 4610

DEDICATÓRIA

Dedico este trabalho primeiramente a deus por permitir minha existência; Aos meus pais, Manuel e Maria, pelo carinho e afeto que toda criança precisa e pelo sacrifício para me proporcionar educação; Ao meu irmão Guilherme, pelo apoio dado no decorrer dessa obra; À minha noiva por compreender as ausências no período dessa obra. Te amo, Débora!

AGRADECIMENTOS

A todos os meus amigos do PPCIC, com os quais pude compartilhar as experiências e desafios da vida acadêmica e científica;
À Carlos Teles e Carlos Vianna, pela amizade e apoio, muitos momentos de apoio, bons amigos;
À Fernanda Fêbs, por me auxiliar em vários momentos ao longo de minha jornada, dentro e fora do mundo acadêmico, uma amiga de peso;
Ao Flávio Matias, por sempre me ajudar com a preocupação e expectativa de sucesso, um bom amigo;
Ao Rafael Guimarães que apontou, em várias oportunidades, qual caminho seguir, apoiando minhas decisões: um grande amigo;
Ao Gabriel Nascimento pela confiança depositada, amizade e dedicação sem as quais eu teria desistido no decorrer do processo, meu muito obrigado!
Aos membros da secretaria, Ivan, Sheila e Bráulio pela presteza e gentileza no atendimento;
A todos os docentes do CEFET-RJ, que se dedicaram a ensinar e transferir seu saber, só tenho a agradecer pelas experiências adquiridas na vivência desse período;
Aos professores Eduardo Ogasawara e Joel dos Santos, pela presteza e gentileza à frente do PPCIC;
Ao professor Jorge Soares que, muito gentilmente, me recebeu como seu orientando quando estava no início de minha jornada, mas, por motivos de força maior, não foi possível prosseguir;
À professora Kele Belloze, pelo carinho, dedicação e paciência despendidos ao longo da jornada do saber e pela ajuda na etapa final desse trabalho;
Ao professor Gustavo Guedes, pela viabilidade dessa dissertação, por seu exemplo pessoal e profissional, sua orientação, amizade, atenção e paciência (muita paciência, por sinal); muito obrigado, foi uma honra ser seu orientando. Exemplo de professor e pessoa;
À banca examinadora, pelo convite aceito para a avaliação deste trabalho.

RESUMO

DETECÇÃO DE SINAIS DE EVENTOS ADVERSOS DE MEDICAMENTOS EM TEXTOS INFORMAIS

A vigilância em saúde, conhecida como farmacovigilância, se define como: “a ciência e as atividades relativas à identificação, avaliação, compreensão e prevenção dos efeitos adversos ou qualquer outro problema relacionado com medicamentos” WHO [2002]. Eventos adversos a medicamentos são responsáveis por aproximadamente 25% dos pacientes internados no atendimento primário, sendo considerados graves em 13% dos casos Meyboom et al. [1999]. A farmacovigilância atua no período de pós-aprovação do medicamento, podendo evitar e atenuar certos eventos adversos. O acesso às várias categorias de dados de saúde no período atual expande a capacidade de análise para pesquisa relacionada a farmacovigilância. Com o advento das técnicas de mineração de texto (MT), processamento de linguagem natural (PLN), aprendizagem de máquina (AM) e extração da informação (EI), houve a possibilidade de extração de conhecimento de textos não estruturados e informais, obtidos de mídias sociais. O objetivo desta dissertação é, ao utilizar a extração da informação, criar um modelo a partir da MT e PLN para detectar sinais de eventos adversos em medicamentos nos textos da mídia social (Twitter) escritos em português do Brasil. A dissertação apresenta extensa revisão bibliográfica sobre os conceitos citados. Guiando o processo, foi desenvolvida uma abordagem baseada na metodologia de MT para identificar possíveis sinais de eventos adversos. Esse processo foi implementado com auxílio do CoreNLP. Para essa dissertação, foi escolhido o idioma português brasileiro, para o qual não existe suporte nativo do CoreNLP, dessa forma, foram implementados o analisador sintático (Pos-Tagger) e o parse de dependência (DEP-PARSER) em português brasileiro. Também foi treinado um modelo de detecção de entidades nomeadas no domínio da farmacovigilância em português Brasileiro utilizando AM em uma abordagem híbrida. Foi proposto um algoritmo para efetiva detecção de sinal de eventos adversos em medicamentos. Complementa-se a metodologia com a experimentação dos modelos criados e do algoritmo desenvolvido. Os resultados representam um esforço inicial na tentativa de atuar sobre o idioma português brasileiro no campo da farmacovigilância. Os experimentos abriram caminho para fomentar o tema e fornecer um instrumental para caminhar em direção ao estado da arte, especificamente para a língua portuguesa.

Palavras-chave: REN; MT; EAM.

ABSTRACT

DETECT SINGS OF ADVERSE DRUG EVENTS IN INFORMAL TEXT

Health surveillance, known as pharmacovigillance, is defined as: “He science and activities of identifying, assessing, understanding and preventing adverse effects or any other drug-related problem” WHO [2002]. Adverse drug events account for approximately 25% of patients admitted to hospitals and clinics in primary care, and are considered severe in 13% of cases Meyboom et al. [1999]. Pharmacovigillance acts in the post-approval period of the drug and can prevent and mitigate certain adverse events. Access to the various categories of health data in the current period expands the analytical capacity for pharmacovigillance-related research. With the advent of text mining (TM), natural language processing (NLP), machine learning (ML) and information extraction (IE) techniques, it was possible to extract knowledge from unstructured and informal texts obtained from social media. The purpose of this dissertation is, by using IE, to create a model from TM and NLP, and to detect signs of adverse events in medicines in social media (Twitter) texts written in Brazilian Portuguese. The dissertation presents an extensive literature review about the concepts mentioned. In order to guide the process, an approach based on the MT methodology was developed to identify possible signs of adverse events. This process was implemented with the help of CoreNLP. For this dissertation, it was chosen the Brazilian Portuguese language, for which there is no support of CoreNLP, so it was implemented the parser (Pos-Tagger) and dependency parse (DEP-PARSER) in Brazilian Portuguese, as well as trained a model of detection of named entities in the domain of pharmacovigillance in Brazilian Portuguese, using ML in a hybrid approach. With the models produced, an algorithm for effective signal detection of adverse events in drugs is proposed. The methodology is complemented with the experimentation of the created models and the developed algorithm. The results demonstrate an innitial effort in trying to act on the Brazilian Portuguese language in the field of pharmacovigillance. The experiments paved the way to promote the theme and provide an instrument to move towards the state of the art, specifically for the Portuguese Language.

Keywords: NER; TM; ADE.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ALGORITMOS

Lista de abreviações

ACE	Extração Automática De Conteúdo
AJAX	Javascript E XML Assíncronos
AM	Aprendizagem De Máquina
ANVISA	Agência Nacional De Vigilância Sanitária
CBOW	Continuous Bag Of Words - ‘‘Saco De Palavras Contínuo”
CIOMS	Conselho De Organizações Internacionais De Ciências Médicas
CORENLP	Biblioteca De Stanford Para Processamento De Linguagem Natural
CRF	Campos Aleatórios Condicionais
CSS	Folhas De Estilo Em Cascata
DCI	Denominação Comum Internacional
DEPPARSER	Parser De Dependência (Analisador Sintático)
EAM	Eventos Adversos Aos Medicamentos
EI	Extração Da Informação
EN	Entidades Nomeadas
EUA	Estados Unidos Da América
HAREM	Avaliação De Reconhecedores De Entidades Mencionadas
HMM	Modelo Oculto De Markovl
HTML	Linguagem De Marcação De Hipertexto
IA	Inteligência Artificial
ISC	Infecções De Sítio Cirúrgico
LAS	Pontuação De Anexo Etiquetado
MD	Mineração De Dados
MT	Mineração De Textos
MUC	Conferência De Compreensão Da Mensagem
OMS	Organização Mundial De Saúde
PLN	Processamento De Linguagem Natural
POSTAGGER	Etiquetador De Parte Do Discurso (Analisador Morfológico)
RAM	Reações Adversas Aos Medicamentos
REN	Reconhecimento De Entidades Nomeadas

RI	Recuperação De Informação
UAS	Pontuação De Anexo Não Etiquetado
WE	Word Embeddings - Vetor De Palavras

SUMÁRIO

Introdução

Os medicamentos são uma importante fonte terapêutica no tratamento e prevenção de diversas doenças. A fim de obter êxito e desenvolver os resultados como esperados, se faz necessário o uso correto dos medicamentos e de sua prescrição, posologia, tempo de uso e, ainda, que o paciente siga todo o processo à risca [Marin et al., 2003].

A repercussão da medicação, estando em uma sociedade, tem muitas peculiaridades [Pfaffenbach et al., 2002]. A partir de um ponto de vista, os medicamentos colaboram para uma maior expectativa de vida ao auxiliar na extinção de doenças e permitir vantagens econômicas e sociais [Pfaffenbach et al., 2002]; sob outra perspectiva, apresentam elevação de despesas com saúde devido sua utilização inadequada, o que acarreta Reações Adversas aos Medicamentos (RAM) [Pfaffenbach et al., 2002].

Ainda que sejam obrigados a seguir um protocolo de desenvolvimento rigoroso e altamente regulamentado antes de serem autorizados para comercialização e que sejam utilizados de forma apropriada, os medicamentos podem provocar eventos indesejáveis ao longo do tratamento. No que tange à segurança de medicamentos, o domínio é conhecido por farmacovigilância. Para a Organização Mundial de Saúde (OMS) em inglês, *WHO*. WHO [2002], ... :

> *"ciência e atividades relativas à identificação, avaliação, compreensão e prevenção de efeitos adversos ou quaisquer problemas relacionados ao uso de medicamentos."*

No Brasil, a entidade responsável pela farmacovigilância é a Agência Nacional de Vigilância Sanitária (ANVISA), sendo sua função identificar, avaliar e monitorar os eventos adversos no uso de medicamentos em solo brasileiro. Os ensaios clínicos controlados que são conduzidos antes da concessão da autorização de mercado envolvem coleta e análises sistemáticas e organizadas de dados de Eventos Adversos aos Medicamentos (EAM), bem como de outros dados relevantes para a segurança do medicamento e são realizados em três distintas fases: I, II e III [Nunes, 2000].

Conforme Nunes [2000], temos três fases para segurança de medicamentos. A fase I, medicamentos em indivíduos saudáveis, que não possuem a doença a qual propõe-

se a tratar. Já a fase II indivíduos que a doença, originando um estudo piloto. Por fim a fase III é amplia a fase II para um grande número de indivíduos por um período de tempo mais longo, em geral comparando com o tempo dos estudos já existentes.

Embora os ensaios clínicos controlados sejam considerados uma característica da demonstração da eficácia de um medicamento, os dados de segurança disponíveis nos estudos de Nunes [2000]; Venulet and Ten [1996]; Pedrós and Tognoni [1993],têm limitações bem reconhecidas. Pode-se citar: o número limitado de participantes inclusos nos estudos (dado o tamanho das populações de pacientes expostos ao medicamento quando no mercado), a duração limitada da exposição do sujeito de estudo ao medicamento (caso particular de medicamento destinado para uso a longo prazo, muito comum em doenças crônicas como asma e diabetes), dados limitados ou inexistentes para pacientes com risco mais alto e populações que são frequentemente excluídas de ensaios clínicos controlados (por exemplo, pacientes com comprometimento de órgãos, pacientes pediátricos, geriátricos, grávidas e lactantes) [Nunes, 2000; Venulet and Ten, 1996; Pedrós and Tognoni, 1993].

Essas limitações exigem que o detentor da autorização de comercialização de medicamentos, continue coletando, analisando e interpretando dados relativos à segurança do paciente. Estes dados, se tornam disponíveis após a introdução do medicamento no mercado, fase conhecida como pós-comercialização.

Em alguns estudos, os dados são obtidos por meio de notificação voluntária. A notificação voluntária é adotada mundialmente como um padrão no âmbito da segurança de pacientes e consiste na coleta e comunicação de EAM [Dias, 2005]. Alguns lugares do mundo utilizam dados de pesquisas realizados pela indústria e por instituições de ensino [Nunes, 2000] para auxiliarem na detecção de EAM em condições clínicas, sendo mais comumente relacionados às RAM graves ou letais, de fácil identificação ou que ocorrem em curto período de tempo [Nunes, 2000].

Conforme a ANVISA [2016], *EAM* são definidas como:

> *"qualquer ocorrência médica desfavorável que pode ocorrer durante o tratamento com um medicamento, mas que não possui, necessariamente, relação causal com esse tratamento. O conceito de evento adverso é amplo, abrangendo uma série de problemas relacionados ao uso dos medicamentos."*

já as *RAM* são:

> *"qualquer resposta prejudicial ou indesejável, não intencional, a um medicamento, que ocorre nas doses usualmente empregadas para profilaxia, diagnóstico ou terapia de doenças. No conceito de RAM pode-se observar a existência de uma relação causal entre o uso do medicamento e a ocorrência do problema."*

Sendo assim todas as RAM são uma forma de EAM, todavia, nem todos os EAM são uma RAM [Stephens, 2011]. Para ilustrar a diferença, os EAM incluem além das RAM, desvios de qualidade do medicamento, uso *"off label"*[1], interações medicamentosas como as que ocorrem na *polifarmácia*[2], ineficiência terapêutica, intoxicação, uso abusivo e erros de medicação [ANVISA, 2009].

Para Hauben and Aronson [2009], a terminologia "sinal" pode auxiliar na clareza e consistência da comunicação sobre segurança de medicamentos a pacientes, profissionais de saúde, laboratórios e entidades reguladoras. Portanto, é essencial estabelecer uma definição comum e clara de um sinal de segurança. Desse modo, o oitavo grupo de trabalho do Conselho de Organizações Internacionais de Ciências Médicas (CIOMS) [Council et al., 2010], definiu o sinal referente a segurança de medicamentos, adaptado da obra de [Hauben and Aronson, 2009], como:

> *"Informações que surgem de uma ou várias fontes (incluindo observações e experimentos), que sugerem uma nova associação potencialmente causal, ou um novo aspecto de uma associação conhecida, entre uma intervenção e um evento ou conjunto de eventos relacionados, seja adverso ou benéfico, que é considerado de probabilidade suficiente para justificar a ação verificadora."*

Os métodos computacionais são os algoritmos utilizados no domínio da farmacovigilância que são conhecidos por *detecção de sinais* que proporcionam aos profissionais de saúde mensurar a segurança dos medicamentos em grande escala para reconhecer potenciais sinais de EAM. Eles provaram, ao longo do tempo, que são um componente importante no domínio da farmacovigilância. Os sinais, sozinhos, não servem para aferir uma relação de causa e efeito entre os elementos então, requerem a intervenção do especialista para descrever essa relação [Harpaz et al., 2012].

As pesquisas em computação têm por um dos objetivos, gerar conhecimento que permita desenvolver e implementar artefato para lidar com problemas anteriormente não

[1] Ainda sem tradução oficial para o português, usa-se o termo "off label" para se referir ao uso diferente do aprovado em bula ou ao uso de produto não registrado no órgão regulatório de vigilância sanitária no país.

[2] polifarmácia - utilização de mais de um medicamento para diagnósticos distintos.

resolvidos ou pouco explorados [Hevner, 2007]. Ao longo do tempo, dados provenientes das redes sociais têm atraído o interesse de pesquisadores de diversas áreas. Entretanto, o extenso volume de material em formato textual presente na internet, potencializado pelo uso massivo de redes sociais, inviabiliza a análise desse conjunto de dados de forma manual, sendo assim, técnicas computacionais são empregadas.

Motivação

Estudos internacionais atestam que as RAM podem relacionar-se às internações hospitalares e são causadoras de $0,5\%$ a $32,9\%$ das hospitalizações em entidades de saúde [Dormann et al., 2003; Pirmohamed et al., 2004; Koh et al., 2005; Van et al., 2006; Patel et al., 2007; Zopf et al., 2008]. Já no Brasil, sabe-se que $0,56\%$ a $54,5\%$ das hospitalizações decorrentes do uso de medicamentos estão, provavelmente, relacionadas a sintomas de RAM [Pfaffenbach et al., 2002; Mastroianni et al., 2009; Fabiana et al., 2011; Noblat et al., 2011] Dessas hospitalizações, 43% dos pacientes estão em risco de adquirir RAM intra-hospitalar [Camargo et al., 2006]. A variação da predominância nos estudos deve-se pelas características do público-alvo nas pesquisas, assim como, pelos diferentes métodos para detecção de causa [Hallas et al., 1992; Beijer and CJ, 2002; Zolezzi and Parsotam, 2005].

Nas últimas décadas, pesquisas demonstraram que tanto a mortalidade quanto a morbidade por utilização de medicamentos são grandes problemas de saúde já conhecidos pela população. Como se sabe, as RAM são a quarta maior causa de mortalidade nos Estados Unidos da América (EUA) [Lazarou et al., 1998]. Além disso, serviços adequados para lidar com as RAM ocasionam gastos com saúde, tendo em vista o tratamento realizado a partir de problemas relacionados a medicamentos [Lazarou et al., 1998]. Alguns lugares no mundo dispõem de 15% a 20% do seu orçamento para lidar com tais complicações [Lazarou et al., 1998].

Cerca de 25% dos pacientes em instituições de saúde do atendimento básico padecem de EAM, sendo este grave em 13% dos casos [Meyboom et al., 1999]. Os EAM são responsáveis por 5% a 10% das internações em hospitais [Mota, 2017] e são capazes de provocar danos temporários e/ou permanentes e, até mesmo, levar a óbito [Griffin and

Resar, 2009], sendo considerados uma das principais causas de adoecimento e morte no mundo [Mota, 2017].

Lazarou et al. [1998] nos trazem um dado muito relevante: outro fator que contribui para os EAM é o fenômeno polifarmácia, responsável por um aumento, entre os anos de 2000 e 2006, de 23% para 26%, respectivamente, nos efeitos adversos associados ao uso de, no mínimo, cinco medicamentos, sem prescrição. Para a população idosa (maiores de 65 anos), foi registrado um aumento, nos mesmo anos, de $6,3\%$ para 12%, associado ao uso de, no mínimo, dois medicamentos prescritos. Estes, fazem maior uso de medicação, com destaque para, aproximadamente, 18% que tomam, ao menos, dez por semana. No Brasil, estima-se que, no mínimo, 35% dos medicamentos são oriundos da automedicação, sendo comum a reutilização de receituário para tal [Barros, 1995].

O problema da polifarmácia é mais elevado na faixa etária da população idosa, a frequência de EAM é maior, aumentando significativamente em função da dificuldade da terapia. O risco aumenta em 13% com dois medicamentos, quando utilizado cinco, o risco sobe para 58% e a partir de sete, 82%. Como consequência ocorre um problema de saúde pública, devido o aumento da taxa de mortalidade. No Brasil a polifarmácia, ocorre indiscriminadamente em diversas cidades [Kennerfalk et al., 2002; Prybys and Gee, 2002; Cassiani et al., 2005].

Dados textuais são recursos muito úteis para obter informações sobre determinado domínio de conhecimento. No campo da farmacovigilância, um grande volume dos dados encontra-se em formato textual como os informes da ANVISA[3], publicações em redes sociais, artigos científicos e bulas de medicamentos. Esses dados, são de grande auxílio na detecção de sintomas de EAM não previstas para o medicamento, assim como para a polifarmácia [Masnoon et al., 2017]. Entretanto para extrair o texto, automaticamente, e estabelecer relações entre os elementos, faz-se necessário identificar as entidades dos domínios existentes no texto (podemos citar, por exemplo, medicamento, sintoma, doença). Sendo assim, uma abordagem computacional é utilizada. Textos informais, comumente empregados em redes sociais, apresentam um desafio em aberto, em geral, pela falta de pontuação, emprego de gírias e erros de ortografia. Isso acaba dificultando o emprego de alguma técnica computacional.

Conforme Amado [2008], o português do Brasil é o quinto idioma mais falado no mundo, com cerca de 230 milhões de falantes, sendo o idioma oficial de 10 países,

[3]ANVISA - Agência Nacional de Vigilância Sanitária - mais detalhes em: `http://portal.anvisa.gov.br`

presente em todos os continentes. A organização das nações unidas para a educação, a ciência e a cultura (UNESCO), proclamou o dia 5 de maio como **Dia Mundial da Língua Portuguesa** [UNESCO, 2019].

Uma língua tão relevante como o português deve ser melhor explorada. Os recursos de criação de conjuntos de dados, anotação dos dados, o Reconhecimento de Entidades Nomeadas (REN) o algoritmo semântico compõe um conjunto de ferramentas presentes em língua inglesa que não existem em língua portuguesa. A criação desses artefatos contribuem para a exploração da língua computacionalmente.

Descrição do problema

O estudo da EMC Corporation®[4] [Gantz and Reinsel, 2012] aponta que a quantidade de dados gerados, gerenciados, armazenados e processados com o uso de tecnologia, no cotidiano das pessoas, produz uma descomunal massa de dados não estruturados. A maior parte desses dados é criada por usuários para usuários, por meio de mídias sociais, televisão digital, transferência de multimídia por *smartphones* e internet e, assim sucessivamente. As empresas são responsáveis por cerca de 80% desses dados, por questões de direitos autorais, legais e privacidade.

Para manipular tais dados, a fim de encontrar sinais de EAM, um processo automatizado deve ser empregado. Por exemplo, o Processamento de Linguagem Natural (PLN). O PLN é um ramo da ciência da computação que relaciona máquinas e linguagem humana, isso é, faz menção a como programar computadores para interpretar e manipular linguagem natural [Chowdhury, 2003]. O PLN faz alusão à outras áreas do saber como, linguística, matemática e computação, tendo várias aplicações que mesclam essas ciências: tradução automática de uma linguagem natural para outra, conversão de texto em fala, sumarização automática e processamento de texto em linguagem natural [Reshamwala et al., 2013].

Existem algumas abordagens para extrair relações entre medicamentos e eventos adversos [Nikfarjam et al., 2015; Gurulingappa et al., 2012a]. O Sistema de Linguagem

[4] EMC Corporation®, líder mundial em armazenamento e gerenciamento da informação - Mais detalhes em: `www.dellemc.com`

Médica Unificada (UMLS)[5] inclui ontologias em português, vocabulário de assuntos médicos (MSHPOR)[6] e dicionário médico para atividades regulatórias em Português (MDRPOR)[7], mantidos pela biblioteca de medicina (NLM)[8] dos EUA [Bodenreider, 2004]. Entretanto, o MDRPOR é a tradução da MeDRA[9] [NLM, 2019] e o MSHPOR [Bireme, 2019b] usa os descritores em ciência da saúde (DeCS) [Bireme, 2019a] que foi criado com base no MeSH[10], sendo também uma tradução.

Bireme [2019a] participa do desenvolvimento da UMLS, sendo a mantedora da MSHPOR. O UMLS foi criado para textos formais (artigos de revistas científicas, livros, anais de congressos, relatórios técnicos, e outros tipos de materiais) [Bodenreider, 2004], ou seja, não seria razoável supor que usuários de redes sociais iriam utilizar essa terminologia complexa e o autor verificou que tanto MDRPOR e MSHPOR não contêm nomes de medicamentos comerciais nacionais (novalgina, buscopam, etc). Aliado ao fato de não obrigação no uso dessa terminologia, como apresenta o artigo [Hasan et al., 2015], seria necessário mapear a Classificação Internacional de Doenças (CID 10)[11], o que demanda uma equipe interdisciplinar. Devido a essas limitações essa abordagem não foi empregada.

Dito isso, no limite do conhecimento do autor, não foram encontrados estudos literários que objetivem extrair relações entre medicamentos e eventos adversos para a língua portuguesa brasileira de textos informais. É um diferencial usar textos em português brasileiro devido à escassez de dados no domínio da farmacovigilância.

Conforme consta no dicionário Aulete [Lexikon, 2015], definimos língua como:

> *"Sistema de comunicação e expressão verbal de um povo, nação, país etc., que permite aos usuários expressar pensamentos, desejos e emoções; IDIOMA."*

Desse modo, podemos concluir que a língua é viva, sendo uma construção social em constante mudança, que ocorre nas ruas, no meio acadêmico, em sermões, palestras, seja falada ou escrita. Para que ela tenha vida são necessários indivíduos que usem seu léxico espontaneamente no mundo real. O inverso, ou seja, uma língua que careça

[5] Do inglês Unified Medical Language System
[6] Do inglês Medical Subject Headings em Português
[7] Do inglês Medical Dictionary for Regulatory activities in Portuguese
[8] Do inglês National Library of Medicine
[9] Do inglês Medical Dictionary for Regulatory activities
[10] Do inglês Medical Subject Headings
[11] Do inglês International Classification of Diseases

de falantes vivos, é dada como língua morta [Lexikon, 2015]. E, por isso, não se deve traduzir dados existentes em outros idiomas e para utilizá-los como base de dados.

Outro motivo para não traduzir dados, agora no âmbito da farmacovigilância, trata da autorização de comercialização de medicamentos, tendo em vista que cada país tem sua legislação. Existem medicamentos proibidos em certas partes do globo e liberados em outras, por exemplo, a Dipirona, proibida nos EUA e na Suécia e permitida no Brasil, ou, ainda, a Sibutramina, proibida na União Europeia, Estados Unidos, Austrália, Uruguai, Paraguai, entre outros [Brasil, 2013].

Objetivos da dissertação

Esta dissertação se concentra em uma abordagem para Extração da Informação (EI) de EAM por meio de um processo desenvolvido ao longo da dissertação em documentos textuais informais no idioma português brasileiro. Dessa forma, foi proposto um modelo de processo bem como o algoritmo *SaúdeAlg*. Para Meystre et al. [2008] algumas técnicas podem ser usadas para extrair informações de textos em linguagem natural, entre elas, tokenização, divisão de sentenças e reconhecimento de entidades nomeadas. Existem algumas ferramentas para auxiliar o processamento da linguagem natural, a literatura destaca a Biblioteca de Stanford Para Processamento de Linguagem Natural (coreNLP) [Manning et al., 2014b], que é altamente citada (cerca de 2.500 citações desde 2014 no Google Acadêmico). A coreNLP é uma ferramenta capaz de extrair a palavra radical, marcar a estrutura da frase e reconhecer algumas entidades como, pessoa, localização, organização, data e dados numéricos [Manning et al., 2014b]. Assim, pode ser útil para identificar entidades nomeadas e a relação entre elas na área de farmacovigilância, não tendo suporte nativo ao idioma português brasileiro.

No modelo proposto a coreNLP foi implementada em português brasileiro no Etiquetador de parte do discurso (Analisador Morfológico) (PosTagger), Parser de Dependência (Analisador Sintático) (DepParser) e REN, sendo esses modelos estatísticos, utilizando Aprendizagem de Máquina (AM). Conforme Levy and Goldberg [2014], o PLN no estado da arte representa seu léxico (palavras) em formato de vetores. Tais vetores são conhecidos por *word embeddings*. Assim, o *SaúdeAlg* utiliza os dados anotados para

criar uma árvore de dependência, com fundamentos da gramática de dependência, para detectar a relação de causa e efeito (analisador semântico).

Nesse cenário, o objetivo principal desta dissertação é contribuir com o processo de detecção de eventos adversos em textos informais publicados em redes sociais, criando um modelo para extrair informações.

Contribuições da dissertação

Esta dissertação apresenta como contribuição: (i) algoritmo *SaúdeAlg* para o reconhecimento de relações semânticas para identificar EAM. (ii) uma abordagem inédita para extração de EAM. (iii) criação de conjunto de dados com base no Twitter em língua portuguesa do Brasil. (iv) criação de conjunto de dados com base em bulas de medicamentos em léxicos em língua portuguesa do Brasil. (v) estender a coreNLP para português brasileiro e aplicá-la para criação de modelos PosTagger e DepParser. (vi) implementação do REN para farmacovigilância em português brasileiro.

As contribuições encontram-se desmembradas nas seguintes publicações realizadas durante o estudo deste trabalho:

- CUNHA, A. M.; SANTOS, G. N. ; GUEDES, G. P. . Uma analise sobre as bulas de medicamentos no Brasil. In: XII Brazilian e-Science Workshop, 2018, Natal. XII Brazilian e-Science Workshop, 2018.
- CUNHA, A. M.; BELLOZE, K. T. ; GUEDES, G. P. . Recognizing pharmacovigilance named entities in Brazilian Portuguese with CoreNLP. In: XIII Brazilian e-Science Workshop, 2019. XIII Brazilian e-Science Workshop, 2019.

Organização da dissertação

Além da presente introdução, esta pesquisa está organizada em outros seis capítulos. No capítulo 1, é apresentado o referencial teórico que serve de base para o

entendimento do leitor, com uma visão holística dos termos e detalhamento das terminologias e conceitos abordados ao longo do trabalho. Continuadamente, no capítulo 2, serão abordadas obras que se relacionam com o tema. No Capítulo 3, por sua vez, é apresentada a proposta para resolver o problema da pesquisa e a relevância do tema. No capitulo 4 são definidos os conjuntos de dados adotados nessa dissertação, bem como a implementação da metodologia no que diz respeito a alteração dos conjuntos de dados. No capítulo 5, é dado continuidade na implementação da metodologia, com maior enfoque nos experimentos produzidos na realização dessa dissertação assim como, as dificuldades encontradas e os estudos de caso. Por fim, no capítulo 5.2.4, teremos a conclusão com os relatos obtidos, os resultados baseados nos objetivos relacionados e sugestões para trabalhos futuros.

1- Referencial Teórico

Neste capítulo são referenciados os conceitos globais utilizados nessa obra, separados por áreas para o melhor proveito do entendimento. São conceitos fundamentais para a compreensão dos próximos capítulos.

1.1- Aspectos Técnicos em Saúde

Esta seção contém os fundamentos teóricos de alguns dos termos utilizados no campo da saúde. São definições relevantes para o entendimento dessa dissertação, tais como: farmacovigilância, medicamento, droga e eventos adversos.

1.1.1 Farmacovigilância

A farmacovigilância consiste na “ciência das atividades relativas à detecção, avaliação, compreensão e prevenção de efeitos adversos ou quaisquer outros possíveis problemas relacionados a medicamentos” [Meyboom et al., 1999]. Posteriormente, seu campo de atuação foi aumentado, anexando, por exemplo, produtos fitoterápicos, medicamentos tradicionais e complementares, hemoterápicos, produtos biológicos, produtos para a saúde e vacinas [Meyboom et al., 1999]. Seu objetivo é prover maior segurança para o paciente, aprimorando o cuidado em relação ao uso do medicamento, melhorar a saúde pública, no que diz respeito à medicação, promover educação para o uso adequado de medicamento e comunicação para o público geral [Bowdler, 1997].

Qualquer medicamento precisa obter a licença de uso, também conhecida por registro, que consiste na autorização concessora da comercialização do mesmo. Além disso, o registro contempla os ensaios pré-clínicos, pesquisas com animais e os ensais clínicos com seres humanos para definir a eficácia da medicação, estabelecendo assim, a

dose segura de administração e possíveis efeitos adversos [Bowdler, 1997]. Porém, esses ensaios têm muitas limitações sendo, geralmente, incapazes de, sozinhos, preverem a segurança do medicamento [WHO, 2002].

Dentre os fatores limitadores têm-se, por exemplo, o fato dos indivíduos testados não excederem os cinco mil, tornando a identificação de um EAM, que tenha ocorrência em população acima de cinco mil, pouco provável [WHO, 2005]; outro fator, é o período de ensaios de curta duração, na sua própria essência, incapaz de prever resultados a longo prazo [Cleophas and Zwinderman, 2000]. Além disso, populações específicas (crianças, gestantes, lactantes, idosos, pacientes enfermos) são excluídas dos ensaios, entretanto, expostas aos medicamentos quando os mesmos são regularizados [WHO, 2005]. Existe, ainda, o problema da polifarmácia, devido a interação entre medicamentos ser desconhecida.

A farmacovigilância, assim como tudo relacionado a segurança de medicamentos, diz respeito a todos que são ou serão submetidos a medicina em algum momento. A qualidade no gerenciamento de segurança de medicamentos e, por consequência, da farmacovigilância, é obrigatória para a detecção inicial de efeitos negativos dos medicamento e está intimamente ligada a tomadas de decisão nos âmbitos local, nacional e internacional [WHO, 2002].

1.1.2 Medicamento e Droga

O termo medicamento significa uma preparação química que pode conter, ou não, um ou mais fármacos com excipientes (veículo farmacologicamente inativo que serve para dar volume a fim de facilitar a distribuição, produção, translado, armazenamento e administração do medicamento). É destinado à prevenção de determinado problema de saúde, com intuito de obter benefícios em prol da cura ou melhora de enfermidades, alterando o funcionamento corporal. Os medicamentos são rotulados pela Denominação Comum Internacional (DCI), ou pelo nome comercial que possui o fármaco. Exemplos de medicamento são *novalgina* e *dorsanol* [Rang et al., 2015; Brunton et al., 2007].

A droga, por sua vez, consiste em uma mistura de componentes brutos que não fazem parte do gênero alimentício ou da dieta e que agem sobre um organismo vivo no

qual, pelo menos um componente, causa alguma reação farmacológica [Rang et al., 2015; Brunton et al., 2007]. Nesse caso, é desconhecido tanto o tipo quanto a composição da mistura, ou seja, a quantidade exata dos componentes e a dose de cada um deles (além da identificação dos elementos e sua concentração) são desconhecidas sempre que a droga é consumida [Rang et al., 2015; Brunton et al., 2007]. Podemos dizer que extratos vegetais, tinturas e extratos de produtos naturais, mesmo que utilizados com propósito terapêutico, são classificados como drogas.

1.1.3 Eventos Adversos

Sabe-se que os medicamentos são formulados sob elevados critérios que visam maximizar a proteção e segurança dos pacientes, entretanto sempre existe risco associado ao seu uso. Motivos diversos deixam os pacientes sujeitos a efeitos indesejados. Conforme WHO [2002]; Pereira [2002]; Edwards and Aronson [2000], eventos adversos em medicamentos é entendido como danos causados aos pacientes relacionados a uso de medicamentos, seja pela utilização apropriada ou inapropriada, não sendo necessário vinculo direto com o medicamento. Podemos citar como exemplo, medicamento injetável, aplicação incorreta causando dor no local. Trata-se de EAM. Para Edwards and Aronson [2000], além disso, eventos adversos permitem identificar e prevenir riscos associados ao modo de tratamento.

1.2- Aspectos Técnicos em Linguística

Nesta seção, estão contidos os fundamentos teóricos acerca de alguns dos termos utilizados no campo da Linguística. São definições importantes para o entendimento dessa obra, tais como: linguística de corpus e gramática.

1.2.1 Linguística de Corpus

Para Sardinha [2004], linguística de *corpus* é um ramo da linguística que consiste na criação e averiguação dos *corpora* (plural em latim de *corpus*) cujo ideal serve para examinar minuciosamente uma língua, compreendendo a sua variação linguística, com o intuito de obter evidências científicas extraídas por computador. Já Dash [2019], tem uma definição mais ampla: para ele, o vocábulo corpus faz alusão ao conjunto de dados linguísticos (formado pelo uso na língua em sua plenitude, o que inclui o idioma falado e escrito), classificados conforme determinados critérios, de magnitude suficiente (o autor não informa valor nem métrica para exprimir tal magnitude) para expressar a variação linguística existente. Sardinha [2004] afirma que a linguística de corpus alterou a forma da pesquisa em linguística em seus níveis, com abordagem computacional e estatística, por tratar-se de uma amostra do idioma.

Para Leech [1992], o volume de dados é um item importante, mas não o principal. O estudo honesto (material autêntico com foco na linguagem, seja ela falada ou escrita) é o fator mais importante. Para o autor, dado o corpus, podemos obter de sua leitura as ocorrências de eventos da língua estudada. Essa atividade é realizada com o auxílio de softwares de computador que lidam com dados em formatos textuais e recursos matemáticos e/ou estatísticos para, dada a amostra (todo corpus é uma amostra), extrair informações que tenham aplicabilidade [Gerber and Vasilévksi, 2007].

O corpus mais antigo que se tem notícia remonta ao século $XVIII$ com o *Vocabulário Português e Latino*, reproduzido entre 1712 -1728 pelo sacerdote Rafael Bluteau. A obra, que foi finalizada no século $XVII$, é tida como o primeiro corpus, que tem cerca de 406 obras, sendo referência para os vocábulos que constam na terminologia dos dicionários [Murakawa, 2006].

1.2.2 Gramática

Não foi encontrado na literatura uma definição única para gramática e sim, concepções. As três principais definições são: descritiva, normativa e internalizada

[Travaglia, 2006]. De acordo com Travaglia [2006], a designada gramática normativa é apresentada como um modelo de regras que norteiam o modo adequado de escrever e falar bem, ou seja, deixar de seguir a regra é tido como um “erro”. A exceção, fica por conta daquilo que é definido como variedade padrão por ter um foco maior na parte escrita da língua que a tradição oral [Travaglia, 2006].

O próprio Travaglia [2006] define gramática descritiva como descrever o modo de operação da língua, dando prioridade a forma oral. Para essa interpretação, após análises teóricas e metodologias, temos criações de enunciados originados de seus falantes; sua diferença em relação à normativa, é que a descritiva deve, ao longo de seu tempo de vida, catalogar essas diversidades [Travaglia, 2006].

Já a gramática internalizada, também conhecida por competência linguística, é o grupo de regras que os adeptos adquirem ao longo da vida. Sendo assim, não existe um erro formalmente definido [Travaglia, 2006]. Para Celso [1985], a gramática de internalização são as regras adquiridas pelos os habitantes em seu meio.

1.2.2.1 Gramática de Dependência

A gramática de dependência é um conjunto de teorias gramaticais recentes, modeladas com base na relação de dependência. Dependência é o entendimento de que as palavras estão relacionadas por ligação direta. Podemos definir a estrutura como o vínculo entre a raiz e seus dependentes: trata-se de uma relação parte para parte, composta por uma raiz que é a unidade lexical (palavras) das quais dependem todas as outras. O *adjetivo* é um modificador do elemento (tendo uma relação de dependência) e o elemento sendo o *sujeito* do verbo (tendo relação de dependência). Nessa relação o verbo é o elemento central da estrutura da oração e os demais elementos, ligações diretas ou indiretas com o verbo. [Hjelmslev, 1975; Partee et al., 2012; Chen and Manning, 2014; De Marneffe et al., 2006].

Usamos a árvore de dependência para representar a estrutura sintática da oração, conforme figura 1. Por exemplo, a oração: *Tomei dipirona e fiquei com sono*

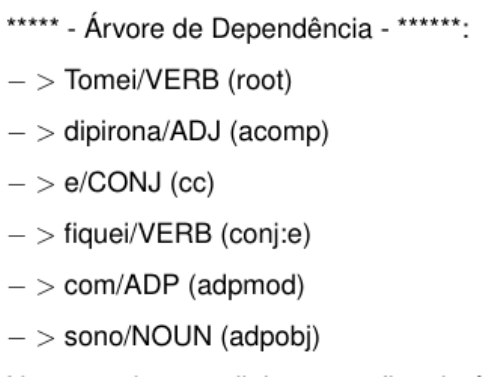

Figura 1 – Exemplo de árvore de dependência

No exemplo, *“e”*,*“dipirona”* e *“fiquei”* têm como raiz (root) *“Tomei”*, desse modo são modificadores do *token “Tomei”*. O *token Tomei* não tem raiz, sendo a própria raiz da árvore. A raiz é, comumente, um verbo. A relação entre dois *tokens* (a dependência) pode ter um rótulo, indicando uma concepção como conjunção, sujeito, objeto direto, etc.

1.3- Aspectos Técnicos em Computação

Esta seção contém os fundamentos teóricos de certos termos utilizados no campo da computação. Tais definições são primordiais para o entendimento dessa dissertação, como EI, Mineração de Textos (MT), PLN, REN, métricas de desempenho e *web scraping*.

1.3.1 Extração da Informação

Atualmente, existe uma grande quantidade de dados no formato textual disponível abertamente para o público, estando boa parte em formato digital (redes sociais, revistas, jornais, blogs, etc.), entretanto muito desse conteúdo é perdido. É humanamente impossível manipular (leia-se, ler, entender, recapitular) o vasto repertório de dados produzido, cotidianamente, ao longo do tempo, fazendo com que oportunidades e conhecimentos sejam perdidos. De modo a colocar alguma ordem nessa imensidão de dados textuais, algumas pesquisas são fomentadas, sendo as abordagens mais comuns a Recuperação

de Informação (RI) e EI.

A RI lida com armazenamento, manipulação e elucidação de elementos de informação, como documentos de interesse, com base nos critérios fornecidos por requisição. Conforme Dixon [1997], EI é uma subárea do PLN utilizada para processar grandes volumes de textos, tendo conhecimento sobre domínio específico [Lewis, 1998]. Seu objetivo é obter, a partir destes, informações [Scarinci, 1997]. Computacionalmente, tem maior desempenho quando comparado com PLN, pois permite ignorar conteúdo que não tem relação com o domínio, ou seja, uma porção do texto é literalmente ignorada no processo, uma vez que não fazem parte do domínio específico [Riloff, 1994].

Para Dixon [1997], EI é o método que une estruturas combinadas de dados encontrados em um ou mais textos. RI com base na consulta do usuário, coleta fontes relevantes (por exemplo, textos, vídeos, sons, imagens, mapas e gráficos). Difere de EI que filtra por conteúdo, ou seja, EI identifica os documentos específicos dentro do corpus. Todavia, não pode obter corpus por intermédio de critérios, conforme ocorre com RI [Scarinci, 1997]. A EI permite obter informações com base em parâmetros específicos, bem como reproduzir as estruturas que não podem ser apresentadas por RI. Não pode-se esquecer que ambas as áreas se complementam, visto que pode-se reduzir o conjunto de dados com RI para redefinir uma busca com EI [Baeza-Yates and Ribeiro-Neto, 2013].

A seguir é apresentado um exemplo para compreender a diferença entre RI e EI. Ao efetuar uma busca em uma base e dados indexada (contendo artigos, revistas, jornais, etc.), é utilizado uma *string* de busca, esse processo é a RI feita de forma manual, ou seja, coleta de dados relevantes com base em critérios estabelecidos pelo usuário, aqui representado pela *string* de busca. O que motivou a busca nessa base de dados foi a investigação de alguma conteúdo específico, não tem sentido pesquisar por pesquisar, o que interessa é o conteúdo das obras. Ao ler e filtrar o conteúdo dessa obras para o fim específico trata-se da EI, ou seja, os fragmentos relevantes são isolados e então extraídas informações desses fragmentos.

A demanda na utilização de EI é explicada pela quantidade, cada vez mais elevada, de dados disponíveis em meios digitais; nasce da necessidade de minimizar o desperdício de oportunidades e, ao mesmo tempo, criá-las, com base nos dados disponíveis. Outros fatores são a aptidão decorrente da tecnologia atual com uso de PLN e a capacidade virtualmente ilimitada de uso. O grande volume de dados existe principalmente na forma de linguagem natural. Para podermos lidar com ele e, posteriormente analisá-lo, preci-

samos estruturá-lo, tornando-os disponível para os computadores. Nessa dissertação, será utilizada a rede social online Twitter. Para RI, foi utilizada a técnica de *Web Scraping* buscando a obtenção dos *posts*, denominados *tweets*, criando, assim, o conjunto de dados. Dessa maneira, a EI permite realizar uma inferência com maior precisão.

Muito trabalhos em MT usam métodos baseados em estatística, considerando documentos como um emaranhado desordenado de palavras ou espaços vetoriais característicos da RI [Mooney and Bunescu, 2005]. Entretanto, ao empregar em Linguística, adquire-se conhecimento obtendo do texto as entidades, bem como propriedades e relacionamentos entre os elementos.

Ainda que o PLN não tenha total entendimento da linguagem natural, a tecnologia atual permite detectar vários elementos no texto e identificar a relação entre as partes [Mooney and Bunescu, 2005]. Logo, a EI desempenha um papel importante na MT. A figura 2 descreve as etapas do processo de extração de informação, as quais são descritas nas seções seguintes:

Figura 2 – Processo de extração da informação, adaptado de Baeza-Yates and Ribeiro-Neto [2013]

1.3.1.1 Analisador Léxico

Nesta etapa, o texto é, basicamente, dividido em orações e em *tokens* (menor unidade léxica). Neste processo está inclusa a remoção de caracteres especiais (emoticons, comentários, delimitadores, etc.) para facilitar a manipulação por analisadores sintáticos.

1.3.1.2 Analisador Morfológico

Esta etapa é aplicada a termos de forma particular. Para cada elemento (palavra) em uma oração, é atribuída sua classe gramatical (em língua portuguesa há dez classes gramaticais, a saber: substantivo, adjetivo, artigo, pronome, numeral, verbo, advérbio, preposição, conjunção e interjeição) e flexão (em gênero, número e grau). O etiquetador morfológico confere uma "marcação" a cada palavra com sua respectiva classe gramatical.

O principal conjunto de etiquetadores morfológicos utilizado é o *Penn Tree bank* [Santorini, 1990] para língua inglesa, não havendo um consenso para língua portuguesa. Nessa dissertação utiliza-se as marcações praticadas pela ferramenta coreNLP [Toutanova et al., 2003], conforme exibido na tabela 1.

Tabela 1 – Definições das dependências de Stanford para analisador morfológico, adaptado de Toutanova et al. [2003].

Etiqueta	Descrição
CC	Conjunção coordenadora
CD	Número cardinal
DT	Determinante
EX	Existencial
FW	Palavra Estrangeira
IN	Preposição ou Conjunção Subordinada
JJ	Adjetivo
JJR	Adjetivo, Comparativo
JJS	Adjetivo, Superlativo
LS	Listar marcador de item

MD	Modal
NN	Substantivo, Singular
NNS	Substantivo, Plural
NNP	Nome próprio, Singular
NNPS	Nome próprio, Plural
PDT	Predeterminado
POS	Final Possessivo
PRP	Pronome Pessoal
PRP	Pronome Possessivo
RB	Advérbio
RBR	Advérbio, Comparativo
RBS	Advérbio, Superlativo
RP	Partícula
SYM	Símbolo
TO	Para
UH	Interjeição
VB	Verbo, forma base
VBD	Verbo, pretérito
VBG	Verbo, gerúndio ou particípio no modo presente
VBN	Verbo, particípio no modo passado
VBP	Verbo, não terceira pessoa do singular no presente
VBZ	Verbo, terceira pessoa do singular no presente
WDT	Wh determinador
WP	WhPronome
WP	WhPronome Possessivo (versão de prólogo WP)
WRB	Wh advérbio

1.3.1.3 Analisador Sintático

Nessa etapa, o foco são as palavras com alguma teoria gramatical, através da criação de uma árvore de análise para cada oração. Com base nessa análise, pode-se

encontrar relações entre cada palavra e outras orações e, também, a função da sentença. A análise sintática é dividida em dois grupos: gramáticas constituintes e de dependência. Nessa obra aborda-se apenas a gramática de dependência, dado todo o histórico com PLN na atividade de REN [Feldman et al., 2007].

A gramática de dependência consiste nas relações diretas entre as palavras (sujeito, predicativo do sujeito, etc.), associando as mesmas conforme sua interação, assim, interligando-as pelo verbo principal - mais detalhes na seção 1.2.2.1. O analisador de dependência vincula palavras em uma frase e atribui rótulos a essas relações apenas em nós terminais. Os etiquetadores morfológicos marcam funções gramaticais (sujeito, predicado, etc.) de acordo com o modelo de frase pensada nos termos da semântica Fregeana[1] com um predicativo central como argumento. O algoritmo é de complexidade linear, tendo pior caso On^3 [Cer et al., 2010]. Podemos conferir na tabela 2 os valores empregados no analisador sintático [Toutanova et al., 2003].

[1] Friedrich Ludwig Gottlob Frege - É tido como um dos pais da lógica moderna. Os itens linguísticos, como termos, predicados e sentenças têm um referente e um sentido. O referente da frase é o sujeito e o predicado simples, sendo determinada a verdade entre os referentes. O referente do sujeito é um objeto e o referente do predicado é uma função que, com base no objeto de entrada, produz um resultado de verdade como saída. Essa relação do sujeito com o predicado é um pensamento (preposição), podendo ser definida por uma função f sendo uma relação tal, que se $x = y$ então $f(x) = f(y)$, constituindo um conjunto de pares ordenados.

Tabela 2 – Definições das dependências de Stanford para analisador sintático, adaptado de Toutanova et al. [2003]

Etiqueta	Descrição	Etiqueta	Descrição
acomp	Adjetivo como Complemento de Verbo	npadvmod	Frase Nominal como Modificador Adverbial
advcl	Modificador de Clausula Adverbial	nsubj	Assunto Nominal
advmod	Modificador de Advérbio Não Clausal	nsubjpass	Assunto Nominal Passivo
agent	Agente	num	Modificador Numérico
amod	Adjetivo Modificador	number	Elemento de Número Composto
ccomp	Complemento Clausulal	parataxis	Parataxe
aux	Verbo Auxiliar	pcomp	Complemento Preposicional
auxpass	Verbo Auxiliar Passivo	pobj	Objeto de uma Preposição
cc	Conjunção Coordenada	poss	Modificador de Posse
ccomp	Complemento Clausulal	possessive	Modificador Possessivo
conj	Conjunção	preconj	Pré-Conjunto
cop	Verbo de Ligação	predet	Predeterminador
csubj	Clausula Sujeito	prep	Modificador Preposicional
csubjpass	Sujeito Passivo Clausulal	prepc	Modificador Clausulal Preposicional
dep	Dependência	prt	Particula de Verbo-Frase
det	Determinante	punct	Pontuação
discourse	Elemento do Discurso	quantmod	Quantificador de Frase Modificada
dobj	Objeto Direto	rcmod	Relacionado a Clausula de Modificação
expl	Existencial	ref	Referência
goeswith	Vai Com	root	Raiz
iobj	Objeto Indireto	tmod	Modificador Temporal
mark	Marcação	vmod	Modificador Verbal Não Finito Reduzido
mwe	Expressão com Várias Palavras	xcomp	Complemento Clausal Aberto
neg	Modificador de Negação	xsubj	Assunto Controlador
nn	Substantivo Modificador Composto		

1.3.1.4 Analisador Semântico

O processo de análise sintática estrutura as orações para representar seu significado, assim sendo, informa sobre o conhecimento cotidiano no mundo [Charniak et al., 2014]. Como estudo, a análise semântica inclui outras atividades, tais como: sinonímia, resolução de ambiguidades, tradução de uma linguagem natural, entre outras [Gabrilovich et al., 2007].

Nessa dissertação será abordado o REN como análise semântica, tendo maiores detalhes na seção 1.3.4 e um algoritmo semântico, aqui definido como filtro semântico será desenvolvido para extrair a relação de causa e efeito entre medicamentos e eventos

mais detalhes na seção 3.1 do capitulo 3.

1.3.2 Mineração de Textos

Existem diversas definições para MT. Para Tan et al. [1999], MT, ou descoberta de conhecimento em textos, é o método para obter padrões intrigantes, não corriqueiros, de documentos textuais. Já para Hearst [1999], a MT é o descobrimento, por meio de tecnologia, de conhecimentos inéditos ou desconhecidos da metodologia de EI, com base em documentos não estruturados. Já na definição de Sullivan [2000], MT é o conhecimento de EI dos textos utilizando a teoria da linguística computacional. Outra visão é a de Weiss et al. [2010], em que as técnicas empregadas em MT são equivalentes às aplicadas em Mineração de Dados (MD), tendo, em suma, a mesma função, divergindo apenas no tipo de dado utilizado (dados textuais para MT e numéricos para MD).

Diante de tantas definições, podemos atingir um consenso de que MT equivale à utilização do conjunto de dados pela investigação dos padrões intrigantes e à obtenção de informações do PLN com emprego das técnicas e algoritmos de PLN, MD e AM. Em síntese, MT e MD têm como diferença o fato de que MT extrai dados de linguagem natural, em sua maioria desestruturados, e MD, dados estruturados em estruturas computacionais conhecidas [Hotho et al., 2005]. A rigor, dados na internet, como os que serão utilizados neste trabalho, são bastantes divergentes por não possuírem uma estrutura definida [Markov and Larose, 2007].

MT e MD são utilizados em grandes conjuntos de dados ao empregarem técnicas de AM. Porém, devido a predominância de linguagem natural em MT, a informação semântica presente pode ser utilizada para extrair informações dos dados em formato textual, sendo este um campo multidisciplinar que contém áreas como RI, EI, PLN e AM [Sumathy and Chidambaram, 2013].

1.3.3 Processamento de Linguagem Natural

O uso de PLN em MT tem o propósito de reconhecer a importância de cada termo, dado o seu contexto, de modo a possibilitar maior qualidade dos resultados obtidos. O PLN é empregado para anexar regras semânticas a o processo de EI [Junior, 2007]. A EI é uma importante tarefa de mineração de texto e tem uso amplamente difundido em várias pesquisas, incluindo PLN. O REN é uma função primária na área de EI [Jiang, 2012].

Entendemos por "linguagem natural" um protocolo para comunicação entre seres humanos de modo habitual (como um idioma, por exemplo, o português). O PLN, também é conhecido como linguística computacional por ser uma abordagem que aprimora o entendimento de computadores sobre a linguagem natural [Kodratoff, 1999]. Ele engloba conteúdo simples, desde frequência de palavras até situações mais complexas, como a ironia, (figura de linguagem[2]) com finalidade de interpretar um determinado contexto e oferecer uma resposta [Bird et al., 2009].

Em PLN temos a atividade de REN, mas existem etapas para serem seguidas antes de obtermos o REN propriamente dito, ou seja, uma forma de cadeia de produção. Esse processo é normalmente realizado em módulos individuais e em ordem, cuja saída de um módulo corresponde à entrada do próximo, em um design conhecido por *pipeline*. Não existe uma norma ou padrão na estrutura, encontrando-se variações na literatura. Nesse trabalho, foram utilizadas etapas comuns encontradas em várias obras, incluindo: tokenização, etiquetagem de análise morfológica (PosTagger), lematização e análise sintática (DepParser) sendo, o REN em si, tratado em outra seção.

1.3.3.1 Tokenização e WordsToSentence

Tokenização, ou atomização[3], é o processo de separar componentes da frase, normalmente léxicos (palavras), em unidades. Essa unidade se chama *token*, e pode ser

[2] A figura de linguagem é um recurso narrativo para utilizar uma interpretação menos recorrente, ou mesmo não literal, do objeto do discurso (nesse caso conhecida por sentido figurado).

[3] Alguns autores de língua portuguesa adotam essa nomenclatura [Linguateca, 2009]

entendida como a estrutura mínima que exprime a mesma semântica do texto original. O coreNLP fornece uma classe para esse recurso em inglês foi utilizado o mesmo para português), denominada *PTBTokenizer*, possuindo agilidade de cerca de $1.000.000$ de *tokens* por segundo, sendo eficiente e implementando o método determinístico. Foi utilizado no trabalho o conjunto de documentos representados em *tokens*, cuja denominação conhecida é “saco de palavras” (em inglês, “*bag of words*”)

WordsToSentence, ou *Splits a sequence*, consiste em dividir uma sequência de *tokens* em frases, dadas as decisões do *tokenizador*. Esse processo é auxiliado por palavras que separam caracteres, como espaços e sinais de pontuação, podendo ser considerados *tokens* delimitadores, caracteres de finalização da frase (.,?!), ou não agrupados a outro *token* [Feldman et al., 2007]. Tal trâmite faz parte da *tokenização* no software coreNLP através da classe *SentencesAnnotation* [Manning et al., 2014a]. Ambas as aplicações pertencem à seção de pré-processamento de dados [Marcus et al., 1993; Junior, 2007].

1.3.3.2 POS-Tagger

O PosTagger é responsável por anotar no texto a análise morfológica atribuindo para cada termo da oração (*token*) uma classe gramatical. Nesta dissertação foi utilizado o conjunto de *tags* do *Penn Tree bank* . O coreNLP foi utilizado para implementar o modelo de máxima entropia [Toutanova et al., 2003].

1.3.3.3 Lematização

Documentos em formato textual contém muitas palavras flexionadas em várias formas. Na língua portuguesa, é possivel flexionar um substantivo em gênero (masculino e feminino), número (singular e plural) e grau, tendo o mesmo significado semântico. O desenvolvimento de palavras é, em maior parte, realizado por procedência de radicais, de modo a terem o mesmo sentido [Cegalla, 1977].

Analisando apenas de modo lexicográfico, o vocábulo na forma de dicionário é designado como lema ou forma canônica. Sendo assim, sua representação consiste da nomenclatura de lemas ou formas canônicas para o caso de verbos no modo infinitivo, e para adjetivos e substantivos no singular e no masculino. Podemos usar como exemplo a palavra *cachorra, cachorras e cachorros*, são substituídos pela sua forma canônica representando *cachorro*. No caso de verbos, as flexões *trabalhará, trabalharemos, trabalharei* são todas formas do verbo *trabalhar* [De Saussure, 1989]. Para o software, esses dados devem vir anotados no corpus de modo a serem utilizados no DepParser.

1.3.3.4 Parse de Dependência - DepParser

O DepParser utiliza predição de palavras, ou seja, cada palavra na oração tem relação com seus dependentes. Com isso a saída do DepParser não é uma árvore e sim uma relação de dependência. A figura 3 apresenta um exemplo de parse de dependência. As relações são listadas em *Dependency Parse*, para cada relação é apresentado o item governador e o dependente. Na relação *root* (raiz), o *token* tomei é dependente de *root*, logo ele mesmo é a raiz. Já na relação dobj, o *token* tomei é governador, tendo o *token* Dipirona como dependente.

```
Tomei Dipirona e fiquei com enjoo

Tokens:
[Text=Tomei PartOfSpeech=VERB]
[Text=Dipirona PartOfSpeech=ADJ]
[Text=e PartOfSpeech=CONJ]
[Text=fiquei PartOfSpeech=VERB]
[Text=com PartOfSpeech=ADP]
[Text=enjoo PartOfSpeech=NOUN]

Dependency Parse :
root(ROOT-0, Tomei-1)
dobj(Tomei-1, Dipirona-2)
cc(Tomei-1, e-3)
conj:e(Tomei-1, fiquei-4)
adpmod(Tomei-1, com-5)
adpobj(com-5, enjoo-6)
```

Figura 3 – Exemplo de Parse de Dependência.

1.3.4 Reconhecimento de Entidade Nomeadas

A atividade de REN foi expressa em Grishman and Sundheim [1996] com objetivo de identificar nomes de pessoas, lugares, organizações e expressões, como hora, moedas e percentuais. Para Bunescu [2007]; Grishman and Sundheim [1996]; Chinchor et al. [1999], é um elemento essencial na EI.

De acordo com Sutton et al. [2012], REN é o problema de reconhecer e agrupar nomes próprios em documentos no formato textual, abrangendo localizações (nomes de cidades, países, etc.), pessoas (presidentes, celebridades, etc.), organizações (empresas, instituições, etc.) e tempo (data, período de duração, etc.). Segundo Nadeau and Sekine [2007], Entidades Nomeadas (EN) são aquelas que tem um ou mais qualificadores intransigentes capazes de referenciar todo o objeto em seu contexto. Conforme Kripke [1972], podemos citar como exemplos, termos que mencionam substâncias, espécimes que não tem detalhamento estabelecido, podendo incluir referências precisas.

Quando se trata da execução do REN, abordagens distintas foram utilizadas no

decorrer do tempo. Nos primórdios, era utilizada a abordagem de dicionário, contando com uma lista de EN [Walker and Amsler, 1986]. Com o tempo de evolução, a abordagem passou a ser baseada em regras inseridas para complementar o uso de dicionários e auxiliar no reconhecimento de nomes próprios [Speck and Ngomo, 2014]. Posteriormente, utilizou-se o AM supervisionado para reduzir o esforço de definir regras manuais e automatizar o processo. Atualmente, utilizam-se abordagens híbridas [Nadeau and Sekine, 2007].

Em sua maioria, dados não estruturados existem em formato textual, como e-mails, notícias jornalísticas, artigos científicos, etc; podem ser formais ou informais [Nassirtoussi et al., 2014]. O texto formal é mais padronizado, com poucos erros gramaticais e de pontuação; costuma ser revisado, e possui uma estrutura correta de sentença [Nassirtoussi et al., 2014]. Podemos exemplificar como texto formal: publicações científicas, artigos de jornais e revistas, etc. Por sua vez, o texto informal não detém padrão na sentença por não seguir regras de pontuação, haver predominância de erros gramaticais e usar abreviações; é o que acorre nos textos publicados em redes sociais [Nassirtoussi et al., 2014].

O REN costuma ser aplicado em ambos os tipos de textos (formais e informais). Entretanto, o texto informal envolve desafios, dada a sua natureza. Na literatura, obtém-se melhor resultado aplicando o REN a textos formais em detrimento de texto informais. Essa dissertação foca no REN em texto informal. O conjunto de dados é composto por tweets da rede social Twitter.

O primeiro evento que se propôs a avaliar o REN foi a Conferência de Compreensão da Mensagem (MUC), sendo a primeira proposta a obter bons resultados em um período pequeno [Grishman and Sundheim, 1996]. Ao longo do tempo, outras metodologias foram ganhando notoriedade e fama, passando a incluir o domínio a ser estudado. Até 2008, era realizado uma Extração Automática de Conteúdo (ACE), com foco em entidades ligadas a domínios de interesse [ACE, 2008]. No idioma português do Brasil é realizada a Avaliação de Reconhecedores de Entidades Mencionadas (HAREM); sua segunda, e mais atual, edição ocorreu em 2008 [Santos et al., 2006; Santos and Cardoso, 2007; Mota and Santos, 2008].

A função de classificação do REN depende do propósito a que a função se destina. Tipicamente, avaliações em grupo são definidas como um conjunto pequeno de classes, todavia, em propósitos para domínios específicos, estas classes são estendidas ou

especializadas de acordo com as necessidades e potencialidades do domínio. Nessa dissertação, a condução é fortemente associada ao domínio da farmacovigilância [Santos and Cardoso, 2007; ACE, 2008; Consortium et al., 2005].

1.3.4.1 Abordagem Baseada em Dicionário

A abordagem baseada em dicionários utiliza dicionários[4] contendo EN sendo, o dicionário, uma lista de palavras já definidas com EN. O método investiga qual categoria consta no dicionário para cada palavra do texto. O problema evidente dessa abordagem é a grande quantidade de palavras, que torna inviável avaliar, manualmente, todas as entidades. Os sistemas precisam ser atualizados, mas essa abordagem não facilita tal processo; além disso, existe o problema das EN não finitas, como números de telefones (são sequência infinitas, desse modo impossível existir uma lista de itens infinitos), sendo assim, as técnicas abordando dicionários não costumam apresentar resultados satisfatórios [Coden et al., 2012; Atkinson and Bull, 2012].

Assim, a abordagem em dicionários tem sido empregada em conjunto com outras abordagens. Sua principal vantagem é a facilidade de implementação em circunstâncias nas quais objetiva-se identificar entidades pequenas em um ambiente pouco dinâmico, por exemplo os nomes de continentes, já que a possibilidade de mudar a lista com o nome dos continentes é remota [Vazquez et al., 2011].

1.3.4.2 Abordagem Baseada em Regras

A abordagem baseada em regras consiste nas regras do objeto de domínio. Tais regras são criadas quando o objeto pode ser identificado com base no texto, e podem ser construídas com base no domínio. Na prática, usa-se um conjunto de instruções do tipo “Se-Então”, que fazem comparações baseada em padrões e expressões regulares; detectado algum deles, uma ação específica é tomada [Gong et al., 2009]. O problema

[4]Conhecido por léxico, ou gazetteer.

de tal abordagem é que ela é específica para o domínio e, em alguns casos, válida para subdomínios, não sendo possível aplicar no mesmo domínio, ou seja, em casos particulares mesmo sendo feita para o domínio a regra pode não se aplicar (por exemplo, em laudos de exames, a nomenclatura não é padronizada, sendo assim, mesmo a regra sendo específica ao domínio “laudo de exames”, pode não ser possível aplicar em outro documento deste domínio). Sendo assim, a necessidade de especialistas no domínio para criar tais regras passa a ser frequente [Esuli et al., 2013].

A abordagem tem bons resultados quando empregada em padrões específicos, como em números de telefone e endereços de *e-mail*. Quando não depende de especialistas no domínio para tal, o REN detecta um número pré-definido de conjuntos semânticos, como definido em [Grishman and Sundheim, 1996]. Essa forma teve sucesso em domínios específicos devido à relação entre conjuntos nesse domínio, como na geologia [Sobhana et al., 2010] e biologia [Campos et al., 2012].

Outra forma manual de uso se dá em regras linguísticas na atividade REN. Nesse processo, com base no domínio, são implementadas regras gramaticais e, para um bom desempenho, é requerido um especialista em linguística e no domínio [Nadeau and Sekine, 2007]. O método não é limitado ao idioma alvo. O uso em léxico raramente é empregado de forma individual e, atualmente, costuma ser empregado com AM.

1.3.4.3 Abordagem Baseada em Aprendizado de Máquina

O AM é uma subárea de alta relevância para a Inteligência Artificial (IA), tendo em vista que a capacidade de aprender é trivial para uma reação inteligente. O AM atua com métodos computacionais para aprender novos saberes e maneiras de organizar o que já se sabe [Mitchell, 1997]. As pesquisas de processos de AM podem favorecer uma melhora em nosso próprio modo de aprendizagem conforme Monard et al. [1997], devido aos modelos de algoritmos de probabilidade com o talento de “aprender”, com base em experimentação.

Para Chakrabarti [2002], o AM analisa a informação com raciocínio lógico e dedutivo para identificar padrões em grandes bases de dados, sendo amplamente empregado no processo de classificação automática de textos. Já para Segaran [2007], são técnicas

computacionais que obtêm conhecimento através de exemplos. Tal conhecimento só é possível devido aos dados não serem aleatórios e possuírem padrões que podem ser obtidos por máquinas; esses dados permitem determinar recursos importantes sobre o volume de dados treinado.

O processo de identificar e classificar as EN é parte integrante de REN, podendo ser automatizado empregando AM. Esse método contém duas etapas: treinamento e teste. O treinamento nada mais é do que ter exemplos de textos com ocorrências de EN já rotuladas. A etapa de teste consiste em dispor de textos rotulados para obter resultados da aptidão do sistema. Os algoritmos são treinados com textos previamente anotados para que aprendam, com base nessas anotações, como determinar qual palavra pertence a qual categoria. O REN em textos tem duas etapas: anotação das classes gramaticais (PosTagger) e anotação das categorias semânticas.

Conforme Nadeau and Sekine [2007], dada a complexidade de criar um sistema REN com base em regras escritas de forma manual, diversas pesquisas foram desenvolvidas na expectativa de desenvolver métodos de aprendizagem automática. Os métodos empregados em REN são divididos em três tipos: aprendizado supervisionado, aprendizado não supervisionado e aprendizado de reforço ou semi-supervisionado.

O aprendizado não supervisionado requer uma amostra do conjunto, todavia, essa amostra não tem os exemplos anotados e sim o conjunto de entrada, ou seja, não há conjunto a ser treinado. O uso mais comum é o agrupamento (clusterização, no inglês). Criar grupos de entidades com base em padrões na amostra, isso é, dados de entrada, tem certa especificação, o que ocorre em maior similaridade a partir do contexto.

A principal vantagem no aprendizado não supervisionado é não precisar de uma base de dados anotada manualmente para efetuar o treinamento. Como desvantagem, ocorre nos algoritmos para REN uma falta de exemplos para treinar e produzir um modelo, dessa forma, seu uso é muito limitado, sendo empregado apenas em cenários específicos com resultados pouco eficientes [Alpaydin, 2009].

Uma atividade isolada tem pouco efeito nesse modelo, visto que o modelo precisa da melhor sequência de atividades corretas. O processo de aprendizagem semi-supervisionado precisa ser capaz de identificar o quão bom é seu processo de aprender e, com isso, aprimorá-lo. Tem como vantagem o equilíbrio entre o aprendizado supervisionado e o não supervisionado. A quantidade de dados de entrada é reduzida, levando menos tempo para preparar os dados e, consequentemente, diminui o custo da operação.

Além disso, tem como benefício o uso de uma base de dados não anotado do aprendizado não supervisionado. As desvantagens ficam por conta da complexidade do processo, visto utilizar um conjunto pequeno de dados, e da complexidade de expandir o aprendizado [Alpaydin, 2009].

Por sua vez, aprendizado supervisionado é a técnica dominante na atividade de REN. O conceito do aprendizado supervisionado é obter do conjunto de entrada (treinamento) as propriedades necessárias para a generalização do processo de classificação de EN em novos textos que não façam parte desse conjunto inicial.

A princípio, o conjunto de textos anotados manualmente tem as entidades rotuladas de modo a criar um conjunto de exemplos. Depois, divide-se em treino e teste. O treino fornece o conhecimento através dos exemplos anotados para o modelo de aprendizado. Já o conjunto de teste serve para validar a generalização, ou seja, determinar se o modelo foi capaz de aprender com base em exemplo e derivar novos casos.

A vantagem do aprendizado supervisionado é a grande quantidade de exemplos pertencentes ao conjunto de entrada, o que possibilita uma generalização mais precisa e, com base no modelo de teste, permite acompanhar a evolução do processo de aprendizagem. Como desvantagem, temos o fato de o conjunto de dados anotados requerer uma base grande e correta de dados, além de mostrar a maior variabilidade possível do domínio alvo [Alpaydin, 2009].

1.3.4.3.1 Word Embeddings

As aplicações de PLN utilizam como elemento comum as palavras. Com isso, precisamos disponibilizá-las em formato acessível para a máquina [Hartmann et al., 2017]. O modelo utilizado é de Word Embeddings - Vetor de Palavras (WE), que é um vetor de números reais que representam palavras em espaços vetoriais [Hartmann et al., 2017], ou seja, um conjunto de técnicas que estruturam a semântica e a sintaxe da linguagem natural em um ambiente usando estatística. Sendo assim, as palavras pertencentes ao corpus são mapeadas como vetores. Esse ambiente criado é definido por *embedding space* [Goldberg, 2017]. Os WE conseguem estruturar palavras em um espaço pequeno e obter similaridade (variando conforme a técnica utilizada). Palavras não usadas no conjunto de treino podem não ser desconhecidas por completo, dada a proximidade vetorial. Desse modo, auxiliam o modelo com um novo vocábulo [Fonseca et al., 2015].

O *Word2Vec* é um método para criar vetores de palavras criando um espaço

vetorial em *n-dimensões*. Tem-se duas arquiteturas: Continuous Bag of Words - “Saco de Palavras contínuo” (CBOW) e *Skip-Gram* (salto contínuo). A arquitetura CBOW, o modelo prevê a palavra atual em “saco de palavras” com base no contexto a sua volta. Para isso, cria-se uma matriz de peso, resultando em um rápido modelo de treinamento [Mikolov et al., 2013]. Neste contexto a ordem das palavras não tem relevância para a previsão [Mikolov et al., 2013]. *Skip-Gram*, o modelo usa a palavra atual em “saco de palavras” para prever as palavras no contexto.

A estratégia *Skip-Gram* é mais eficiente que CBOW em contextos de mais distantes. Entretanto, CBOW tem um custo computacional menor, quando comparado com *Skip-Gram* [Mikolov et al., 2013]. A figura 4 apresenta um modelo simplificado do processo de *Word2Vec*. Ao mapear esse espaço de alta dimensão é possível detectar palavras que compartilham espaços muito próximos entre si. O uso de WE tem mais importância ao longo do tempo no processo de AM, visto que o seu uso eleva substancialmente a precisão de atividades de PLN [Mikolov et al., 2013].

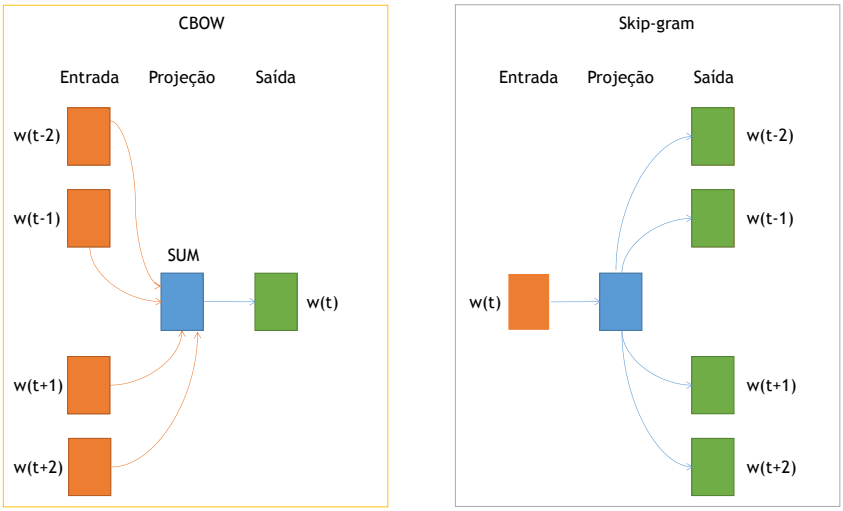

Figura 4 – Representação simplificada do modelo de *Word2Vec*, adaptado de Mikolov et al. [2013].

Tem-se, também, a técnica *Skip-Gram*, que objetiva minimizar o processo comparado ao CBOW. Ela é inversa ao CBOW (que prevê palavras próximas a partir de uma palavra), ou seja, ao definir uma palavra como entrada, um classificador faz as projeções. A previsão consiste nas palavras anteriores e posteriores. Assim, quanto maior o conjunto, maior o esforço computacional [Mikolov et al., 2013]. O número de dimensões do vetor tem impacto direto no desempenho, logo, quanto maior o número de dimensões mais demorado é o processo [Mikolov et al., 2013].

Na literatura, existem muitos métodos de aprendizagem supervisionada para REN. Nessa dissertação, foi utilizado método com fundamento em modelos probabilísticos. Tal método foi escolhido devido ao aprendizado ser realizado de forma estatística, de modo a não precisar reproduzir conceitos semânticos inerentes ao idioma de estudo. Será adotado o modelo Campos Aleatórios Condicionais (CRF), que é derivado do Modelo Oculto de Markovl (HMM). Dessa maneira, ambos serão mencionados com maior foco no CRF. O coreNLP adota o modelo *Linear-Chain Conditional Random Field*, sendo este um caso particular de CRF.

1.3.4.3.2 CRF

O modelo HMM é um complemento ao modelo da cadeia de Markov, com incremento de mais um processo estocástico. Em suma, o HMM é uma cadeia de Markov com um processo estocástico oculto. Diferente do original, no HMM as observações do modelo têm uma função de probabilidade [Rabiner, 1989], tendo como propósito replicar uma fonte de sinais com base em um grupo de observações. Essa fonte é um sistema qualquer, sendo o grupo de observações $O = \{O_1, O_2, O_3, \cdots, O_n\}$ e sinais S, $S = \{S_1, S_2, S_3, \cdots, S_n\}$. Simplificando, o modelo é regido pelas observações e por funções probabilísticas (instante inicial, alteração entre estados e envio de sinal pelo estado) [de Oliveira and Morita, 2000; Sutton and McCallum, 2006]. O HMM é usado em larga escala no PLN, no PosTagger e em REN. Em PLN, as observações são os vocábulos presentes no documento e, os sinais, os aspectos semânticos desses vocábulos observados [de Oliveira and Morita, 2000].

Para Sutton et al. [2012], CRF é uma distribuição condicional $P(M|C)$ com associação a um modelo gráfico. A variável C trata-se de um vetor com variáveis aleatórias de entrada, e a variável M é um outro vetor de variáveis aleatórias, porém de saída. Assim, $P(M|C)$ é a probabilidade de ser ofertado como entrada o vetor C e ter como saída o vetor M. CRF é implementado em vários segmentos, podendo-se citar: visão computacional [He et al., 2004], [Kumar and Hebert, 2004], processamento de texto [Taskar et al., 2002; Sha and Pereira, 2003; Peng and McCallum, 2006] e bioinformática [Liu et al., 2005; Sato and Sakakibara, 2005].

CRF são modelos criados utilizando definições de campos aleatórios de markovianos [Hammersley and Clifford, 1971]. A diferença entre CRF e campos aletórios, consiste nas variávei aleatóriasm que no CRF, estão sujeitas à determinado conjunto de

observações [Lafferty et al., 2001]. Sendo um grafo $G = (V, E)$ um campo aleatório sobre as variáveis aleatórias $V = v1, v2, v3, ..., vn$ tal que y um evento do espaço amostral de V e x sendo uma sequência fixada de símbolos observados. Então $p(y|x)$ é um CRF se $p(y|x)$ é fatorada de acordo com G, $Z(x)$ uma normatização para p tal que p seja válida, para cada clique C possui K_c funções reais arbitrárias f_k sobre vértices de C, dada a sequência de observações x e parametrizadas por λ [Bonadio, 2018]. Podemos representar por meio da equação 1:

$$p(y|x) = \frac{1}{Z(x)} exp\{ \sum_{C \in \mathcal{C}(G)} \sum_{k=1}^{K_c} \lambda_k f_k(yc, x)\} \tag{1}$$

Normalizado por $Z(x)$ temos a equação 2:

$$Z(x) = \sum_{y} exp\{ \sum_{C \in \mathcal{C}(G)} \sum_{k=1}^{K_c} \lambda_k f_k(yc, x)\} \tag{2}$$

O caso particular de campo aleatório condicional de cadeias lineares (LCCRF, do inglês linear-chain conditional random field) consistem em cliques de tamanho 2 e 1, desse modo criando a cadeia linear [Bonadio, 2018]. Para Bonadio [2018], após normalizações e uniformizações temos a equação 3, na qual o primeiro somatório representa a fatoração dos cliques.

$$p(y|x) = \frac{1}{Z(x)} exp\{ \sum_{t=1}^{N} \sum_{k=1}^{K} \lambda_k f_k(y_t, y_t - 1, x)\} \tag{3}$$

1.3.4.4 Abordagem Híbrida

O REN em abordagem híbrida, conforme Srihari [2000], consiste em AM e é baseado em regras e em dicionários. Tem por intuito aprimorar o resultado combinando as vantagens de cada método para suprir suas desvantagens. A implementação do REN que use AM com dicionário léxico contendo elementos do domínio, além de nomes próprios, tem maior potencial de alcançar melhores resultados por adotar mais de uma estratégia para identificar as EN.

Muitas obras abordam essa estratégia, podendo ser citado o trabalho de Rocktäs-

chel et al. [2012]; Tkachenko and Simanovsky [2012], cujo sistema faz o REN usando CRF com uso de dicionário. O CRF é treinado para identificar as EN e o dicionário aprimora os resultados melhorando a correta identificação de termos. Os resultados são superiores aos obtidos quando utilizado apenas uma abordagem, empregando uma única estratégia.

Mesmo no aprendizado supervisionado, o uso de dicionários tem benefícios. Com o objetivo de fazer uso desses resultados aprimorados, essa dissertação emprega a estratégia híbrida para o idioma português brasileiro.

1.3.5 Métricas de Desempenho

Os testes são realizados em modelos estatísticos, obtendo, assim, a probabilidade dos eventos ocorrerem. Os resultados desses testes são valores feitos para interpretação, de modo a identificar o verdadeiro estado do modelo. A cobertura, ou abrangência, e precisão derivam desses valores. Conforme a tabela 3, tem-se uma analogia com o modelo de Long [2002] sobre o modo que nossas mentes funcionam para aparência. São quatro tipos:

- **As coisas são o que aparentam ser**
- **São e não aparentam ser**
- **Não são, mas aparentam ser**
- **Não são nem aparentam ser**

Tabela 3 – Relação entre Aparentar e Ser adaptado de [Long, 2002]

		SER	
		SIM	NÃO
APARENTAR	SIM	A realidade é o que aparenta ser	A realidade não é, mas aparenta ser
	NÃO	A realidade é, mas não aparenta ser	A realidade não é, e nem aparenta ser

Na tabela 4 tem-se as EN anotadas, ou seja, definidas pelo criador da base de treino. O REN são as EN detectadas pelo modelo, com isso tem-se os valores:

- **VP (Verdadeiro Positivo)** – O REN detectou EN e tratava-se de fato de EN;
- **FP (Falso Positivo)** – O REN detectou EN, entretanto não tratava-se de EN;
- **FN (Falso Negativo)** – O REN não detectou e tratava-se, de fato, de EN.
- **VN (Verdadeiro Negativo)** – O REN não detectou e não tratava-se de EN.

Tabela 4 – Certeza do Reconhecimento de Entidade Nomeada

		EN Anotada	
		EN Positiva	EN Negativa
REN	REN Positivo	VP	FP
	REN Negativo	FN	VN

As definições para validar os resultados de comparação foram incluídas, sendo elas: medidas de precisão, cobertura, F$_\beta$ e, de forma resumida, a validação cruzada aplicada no Capítulo 5.

As obras que abordam o tema de classificadores utilizam três medidas para comparação com intuito de mensurar a assertividade do modelo treinado, são: cobertura (recall), precisão e medida F$_\beta$ [Davis and Goadrich, 2006]. Para o DepParser, existem duas medidas, Pontuação de Anexo Não Etiquetado (UAS) e Pontuação de Anexo Etiquetado (LAS) [Candito et al., 2010; Buchholz and Marsi, 2006].

A precisão é a porcentagem de acertos ao classificarmos os elementos e tem por objetivo auxiliar no processo de avaliação de classificação. A equação 4 apresenta a fórmula para o cálculo da precisão [Davis and Goadrich, 2006].

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{4}$$

A cobertura, também chamada de abrangência, é a porcentagem de itens selecionados de forma correta. A equação 5 apresenta a fórmula para o cálculo da cobertura [Davis and Goadrich, 2006].

$$Cobertura\ ou\ \text{Abrangência} = \frac{VP}{VP + FN} \tag{5}$$

A medida F$_\beta$ é a média harmônica entre a precisão e a cobertura (equação 6 Clark et al. [2013]). Por convenção, $_\beta = 1$. A equação 7 apresenta a fórmula para o cálculo de

F$_\beta$ [Davis and Goadrich, 2006].

$$F0 = \frac{(\beta^2 + 1) * \text{Precisão} * Cobertura}{\text{Precisão} + Cobertura} \tag{6}$$

$$F1 = \frac{2 * \text{Precisão} * Cobertura}{\beta * \text{Precisão} + Cobertura} \tag{7}$$

De forma intuitiva, tendo 100 itens de REN com rótulo de doença, o modelo identifica 80 itens, sendo 50, verdadeiros positivos. Nesse cenário, tem-se uma cobertura com valor de 50%, ou seja, dos 100 itens, 50 são classificados e identificados de modo apropriado. A precisão fica próxima dos 62%, sobre os 80 classificados como doença.

As medidas *tag*, *sentenças corretas* e *palavras desconhecidas* são aplicadas para o PosTagger. *Tags*, consistem no ato de rotular o *token*, apresenta o total de itens "etiquetados" como certo ou errado. *Sentenças corretas* apresentam sentenças com todas as *tag* corretas. *palavras desconhecidas* são as palavras que não foram usadas no conjunto de treinamento.

As métricas UAS e LAS são aplicadas apenas para o DepParser. A UAS diz respeito à porcentagem de *tokens* que tiveram sua raiz corretamente identificada. Já o LAS, toma os *tokens* com a raiz correta (UAS) e caracteriza os que tiveram a etiquetagem efetuada de forma correta [Candito et al., 2010; Buchholz and Marsi, 2006].

No processo de avaliação de classificadores, comparamos corpora distintas no classificador. Para apresentar a medida de precisão, cobertura e F1, o conjunto de dados precisa ter corpora de treino e corpora de teste. A proporção entre conjunto de dados de teste e treino é conhecido por *holdout* [Kohavi et al., 1995].

O método *holdout* divide o conjunto de dados em dois, de forma exclusiva, que são definidos como teste e treino [Kohavi et al., 1995]. No emprego desse método é comum o uso de $2/3$ dos dados para o conjunto de treino e $1/3$ para o conjunto de teste [Kohavi et al., 1995]. A separação desses conjuntos de dados será apresentada no capítulo 5.

A avaliação *holdout* não é a única utilizada, sendo também empregado o método de validação cruzada de Bailey and Elkan [1993], denominado $k - fold$ (k-subconjuntos) [Kohavi et al., 1995]. No método $k - fold$, o conjunto de dados, chamado de $Base$, é dividido em k subconjuntos (base$_1$, base$_2$, base$_3$ ... base$_k$) com aproximadamente o mesmo tamanho [Kohavi et al., 1995]. O conceito consiste em testar o modelo com base$_k$, e os demais para treinamento; o processo é repetido trocando o subconjunto a cada

execução.

Não existe um padrão na literatura para o uso do processo de validação cruzada. Nessa dissertação, utilizou-se, para o PosTagger e REN, 10 sub-conjuntos que corresponde ao máximo encontrado nas obras. E, devido à restrição de tempo e combinação com uso de WE, foi adotado o processo de 5 sub-conjuntos para o DepParser. A figura 5 exemplifica o processo com 10 sub-conjuntos.

1ª Execução	2ª Execução	3ª Execução	4ª Execução	5ª Execução	6ª Execução	7ª Execução	8ª Execução	9ª Execução	10ª Execução
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
1-fold	2-fold	3-fold	4-fold	5-fold	6-fold	7-fold	8-fold	9-fold	10-fold
Modelo de Treino									
Modelo de Teste									

Figura 5 – Exemplo do processo de validação cruzada com 10 sub-conjuntos.

1.3.6 Web Scraping

Web Scraping consiste em obter um conjunto de dados através da web. Para melhor compreensão, diferencia-se *Web Scraping* de *Web Crawling*, com base em várias obras [Maylawati and Saptawati, 2017; Sandi et al., 2016; Vasani, 2014; Turland, 2010; Marres and Weltevrede, 2013].

Web Scraping consiste na coleta de dados, de forma automática, de uma página na web e na extração de informações peculiares dela. Tais informações podem ser acomodadas em diversos meios (banco de dados, arquivos, etc) e, em sua maioria, usam robôs (os *scrapers*) para a tarefa de buscar os dados.

Já *Web Crawling* é definida como a coleta de dados automática, de uma página na web com extração dos links, assim, pode-se voltar a visitá-los e obter novos links de forma recursiva. Os dados obtidos criam um índice (de onde deriva o outro nome do *Web Crawling*: indexação). O Google, por exemplo, faz a indexação de sites (*Web Crawling*); o

mesmo acontece com outros motores de busca.

A extração automatizada dos dados na internet, devido ao formato Linguagem de Marcação de Hipertexto (HTML), é estruturada de forma hierárquica em relação a seu conteúdo [Antoniou and Harmelen, 2008]. Quanto mais o scraping requer estrutura de dados, mais fácil se torna o uso da técnica visando a estrutura em si. Por exemplo, olhando o conteúdo HTML usado nos tweets, pode-se observar, no código fonte da página Web utilizada para criar o documento em formato HTML, o Folhas de Estilo Em Cascata (CSS) sendo utilizado para estilização da página. Em outras palavras, o HTML marca o conteúdo e o CSS diz como deve ser apresentado.

O formato HTML5 e CSS3 são preferíveis para efetuar a análise e extração com a técnica de scrapper. Entretanto, quando se trata do Twitter, se faz necessário construir um web scrapper exclusivo devido à estrutura de seu código. O scrapper simula a atividade de um usuário para obter os dados, de forma automática.

Para uso do Web Scraping é necessário definir os parâmetros que serão utilizados no Twitter com base na lista de medicamentos e sintomas [da Cunha et al., 2018]. O twitter tem um campo de busca textual permitindo criar uma *query* para busca. *Devido o fator tempo, foi selecionado aleatoriamente alguns eventos e utilizados todos os medicamentos não comerciais. Para maiores detalhes seção 4.3 do capítulo 4* A utilizada nesse trabalho teve o seguinte formato: **(medicamento OR medicamento OR ...) AND (evento OR evento OR . . .) since: Data Inicial until: Data Final**.

O elemento *OR* faz menção lógica *OU*, podendo ser entendido como sendo um medicamento ou outro, isso é, basta ter um dos medicamentos para ser válido. O mesmo se aplica aos eventos. Os parênteses servem para isolar medicamento de eventos. Já o elemento *AND* faz menção a lógica *E*, ou seja, precisa ter pelo menos um medicamento e evento.

Analisando a URL do Twitter (**https://twitter.com/search?l=pt&f=tweets&vertical=default&q=**, cada parâmetro tem sua função, sendo:

- **https**[5]**://** ;
- **twitter.com** é o site do Twitter;
- /**search** indica tratar-se de uma consulta;

[5]HTTPS - Protocolo de Transferência de Hipertexto Seguro, do inglês Hyper Text Transfer Protocol Secure, é o protocolo de transferência de dados entre redes de computadores na internet.

- **?** é o símbolo que indica o início dos dados passados através da URL, ou seja, o método GET;
- Para incluir mais de uma informação, acrescenta-se o símbolo **&** para concatenar os valores;
- **l=** indica o idioma utilizado na pesquisa. O **pt** indica português;
- **f=tweets** indica o objetivo da busca;
- **vertical=default** força o foco nos tweets em formato vertical;
- **q=** por fim indica as queries desejadas (nesse caso, as duas criadas anteriormente).

Os desafios ficam em como descobrir em qual momento ocorre o processo de término da requisição Javascript e XML Assíncronos (AJAX), ou seja, quando não há mais tweets para carregar e quando a requisição inicial é nula, ou seja, nenhum tweet foi encontrado. Analisando o código fonte da página, pode-se perceber que uma imagem é carregada quando não há novos tweets (a figura 6 apresenta essa imagem). Continuando com a análise, ao carregar a imagem da figura 6, uma nova propriedade CSS é criada (*class = 'btn-link back-to-top hidden'*), dessa forma, pode-se utilizar a presença dessa propriedade para finalizar a rolagem da tela.

A figura 7 representa a tela quando o retorno da requisição é nulo. Novamente, foi necessário analisar o código fonte. A propriedade do HTML denominada *class = 'SearchEmptyTimeline-emptyDescription'* apresenta um valor diferente de nulo quando não há retorno, ou seja, existe um nulo nessa propriedade e, apenas no momento que o retorno é nulo, essa propriedade muda de valor para um valor aleatório. Os tweets no campo de texto são utilizados pela propriedade do CSS *'class = TweetTextSize'*.

Voltar ao topo da página ↑

Figura 6 – Fim das requisições AJAX para obtenção de novos tweets.

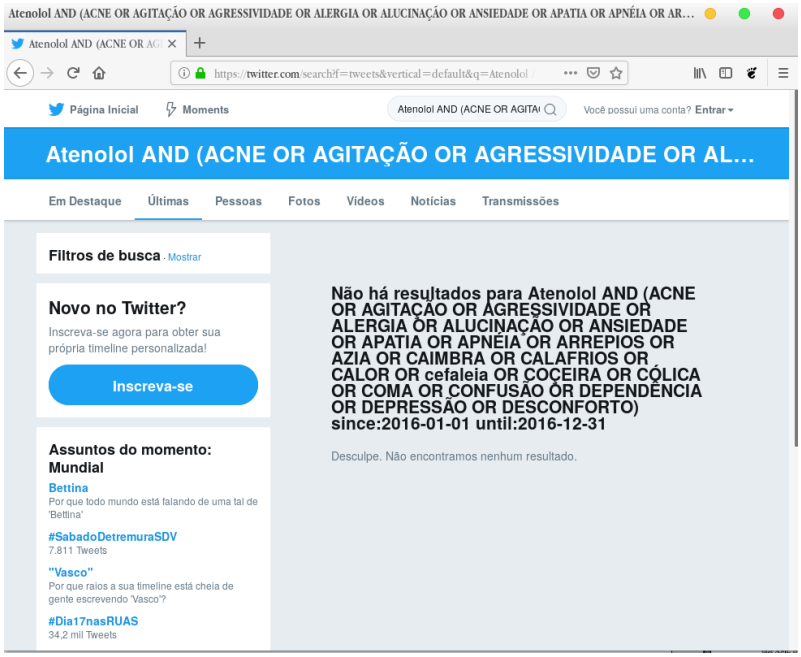

Figura 7 – Busca nos tweets por medicamento e sintomas cujo resultado retornou vazio.

1.4- Recursos e Ferramentas

Nesta seção, são referenciados os recursos e ferramentas mencionados ao longo dessa dissertação, com intuito de alcançar os objetivos propostos. E, por isso, entendê-los é fundamental para o prosseguimento do trabalho.

1.4.1 Universal Dependence

O projeto consiste em um framework para anotações linguisticas no formato de árvore sintática, em constante desenvolvimento, com suporte a muitos idiomas e, tem por intuito facilitar o desenvolvimento de parse linguístico. O padrão de anotação utiliza o modelo de dependências universais de Stanford (este em constante evolução), tags de fala do Google e tags morfossintáticas. O projeto é livre, possibilitando qualquer um usar os dados, bem como contribuir com o projeto [Berzak et al., 2016].

Em resumo, a ideia é criar uma uniformização dos níveis de sintaxe e morfossintaxe para diferentes idiomas [Nivre et al., 2016]. Esse framework tem dados anotados em português, com um subgrupo em português brasileiro, sendo utilizado como fonte de dados para essa dissertação.

1.4.2 Twitter

Originalmente, o Twitter é uma rede social em que ocorre troca de mensagens curtas, de no máximo 140 caracteres, denominados *tweets* [Twitter, 2006]. Atualmente, o limite foi ampliado para 280 caracteres [Rosen, 2017; Twitter, 2018]. O *tweet* fica visível para qualquer pessoa que tenha acesso ao serviço, não sendo necessário, para tal, um acesso autenticado [Twitter, 2006]. O Twitter conta com 330 milhões de usuários ativos por mês, figurando entre uma das maiores redes sociais do mundo, com um aumento anual de 4% nesses números [Singapore and hootsuite, 2018]. Esse perfil de usuários gera um grande volume de *tweets* por minuto. A estimativa de junho de 2018 foi de $12,986,111$ [Statistics Portal, 2018]. O Brasil está em sexto colocado no ranque da rede social no mundo [Singapore and hootsuite, 2018].

O Twitter tem um modo próprio, no qual os usuários seguem e são seguidos, diferente de redes sociais tradicionais (como o Facebook), não é necessário reciprocidade. Desse modo, pode-se seguir usuários sem ser seguido pelos mesmos. Ao seguir um usuário, ganha-se acesso a todos os *tweet* do mesmo. Com a evolução da rede, a resposta a um *tweet* (conhecida por *retweet*), deu origem a um modelo de divulgação de informações de amplo espectro e de debate sobre temas da atualidade, sendo uma fonte de informações para pesquisa em diversas áreas [Kwak et al., 2010].

O Twitter faz uso do AJAX. Esse recurso é utilizado no processo de rolagem da tela no qual, a requisição AJAX, traz novos *tweets*, sendo muito útil por permitir que novos *tweets* sejam carregados por demanda ao invés de carregar a quantidade total de uma única vez. Outra vantagem é preservar os componentes já existentes da página, ou seja, acrescenta novos *tweets* e não modifica outros elementos da página (menus, imagens, etc). O Twitter emprega o padrão HTML5 o que facilita a interpretação do código, devido a este padrão ser semântico.

1.4.3 CoreNLP

O coreNLP é um framework em código aberto, escrito em Java, com suporte à maioria das tarefas rotineiras de PLN, no qual é possível marcar a estrutura de orações sintática e morfologicamente, dividir orações em termos, estabelecer dependência entre termos, extrair sentimentos, etc. Possui um uso prático e confiável em documentos no formato textual, realiza uma análise de alta qualidade, tem suporte para vários idiomas, permitindo expandi-lo quando necessário (como é caso dessa dissertação, que utiliza o português brasileiro) e fornece interfaces para diversas linguagens de programação atuais.

Foi desenvolvido para ser um software de análise linguística contendo um pipeline simples. Com uma única opção é possível selecionar módulos, podendo incluir anotadores externos [Manning et al., 2014a]. A figura 8 descreve os principais módulos do programa, que serão explorados ao longo dessa dissertação, conforme a seção 3.

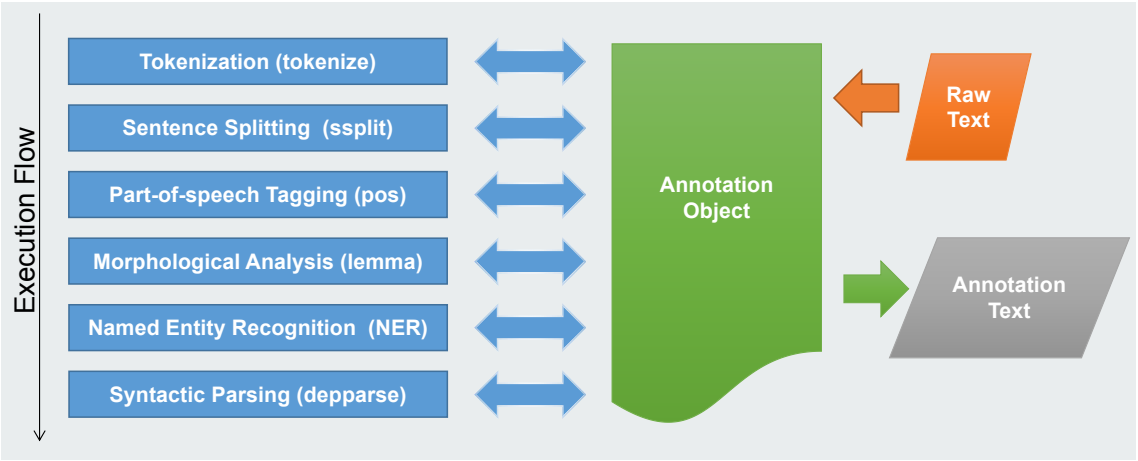

Figura 8 – Arquitetura do CoreNLP adaptado de Manning et al. [2014a].

O coreNLP adota os arquivos no formato *CoNLL-U*, definido por McDonald et al. [2013]. Este é um formato que utiliza dez campos separados por tabulação (tsv) e possuem a codificação *UTF8*. O caractere "*LF*" (linha em branco) representa o término de uma sentença, já o sublinhado "_", define o valor como não especificado. A figura 9 apresenta um exemplo de arquivo nesse formato, com as seguintes informações:

- **ID** - Identificador único. Começa em 1 para cada oração;

- **FORM** - Formulário para termos ou símbolos;
- **LEMMA** - Lema da palavra (essa anotação é utilizada no processo de lematização);
- **UPOS** - Tag para fala universal (diferença entre morfemas lexicais e gramaticais);
- **XPOS** - Tag do PosTagger;
- **FEATS** - Característica morfológica (caracteres específicos do idioma);
- **HEAD** - Raiz da palavra atual;
- **DEPREL** - Relação entre dependências;
- **DEPS** - Gráfico de dependência;
- **MISC** - Anotação diversa;

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Ver	ver	VERB	VERB	_	0	root	_	_
2	também	também	ADV	ADV	_	1	advmod	_	_
3	a	_	DET	DET	_	4	det	_	_
4	lista	lista	NOUN	NOUN	_	1	obj	_	_
5	de	_	ADP	ADP	_	6	case	_	_
6	entidades	entidade	NOUN	NOUN	_	4	nmod	_	_
7	que	_	PRON	PRON	_	9	nsubj	_	_
8	tenham	ter	AUX	AUX	_	9	aux	_	_
9	emitido	emitir	VERB	VERB	_	6	acl:relcl	_	_
10	selos	selo	NOUN	NOUN	_	9	obj	_	_
11	postais	postal	ADJ	ADJ	_	10	amod	_	paceAfter=No
12	.	_	PUNCT	.	_	1	punct	_	_

Figura 9 – Exemplo de arquivo no formato CoNLL-U.

1.4.3.1 Modelo implementado de analisador sintático

Para Hladka and Holub [2015], o computador aprende o modelo com base nos dados de um treinamento denominado *preditor*, a fim de representar o conhecimento essencial. Sendo muito útil para a análise, auxilia na adivinhação correta da árvore de dependência para determinada frase. Uma vez que a análise de dependência é uma tarefa de classificação, os valores alvo a serem adivinhados são do tipo discretos. A ilustração 10 mostra uma representação simplificada do modelo, sendo:

- *Objetos Reais* – Conjunto de dados (Treino e teste);

- *Vetores de Recursos* – Recursos são propriedades e vetores de características são listas ordenadas dessas características (lemas, valores sintáticos, etc.);
- *Valor Alvo* – raiz e recursos;
- *Preditor* – Representa os conhecimentos essenciais;
- *Predições Verdadeiras* – Predições verdadeiras;
- *Extração de Recursos* – Extrair os recursos.

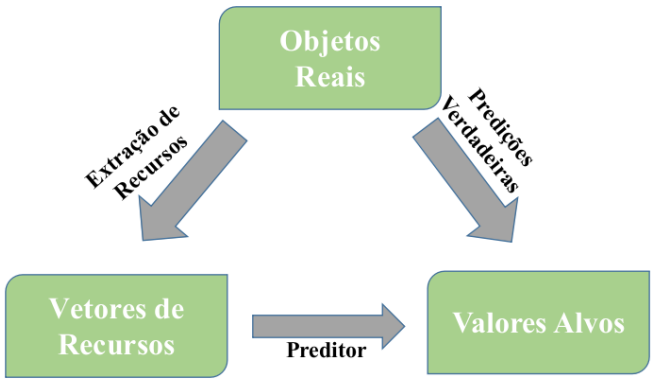

Figura 10 – Modelo de treino do DepParser adaptado de Hladka and Holub [2015].

A etiquetagem (PosTagger) é realizada de modo determinístico conforme modelo de recurso. Quanto melhor o modelo, melhor a previsibilidade. Ao lidar com linguagens complexas podemos, também, lidar com linguagens em formato textual. Desse modo, a construção do preditor envolve a escolha dos algoritmos, parâmetros e recursos, assim como o fornecimento de dados de treinamento para a máquina com a finalidade de posterior aferição junto ao conjunto de dados de teste. Com o término da avaliação, é efetuada a comparação entre os valores reais e os dados testados, utilizando duas métricas para isso:

- UAS – Relação entre tokens que tiveram correta identificação da raiz, do total existente;
- LAS – Relação dos UAS que foram devidamente etiquetados.

O modelo de transição de máxima entropia tem por intuição o processo. O analisador cria uma análise de tempo linear sobre as palavras de uma frase, mantendo, em cada passo, uma análise parcial, ou seja, uma estrutura de pilha de palavras sendo processadas a partir de um cache de palavras ainda não processadas. O analisador prossegue com as transições, em seu estado, até o cache esvaziar. O estado inicial é ter todas as palavras em ordem no cache, com uma única raiz fictícia na pilha de modo a fazer a transição de três maneiras. Com esses três tipos de transições, o analisador pode efetuar qualquer análise de dependência:

- **LEFT-ARC** – Marca o segundo item da pilha como dependente do primeiro item e remove o segundo item (Requer pelo menos dois itens);
- **RIGHT-ARC** – Marca o primeiro item da pilha como dependente do segundo item e remove o primeiro item (Requer pelo menos dois itens);
- **SHIFT** – Remove do cache uma palavra e coloca na pilha (Requer um elemento).

2- Trabalhos Relacionados

Neste capítulo são apresentadas as principais obras que serviram de base para esse estudo. Foi feita uma abordagem resumida de tais obras. Uma vez que essa dissertação consiste em uma combinação das atividades de PLN, detecção de EAM que incluem as RAM e farmacovigilância nas redes sociais, foram relacionados trabalhos dessas três áreas. As seções que se seguem apresentam estes trabalhos.

2.1- Processamento de Linguagem Natural

A análise de textos de bases heterogêneas, no domínio da farmacovigilância, utilizando PLN, tem ganhado interesse na área da computação. Pushpa and Kamakshi [2018] fizeram uma revisão metodológica sobre uso de técnicas de PLN para identificar medicamentos e eventos adversos. Já Luo et al. [2017] optaram por uma revisão estruturada do uso de PLN em registros eletrônicos de narrativas médicas para a farmacovigilância, abrangendo o caso da polifarmácia e uso de medicamento fora das especificações. Harpaz et al. [2014], por sua vez, fizeram uso de PLN e MT em fonte de dados distintas, como literatura biomédica, narrativas clínicas, redes sociais, registros de saúde, concluindo com uma discussão acerca dos desafios no uso de MT em EAM. As fontes de informações no domínio da farmacovigilância envolvem redes sociais, documentação técnica e a vivência clínica de acordo com [Harpaz et al., 2014].

Em Ong et al. [2012], foi explorada a possibilidade do emprego de classificação estatística em relatórios de incidentes clínicos a fim de identificar, automaticamente, EAM de risco elevado. De modo geral, os autores obtiveram melhores resultados de previsão com o conjunto de dados especializados (erros apenas na identificação dos pacientes) em detrimento do conjunto de dados generalista (gama diversificada de tipos de incidentes).

Outros estudos embasam a utilização de PLN no processo de identificação dos eventos adversos em medicamentos e na extração da relação entre elas. As obras de Wang et al. [2009] e Benton et al. [2011] utilizam de prontuários médicos digitais e narrativa

clínica utilizando análise estatística para evidenciar possíveis EAM. Gurulingappa et al. [2012a] e Maitra et al. [2014] fazem análises a partir de relatórios de casos médicos. O primeiro utiliza um sistema de aprendizagem de máquina para reconhecer as relações entre eventos adversos em medicamentos. O segundo cria um ambiente de exploração de texto para, com isso, melhorar o indicador de acurácia no reconhecimento de requisitos diversificados na situação de enfermidade e revelar as RAM, com evidências científicas correlacionadas reconhecidas.

2.2- Detecção de Eventos Adversos

Os registros de anestesia que fornecem dados sobre a aplicação de antibióticos no decorrer de cirurgias pertencem a um ranque oriundo da sociedade americana de anestesiologistas para identificar EAM. Em sua obra, Michelson et al. [2014] utilizaram um prontuário eletrônico e conseguiram identificar 59 EAM do tipo Infecções de Sítio Cirúrgico (ISC). Em contrapartida, o modelo de vigilância manual detectou apenas 22 EAM do tipo ISC.

No estudo conduzido por Ross et al. [2013], foi utilizado um banco de dados de genótipos e fenótipos, denominado *dbGaP*, desenvolvido com base no repositório *PubMed*, com o intuito de analisar técnicas de classificação em textos, no âmbito da recuperação de informações [Ross et al., 2013]. No decorrer do processo, foram empregados algoritmos de AM para diferenciar os metadados sobre pulmão, exames de sangue e coração, podendo ser aplicado na EAM.

No processo de detecção de RAM, a complexidade ocorre na relação de causa e efeito entre o medicamento e o sintoma, ou seja, identificar causalidade entre si. Nessa linha, Haerian et al. [2012] criaram uma metodologia para verificar a correspondência entre o medicamento e a ocorrência de eventos adversos. Obtiveram êxito, mas seus resultados foram limitados pela sensibilidade do método. A proposta foi aplicada a 119.920 registros de internação e a duas RAMs, rabdomiólise e agranulocitose, sendo capaz, em 75% dos casos, de identificar essas intercorrências com uma economia real: para cada uma hora investida, economizou-se vinte horas na revisão manual.

Na obra de Kuhn et al. [2010], os autores criaram uma base de dados denominada

Side Effect Resource, que serviu de alicerce para vários estudos posteriores em RAM. O conjunto de dados viabilizou uma combinação de 1450 RAM com 888 medicamentos sendo, para 200 desses medicamentos, possível extrair RAMs de placebo. Os dados foram distribuídos de forma gratuita. Nessa mesma linha, a obra de Duke and Friedlin [2010] apresentou uma base de dados de apoio ao negócio, com intuito de ampliar as informações sobre EAM. A base de conhecimento, denominada *SPLICER*, foi criada e mantida utilizando PLN para extrair informações de eventos adversos de rótulos estruturados de produtos. A base de dados conta com 534.125 eventos adversos e 5.602 rótulos e produtos. Foi feita uma síntese do processo de avaliação de desempenho de apoio ao negócio, de modo que o regresso dos dados de EAM ocorra após a entrada de um documento de continuidade assistencial.

Gurulingappa et al. [2012b] tiveram por intuito facilitar a validação de pesquisas que utilizem MT e EAM. Para isso, criaram um corpus e o converteram para um *treebank* com auxílio da base de dados da *Medline*, originada de relatório de casos médicos. Esse *treebank* é categorizado com o padrão ouro para medicamentos, dosagens, efeitos colaterais e a relação entre eles, buscando facilitar a criação de métodos automáticos de detecção de RAM.

Xu and Wang [2014] utilizam um amplo conjunto de dados: quatro milhões de registros do *Food and Drug Administration Adverse Event Report System - FAERS* e vinte e um milhões de artigos biomédicos da *Medline*. Foram obtidos $2.787.797$ pares de medicamentos e eventos da FAERS, com uma baixa precisão, de $0,025$. Ao incluir sinais da Medline, combinando-os com os da FAERS, ocorreu uma melhora para $0,3371$. O sistema tem sido base para várias outras pesquisas na identificação de EAM, sendo um conjunto de dados distribuído de forma gratuita no formato aberto.

Ramesh et al. [2014] também utilizou a *Food and Drug Administration Adverse Event Report System - FAERS* para, com base nos relatórios estruturados, obter os dados de narrativas contidas nesses relatórios que não são estruturados. Tais dados contém a avaliação de gravidade e a descrição dos EAM. Assim, foi desenvolvido um corpus anotado e criado de um REN com base em AM. O REN teve média F1 de $0,73$ para as entidades nomeadas. Desse modo, além do corpus, o sistema conseguiu extrair automaticamente informações sobre medicação e eventos adversos das narrativas.

Muitos trabalhos abordam detecção de RAM relacionados à vacinas aplicando técnicas de PLN e MT em conjuntos de dados como o *Vaccine Adverse Event Reporting*

System [Baer et al., 2016] e [Courtot et al., 2014]. Para Hur et al. [2012], a *febre* é uma das RAM mais comuns. *Febre* é o processo de aumento da temperatura corporal acima do normal e acontece, em sua maioria, devido à alguma enfermidade Hur et al. [2012].

Na obra de Hur et al. [2012], os autores buscaram compreender a *febre* como RAM mais comum, por meio de interações genéticas, utilizando a base de dados *PubMed*. O resultado foi uma rede de 403 genes e 577 relações entre genes e febre após vacinação. Hazlehurst et al. [2009], trabalharam na identificação de RAM, também após vacinação, modificando o sistema *Mediclass* proposto originalmente por Hazlehurst et al. [2005]. O grupo obteve resultados satisfatórios empregando PLN.

2.3- Farmacovigilância no Âmbito das Redes Sociais

O advento das redes sociais trouxe consigo uma nova e valiosa fonte de dados, uma vez que o público compartilha experiências médicas pessoais uns com os outros por meio delas. Dessa maneira, os dados das redes sociais se apresentam como um novo desafio às técnicas de PLN no processo de identificar RAM.

Cocos et al. [2017] utilizou um processo de aprendizagem profundo de máquina para identificar as RAM em postagens no Twitter. Já Batista and Figueira [2017] fez uso de quatro ferramentas (coreNLP, *GATE*, *OpenNLP* e *Twitter NLP*) para compreender como tais ferramentas identificam entidades nomeadas e se combiná-las pode ter um melhor resultado no REN em postagens no Twitter. Alvaro et al. [2015], por sua vez, analisou a extração de dados do Twitter usando aprendizagem de máquina com colaboração coletiva, além de anotadores leigos para extrair a relação entre medicamentos e eventos adversos. Rajapaksha and Weerasinghe [2015] propõem e avaliam um modelo automático de extração onde cada EAM mencionado em postagens do Twitter é classificando em efeito adverso ou outro efeito. Após essa etapa, é classificado novamente para que se possa ser identificado como *efeito conhecido* ou *efeito desconhecido*. O'Connor et al. [2014] analisou os posts do Twitter utilizando o tamanho do efeito com Cohen e kappa para calcular a relação entre medicamentos e efeitos adversos.

No âmbito das redes sociais, existem outras obras que abordam o *DailyStrength Twitter*. O *DailyStrength* é uma rede social que atua como grupo de apoio, no qual

os usuários narram suas experiências de vida e situação médica[1]. Nikfarjam et al. [2015] desenvolveram o *ADRMine*, um sistema de extração de conceitos que utiliza aprendizagem de máquina e CRF, um tipo de modelagem estatística de predição [Settles, 2004].

Outras obras tratam do desempenho da extração de conceitos utilizando uma abordagem com ganho de desempenho. Sarker and Gonzalez [2015] fez uso das duas redes sociais, (Twitter e DailyStrength) e criou dois conjuntos de dados. Gurulingappa et al. [2012a] criou o corpus *ADE* para detecção de RAM, incluindo recursos semânticos, como sentimentos e polaridades. Ambos os trabalhos também fizeram uso do aprendizado de máquina para a classificação automática de trechos do texto de RAM e apresentaram como resultado que os métodos usados, bem como o incremento de características semânticas, aprimoram o desempenho da classificação.

Pesquisas anteriores sobre farmacovigilância em redes sociais mostraram resultados favoráveis na identificação de EAM com base nos debates ocorridos nas mesmas, ressaltando a importância científica das mídias sociais na detecção de sinais de segurança para novos medicamentos, evidenciado por Bian et al. [2012] e, para identificar medicamentos perigosas, como explicitado por Chee et al. [2011]. Uma grande dificuldade para a pesquisa de farmacovigilância em redes sociais é a falta de dados anotados [Mintz et al., 2009].

2.4- Considerações

Existem ainda, trabalhos que discutem a identificação de entidades nomeadas com auxílio de vocabulários controlados para a anotação de corpora. Tais vocabulários predominam no UMLS [Bodenreider, 2004] e MedDRA[2]. A Unified Medical Language System) (UMLS) é utilizada nos seguintes trabalhos Sarker and Gonzalez [2015], Nikfarjam et al. [2015], Liu and Chen [2015], Ramesh et al. [2014] e a MedDRA nas obras Ly et al. [2018], Rajapaksha and Weerasinghe [2015], Maitra et al. [2014] , Yeleswarapu et al. [2014].

Todas as obras citadas até o presente momento utilizaram como fonte primária de

[1] `https://www.dailystrength.org`
[2] https://www.meddra.org

dados, informações na língua inglesa. No entanto, existem pesquisas no ramo do PLN implementadas em outros idiomas, por exemplo na língua espanhola Oronoz et al. [2015], [Segura-Bedmar et al., 2015]. E na língua francesa Chen et al. [2018], [Morlane-Hondère et al., 2016].

Oronoz et al. [2015] desenvolveu um corpus padrão ouro, anotado manualmente por especialistas a partir de registros eletrônicos de saúde com foco em medicamento e doenças, cujo nome é *IxaMed-GS*. Já Segura-Bedmar et al. [2015] criou um corpus para ser utilizado na extração de EAM em abordagens que utilizem aprendizagem de máquina; também apresenta um sistema para detectar EAM e indicações de medicamentos a partir de mensagens de usuários extraídas de um fórum de saúde espanhol. Os autores desenvolveram o primeiro banco de dados espanhol construído automaticamente, *SpanishDrugEffectDB*, não utiliza dados rotulados, o estudo emprega a obra de Min et al. [2013] de Distant Supervision (DS). Basicamente a supervisão distante (DS) tem sido amplamente utilizada para treinar extratores de relações sem usar dados marcados manualmente. Ele cria automaticamente exemplos de treinamento, rotulando as menções de relação (Uma ocorrência de um par de entidades com a sentença de origem.) [Deng and Sun, 2019].dd Dadas duas entidades, a supervisão distante explora sentenças que as mencionam diretamente para prever sua relação semântica [Deng and Sun, 2019].

Chen et al. [2018] exibiu, em seu trabalho, uma ferramenta de visualização e mineração de textos que é implementada pelas perspectivas dos usuários de cinco fóruns de saúde franceses populares, que fornecem os dados para o modelo *aberto*. Os autores fizeram uso de métodos de REN e extração das relações utilizando, como corpus, o *RacinePharma*, que abrange os medicamentos existentes no mercado francês, e o Anatomical Therapeutic Chemical (ATC) para a extração das relações[3] e sistema de classificação de medicamentos e, por fim, o MedDRA para os distúrbios. Morlane-Hondère et al. [2016] descreveu, por sua vez, um sistema que extrai entidades médicas de revisões de medicamentos francesas escritas por usuários. Dois classificadores são usados: o primeiro realiza a extração de entidades médicas mínimas e o segundo, os combina para reconhecer entidades mais complexas.

As obras são listadas por categorias, conforme apresentadas. As categorias não são, no entanto, mutuamente excludentes. Desse modo, uma obra pode pertencer a mais de uma categoria, sendo elas: *Processamento de Linguagem Natural*, *Detecção de*

[3]https://www.whocc.no/atc

Eventos Adversos em inglês, *Farmacovigilância no Âmbito das Redes Sociais* e *Detecção de Eventos Adversos em outros idiomas*. A tabela 5 apresenta os referidos trabalhos divididos por categorias.

Processamento de Linguagem Natural	Detecção de Eventos Adversos em inglês	Farmacovigilância no Âmbito das Redes Sociais	Detecção de Eventos Adversos em outros idiomas
Pushpa and Kamakshi [2018]	Michelson et al. [2014]	Cocos et al. [2017]	Sarker and Gonzalez [2015]
Luo et al. [2017]	Ross et al. [2013]	Batista and Figueira [2017]	Nikfarjam et al. [2015]
Harpaz et al. [2014]	Clark et al. [2014]	Alvaro et al. [2015]	Liu and Chen [2015]
Ong et al. [2012]	Hazlehurst et al. [2005]	Rajapaksha and Weerasinghe [2015]	Ramesh et al. [2014]
Wang et al. [2009]	Haerian et al. [2012]	O'Connor et al. [2014]	Ly et al. [2018]
Benton et al. [2011]	Kuhn et al. [2010]	Nikfarjam et al. [2015]	Rajapaksha and Weerasinghe [2015]
Gurulingappa et al. [2012a]	Duke and Friedlin [2010]	Settles [2004]	Maitra et al. [2014]
Maitra et al. [2014]	Gurulingappa et al. [2012b]	Sarker and Gonzalez [2015]	Yeleswarapu et al. [2014]
	Xu and Wang [2014]	Gurulingappa et al. [2012a]	Oronoz et al. [2015]
	Ramesh et al. [2014]	Bian et al. [2012]	Segura-Bedmar et al. [2015]
	Baer et al. [2016]	Chee et al. [2011]	Chen et al. [2018]
	Botsis et al. [2011]	Mintz et al. [2009]	Morlane-Hondère et al. [2016]
	Botsis et al. [2012]		Oronoz et al. [2015]
	Botsis et al. [2013]		Segura-Bedmar et al. [2015]
	Courtot et al. [2014]		Chen et al. [2018]
	Hur et al. [2012]		Morlane-Hondère et al. [2016]
	Hazlehurst et al. [2009]		
	Hazlehurst et al. [2005]		

Tabela 5 – Classificação dos Trabalhos Relacionados

Este trabalho difere dos supracitados por: (i) criar um modelo para extrair informação em português brasileiro. (ii) se concentrar em documentos informais. (iii) possuir termos obtidos do bulário online *Bulario.com* e da rede social Twitter, que estão no idioma supracitado. (iv) criar PosTagger, DepParser e REN em português brasileiro por meio da coreNLP. (v) usar filtro semântico para flexibilização de regras na detecção de relação de eventos adversos. E, assim, iniciar o suporte no domínio da farmacovigilância em português do Brasil.

3- Proposta

No capítulo 1 foi apresentado o tema, que tem por objetivo identificar eventos adversos de medicamentos em textos. Neste capítulo, é apresentada uma proposta para cumprir o objetivo proposto na língua portuguesa brasileira em textos informais empregando modelo de EI e PLN com AM. Este capítulo é dedicado a explicitar a proposta e os preceitos metodológicos que norteiam essa dissertação. Trata-se de um estudo exploratório de produção tecnológica. Com abordagem quantitativa, no que diz respeito aos modelos criados, e qualitativa, na detecção de eventos adversos.

A pesquisa metodológica aprimora ferramentas e envolve, em sua maioria, métodos complexos e refinados, bem com uma condução rigorosa da pesquisa, sendo responsável pela elaboração, validação e avaliação de ferramentas e métodos de pesquisa. O estudo metodológico tem como alvo a criação de instrumento confiável, funcional e preciso, de modo a ser utilizado, em sua maioria por outros pesquisadores [Pollit et al., 2004].

Até o presente momento não existem obras que abordem a extração de eventos adversos em língua portuguesa brasileira, dadas as características próprias da língua. Não sendo possível, tampouco, uma transição direta entre outras línguas. Ciente de tais limitações e visando contribuir para a pesquisa na área, esse trabalho objetiva iniciar os estudos sobre o tema.

Nesta dissertação é utilizado um modelo de extração da informação conhecido por *template*, criado para extração EAM. Emprega-se PLN na implementação dos modelos (PosTagger, DepParser e REN), fazendo uso de AM. Emprega-se a arquitetura da figura 11 como ferramenta informatizada, de forma a contribuir para o processo de notificação de eventos adversos em saúde.

A figura 11 exemplifica a elaboração do modelo proposto. Inicialmente, a fonte de dados *tweets* (postagens da rede social twitter). Com o conjunto de dados definido e tratado, foi feita sua submissão ao sistema de extração de informação (sendo a proposta desta dissertação), de modo a analisar seu conteúdo em busca de medicamentos e eventos (nessa dissertação, eventos devem ser entendido como sintomas ou doenças). Um *tweet* sem medicamento, não auxilia no processo de detecção de sinal, afinal, não seria

possível saber a qual medicamento relacioná-lo, de modo o inverso, um medicamento sem evento relatado, não pode ser associado a uma queixa. Resumidamente, a relação de causa e efeito entre eles precisa ser estabelecida, de modo que, dado o medicamento ou evento, haja a verificação da ocorrência de eventos adversos.

Pode-se tomar como exemplo o seguinte *tweet*: “Dor de cabeça . . . Tomei dipirona mas tá dando enjoo”. O resultado esperado desse *tweet* é a identificação dos eventos *Dor de cabeça* e *enjoo*, sendo que o evento *Dor de cabeça* está associado ao motivo de usar o medicamento e o evento *enjoo*, ao uso do medicamento, nesse caso, a *dipirona*. Desse modo, tem-se um evento ocorrido do uso do medicamento, o que caracteriza um evento adverso. O modelo de sistema de informação proposto deve ser capaz de alcançar esse resultado.

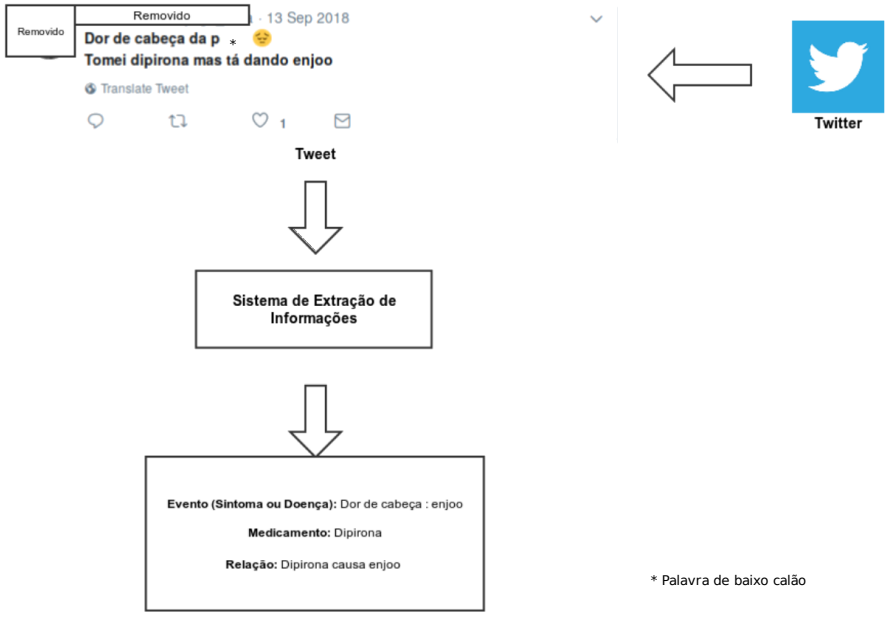

Figura 11 – Modelo teórico proposto.

Com o modelo proposto da solução definida, foi verificada sua viabilidade do ponto de vista prática. Para tal, a tecnologia adotada foi o PLN, juntamente com a tarefa de REN, tarefa esta, amplamente difundida que atende a proposta de identificar os medicamentos e eventos existentes no texto e, com isso, obtém-se a viabilidade de parte do modelo. A outra parte consiste na detecção de relação entre os elementos (sintáticos e morfológicos) na língua portuguesa brasileira. Não foram encontradas obras para relações em língua portuguesa brasileira no domínio da farmacovigilância. Entretanto, essa etapa se trata de um estudo exploratório, que não tem a intenção de atingir o estado da arte existente em

outros idiomas, apenas iniciar os estudos sobre o tema no idioma alvo. O filtro semântico, é responsável por atribuir a relação de causa e efeito, ou seja, associar medicamento causa evento. O filtro semântico abordado com mais detalhes na subseção 3.1.

A figura 12 exemplifica os processos de REN e o filtro semântico. Dada a frase "fui tomar uma dipirona por causa da dor nas costas e fiquei com dor de cabeça aff", a tarefa de REN deve ser capaz de reconhecer *dipirona* como medicamento e, *dor nas costas* e *dor de cabeça* como eventos. O filtro semântico deve associar que o evento *dor nas costas* motivou o uso do medicamento *dipirona* e que o evento *dor de cabeça* foi ocasionado pelo uso do medicamento, marcando o *tweet* como 'com evento adverso'.

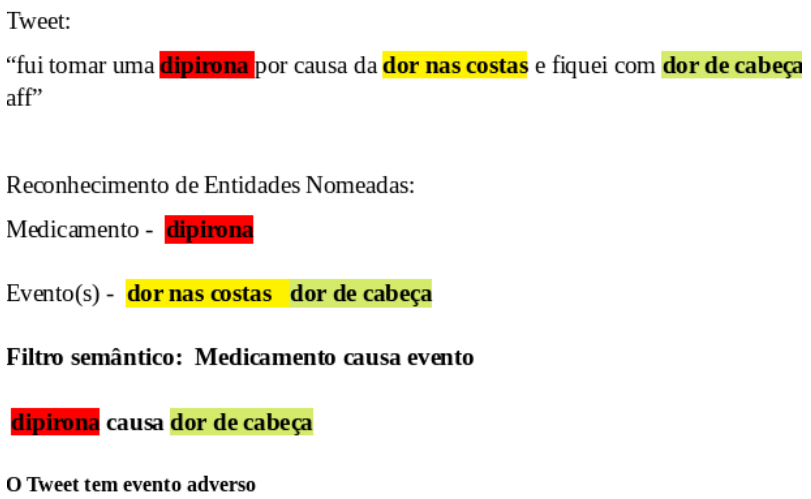

Figura 12 – Exemplo de utilização do PLN.

Do ponto de vista operacional, o software coreNLP não tem suporte nativo para a língua portuguesa brasileira, sendo necessário expandir sua capacidade e treinar o REN para detectar medicamentos e eventos, sendo assim precisamos dos dados disponíveis em formato adequado para treinamento. A tarefa de REN requer o uso de PosTagger. Basicamente, ele atua como etiquetador atribuindo a classe gramatical ao elemento.

O REN e o PosTagger são modelos estatísticos que devem ser treinados. O DepParser utiliza a teoria gramatical de dependência para apresentar as relações gramaticais. É, também, um modelo estatístico que deve ser treinado e faz uso do PosTagger. Uma vez que há três modelos distintos, foi priorizado o treinamento e teste destes para posterior implementação do filtro semântico.

A figura 13 apresenta o processo de criação dos modelos estatísticos utilizados. Para treinar os modelos se faz necessário o uso de dados, então duas etapas anteriores de coleta de dados e pré-processamento foi realizada. Nessa etapa, como fonte de dados,

foram utilizados os dicionários léxicos ou toponímicos criados ao longo dessa dissertação e o projeto *Universal Dependencies*. Após a coleta e processamento dos dados, o software coreNLP foi utilizado para treinar o PosTagger com os dados pré-processados do *Universal Dependencies*. Em seguida, foi feito o treinamento do DepParser fornecendo o PosTagger, os vetores de palavras e os dados do projeto *Universal Dependencies* como parâmetros. O próximo passo consistiu no treinamento do REN, isso foi feito fornecendo os dados do Twitter e dos Dicionários toponímicos. Por fim, efetuou-se a validação dos modelos. Os detalhes do processo são abordados na seção 5.

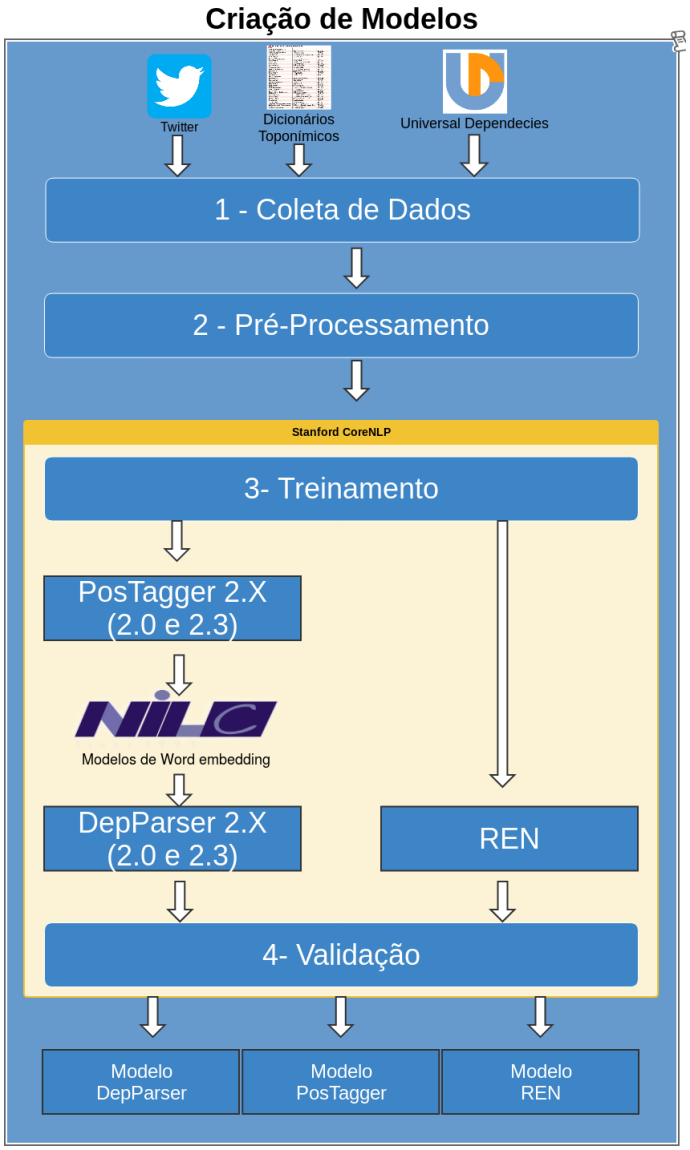

Figura 13 – Modelo de criação dos modelos estatísticos.

Concluindo, a figura 14 apresenta a estrutura da aplicação criada para identificação

de eventos adversos. O começo se dá pela coleta de dados do Twitter e seu pré-processamento. Posteriormente, utiliza-se o software coreNLP com os modelos estatísticos criados (PosTagger, DepParser e REN), exportando os dados em um *XML* em seguida é executado o algoritmo SaúdeAlg (que contém o filtro semântico), a fim de detectar a presença de possível eventos adversos.

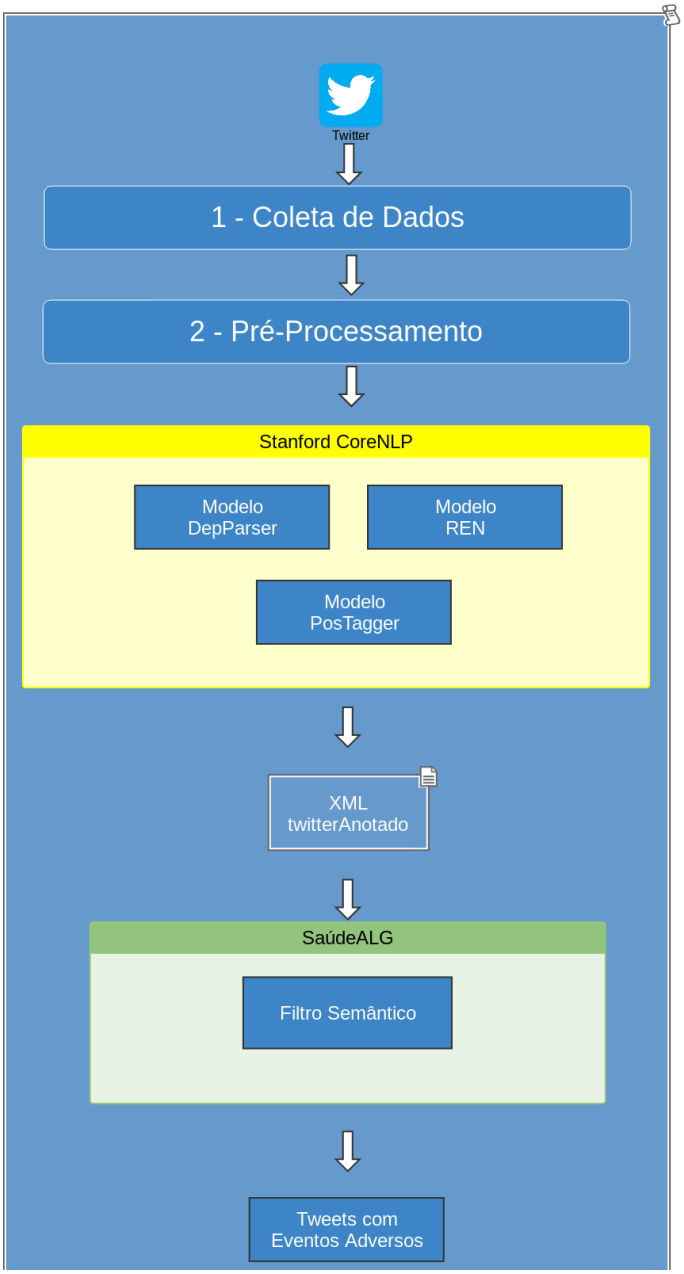

Figura 14 – Proposta da aplicação empregada nessa dissertação.

3.1- Filtro Semântico

Inicialmente, foi realizada uma abordagem exploratória dos dados, pela falta de recursos para exploração da língua portuguesa brasileira. Essa abordagem, não tem por objetivo esgotar o tema, mas mostrar que é possível usar outras estratégias e apresentar uma solução viável para o problema proposto.

A seguir, foi verificada a relação dos dados já registrados pelo REN com a anotação de medicamentos e eventos dos *tweets*. Também foi estabelecida a relação de dependência entre os termos existentes. Essa relação é criada através da anotação do DepParser: uma relação de dependência é formada por dois elementos, o governador e o dependente. Cada elemento da relação contém o id do token ao qual faz relação.

Conforme visto na figura 14, o software coreNLP faz anotação de todo o processo e ao término exporta um arquivo XML com os dados anotados (denominado *twitterAnotado*). Esse arquivo é, então, utilizado como entrada para o algoritmo SaúdeAlg, que consiste do filtro semântico.

A partir da proposta descrita nesse capítulo, inicialmente, foi feito um filtro para identificar a presença dos elementos *medicamento*, *evento*, *advérbio* e *verbo* e, na ausência de algum deles, os *tweets* devem ser considerados sem evento adverso (as linhas do algoritmo SaúdeAlg de 5 até 10 realizam essa atividade). Da linha 12 em diante, é realizada a chamada da função *buscaRegras* que implementa as seguintes regras:

Com base na gramática de dependência, foi realizado um processo de observação, buscando encontrar, manualmente, um padrão. A relação de dependência, nomeada como *adpmod* e *adpobj* tem correlação com alguns casos de *tweets* que relatam eventos adversos, em condições determinadas, descritas abaixo.

Primeira regra (relação entre *adpmod* e *adpobj*):

- Relação *adpmod*
  - O elemento governador deve ser um verbo (POS=VERB);
  - Esse verbo deve ter seu id posterior ao do medicamento (REN=medicamento), ou seja, verbo.id $>$ medicamento.id;
  - O elemento dependente deve ser um advérbio (POS=ADP).
- Relação *adpobj*

- O elemento governador deve ser o mesmo advérbio utilizado na relação *adpmod*;
- O elemento dependente deve ser o evento (REN=evento).

Os exemplos das regras irão conter a saída do programa para facilitar o entendimento. Podemos notar quatro seções nas figuras que representam as regras *Sentence*, *Tokens*, *Dependency Parse* e *Extracted the following NER entity mentions*, vai ser abreviada para *NER*. A seção *Sentence* apresenta a senteça utilizada, no exemplo “tomar o remédio paracetamol e não ficar com azia é impossível”. A seção *Tokens* lista todos os *tokens* da sentença, *CharacterOffsetBegin* e *CharacterOffsetEnd* representam a posição inicial e final não sendo utilizado, *PartOfSpeech* é o valor do PosTagger, *Lemma* forma lamtizada do verbo não sendo utilizado e *NamedEntityTag* que representa o REN. A seção *Dependency Parse* lista as relações de dependências, sendo o primeiro elemento o governador e o segundo o dependente.

A figura 15, apresenta um exemplo da aplicação da primeira regra. Conforme a regra 1, analisamos as relações *adpmod* e *adpobj*. São elas adpmod(ficar-7, com-8) e adpobj(com-8, azia-9). Na relação adpmod podemos ver o *token* ficar como verbo [Text=ficar PartOfSpeech=VERB NamedEntityTag=O], depois o medicamento posterior ao verbo, no caso o *token* paracetamol [Text=paracetamol PartOfSpeech=ADJ NamedEntityTag=DRUG], ainda temos posição (id) do verbo maior que a do medicamento *token* ficar na 7 posição e *token* paracetamol na 4 posição, posteriormente o elemento dependente deve ser um advérbio (ADP) aqui representado pelo *token* com [Text=com PartOfSpeech=ADP NamedEntityTag=O] fim da relação *adpmod*. Na relação *adpobj* temos que ter o mesmo advérbio utilizado na relação anterior como governador adpobj(com-8, azia-9) e por fim o evento como item dependente [Text=azia PartOfSpeech=NOUN NamedEntityTag=EVENT], tendo todos os requisitos, temos um evento adverso.

```
Sentence #1 (12 tokens):
tomar o remédio paracetamol e não ficar com azia é impossível.

Tokens:
[Text=tomar CharacterOffsetBegin=1 CharacterOffsetEnd=6 PartOfSpeech=VERB
Lemma=tomar NamedEntityTag=O]
[Text=o CharacterOffsetBegin=7 CharacterOffsetEnd=8 PartOfSpeech=DET Lemma=o
NamedEntityTag=O]
[Text=remédio CharacterOffsetBegin=9 CharacterOffsetEnd=16 PartOfSpeech=NOUN
Lemma=remédio NamedEntityTag=O]
[Text=paracetamol CharacterOffsetBegin=17 CharacterOffsetEnd=28
PartOfSpeech=ADJ Lemma=paracetamol NamedEntityTag=DRUG]
[Text=e CharacterOffsetBegin=29 CharacterOffsetEnd=30 PartOfSpeech=CONJ Lemma=e
NamedEntityTag=O]
[Text=não CharacterOffsetBegin=31 CharacterOffsetEnd=34 PartOfSpeech=ADV
Lemma=não NamedEntityTag=O]
[Text=ficar CharacterOffsetBegin=35 CharacterOffsetEnd=40 PartOfSpeech=VERB
Lemma=ficar NamedEntityTag=O]
[Text=com CharacterOffsetBegin=41 CharacterOffsetEnd=44 PartOfSpeech=ADP
Lemma=com NamedEntityTag=O]
[Text=azia CharacterOffsetBegin=45 CharacterOffsetEnd=49 PartOfSpeech=NOUN
Lemma=azia NamedEntityTag=EVENT]
[Text=é CharacterOffsetBegin=50 CharacterOffsetEnd=51 PartOfSpeech=AUX Lemma=é
NamedEntityTag=O]
[Text=impossível CharacterOffsetBegin=52 CharacterOffsetEnd=62
PartOfSpeech=VERB Lemma=impossível NamedEntityTag=O]
[Text=. CharacterOffsetBegin=62 CharacterOffsetEnd=63 PartOfSpeech=. Lemma=.
NamedEntityTag=O]

Dependency Parse (enhanced plus plus dependencies):
root(ROOT-0, impossível-11)
csubjpass(impossível-11, tomar-1)
det(remédio-3, o-2)
dobj(tomar-1, remédio-3)
amod(remédio-3, paracetamol-4)
cc(tomar-1, e-5)
neg(ficar-7, não-6)
conj:e(tomar-1, ficar-7)
csubjpass(impossível-11, ficar-7)
adpmod(ficar-7, com-8)
adpobj(com-8, azia-9)
auxpass(impossível-11, é-10)
p(impossível-11, .-12)

Extracted the following NER entity mentions:
paracetamol     DRUG
azia            EVENT
```

Figura 15 – Regra 1 do algoritmo SaúdeAlg.

Segunda regra (aplicada apenas para o evento “alergia” (REN=evento) e (evento=alergia): Verbo anterior ao evento e medicamento.

- Evento = Alergia (REN=evento=alergia);
- Evento (REN=evento=alergia) e medicamento (REN=medicamento) posteriores ao verbo.

A fim apresentar a flexibilidade e possibilidades do filtro semântico, foi selecionado

apenas um evento (alergia) e criado um regra específica. Essa possibilidade permite criar e manipular conjuntos de regra conforme o propósito. Por exemplo um laboratório farmacêutico tem uma visão diferente da ANVISA.

A figura 16, apresenta um exemplo da aplicação da segunda regra. Temos a sentença “ah hoje foi um dia incrível cheguei a conclusão que eu realmente tenho alergia de dipirona.”. Primeiramente temos que identificar o evento como alergia (*NER* = alergia), em seguida o evento e medicamento posterior ao verbo. *Token* foi na 3 posição como verbo, *dipirona* como medicamento na 16 posição e *token* alergia como evento na posição 14. Satisfeitos os requisitos temos um evento adverso.

```
Sentence #2 (17 tokens):
ah hoje foi um dia incrível cheguei a conclusão que eu realmente tenho alergia de dipirona.

Tokens:
[Text=ah CharacterOffsetBegin=63 CharacterOffsetEnd=65 PartOfSpeech=ADP Lemma=ah NamedEntityTag=O]
[Text=hoje CharacterOffsetBegin=66 CharacterOffsetEnd=70 PartOfSpeech=ADV Lemma=hoje NamedEntityTag=O]
[Text=foi CharacterOffsetBegin=71 CharacterOffsetEnd=74 PartOfSpeech=VERB Lemma=foi NamedEntityTag=O]
[Text=um CharacterOffsetBegin=75 CharacterOffsetEnd=77 PartOfSpeech=DET Lemma=um NamedEntityTag=O]
[Text=dia CharacterOffsetBegin=78 CharacterOffsetEnd=81 PartOfSpeech=NOUN Lemma=dia NamedEntityTag=O]
[Text=incrível CharacterOffsetBegin=82 CharacterOffsetEnd=90 PartOfSpeech=ADJ Lemma=incrível NamedEntityTag=O]
[Text=cheguei CharacterOffsetBegin=91 CharacterOffsetEnd=98 PartOfSpeech=AUX Lemma=cheguei NamedEntityTag=O]
[Text=a CharacterOffsetBegin=99 CharacterOffsetEnd=100 PartOfSpeech=ADP Lemma=a NamedEntityTag=O]
[Text=conclusão CharacterOffsetBegin=101 CharacterOffsetEnd=110 PartOfSpeech=NOUN Lemma=conclusão
NamedEntityTag=O]
[Text=que CharacterOffsetBegin=111 CharacterOffsetEnd=114 PartOfSpeech=PRON Lemma=que NamedEntityTag=O]
[Text=eu CharacterOffsetBegin=115 CharacterOffsetEnd=117 PartOfSpeech=PRON Lemma=eu NamedEntityTag=O]
[Text=realmente CharacterOffsetBegin=118 CharacterOffsetEnd=127 PartOfSpeech=ADV Lemma=realmente
NamedEntityTag=O]
[Text=tenho CharacterOffsetBegin=128 CharacterOffsetEnd=133 PartOfSpeech=AUX Lemma=tenho NamedEntityTag=O]
[Text=alergia CharacterOffsetBegin=134 CharacterOffsetEnd=141 PartOfSpeech=VERB Lemma=alergia
NamedEntityTag=EVENT]
[Text=de CharacterOffsetBegin=142 CharacterOffsetEnd=144 PartOfSpeech=ADP Lemma=de NamedEntityTag=O]
[Text=dipirona CharacterOffsetBegin=145 CharacterOffsetEnd=153 PartOfSpeech=NOUN Lemma=dipirona
NamedEntityTag=DRUG]
[Text=. CharacterOffsetBegin=153 CharacterOffsetEnd=154 PartOfSpeech=. Lemma=. NamedEntityTag=O]

Dependency Parse (enhanced plus plus dependencies):
root(ROOT-0, foi-3)
adpmod(foi-3, ah-1)
adpcomp(ah-1, hoje-2)
det(dia-5, um-4)
attr(foi-3, dia-5)
amod(dia-5, incrível-6)
aux(conclusão-9, cheguei-7)
adp(conclusão-9, a-8)
attr(foi-3, conclusão-9)
dobj(alergia-14, que-10)
nsubj(alergia-14, eu-11)
advmod(alergia-14, realmente-12)
aux(alergia-14, tenho-13)
rcmod(conclusão-9, alergia-14)
adpmod(alergia-14, de-15)
adpobj(de-15, dipirona-16)
p(foi-3, .-17)

Extracted the following NER entity mentions:
alergia EVENT
dipirona        DRUG
```

Figura 16 – Regra 2 do algoritmo SaúdeAlg.

Terceira regra (aplicada apenas para o evento “alergia” (REN=evento) e (evento=alergia): Evento sucedido com o artigo “a” (Posição do REN=evento + 1); medicamentos (REN=medicamento) posteriores a eles.

- Evento = Alergia (REN=evento=alergia);

- Evento sucedido pelo artigo “a”;
- Medicamento posterior a eles.

A terceira regra é uma variação da segunda, com objetivo de melhorar a precisão. Analisando os dados o evento “alergia” em sua maioria é sucedido do artigo “a”. Essa nova regra seleciona essa sequência exata de “alergia a”.

A figura 17, apresenta um exemplo da aplicação da terceira regra. Temos a sentença “alergia a dipirona agora parabéns eu conseguiu me superar um bjooo.”. Primeiramente temos que identificar o evento como alergia, *NER* alergia EVENT sendo o próximo *token* necessariamente o artigo a, o *token* alergia ocupa a 1 posição posterior o *token* a na posição 2 e o medicamento na posição 3.

```
Sentence #3 (12 tokens):
alergia a dipirona agora parabéns eu conseguiu me superar um bjooo.

Tokens:
[Text=alergia CharacterOffsetBegin=155 CharacterOffsetEnd=162 PartOfSpeech=VERB Lemma=alergia NamedEntityTag=EVENT]
[Text=a CharacterOffsetBegin=163 CharacterOffsetEnd=164 PartOfSpeech=ADP Lemma=a NamedEntityTag=O]
[Text=dipirona CharacterOffsetBegin=165 CharacterOffsetEnd=173 PartOfSpeech=VERB Lemma=dipirona NamedEntityTag=DRUG]
[Text=agora CharacterOffsetBegin=174 CharacterOffsetEnd=179 PartOfSpeech=ADV Lemma=agora NamedEntityTag=O]
[Text=parabéns CharacterOffsetBegin=180 CharacterOffsetEnd=188 PartOfSpeech=NOUN Lemma=parabéns NamedEntityTag=O]
[Text=eu CharacterOffsetBegin=189 CharacterOffsetEnd=191 PartOfSpeech=PRON Lemma=eu NamedEntityTag=O]
[Text=conseguiu CharacterOffsetBegin=192 CharacterOffsetEnd=201 PartOfSpeech=VERB Lemma=conseguiu NamedEntityTag=O]
[Text=me CharacterOffsetBegin=202 CharacterOffsetEnd=204 PartOfSpeech=PRON Lemma=I NamedEntityTag=O]
[Text=superar CharacterOffsetBegin=205 CharacterOffsetEnd=212 PartOfSpeech=VERB Lemma=superar NamedEntityTag=O]
[Text=um CharacterOffsetBegin=213 CharacterOffsetEnd=215 PartOfSpeech=DET Lemma=um NamedEntityTag=O]
[Text=bjooo CharacterOffsetBegin=216 CharacterOffsetEnd=221 PartOfSpeech=NOUN Lemma=bjooo NamedEntityTag=O]
[Text=. CharacterOffsetBegin=221 CharacterOffsetEnd=222 PartOfSpeech=. Lemma=. NamedEntityTag=O]

Dependency Parse (enhanced plus plus dependencies):
root(ROOT-0, alergia-1)
adpmod(conseguiu-7, a-2)
adpobj(a-2, dipirona-3)
advmod(dipirona-3, agora-4)
dobj(dipirona-3, parabéns-5)
nsubj(conseguiu-7, eu-6)
ccomp(alergia-1, conseguiu-7)
dobj(superar-9, me-8)
xcomp(conseguiu-7, superar-9)
det(bjooo-11, um-10)
dobj(superar-9, bjooo-11)
p(alergia-1, .-12)

Extracted the following NER entity mentions:
alergia EVENT
dipirona        DRUG
```

Figura 17 – Regra 3 do algoritmo SaúdeAlg.

Algoritmo 1 – SaúdeAlg (twittersAnotados)

Entrada: twittersAnotados

Saída: twittersFinais

```
1  início
2      // Total de frases após o Parse do XML
3      tamanho ← twittersAnotados.sentences.size()
4      i = 0
5      while i ⩽ tamanho do
6          Med ← buscaNER_Medicamento(twittersAnotados[NER][i]);
7          Event ← buscaNER_Evento(twittersAnotados[NER][i]);
8          ADP ← buscaPOS_Advérbio(twittersAnotados[POS][i]);
9          VERB ← buscaPOS_Verbo(twittersAnotados[POS][i]);
10         if Med OR Event = NULL then
11             break;
12         else
13             lista ← buscaRegras(twittersAnotados[i]);
14             if lista <> NULL then
15                 twittersFinais ← twittersAnotados[i].text =
                    "TemEventoAdverso";
16             else
17                 twittersFinais ← twittersAnotados[i].text =
                    "NãoTemEventoAdverso";
18     return twittersFinais;
```

3.2- Execução da Metodologia

Nesta seção, são explicadas as etapas da metodologia para identificação dos sinais em farmacovigilância que norteiam essa dissertação. A figura 18 lista as cinco etapas e a principal atividade realizada em cada uma sendo, nesta ordem: *coleta de dados*, *pré-processamento*, *treinamento*, *validação* e *classificação*.

A execução da metodologia descreve como, dado o conjunto de dados é possível, identificar sinais de eventos adversos. Por utilizar bases de dados textuais, pode-se empregar duas abordagens: estatística e semântica.

A abordagem estatística é fundamentada em AM, e tem como base a frequência em que os termos aparecem. Por sua vez, a abordagem semântica se baseia em gramática, que consiste no significado dos termos.

A fim de solucionar o enunciado, utilizou-se uma abordagem mista. Para o REN, o PosTagger e DepParser foi escolhida a abordagem estatística e para o filtro semântico, a semântica.

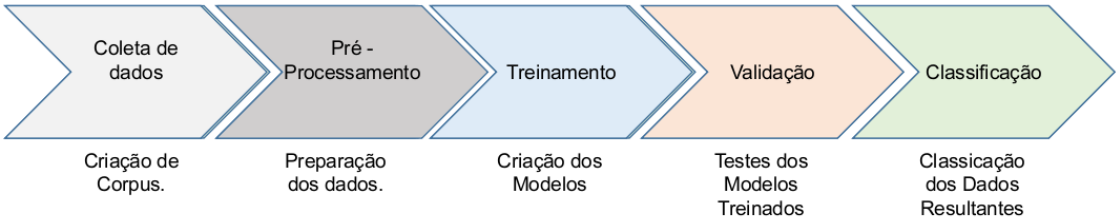

Figura 18 – Cronologia das etapas dessa dissertação adaptado de Aranha et al. [2007].

A primeira etapa é a coleta de dados que abrange a seleção de documentos que irão constituir os corpora, essencial para outras etapas. Pode-se destacar a relevância que os documentos devem ter com o segmento do conhecimento a ser extraído. A seleção de documentos irrelevantes poderia comprometer a proposta desta dissertação pois os corpora foram utilizados para criação do DepParser, do PosTagger e do REN.

Para DepParser e o PosTagger utilizou-se o projeto *UniversalDependencies* [Bosco et al., 2013; De Marneffe et al., 2006; McDonald et al., 2013; Haverinen et al., 2014], mais precisamente, o corpus em português do Brasil denominado *UD_Portuguese-GSD* nas versões 2.0, utilizada no início deste trabalho, e 2.3, versão atual. E, para o REN utilizou-se os dicionários toponímicos compostos dos dados de medicamentos, substâncias ativas e eventos, criados com base em uma pesquisa manual nos principais laboratórios farmacêuticos e no site da ANVISA [da Cunha et al., 2018].

A segunda etapa é conhecida por pré-processamento. Para Gonçalves et al. [2006], essa etapa é encarregada da tarefa de definir uma representação estruturada para os dados, de modo que possam ser interpretados por computadores. Os sistemas de MT não agem diretamente, com seus algoritmos, sobre conjuntos de dados despreparados [Corrêa et al., 2003]. Após a realização da coleta de dados, precisa-se adequar o conjunto

de modo que os dados possam ser manipulados.

Pré-processamento é, em muitos casos, o processo mais custoso de uma metodologia baseada em MT, tendo em vista que não existe uma única abordagem que seja aplicável a todos os domínios. Dessa forma, são necessários muitos experimentos, baseados em tentativa e erro, para chegar a um modelo adequado [Feldman et al., 2007]. O objetivo principal da etapa de pré-processamento é a limpeza de dados, preenchimento dos valores faltantes, remoção de dados ruidosos e discrepantes e, resolução de inconsistências no domínio do qual se deseja extrair os dados [Han et al., 2011].

A figura 19 exemplifica a tarefa de tokenização da etapa do pré-processamento. Na coluna da esquerda se encontra o *tweet* e, na da direita, a tokenização. Pode-se notar que o termo "dor de cabeça", mesmo com espaçamento entre os elementos, é reconhecido como uma única unidade.

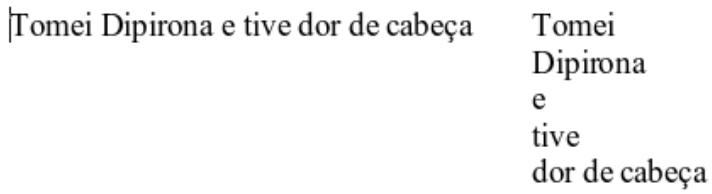

Figura 19 – Exemplo do processo de Tokenização.

A figura 20 apresenta um modelo de arquivo lematizado. No primeiro exemplo, "*Os remédios me fizeram mal.*", os termos "**Os**" e "**remédios**" são flexionados em número para, respectivamente, "**O**" e "**remédio**", e o verbo "**fizeram**" foi posto no infinitivo como "**fazer**". No próximo exemplo, "*Tomei a segunda dose de dipirona hoje*", o verbo "**Tomei**" também foi posto no infinitivo como "**Tomar**", e o substantivo feminino "**segunda**" foi flexionado em gênero para "**segundo**".

Os remédios me fizeram mal.	O remédio me fazer mal .
Tomei a segunda dose de dipirona hoje.	Tomar a segundo dose de dipirona hoje .

Figura 20 – Exemplo do processo de Lematização.

Para o REN, os dados precisaram ser anotados manualmente. Com isso, adotou-se o formato de dados *BO*, sendo *B* (Begin) equivalente a EN e *O* (Outside) não sendo EN, um modelo mais simples quando comparado ao BIO em inglês IOB. O *I* (inside) equivalente a EN composta. Quando o elemento for único adota-se o formato I-EN, em caso de elemento composto, a parte inicial utiliza o formato B-EN e nas partes posteriores I-EN.

A figura 21 apresenta o formato BIO, o medicamento Dipirona é representado em uma única parte *I-DRUG*, por ter uma palavra simples. Já o evento dor de cabeça é divido em três partes dor *B-EVENT*, de *I-EVENT* e cabeça *I-EVENT* [Ratinov and Roth, 2009]. Entretanto, essa anotação requer um maior tempo devido aos detalhes de palavras compostas, desse modo foi o adotado o formato *BO* conforme a figura 22 apresenta um exemplo utilizando a notação *DISEASE* (doença), *EVENT* (sintoma), *DRUG* (medicamento) e *O* (Outside), que corresponde a letra *B*.

Tomou	O
Dipirona	I-DRUG
e	EVENT
teve	O
dor	B-EVENT
de	I-EVENT
cabeça	I-EVENT

Figura 21 – Exemplo de arquivo no formato BIO

Ela	O
tem	O
febre	EVENT
e	O
sempre	O
toma	O
Dipirona	DRUG
mas	O
teve	O
pressão	DISEASE
alta	DISEASE

Figura 22 – Exemplo de arquivo no formato BO

A terceira etapa consistiu no treinamento dos modelos REN, PosTagger e DepParser. Tais modelos fazem uso da abordagem estatística, assim é necessário treinar os mesmos. Obteve-se os dados que passaram por pré-processamento, sendo configurados no formato adequado e anotados. Nessa etapa, o software coreNLP foi utilizado para treinar os modelos. Ao término da mesma, obteve-se um arquivo XML com todos os *tweets* processados e devidamente anotados.

O quarto procedimento foi a validação dos dados. Para os modelos estatísticos é utilizado o padrão *holdout* e validação cruzada de modo a medir sua aptidão em relação à tarefa. Ao término dessa fase, os modelos alcançaram desempenho suficiente para realizar os devidos processos.

Concluindo, a quinta etapa tratou da criação do algoritmo para identificar a relação semântica entre os elementos identificados no REN (medicamento e eventos). Assim, foi possível utilizar a árvore sintática, fornecida pelo DepParser, para identificar os possíveis eventos adversos. Para tal, o sistema importou o arquivo XML, criado na etapa anterior, contendo os dados existentes. Ao término do processo, o sistema exportou uma planilha com os tweets e as marcações de eventos adversos ou não. Para o filtro semântico, uma vez que opera de modo semântico, considerou-se um classificador binário.

4- Conjuntos de Dados

Neste capítulo, são definidos os conjuntos de dados criados nessa dissertação, com uma breve descrição e o processo de aquisição dos dados. Uma vez que manipulam os conjuntos de dados, modificando-os, as etapas relativas da metodologia apresentada no capítulo 3, são abordadas nesse capítulo.

Este capítulo está organizado como segue: a seção 4.1 apresenta como foram obtidos o conjunto de dado UD_Portuguese-GSD. A seção 4.2 apresenta como foram obtidos o conjunto de dado dic_lexico. A seção 4.3 apresenta como foram obtidos o conjunto de dado *scraping*. Por fim, a seção **??** inicia o processo de implementação da metodologia no que tange a alteração dos conjunto de dados.

4.1- Conjunto de dados UD_Portuguese-GSD

Esse conjunto de dados é a conversão do conjunto de dados do projeto Google Universal Dependency (em português Google parse de dependência multiligual) ou *uni-dep-tb*. O projeto *uni-dep-tb*, atualmente está descontinuado (legado). Devido ao estado do *uni-dep-tb* muito pouca informação sobre a primeira versão é encontrada nos dias de hoje. O projeto foi migrado para o formato CoNLL sendo nomeado para *UD_Portuguese-GSD*, no qual também foram incorporadas as anotações de dependência de Stanford. Seu uso é recorrente e sofre constantes atualizações, a versão mais recente denominada *UD_Portuguese-GSD* 2.3 é de Abril de 2018 [McDonald et al., 2013].

4.1.1 Versão 2.0

O conjuntos de dados *UD_Portuguese-GSD* na versão 2.0 contém as seguintes características:

- 12.078 sentenças;
- 297.478 tokens;
- 319.380 anotações sintáticas;
- 21.902 tokens de múltiplas palavras, existindo 38 tipos (pela, nos, as, nas, etc);
- 15 tags UPOS (ADJ, ADP, ADV, AUX, CCONJ, DET, NOUN, NUM, PART, PRON, PROPN, PUNCT, SYM, VERB, X) de 17 possíveis (não implementa SCONJ e INTJ);
- Não contém palavras com espaços em branco;
- Contém 36 verbos de ligação lematizados;
- Contém 36 verbos auxiliares lematizados;
- Contém 74 palavras do tipo partículas morfológicas;
- Contém 17 possibilidades morfológicas.

4.1.2 Versão 2.3

Já a versão *UD_Portuguese-GSD* 2.3, contém as seguintes atualizações e mudanças em relação a versão *UD_Portuguese-GSD* 2.0, são elas:

- Presença de lemas do projeto MorphoBr[1];
- Contém 62 verbos auxiliares;
- Contém 58 ora como verbos de ligação, ora como verbo auxiliar;
- Corrigido anotação *XPOS* (tag referente a fala específica do idioma);
- Corrigido anotação *UPOS* (tag referente a fala).

[1] MorphoBr - recursos para análise morfológica do português - `https://github.com/LFG-PTBR/MorphoBr`

4.2- Conjunto de dados dic_lexico

A elaboração do conjunto de dados foi obtido de bulas de medicamentos da internet. Foi conduzida uma busca por “bulas online” no motor de busca Google, com intuito de simular um usuário leigo. Como resultado foi tomada a ação padrão de acesso a primeira opção disponibilizada, sendo essa o site *bulário.com* o qual foi extraído os dados das bulas de medicamentos. Para incrementar a versão final foram obtidas bulas de medicamentos pela Autoridade Nacional de Medicamentos e Produtos de Saúde[2].

O conjunto de dados *dic_lexico*, é uma coletânea de eventos e nome de medicamentos. Apresentado em formato *txt*, denominados *EventosAdversos.txt*, *Remedios.txt*,*SubstânciasAtivas.txt*. Esses arquivos totalizam:

- 20.103 nomes de medicamentos comerciais;
- 1.922 nomes de fármacos;
- 325 nomes de laboratórios;
- 1.629 nomes de fobias;
- 11.560 nomes de eventos (doenças ou sintomas);
- 8.124 sobrenomes Utilizado para facilitar o modelo de CRF;
- 7.421 nomes próprios Utilizado para facilitar o modelo de CRF.

4.3- Conjunto de dados *scraping*

O conjunto de dados *scraping* faz uso da técnica de *web scraping*. Inicialmente foi preterido a interface para programação de aplicação[3], nativa do Twitter. Entretanto, foi inviável seu uso para os objetivos propostos, devido a uma série de limitações presentes API, entre elas:

[2] `https://www.atlasdasaude.pt/lista-de-medicamentos-infarmed`
[3] Do inglês Application Programming Interface

- Os dados são coletados em tempo real, sendo necessária uma autorização especial do Twitter para dados de períodos anteriores;
- O volume máximo de dados é limitado por dia;
- O campo de texto é limitado e uma quantidade grande de elementos (nomes de medicamentos, sintomas e doenças) não são aceitos.

Para contornar este problema, foi utilizada a técnica *web scraping* (maiores detalhes na subseção 4.3.1). O conjunto de dados *scraping* está em formato *csv*, denominado *tweet2016.csv* e a lista de sintomas (*SintomasUsados.txt*) e medicamentos (*SubstânciasUsados.txt*), utilizados em formato *txt*, podem ser encontrados em `https://github.com/AlexandreMC/SBC/tree/master/2019`. Estes arquivos têm as seguintes características:

- 6.662 tweets;
- Dados coletados no período de 01 de janeiro de 2016, até 31 de dezembro de 2016 (mais detalhes na subseção 4.3.1);
- Criado utilizando 40 sintomas (escolhidos aleatoriamente, mais detalhes na subseção 4.3.1);
- Criado utilizando 841 medicamentos (escolhidos aleatoriamente, mais detalhes na subseção 4.3.1);
- 6.587 sentenças;
- 119.610 tokens.

4.3.1 *Web Scraping*

Para a data, inicial e final, foi escolhido o ano de 2016. A pesquisa começou em 2018 e o ano ainda estava em curso, com isso o período mais propício seria o ano de 2017, todavia, devido ao processo turbulento da eleição para presidente, o Twitter, em 2017, estava com foco nas eleições presidenciais, sendo assim foi preferido o ano anterior 2016.

Utilizou-se o ano inteiro, de $01/01/2016$ até $31/12/2016$, para abranger os eventos sazonais (datas festivas, feriados, surtos, epidemias, etc) e, assim maximizar a variação dos dados. O formato utiliza o carácter hífen - ao invés da barra / como separador da data e adota o padrão americano de notação, ou seja, 2016-01-01 e 2016-12-31.

No entanto, o campo de busca apresentou limitações. A combinação de eventos e medicamentos resultou em milhares de queries, então foi selecionado um grupo de eventos de forma aleatória para todos os medicamentos presentes (não sendo incluído na busca os nomes comerciais dos medicamentos, os mesmos foram utilizados apenas no conjunto de dados *dic_lexico*). O modelo de query ficou da seguinte forma: **medicamento AND (evento OR evento OR evento OR . . .)**. O medicamento foi alterado em cada execução, resultado em 1.682 execuções, contendo 40 sintomas e 841 medicamentos. Com isso foram criadas duas queries para todos os medicamentos utilizados:

[MEDICAMENTO] AND (ACNE OR AGITAÇÃO OR AGRESSIVIDADE OR ALERGIA OR ALUCINAÇÃO OR ANSIEDADE OR APATIA OR APNÉIA OR ARREPIOS OR AZIA OR CAIMBRA OR CALAFRIOS OR CALOR OR CEFALEIA OR COCEIRA OR CÓLICA OR COMA OR CONFUSÃO OR DEPENDÊNCIA OR DEPRESSÃO OR DESCONFORTO) since:2016-01-01 until:2016-12-31.

[MEDICAMENTO] AND (DIARRÉIA OR DISFUNÇÃO OR DOENÇA OR DOR OR EDEMA OR ENJÔO OR ENXAQUECA OR FADIGA OR FEBRE OR FRAQUEZA OR HEMORRAGIA OR INCHAÇO OR MORTE OR NAUSEA OR NÁUSEA OR RAIVA OR SUOR OR TONTURA OR VÔMITO) since:2016-01-01 until:2016-12-31.

O Twitter usa o método GET [4], ou seja, é necessário formatar a consulta anterior para o formato adequado. Assim, as duas queries anteriores foram modificadas para esse formato, substituindo o espaço em branco e acentos por outros símbolos (o espaço em branco é substituído por *%20*, o caractere *Ã* é divido em dois elementos, a letra *A* representado por *%C3* e o til representado por *%830* logo o caractere *Ã* é representando pela união dos dois elementos *%C3%83O*.

As querys criadas são utilizadas no *web scraping*. A figura 23 apresenta o processo de execução do *web scraping*, o software *RStudio* em segundo plano, com o código em execução do *web scraping* e o navegador Mozila Firefox, em primeiro plano, sendo gerido pelo *web scraping*.

O ícone de robô ao lado da barra de endereço do navegador indica que o mesmo

[4]Método GET é utilizado quando se quer passar poucas informações para realizar uma pesquisa ou passar informações para outra página através da própria URL.

opera de forma automática. Um fato importante é que não estamos autenticados no Twitter, ou seja, não precisamos estar cadastrados na plataforma para acessar os dados, visto que os mesmos estão dispostos ao público de forma direta, gratuita e sem restrição.

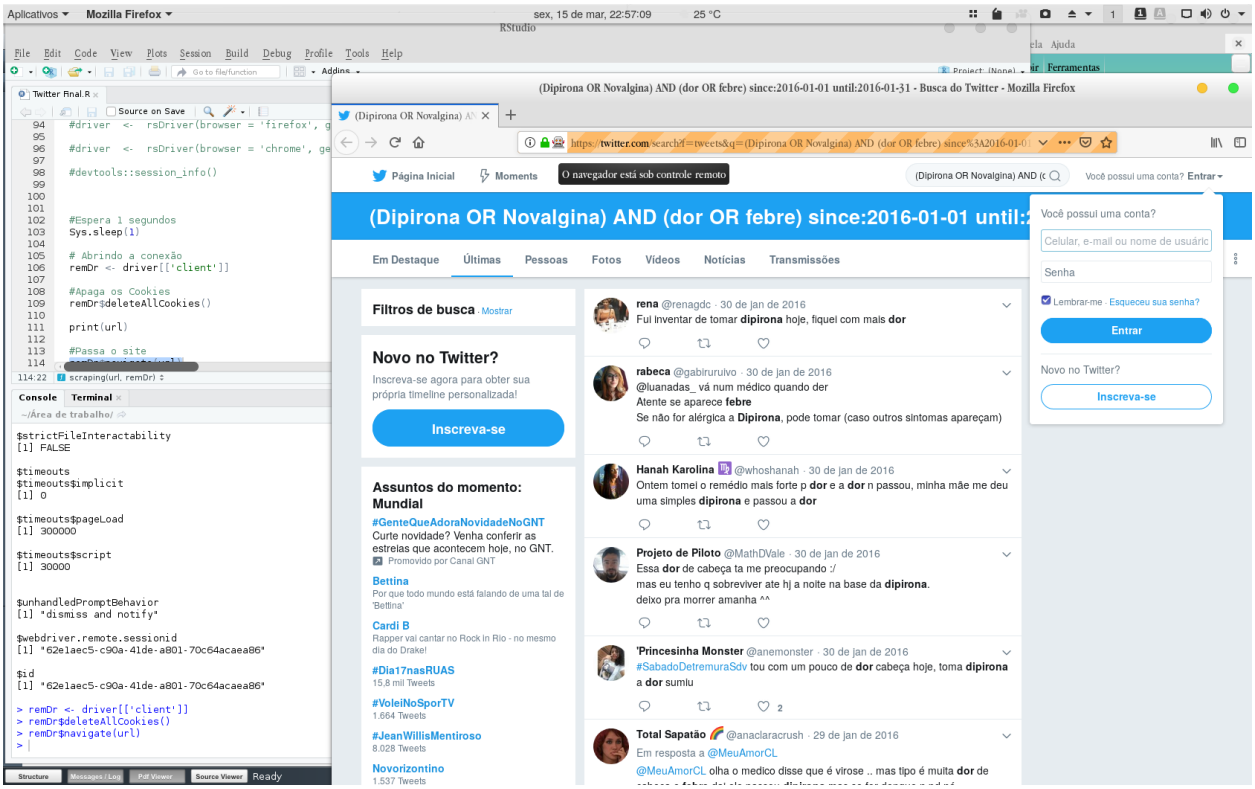

Figura 23 – Processo de scrap, em execução, coletando dados do Twitter.

4.4- Pré-Processamento

Terminada a análise exploratória, foi efetuado o pré-processamento de cada conjunto de dados, individualmente. Os conjuntos de dados denominados *UD_Portuguese-GSD* 2.0 e *UD_Portuguese-GSD* 2.3 foram obtidos pelo projeto Universal Dependence [Bosco et al., 2013], [De Marneffe et al., 2006], [McDonald et al., 2013], [Haverinen et al., 2014]. Esse projeto contém um *treebank*, em português brasileiro, com tokens, segmentação de palavras e anotações morfológicas, tendo a subseção 1.4.1 mais informações sobre o projeto. Os conjuntos de dados *UD_Portuguese-GSD*, nas versões 2.0 e 2.3, já se encontravam no formato CoNLL-U, tokenizado e lematizado, sendo necessário apenas remover as linhas comentadas.

O conjunto de dados *dic_lexico* utilizado para os experimentos é denominado *dic_lexico*, sendo esse conjunto de dados extraído de sites e bulas de medicamentos e publicado no artigo [da Cunha et al., 2018]; novos dados de outras fontes de dados foram incrementados ao projeto, originando uma atualização ao longo da dissertação. O mesmo foi convertido para o formato de duas colunas, separado por tabulação. Na primeira coluna é informado o tipo do dado (medicamento ou evento) e na segunda, o dado em si (nome do medicamento, ou nome do fármaco, ou nome do sintoma, ou nome da doença). Depois, foram removidos os valores duplicados e, por fim, efetuada a anotação dos dados.

O último, conjunto de dados denominado *scraping*, foi extraído da rede social Twitter (mais detalhes sobre a rede estão disponíveis na subseção 1.4.2). Esse conjunto de dados e tido como de maior relevância, por fazer parte do REN. Os dados foram obtidos de forma automática com uso da técnica *web scraping*. Foram mantidos os possíveis erros de digitação e abreviações de modo a não descaracterizá-los. No entanto, foram removidos todos os caracteres não alfanuméricos, caracteres acentuados do idioma e espaços em branco (exclui pontuação, emoticon, etc). E, assim, efetuou-se a anotação dos dados. Em seguida, foram removidos os espaços extras criados com a remoção dos caracteres (espaços no meio, no início e no fim) e o caractere ponto *(.)* foi inserido ao final de cada linha. Por fim, o conjunto de dados *scraping* passou pelos processos de tokenização e divisão de sentenças através do software coreNLP.

5- Avaliação Experimental

Neste capítulo, é dado continuidade a metodologia implementada no capítulo 4, evidenciando os experimentos realizados para apresentar a proposta de modo prático. Os experimentos consistem no treinamento e validação dos modelos, sendo o maior foco dessa dissertação, e na implementação de regra no filtro semântico, apresentando sua viabilidade. Os experimentos se utilizam dos conjuntos de dados obtidos do Twitter, dos dicionários toponímicos e dos conjuntos do projeto Universal Dependence, nas versões 2.0 e 2.3.

Este capítulo está organizado como segue: A seção 5.1 aborda o processo de treinamento dos modelos estatísticos. A seção 5.2 trata da validação dos modelos. Por fim, a seção **??** discute os resultados obtidos.

5.1- Treinamento

Com os dados pré-processados na seção 4.4, os conjuntos de dados estão aptos a serem utilizados como material para produzir os modelos de PosTagger, DepParser e REN. O equipamento utilizado para o processo de treinamento foi um PC do tipo Desktop, equipado com CPU Intel Core I5 7500 com 16 GB de RAM DDR4 2400 MHz e HD 3 TB 7200 RPM. O sistema operacional utilizado foi GNU/Linux Debian 9.9 (Stretch). O PosTagger foi processado em cerca de 4 horas para cada modelo, no total de 22 modelos. O DepParser levou cerca de 22 horas para cada modelo sem WE, no total de 12 modelos e aproximadamente 49 horas para cada modelo com WE, no total de 24. O REN consumiu 2 horas para cada modelo, no total de 11 modelos.

5.2- Validação

Uma vez que os modelos estatísticos, PosTagger, DepParser e REN, estavam criados, se fez necessário validá-los. Nesta dissertação, utilizou-se o processo de holdout e validação cruzada, descrito no capítulo 1, seção 1.3.5.

5.2.1 Pos-Tagger

Para o PosTagger, as métricas de avaliação são *sentenças*, *tags (ato de "etiquetar" os tokens)* e *palavras desconhecidas*. As *sentenças corretas* representam as sentenças por inteiro no qual, uma única tag errada, invalida toda a sentença. Já as *tags* representam a abrangência geral do modelo. As *palavras desconhecidas*, por sua vez, representam as palavras que não se encontravam no conjunto de dados. Tais dados foram classificados em *acertos* e *erros*, em duas bases de dados distintas.

A tabela 6 apresenta os resultados obtidos no treinamento do modelo do PosTagger, para o qual se utilizou o processo de *holdout* no conjunto de dados *UD_Portuguese-GSD* 2.0. O conjunto de dados foi executado com 1.198 sentenças contendo 26.460 palavras, sendo 2.273 desconhecidas.

Tabela 6 – Resultado do processo de holdout no Pos-Tagger: conjunto de dados *UD_Portuguese-GSD* 2.0

Métricas	Acertos	Erros
Sentenças	$679(56,68\%)$	$519(43,32\%)$
Tags	$28.552(96,92\%)$	$908(3,08\%)$
Palavras desconhecidas	$2.051(90,23\%)$	$222(9,77\%)$

O processo de validação cruzada do conjunto de dados *UD_Portuguese-GSD* 2.0 é apresentada na tabela 8 no treinamento do modelo do PosTagger. Foi divido em 10 subconjuntos, conforme tabela 7:

Tabela 7 – Subconjuntos criados no processo de validação cruzada no Pos-Tagger: conjunto de dados *UD_Portuguese-GSD* 2.0

Subconjuntos	Sentenças	Palavras	Palavras Desconhecidas
1	1.088	26.879	2.070
2	1.087	26.853	1.991
3	1.115	26.834	2.014
4	1.082	26.914	1.981
5	1.078	26.927	1.998
6	1.078	26.860	1.935
7	1.055	26.847	1.955
8	1.090	26.864	2.011
9	1.076	26.878	1.965
10	1.049	26.811	2.107

Tabela 8 – Resultado do processo de validação cruzada no Pos-Tagger: conjunto de dados *UD_Portuguese-GSD* 2.0

Fold	Métricas	Acertos	Erros
1	Sentenças	$613(56,34\%)$	$475(43,65\%)$
	Tags	$26.053(96,93\%)$	$826(3,07\%)$
	Palavras desconhecidas	$1.869(90,29\%)$	$201(9,71\%)$
2	Sentenças	$642(59,06\%)$	$445(40,94\%)$
	Tags	$26.102(97,20\%)$	$751(2,79\%)$
	Palavras desconhecidas	$1.837(92,26\%)$	$154(7,73\%)$
3	Sentenças	$651(58,39\%)$	$464(41,61\%)$
	Tags	$26.006(96,91\%)$	$828(3,09\%)$
	Palavras desconhecidas	$1.853(92,01\%)$	$161(7,99\%)$
4	Sentenças	$612(56,56\%)$	$470(43,44\%)$
	Tags	$26.083(96,99\%)$	$831(3,09\%)$
	Palavras desconhecidas	$1.804(91,06\%)$	$177(8,93\%)$

	Sentenças	$606(56,21\%)$	$472(43,78\%)$
5	Tags	$26.078(96,85\%)$	$849(3,15\%)$
	Palavras desconhecidas	$1.809(90,54\%)$	$189(9,46\%)$
	Sentenças	$625(57,98\%)$	$453(42,02\%)$
6	Tags	$26.075(97,08\%)$	$785(2,92\%)$
	Palavras desconhecidas	$1.758(90,85\%)$	$177(9,15\%)$
	Sentenças	$606(57,44\%)$	$449(42,56\%)$
7	Tags	$26.055(97,05\%)$	$792(2,95\%)$
	Palavras desconhecidas	$1.771(90,59\%)$	$184(9,41\%)$
	Sentenças	$620(56,88\%)$	$470(43,12\%)$
8	Tags	$26.015(96,84\%)$	$849(3,16\%)$
	Palavras desconhecidas	$1.839(91,45\%)$	$172(8,55\%)$
	Sentenças	$594(55,20\%)$	$482(44,79\%)$
9	Tags	$26.050(96,92\%)$	$828(3,08\%)$
	Palavras desconhecidas	$1.768(89,97\%)$	$197(10,02\%)$
	Sentenças	$608(57,96\%)$	$441(42,04\%)$
10	Tags	$25.999(96,97\%)$	$812(3,03\%)$
	Palavras desconhecidas	$1.909(90,60\%)$	$98(9,40\%)$

A tabela 9 apresenta os resultados obtidos no treinamento do modelo do PosTagger do conjunto de dados *UD_Portuguese-GSD* 2.0. A média é representada por $\bar{X}$, desvio padrão representado por σ e erro padrão da média σ_x.

Tabela 9 – Síntese da validação cruzada no Pos-Tagger: conjunto de dados *UD_Portuguese-GSD* 2.0

Medidas de Dispersão	$\bar{X}$		σ		σ_x	
Métricas:	Acertos	Erros	Acertos	Erros	Acertos	Erros
Sentenças	$617,7$	$461,8$	$16,56$	$13,33$	$5,24$	$4,21$
Tags	$26.051,6$	$815,1$	$33,12$	$29,26$	$10,47$	$9,25$
Palavras desconhecidas	$1821,7$	171	$46,21$	$28,04$	$14,61$	$8,87$

A tabela 10 apresenta os resultados obtidos no treinamento do modelo do PosTagger, no qual utilizou o processo de holdout no conjunto de dados *UD_Portuguese-GSD* 2.3. O conjunto de dados foi executado com 1.204 sentenças, contendo 33.638 palavras, sendo 2.273 desconhecidas.

Tabela 10 – Resultado do processo de holdout no Pos-Tagger: conjunto de dados *UD_Portuguese-GSD* 2.3

Métricas	Acertos	Erros
Sentenças	$658(54,65\%)$	$546(45,35\%)$
Tags	$32.680(97,15\%)$	$958(2,85\%)$
Palavras desconhecidas	$2.038(89,66\%)$	$235(10,34\%)$

O processo de validação cruzada do conjunto de dados *UD_Portuguese-GSD* 2.3 é apresentado na tabela 12, no treinamento do modelo do PosTagger. Foi divido em 10 subconjuntos, conforme tabela 11:

Tabela 11 – Subconjuntos criados no processo de validação cruzada no Pos-Tagger: conjunto de dados *UD_Portuguese-GSD* 2.3

Subconjuntos	Sentenças	Palavras	Palavras Desconhecidas
1	1.094	30.693	2.070
2	1.120	30.626	2.026
3	1.115	26.834	2.014
4	1.094	30.790	1.983
5	1.084	30.653	1.996
6	1.053	30.620	1.938
7	1.055	26.847	1.955

8	1.091	30.601	2.015
9	1.092	30.741	1.972
10	1.066	30.724	2.119

Tabela 12 – Resultado do processo de validação cruzada no Pos-Tagger: conjunto de dados *UD_Portuguese-GSD* 2.3

Fold	Métricas	Acertos	Erros
	Sentenças	603(55, 19%)	491(44, 88%)
1	Tags	29.829(97, 18%)	864(2, 81%)
	Palavras desconhecidas	1.857(89, 71%)	213(10, 29%)
	Sentenças	636(58, 62%)	449(41, 38%)
2	Tags	29.901(97, 52%)	760(2, 48%)
	Palavras desconhecidas	1.834(92, 58%)	147(7, 42%)
	Sentenças	651(58, 39%)	464(41, 61%)
3	Tags	29.006(96, 91%)	828(3, 09%)
	Palavras desconhecidas	1.853(92, 01%)	160(7, 90%)
	Sentenças	602(55, 027%)	492(44, 97%)
4	Tags	29.944(97, 25%)	846(2, 75%)
	Palavras desconhecidas	1.794(90, 47%)	189(9, 53%)
	Sentenças	610(56, 27%)	474(43, 73%)
5	Tags	29.814(97, 26%)	839(2, 74%)
	Palavras desconhecidas	1.807(90, 53%)	189(9, 47%)
	Sentenças	619(56, 89%)	469(43, 11%)
6	Tags	29.860(97, 29%)	833(2, 71%)
	Palavras desconhecidas	1.748(90, 34%)	187(9, 66%)
	Sentenças	589(55, 93%)	464(44, 06%)
7	Tags	29.801(97, 32%)	819(2, 67%)

	Palavras desconhecidas	$1.753(90,45\%)$	$185(9,54\%)$
	Sentenças	$605(55,45\%)$	$486(44,55\%)$
8	Tags	$29.717(97,11\%)$	$884(2,89\%)$
	Palavras desconhecidas	$1.840(91,31\%)$	$175(8,68\%)$
	Sentenças	$602(55,13\%)$	$490(44,87\%)$
9	Tags	$29.907(97,29\%)$	$834(2,71\%)$
	Palavras desconhecidas	$1.789(90,72\%)$	$183(9,28\%)$
	Sentenças	$606(56,85\%)$	$460(43,15\%)$
10	Tags	$29.902(97,32\%)$	$822(2,67\%)$
	Palavras desconhecidas	$1.925(90,84\%)$	$194(9,15\%)$

A tabela 13 apresenta os resultados obtidos no treinamento do modelo do PosTagger no conjunto de dados *UD_Portuguese-GSD* 2.3. A média é representada por $\bar{X}$, desvio padrão representado por σ e erro padrão da média σ_x.

Tabela 13 – Síntese da validação cruzada no Pos-Tagger: conjunto de dados *UD_Portuguese-GSD* 2.3

Medidas de Dispersão	$\bar{X}$		σ		σ_x	
Métricas:	Acertos	Erros	Acertos	Erros	Acertos	Erros
hline Sentenças	$612,3$	$473,9$	$17,45$	$14,35$	$5,52$	$4,54$
Tags corretas	$29.768,1$	$712,9$	$261,64$	$187,51$	$82,74$	$59,3$
Palavras desconhecidas	$1819,7$	$182,2$	$50,64$	$17,31$	$16,01$	$5,47$

No conjunto de dados *UD_Portuguese-GSD* 2.0 o modelo teve 96,9178% de tags em língua portuguesa, o que representa um resultado bastante elevado, próximo ao 100%. Já a versão *UD_Portuguese-GSD* 2.3, o modelo teve 97,1520% de tags em língua portuguesa, o que representa também um resultado bastante elevado, próximo ao 100%. O conjunto de dados *UD_Portuguese-GSD* 2.3 teve o resultado final melhor, o que era previsto por ser uma versão aprimorada da versão anterior *UD_Portuguese-GSD* 2.0.

Pode-se concluir que, em ambos os conjuntos de dados, a acurácia teve um desempenho satisfatório. Entretanto, mesmo com o processo de validação cruzada, não

se pode garantir que a acurácia apresentada se replique em textos diversos. Os conjuntos de dados têm 297.478 tokens, desse modo, não estabelece um conteúdo representativo da língua portuguesa, com seus vários gêneros textuais.

5.2.2 Parser de Dependência

Para o DepParser, as métricas de avaliação são UAS e LAS. UAS não considera relação semântica para criação da relação de dependência entre os elementos, ou seja, considera apenas a raiz identificada corretamente. O LAS exige o rótulo semântico para criação da relação de dependência entre os elementos, em outras palavras, a relação semântica nesse caso é a análise sintática (exemplo: sujeito, predicado, etc) que, dada a raiz corretamente identificada, tiveram a relação rotulada adequadamente. Já o uso de WE, fornece ao DepParser conhecimento inicial, de modo que o treinamento do modelo não comece sem conhecimento algum. WE de 300 dimensões do tipo CBOW e *Skip-gram* foram aplicados para comparação em ambos os modelos de dados. Para mais detalhes, no capítulo 1, seção 1.3.5.

A tabela 14 apresenta os resultados obtidos no treinamento do modelo do DepParser, utilizando o processo de holdout no conjunto de dados *UD_Portuguese-GSD* 2.0. O conjunto de dados foi executado com três configurações: sem usar WE, utilizando WE do tipo CBOW com arquivo de 300 dimensões, e utilizando WE do tipo *Skip-Gram* com arquivo de 300 dimensões. O conjunto de dados foi executado com 1.198 sentenças contendo 29.460 palavras, sendo 2.273 $(7,72\%)$ desconhecidas.

Tabela 14 – Resultado do processo de holdout no DepParser: conjunto de dados *UD_Portuguese-GSD* 2.0

Métricas	Sem *Word Embedding*	Com *Word Embedding* *CBOW* 300 Dimensões	Com *Word Embedding* *Skip-Gram* 300 Dimensões
UAS	$84,36\%$	$86,60\%$	$86,38\%$
LAS	$82,20\%$	$84,74\%$	$84,57\%$

O processo de validação cruzada no conjunto de dados *UD_Portuguese-GSD* 2.0 é apresentada na tabela 8, no treinamento do modelo do DepParser. Os 5 subconjuntos

foram conduzidos da seguinte forma, conforme tabela 15:

Tabela 15 – Subconjuntos criados no processo de validação cruzada no DepParser: conjunto de dados *UD_Portuguese-GSD* 2.0

Subconjuntos	Sentenças	Palavras	Palavras Desconhecidas
1	2.173	53.698	4.300 $(8,01\%)$
2	2.195	53.695	4.228 $(7,87\%)$
3	2.154	53.718	4.173 $(7,77\%)$
4	2.143	53.692	4.212 $(7,84\%)$
5	2.123	53.650	4.258 $(7,94\%)$

:

Tabela 16 – Resultado do processo de validação cruzada no DepParser: conjunto de dados *UD_Portuguese-GSD* versão 2.0

Fold	Métricas	Sem Word Embedding	Com *Word Embedding* *CBOW* 300 Dimensões	Com *Word Embedding* *Skip-Gram* 300 Dimensões
1	UAS	$82,80\%$	$86,22\%$	$86,74\%$
	LAS	$80,74\%$	$84,25\%$	$84,84\%$
2	UAS	$82,76\%$	$86,61\%$	$87,00\%$
	LAS	$80,78\%$	$84,66\%$	$84,66\%$
3	UAS	$82,86\%$	$86,16\%$	$86,61\%$
	LAS	$80,91\%$	$84,36\%$	$84,73\%$
4	UAS	$82,53\%$	$86,37\%$	$86,33\%$
	LAS	$80,60\%$	$84,46\%$	$84,46\%$
5	UAS	$82,62\%$	$86,43\%$	$86,56\%$
	LAS	$80,62\%$	$84,42\%$	$84,87\%$

A tabela 17 apresenta os resultados obtidos no treinamento do modelo do DepParser no conjunto de dados *UD_Portuguese-GSD* 2.0. O qual, a média é representada por $\bar{X}$,

desvio padrão representado por σ e erro padrão da média σ_x.

Tabela 17 – Síntese da validação cruzada no DepParser: conjunto de dados *UD_Portuguese-GSD* 2.0.

Métricas	Medidas de Dispersão	Sem *Word Embedding*	Com *Word Embedding* *CBOW* 300 Dimensões	Com *Word Embedding* *Skip-Gram* 300 Dimensões
UAS	$\bar{X}$	$82,71\%$	$86,36\%$	$86,65\%$
	σ	$0,12\%$	$0,16\%$	$0,22\%$
	σ_x	$0,05\%$	$0,07\%$	$0,10\%$
LAS	$\bar{X}$	$80,73\%$	$84,43\%$	$84,71\%$
	σ	$0,11\%$	$0,13\%$	$0,15\%$
	σ_x	$0,05\%$	$0,06\%$	$0,07\%$

A tabela 18 apresenta os resultados obtidos no treinamento do modelo do DepParser, que fez uso do processo de holdout no conjunto de dados *UD_Portuguese-GSD* 2.3. O conjunto de dados foi executado com 1.204 sentenças com 31.496 palavras, sendo $2.276(7,23\%)$ desconhecidas.

Tabela 18 – Resultado do processo de holdout no DepParser: conjunto de dados *UD_Portuguese-GSD* 2.3

Métricas	Sem *Word Embedding*	Com *Word Embedding* *CBOW* 300 Dimensões	Com *Word Embedding* *Skip-Gram* 300 Dimensões
UAS	$87,37\%$	$88,91\%$	$89,03\%$
LAS	$84,94\%$	$86,65\%$	$86,73\%$

O processo de validação cruzada no conjunto de dados *UD_Portuguese-GSD* 2.3 é apresentada na tabela 20 no treinamento do modelo do DepParser. Foi divido em 5 subconjuntos: com 3 configurações: sem usar WE, utilizando WE do tipo CBOW, com arquivo de 300 dimensões, e utilizando WE do tipo *Skip-Gram*, com arquivo de 300 dimensões. Os 5 subconjuntos foram executados da seguinte maneira, conforme tabela 19:

Tabela 19 – Subconjuntos criados no processo de validação cruzada no DepParser: conjunto de dados *UD_Portuguese-GSD* 2.3

Subconjuntos	Sentenças	Palavras	Palavras Desconhecidas
1	2.177	57.372	4.294
2	2.212	57.465	4.241
3	2.172	57.437	4.178
4	2.142	57.239	4.196
5	2.158	57.629	4.286

:

Tabela 20 – Resultado do processo de validação cruzada no DepParser versão 2.3

Fold	Métricas	Sem *Word Embedding*	Com *Word Embedding* *CBOW* 300 Dimensões	Com *Word Embedding* *Skip-Gram* 300 Dimensões
1	UAS	$86,07\%$	$88,74\%$	$88,81\%$
	LAS	$83,80\%$	$86,38\%$	$86,53\%$
2	UAS	$86,37\%$	$89,36\%$	$89,42\%$
	LAS	$84,08\%$	$86,97\%$	$87,05\%$
3	UAS	$86,14\%$	$89,17\%$	$88,94\%$
	LAS	$83,82\%$	$86,77\%$	$86,61\%$
4	UAS	$85,83\%$	$89,01\%$	$89,30\%$
	LAS	$83,51\%$	$86,60\%$	$86,86\%$
5	UAS	$86,29\%$	$88,86\%$	$89,05\%$
	LAS	$84,11\%$	$86,64\%$	$86,79\%$

A tabela 21 apresenta os resultados obtidos no treinamento do modelo do DepParser no conjunto de dados *UD_Portuguese-GSD* 2.3. A média é representada por $\bar{X}$, desvio padrão representado por σ e erro padrão da média σ_x.

Tabela 21 – Síntese da validação cruzada no DepParser: conjunto de dados *UD_Portuguese-GSD* 2.3

Métricas	Medidas de Dispersão	Sem *Word Embedding*	Com *Word Embedding* *CBOW* 300 Dimensões	Com *Word Embedding* *Skip-Gram* 300 Dimensões
UAS	$\bar{X}$	$86,14\%$	$89,03\%$	$89,10\%$
	σ	$0,19\%$	$0,22\%$	$0,22\%$
	σ_x	$0,08\%$	$0,10\%$	$0,10\%$
LAS	$\bar{X}$	$83,85\%$	$86,67\%$	$86,77\%$
	σ	$0,20\%$	$0,20\%$	$0,18\%$
	σ_x	$0,09\%$	$0,08\%$	$0,08\%$

O conjunto de dados *UD_Portuguese-GSD* 2.0, apresenta resultado equilibrados entre UAS e LAS. Fica evidente que o uso do WE melhorou o modelo, com uma ligeira vantagem para o modelo WE do tipo CBOW. A comparação desses subconjuntos foi sintetizada após o término da validação cruzada. Vê-se uma variação ínfima entre eles, com vantagem sutil para o modelo WE do tipo *Skip-gram*.

Já o conjunto de dados *UD_Portuguese-GSD* 2.3, apresenta resultado equilibrados entre UAS e LAS. Fica evidente que o uso do WE melhorou o modelo, com uma ligeira vantagem para o modelo WE do tipo *Skip-gram*. A comparação desses subconjuntos foi sintetizada após o término da validação cruzada. Pode ser percebida uma variação ínfima entre eles, com vantagem sutil para o modelo WE do tipo *Skip-gram*.

O conjunto de dados *UD_Portuguese-GSD* 2.0 teve como melhor resultado UAS de $86,60\%$ e LAS de $84,74\%$. Já a versão *UD_Portuguese-GSD* 2.3 teve como melhor resultado UAS de $89,03\%$ e LAS de $86,73\%$. O conjunto de dados *UD_Portuguese-GSD* 2.3 teve o resultado final melhor, o que era previsto por ser um versão aprimorada da versão anterior *UD_Portuguese-GSD* 2.0.

5.2.3 Reconhecimento de Entidade Nomeada

Para o REN, as métricas de avaliação são *F1*, *abrangência* e *precisão*. Foram aplicadas tais métricas em duas entidades, *medicamentos* e *eventos*, e em duas bases de dados. A tabela 22 apresenta os resultados obtidos no treinamento do modelo do REN, no qual utilizou o processo de holdout com os conjuntos de dados *scraping* e *dic_lexico*. O conjunto de dados foi executado com 85.479 palavras e 4.634 sentenças.

Tabela 22 – Resultado do processo de holdout no REN: conjuntos de dados *scraping* e *dic_lexico*.

Entidades \Métricas	Precisão	Abrangência	F1
Medicamentos	$0,99$	$0,99$	$0,99$
Eventos	$0,98$	$0,99$	$0,98$

A tabela 24 apresenta os resultados obtidos no treinamento do modelo do REN, o qual se fez uso do processo de validação cruzada com os conjuntos de dados *scraping* e *dic_lexico*. Foi dividido em 10 subconjuntos, conforme tabela 23:

Tabela 23 – Subconjuntos criados no processo de validação cruzada no REN: conjuntos de dados *scraping* e *dic_lexico*

Subconjuntos	Palavras	Sentenças
1	12.824	694
2	12.807	710
3	12.812	706
4	12.853	664
5	12.815	702
6	12.817	701
7	12.794	724
8	12.818	700
9	12.762	756
10	12.913	603

Tabela 24 – Resultado do processo de validação cruzada no REN: conjuntos de dados *scraping* e *dic_lexico*.

Fold	Entidades \Métricas	Precisão	Abrangência	F1
1	Medicamento	$0,99$	$0,99$	$0,99$
	Evento	$0,98$	$0,98$	$0,98$
2	Medicamento	$0,99$	$0,99$	$0,99$
	Evento	$0,98$	$0,99$	$0,98$
3	Medicamento	$0,99$	$0,99$	$0,99$
	Evento	$0,99$	$0,98$	$0,99$
4	Medicamento	$0,99$	$0,99$	$0,99$
	Evento	$0,98$	$0,98$	$0,97$
5	Medicamento	$0,98$	$0,99$	$0,98$
	Evento	$0,99$	$0,99$	$0,99$
6	Medicamento	$0,99$	$0,99$	$0,99$
	Evento	$0,98$	$0,98$	$0,98$
7	Medicamento	$0,99$	$0,99$	$0,99$
	Evento	$0,99$	$0,99$	$0,99$
8	Medicamento	$0,99$	$0,99$	$0,99$
	Evento	$0,99$	$0,99$	$0,98$
9	Medicamento	$0,99$	$0,99$	$0,99$
	Evento	$0,98$	$0,99$	$0,98$
10	Medicamento	$0,97$	$0,98$	$0,98$
	Evento	$0,93$	$0,98$	$0,96$

A tabela 25 apresenta os resultados obtidos no treinamento do modelo do REN. O qual a média é representada por $\bar{X}$, desvio padrão representado por σ e erro padrão da média σ_x.

Tabela 25 – Síntese da validação cruzada no REN: conjuntos de dados *scraping* e *dic_lexico*.

Entidades	Medidas de Dispersão	Precisão	Abrangência	F1
Medicamento	$\bar{X}$	$0,99$	$0,99$	$0,99$
	σ	$0,007$	$0,004$	$0,005$
	σ_x	$0,00$	$0,00$	$0,00$
Evento	$\bar{X}$	$0,98$	$0,98$	$0,98$
	σ	$0,02$	$0,004$	$0,008$
	σ_x	$0,01$	$0,00$	$0,00$

As bases de dados *scraping* e *dic_lexico* foram combinadas para criar o modelo desejado. A tabela 22 apresenta os resultados do método *holdout*. A detecção de entidades *medicamentos* e *eventos* tem todas as métricas acima de 99% para ambas entidades. O resultado já era esperando, tendo em vista que a base de dados *dic_lexico* contém uma lista com a relação dos medicamentos e eventos.

Conclui-se que o modelo criado obteve um resultado muito próximo de 100%. Entretanto, mesmo com processo de validação cruzada, o modelo pode não ser adequado em determinado momento, dada a evolução natural da língua que ocorre com os dados informais.

5.2.4 Filtro Semântico

O filtro semântico consiste no algoritmo SaúdeAlg, este por usa vez contém três regras e a combinação delas. As métricas de avaliação empregadas foram *F1*, *abrangência* e *precisão*, com uso da base de dados *scraping* e os modelos PosTagger, DepParser e REN. Logo que os dados foram anotados pelo software coreNLP, com os modelos criados as regras semânticas foram aplicadas.

Nessa etapa é efetuada a análise do filtro semântico. O conjunto de dados *scraping* foi utilizado e realizado a validação egra a regra e, posteriormente, com as três regras em simultâneo. A tabela 26 apresenta os resultados obtidos na implementação do filtro

semântico.

Tabela 26 – Resultado do Filtro Semântico

Regras	Precisão	Abrangência	F1
Regra1	$0,21$	$0,32$	$0,25$
Regra2	$0,24$	$0,52$	$0,32$
Regra3	$0,012$	$0,80$	$0,024$
Todas	$0,46$	$0,41$	$0,43$

O conjunto de dados *UD_Portuguese-GSD* 2.3 gerou um novo conjunto de relações semânticas, incompatíveis com a versão do conjunto de dados anterior o conjunto de dados *UD_Portuguese-GSD* 2.0, que foi utilizado no início desta dissertação. Ao explorar essa nova versão não foi encontrado regras dessas novas relações. Não é dito que não é possível extrair regras dessa nova versão e sim que o autor não é especialista no domínio, e dada sua limitação não conseguiu deduzir tais regras. Dessa maneira não utiliza o conjunto de dados *UD_Portuguese-GSD* 2.3 nesta etapa.

A primeira regra é genérica e tem por intenção identificar a maior quantidade possível de efeitos adversos. Já a segunda regra é um exemplo de uso em apenas um evento: *alergia*. Por fim, a terceira regra é um complemento da segunda, que objetiva apresentar adaptabilidade à regra.

O cálculo das regras foi baseado no total de eventos, mesmo quando a regra era destina a um evento específico (regra 2 e regra 3), o que impactou negativamente o resultado. No entanto, o modelo é evolutivo e adaptativo e uma nova regra pode ser inserida a qualquer momento (seja para um grupo de eventos, para todos ou para um específico), sendo esse seu maior potencial. Com as simples regras atuais, realizadas sem auxílio de um especialista no domínio, foi possível verificar um resultado próximo da metade dos eventos existentes. De maneira quantitativa, dos 6.662 *tweets* no ano de 2016 foram encontrados 481 evento adversos resultando em 13 sinais.

Considerações Finais

Essa dissertação abordou o desafio de detectar sinais de eventos adversos a medicamentos em textos informais, que são um desafio para qualquer técnica e estratégia existente nos dias de hoje, em qualquer idioma. Para alcançar este objetivo, foi proposto um modelo de EI. O estudo proposto teve caráter exploratório, possuindo uma abordagem inédita em seu algoritmo, quando empregado na língua portuguesa brasileira. Assim sendo, a proposta não era alcançar o estado da arte existente em outros idiomas, principalmente em língua inglesa, que detêm todo um arcabouço de recursos, mas promover o fomento do tema e propor um modelo viável para execução do objetivo apresentado, que pode ser melhorado de forma incremental.

Mesmo diante das limitações do idioma escolhido, a abordagem adotada atendeu aos desígnios do método científico e contribuiu de forma relevante para a pesquisa e para a sociedade, podendo torná-la mais segura. Todavia, o modelo proposto fornece elementos suficientes, com um resultado parcial, dentro do prazo e escopo estipulados.

Tal modelo atua em toda a cadeia de processo (EI, MT e PLN), que por sua vez, contém a obtenção de conjunto de dados, anotação dos dos dados, pré-processamento dos dados, tokenização, lematização, REN, até a criação de um algoritmo para solucionar a relação semântica entre os eventos adversos e as entidades nomeadas. As obras que utilizam língua portuguesa, se resumem a criação do REN "genérico", ou seja, para qualquer tipo de dados, apresentando os resultados no domínio de *pessoas*, *lugares* e *organizações*. A proposta aqui foi além, o modelo de EI proposta emprega o REN como meio para alcançar o objetivo e não como fim, também desenvolve o algoritmo; mesmo inicial e não contemplando todos os cenários possíveis, promove uma reflexão sobre a ausência de material acerca do assunto.

De modo a dar visibilidade ao tema e auxiliar nas tarefas de PLN, a literatura destaca o coreNLP que é altamente citado (mais de 2.000 citações), um software livre, mantido pela universidade de *Stanford*. As contribuições feitas foram extensões dos modelos de DepParser e PosTagger em português do Brasil, sendo essa abordagem escolhida em detrimento da criação de um modelo próprio, que teria pouca ou nenhuma visibilidade. Mesmo com a ascensão do *GitHub*, a falta de documentação, atualização

constante e melhorias, possivelmente faria dele um *"abandonware"*[1].

Ao longo do processo, a limitação para obtenção e anotação dos dados ficou evidente. Para anotar o conjunto de dados não foi encontrado colaboradores (mediante remuneração ou não), o que impactou na duração da pesquisa diretamente, tendo o autor que anotar todas os conjuntos de dados de forma manual e metódica, com base nos dicionários toponímicos criados para esse fim.

Os modelos de DepParser, PosTagger e REN, mostraram-se satisfatórios, sendo o primeiro trabalho a debater textos informais em português do Brasil para detecção de eventos adversos. As bases de dados e as propriedades utilizadas no treinamento foram disponibilizadas, permitindo a reprodutividade do estudo e para auxiliar a comunidade científica. Os conjuntos de dados criados ao longo de trabalho também são inéditos (o autor, até o presente momento, não tem ciência sobre dados dessa espécie em língua portuguesa do Brasil), e contribuem para outras abordagens relativas ao tema.

O filtro semântico apresenta a possibilidade de extrair dados com maior riqueza de informações, podendo incluir o nome do medicamento e o nome do evento adverso, o que caracteriza um filtro mais elaborado, visto que regras podem ser criadas por demanda e combinadas para situações adversas. Como exemplo tem-se o evento adverso febre, que por ser o mais comum, leva à preferência por regras com maior precisão que facilitam a posterior análise, ou seja, por ser um evento adverso comum, apenas em casos com maior certeza (casos positivos), o esforço de observar os dados é justificado; oposto se observa com o evento adverso morte que, por sua complexidade, necessita de maior abrangência, ou seja, pelo alto grau do impacto (resultante em óbito), a quantidade de casos em detrimento da certeza é prioritário, preferível analisar o máximo possível (abrangência), independente da certeza.

A flexibilidade obtida pelo modelo proposto oferece novas formas de explorar as informações. Pode-se optar selecionar os dados por: frequência, raridade, dano (morte tem impacto maior que febre), grupos de interesse (apenas X medicamentos ou N eventos são necessários), evolução da língua de forma natural (palavras são adicionas ao vocabulário, gírias podem ser incorporadas, podendo-se citar "legal" que, em um passado não muito distante era tida como gíria e hoje está integrada à norma culta).

Como resultado obtido nesta dissertação, verificamos que o auxilio de profissionais de saúde, laboratórios e governos no importante processo de farmacovigilância

[1] Software que foi descontinuado, no sentido de "abandonado".

pode extrair *insights* das percepções de sinais. Por exemplo a maior reclamação do medicamento Dipirona foi o seu sabor desagradável, o laboratório pode modificar o incipiente[2] melhorando sua palatabilidade, o que poderia levar a um aumento da receita, uma vez que se trata de um medicamento comum, produzido em larga escala, com pouca margem para lucratividade. Os profissionais de saúde com acesso a recursos (ferramentas, modelos e dados), podem obter novos mais dados e novos resultados. A nível governamental, pode-se agregar a plataforma de farmacovigilância existente, com a possibilidade de atribuir o evento adverso ao medicamento, auxiliando no processo de *recall* de medicamentos, seja pela suspensão do lote ou pelo próprio fabricante. O processo de melhoria contínua de malefícios e benefícios auxiliará a disponibilidade de tratamentos mais seguros e mais eficientes aos pacientes.

Limitações do estudo

Ao longo do trabalho a dificuldade de obtenção de conjuntos de dados em português brasileiro foi muito elevada. Para contornar esse cenário foi criado um *scrapy* para obtenção do conjunto de dados. Outra dificuldade encontrada está diretamente relacionada ao tamanho do conjunto de dados: um conjunto maior iria requerer mais tempo em todas as etapas.

Em língua portuguesa brasileira não foi identificado, até o presente momento, pelo autor, rede social exclusiva para saúde, elevando o nível de dificuldade em obter e tratar os dados, visto que a postagem desse tipo de conteúdo não é predominante. Por conta dessa dificuldade adicional, utilizou-se apenas uma rede, o Twitter. Outras redes sociais ou uma nova, criada para esse fim, seria muito produtivo.

A multidisciplinaridade (especialistas no domínio da linguística para as relações gramaticais e especialistas no domínio da saúde para anotar os modelos e validá-los) constituiu uma forte limitação. O ideal seria trabalhar com três anotadores, para poder gerar comparações entre seus resultados. Infelizmente, não foi possível. O autor tentou contatos em universidades parceiras oferecendo renumeração e/ou a participação em artigos científicos, todavia não ocorrem interessados. Isso resultou na anotação dos

[2] O veículo empregado para transportar o fármaco.

dados de forma manual única e exclusivamente pelo próprio. O formato da anotação simplificado “BIO” em detrimento ao “BILOU” mais completo, impactando nas palavras compostas.

A inexistência de conjunto de dados (e.g., *WordNet*[3]), para a desambiguação lexical influenciou na obra. Como exemplo podemos citar o mineral *zinco*, cujo sua dose pode acarretar em calor, em contrapartida temos a telha de moradias, produzidas com um material zinco, que trás calor ao ambiente. Para substantivos inexistem *WordNet* em português brasileiro.

Trabalhos futuros

Anotação dos conjuntos de dados por especialistas do domínio. Realizar o treinamento dos modelos novamente e implementar variações de WE. Nessa dissertação foi utilizado o modelo *Word2Vec*, além da variação de dimensões do arquivos, nessa dissertação adotamos o valor de 300 dimensões.

Desenvolvimento de uma rede social, dedicada a saúde. Com foco em EAM com suporte para lidar com as diversidades que o tema requer, com apoio da comunidade científica e de instituições com interesse no tema.

Ampliação do conjunto de dados, incluindo novas redes sociais, nessa dissertação, foi utilizado o *Twitter*. A rede *Instagram* é um desafio, pelo foco em imagens.

[3]WordNet são repositórios de dados léxicos, contém as classes gramaticais e são agrupados com o sinônimos cognitivos, expressando um conceito.

Referências Bibliográficas

ACE, E. N. (2008). Automatic content extraction 2008 evaluation plan (ace08). *April, Disponível em: https://my.eng.utah.edu/~cs6961/papers/ACE-2008-description.pdf*.

Alpaydin, E. (2009). *Introduction to machine learning*. MIT press.

Alvaro, N., Conway, M., Doan, S., Lofi, C., Overington, J., and Collier, N. (2015). Crowdsourcing twitter annotations to identify first-hand experiences of prescription drug use. *Journal of biomedical informatics*, 58:280–287.

Amado, R. d. S. (2008). O ensino e a pesquisa de português para falantes de outras línguas. *Patrícia galvão: The private autobiography of a brazilian feminist writer*, page 66.

Antoniou, G. and Harmelen, F. (2008). A semantic web primer, 2nd edn, cooperative information systems.

ANVISA (2009). Resolução de diretoria colegiada número 4. Disponível em: http://bvsms.saude.gov.br/bvs/saudelegis/anvisa/2009/res0004_10_02_2009.html - Acessado em: 12 nov. 2017.

ANVISA (2016). Perguntas frequentes – farmacovigilância. Disponível em: http://portal.anvisa.gov.br/documents/33868/2895429/Perguntas+frequentes+%E2%80%93+Farmacovigil%C3%A2ncia/f8935efb-7ba4-404e-96a7-271871f5f9d2?version=1.0&download=true - Acessado em: 12 nov. 2017.

Aranha, C. N., Vellasco, M., and Passos, E. (2007). Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional. *Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, RJ*.

Atkinson, J. and Bull, V. (2012). A multi-strategy approach to biological named entity recognition. *Expert Systems with Applications*, 39(17):12968–12974.

Baer, B., Nguyen, M., Woo, E., Winiecki, S., Scott, J., Martin, D., Botsis, T., and Ball, R. (2016). Can natural language processing improve the efficiency of vaccine adverse event report review? *Methods of information in medicine*, 55(02):144–150.

Baeza-Yates, R. and Ribeiro-Neto, B. (2013). *Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora.

Bailey, T. L. and Elkan, C. (1993). Estimating the accuracy of learned concepts.". In *Proc. International Joint Conference on Artificial Intelligence*.

Barros, J. A. C. d. (1995). Propaganda de medicamentos: atentado à saúde?

Batista, F. and Figueira, Á. (2017). The complementary nature of different nlp toolkits for named entity recognition in social media. In *Portuguese Conference on Artificial Intelligence*, pages 803–814. Springer.

Beijer, H. and CJ, d. B. (2002). Hospitalisations caused by adverse drug reactions (adr): a meta-analysis of observational studies. *Pharmacy World and Science*, 24(2):46–54.

Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., Leonard, C. E., and Holmes, J. H. (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of biomedical informatics*, 44(6):989–996.

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., and Katz, B. (2016). Universal dependencies for learner english. *arXiv preprint arXiv:1605.04278*.

Bian, J., Topaloglu, U., and Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

Bireme (2019a). Decs - descritores em ciências da saúde. Disponível em: `http://decs.bvs.br/P/decsweb2019.htm`. Data do Acesso: 15 jan. 2019.

Bireme (2019b). This is the portuguese translation of mesh. Disponível em: `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHPOR/index.html`. Data do Acesso: 15 jan. 2019.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bonadio, Í. (2018). *Algoritmos eficientes para análise de campos aleatórios condicionais semi-markovianos e sua aplicação em sequências genômicas*. PhD thesis, Universidade de São Paulo.

Bosco, C., Montemagni, S., and Simi, M. (2013). Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69.

Botsis, T., Buttolph, T., Nguyen, M. D., Winiecki, S., Woo, E. J., and Ball, R. (2012). Vaccine adverse event text mining system for extracting features from vaccine safety reports. *Journal of the American Medical Informatics Association*, 19(6):1011–1018.

Botsis, T., Nguyen, M. D., Woo, E. J., Markatou, M., and Ball, R. (2011). Text mining for the vaccine adverse event reporting system: medical text classification using informative feature selection. *Journal of the American Medical Informatics Association*, 18(5):631–638.

Botsis, T., Woo, E., and Ball, R. (2013). The contribution of the vaccine adverse event text mining system to the classification of possible guillain-barré syndrome reports. *Applied clinical informatics*, 4(01):88–99.

Bowdler, J. (1997). Effective communications in pharmacovigilance: the erice report. *Birmingham, England: W Lake Limited*.

Brasil, C. d. D. (2013). Informações sobre a comercialização dos medicamentos neosaldina, sibutramina diane 35, avastin e hormônio do crescimento. Disponível em: `http://www.camara.gov.br/proposicoesWeb/prop_mostrarintegra?codteor=1059367`- Acessado em: 15 jul. 2018. Data do Acesso: 15 jul. 2018.

Brunton, L., Chabner, B., and Knollmann, B. (2007). As bases farmacológicas da terapêutica goodman & gilman. porto alegre: Artmed.

Buchholz, S. and Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Bunescu, R. C. (2007). *Learning for information extraction: from named entity recognition and disambiguation to relation extraction*. PhD thesis.

Camargo, A. L., Ferreira, M. B. C., and Heineck, I. (2006). Adverse drug reactions: a cohort study in internal medicine units at a university hospital. *European journal of clinical pharmacology*, 62(2):143–149.

Campos, D., Matos, S., and Oliveira, J. L. (2012). Biomedical named entity recognition: a survey of machine-learning tools. In *Theory and Applications for Advanced Text Mining*. IntechOpen.

Candito, M., Nivre, J., Denis, P., and Anguiano, E. H. (2010). Benchmarking of statistical dependency parsers for french. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 108–116. Association for Computational Linguistics.

Cassiani, S. H. D. B. et al. (2005). A segurança do paciente e o paradoxo no uso de medicamentos. *Rev Bras Enferm*, 58(1):95–99.

Cegalla, D. P. (1977). *Novíssima gramática da língua portuguesa:(com numerosos exercícios)*. Companhia Editora Nacional.

Celso, P. L. (1985). Língua e liberdade: por uma nova concepção de língua materna e seu ensino. *Porto Alegre: L&PM*.

Cer, D. M., De Marneffe, M.-C., Jurafsky, D., and Manning, C. D. (2010). Parsing to stanford dependencies: Trade-offs between speed and accuracy. In *LREC*. Floriana, Malta.

Chakrabarti, S. (2002). *Mining the Web: Discovering knowledge from hypertext data*. Elsevier.

Charniak, E., Riesbeck, C. K., McDermott, D. V., and Meehan, J. R. (2014). *Artificial intelligence programming*. Psychology Press.

Chee, B. W., Berlin, R., and Schatz, B. (2011). Predicting adverse drug events from personal health messages. In *AMIA Annual Symposium Proceedings*, volume 2011, page 217. American Medical Informatics Association.

Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Chen, X., Faviez, C., Schuck, S., Louët, L.-L., Texier, N., Dahamna, B., Huot, C., Foulquié, P., Pereira, S., Leroux, V., et al. (2018). Mining patients' narratives in social media for pharmacovigilance: adverse effects and misuse of methylphenidate. *Frontiers in pharmacology*, 9:541.

Chinchor, N., Brown, E., Ferro, L., and Robinson, P. (1999). 1999 named entity recognition task definition. *MITRE and SAIC*.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

Clark, A., Fox, C., and Lappin, S. (2013). *The handbook of computational linguistics and natural language processing*. John Wiley & Sons.

Clark, K., Sharma, D., Qin, R., Chute, C. G., and Tao, C. (2014). A use case study on late stent thrombosis for ontology-based temporal reasoning and analysis. *Journal of biomedical semantics*, 5(1):49.

Cleophas, T. J. and Zwinderman, A. H. (2000). Limitations of randomized clinical trials. proposed alternative designs. *Clinical chemistry and laboratory medicine*, 38(12):1217–1223.

Cocos, A., Fiks, A. G., and Masino, A. J. (2017). Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. *Journal of the American Medical Informatics Association*, 24(4):813–821.

Coden, A., Gruhl, D., Lewis, N., Tanenblatt, M., and Terdiman, J. (2012). Spot the drug! an unsupervised pattern matching method to extract drug names from very large clinical corpora. In *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, pages 33–39. IEEE.

Consortium, L. D. et al. (2005). Ace (automatic content extraction) english annotation guidelines for entities. *Version*, 5(6):05–08.

Corrêa, A. C. G. et al. (2003). Recuperação de documentos baseados em informação semântica no ambiente ammo.

Council, f. I. O. o. M. S. et al. (2010). Practical aspects of signal detection in pharmacovigilance: report of cioms working group viii. *Geneva: Council for International Organizations of Medical Sciences*.

Courtot, M., Brinkman, R. R., and Ruttenberg, A. (2014). The logic of surveillance guidelines: an analysis of vaccine adverse event reports from an ontological perspective. *PloS one*, 9(3):e92632.

da Cunha, A. M., Nascimento, G., and Guedes, G. P. (2018). Uma análise sobre as bulas de medicamentos no brasil. *Brazilian e-Science Workshop*, 12.

Dash, N. (2019). Corpus linguistics: An introduction.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

De Marneffe, M.-C., MacCartney, B., Manning, C. D., et al. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454. Genoa Italy.

de Oliveira, L. E. S. and Morita, M. E. (2000). Introdução aos modelos escondidos de markov (hmm).

De Saussure, F. (1989). *Cours de linguistique générale: Édition critique*, volume 1. Otto Harrassowitz Verlag.

Deng, X. and Sun, H. (2019). Leveraging 2-hop distant supervision from table entity pairs for relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 410–420, Hong Kong, China. Association for Computational Linguistics.

Dias, M. (2005). The brazilian pharmacovigilance programme uppsala reports. edição. 18, abril 2002. *Publicado por the Uppsala Monitoring Centre.*, 9.

Dixon, M. (1997). An overview of document mining technology. *Unpublished paper*.

Dormann, H., Criegee-Rieck, M., Neubert, A., Egger, T., Geise, A., Krebs, S., Schneider, T. H., Levy, M., Hahn, E. G., and Brune, K. (2003). Lack of awareness of community-acquired adverse drug reactions upon hospital admission. *Drug safety*, 26(5):353–362.

Duke, J. D. and Friedlin, J. (2010). Adessa: a real-time decision support service for delivery of semantically coded adverse drug event data. In *AMIA Annual Symposium Proceedings*, volume 2010, page 177. American Medical Informatics Association.

Edwards, I. R. and Aronson, J. K. (2000). Adverse drug reactions: definitions, diagnosis, and management. *The lancet*, 356(9237):1255–1259.

Esuli, A., Marcheggiani, D., and Sebastiani, F. (2013). An enhanced crfs-based system for information extraction from radiology reports. *Journal of biomedical informatics*, 46(3):425–435.

Fabiana, R. V., Marcos, F., Galduroz, J. C., and Mastroianni, P. C. (2011). Adverse drug reaction as cause of hospital admission of elderly people: a pilot study. *Lat. Am. J. Pharm*, 30(2):347–53.

Feldman, R., Sanger, J., et al. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2.

Gabrilovich, E., Markovitch, S., et al. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Gantz, J. and Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012):1–16.

Gerber, R. M. and Vasilévksi, V. (2007). Um percurso para pesquisas com base em corpus.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Gonçalves, T., Silva, C., Quaresma, P., and Vieira, R. (2006). Analysing part-of-speech for portuguese text classification. In *International Conference on Intelligent Text Processing AND Computational Linguistics*, pages 551–562. Springer.

Gong, L.-J., Yuan, Y., Wei, Y.-B., and Sun, X. (2009). A hybrid approach for biomedical entity name recognition. In *2009 2nd International Conference on Biomedical Engineering and Informatics*, pages 1–5. IEEE.

Griffin, F. A. and Resar, R. K. (2009). Ihi global trigger tool for measuring adverse events. *Institute for Healthcare Improvement Innovation Series White Paper*.

Grishman, R. and Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, volume 1.

Gurulingappa, H., Mateen-Rajpu, A., and Toldo, L. (2012a). Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):15.

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., and Toldo, L. (2012b). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Haerian, K., Varn, D., Vaidya, S., Ena, L., Chase, H., and Friedman, C. (2012). Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clinical Pharmacology & Therapeutics*, 92(2):228–234.

Hallas, J., Gram, L. F., Grodum, E., Damsbo, N., Brosen, K., Haghfelt, T., Harvald, B., Beck-Nielsen, J., Worm, J., and Jensen, K. B. (1992). Drug related admissions to medical wards: a population based survey. *British journal of clinical pharmacology*, 33(1):61–68.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. *Unpublished manuscript*, 46.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., and Shah, N. H. (2014). Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug safety*, 37(10):777–790.

Harpaz, R., Vilar, S., DuMouchel, W., Salmasian, H., Haerian, K., Shah, N. H., Chase, H. S., and Friedman, C. (2012). Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. *Journal of the American Medical Informatics Association*, 20(3):413–419.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Hasan, S. A., Zhu, X., Liu, J., Barra, C. M., Oliveira, L., and Farri, O. (2015). Ontology-driven semantic search for brazilian portuguese clinical notes. *Studies in health technology and informatics*, 216:1022–1022.

Hauben, M. and Aronson, J. K. (2009). Defining signal and its subtypes in pharmacovigilance based on a systematic review of previous definitions. *Drug safety*, 32(2):99–110.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., Ojala, S., Salakoski, T., and Ginter, F. (2014). Building the essential resources for finnish: the turku dependency treebank. *Language Resources and Evaluation*, 48(3):493–531.

Hazlehurst, B., Frost, H. R., Sittig, D. F., and Stevens, V. J. (2005). Mediclass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *Journal of the American Medical Informatics Association*, 12(5):517–529.

Hazlehurst, B., Naleway, A., and Mullooly, J. (2009). Detecting possible vaccine adverse events in clinical notes of the electronic medical record. *Vaccine*, 27(14):2077–2083.

He, X., Zemel, R. S., and Carreira-Perpiñán, M. Á. (2004). Multiscale conditional random fields for image labeling. In *Computer vision AND pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on*, volume 2, pages II–II. IEEE.

Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 3–10. Association for Computational Linguistics.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4.

Hjelmslev, L. (1975). *Prolegômenos a uma Teoria da Linguagem. Estudos*. São Paulo: Perspectiva.

Hladka, B. and Holub, M. (2015). A gentle introduction to machine learning for natural language processing: How to start in 16 practical steps. *Language and Linguistics Compass*, 9(2):55–76.

Hotho, A., Nürnberger, A., and Paaß, G. (2005). A brief survey of text mining. In *Ldv Forum*, volume 20, pages 19–62. Citeseer.

Hur, J., Özgür, A., Xiang, Z., and He, Y. (2012). Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining. *Journal of biomedical semantics*, 3(1):18.

Jiang, J. (2012). Information extraction from text. In *Mining text data*, pages 11–41. Springer.

Junior, J. R. C. (2007). Desenvolvimento de uma metodologia para mineração de textos. *Departamento de Engenharia Elétrica, Pontífica Universidade Católica do Rio de Janeiro*.

Kennerfalk, A., Ruigómez, A., Wallander, M.-A., Wilhelmsen, L., and Johansson, S. (2002). Geriatric drug therapy and healthcare utilization in the united kingdom. *Annals of Pharmacotherapy*, 36(5):797–803.

Kodratoff, Y. (1999). Knowledge discovery in texts: a definition, and applications. In *International Symposium on Methodologies for Intelligent Systems*, pages 16–29. Springer.

Koh, Y., Kutty, F. B. M., and Li, S. C. (2005). Drug-related problems in hospitalized patients on polypharmacy: the influence of age and gender. *Therapeutics and clinical risk management*, 1(1):39.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Kripke, S. A. (1972). Naming and necessity. In *Semantics of natural language*, pages 253–355. Springer.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L. J., and Bork, P. (2010). A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1):343.

Kumar, S. and Hebert, M. (2004). Discriminative fields for modeling spatial dependencies in natural images. In *Advances in neural information processing systems*, pages 1531–1538.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lazarou, J., Pomeranz, B. H., and Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama*, 279(15):1200–1205.

Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in corpus linguistics*, pages 105–122.

Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 302–308.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer.

Lexikon, E. D. (2015). Dicionário aulete. Disponível em: `http://www.aulete.com.br/lingua-` Acessado em: 15 jul. 2018.

Linguateca (2009). Comparação global da atomização e das classificações. Disponível em: `https://www.linguateca.pt/aval_conjunta/morfolimpiadas/atomizacao.html`. Data do Acesso: 15 jul. 2018.

Liu, X. and Chen, H. (2015). A research framework for pharmacovigilance in health social media: identification and evaluation of patient adverse drug event reports. *Journal of biomedical informatics*, 58:268–279.

Liu, Y., Carbonell, J., Weigele, P., and Gopalakrishnan, V. (2005). Segmentation conditional random fields (scrfs): A new approach for protein fold recognition. In *Annual International Conference on Research in Computational Molecular Biology*, pages 408–422. Springer.

Long, A. A. (2002). *Epictetus: A Stoic and Socratic guide to life*. Clarendon Press.

Luo, Y., Thompson, W. K., Herr, T. M., Zeng, Z., Berendsen, M. A., Jonnalagadda, S. R., Carson, M. B., and Starren, J. (2017). Natural language processing for ehr-based pharmacovigilance: a structured review. *Drug safety*, 40(11):1075–1089.

Ly, T., Pamer, C., Dang, O., Brajovic, S., Haider, S., Botsis, T., Milward, D., Winter, A., Lu, S., and Ball, R. (2018). Evaluation of natural language processing (nlp) systems to annotate drug product labeling with meddra terminology. *Journal of biomedical informatics*.

Maitra, A., Annervaz, K., Jain, T. G., Shivaram, M., and Sengupta, S. (2014). A novel text analysis platform for pharmacovigilance of clinical drugs. *Procedia Computer Science*, 36:322–327.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014a). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014b). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Marin, N., Luiza, V. L., Osorio-de Castro, C. G. S., and Machado-dos Santos, S. (2003). Assistência farmacêutica para gerentes municipais. In *Assistência farmacêutica para gerentes municipais*.

Markov, Z. and Larose, D. T. (2007). *Data mining the Web: uncovering patterns in Web content, structure, and usage*. John Wiley & Sons.

Marres, N. and Weltevrede, E. (2013). Scraping the social? issues in live social research. *Journal of cultural economy*, 6(3):313–335.

Masnoon, N., Shakib, S., Kalisch-Ellett, L., and Caughey, G. E. (2017). What is polypharmacy? a systematic review of definitions. *BMC geriatrics*, 17(1):230.

Mastroianni, P. d. C., Varallo, F. R., Barg, M. S., Noto, A. R., and Galduróz, J. C. F. (2009). Contribuição do uso de medicamentos para a admissão hospitalar. *Brazilian Journal of Pharmaceutical Sciences*, 45(1):163–170.

Maylawati, D. S. and Saptawati, G. P. (2017). Set of frequent word item sets as feature representation for text with indonesian slang. In *Journal of Physics: Conference Series*, volume 801, page 012066. IOP Publishing.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., et al. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 92–97.

Meyboom, R. H., Egberts, A. C., Gribnau, F. W., and Hekster, Y. A. (1999). Pharmacovigilance in perspective. *Drug safety*, 21(6):429–447.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144.

Michelson, J. D., Pariseau, J. S., and Paganelli, W. C. (2014). Assessing surgical site infection risk factors using electronic medical records and text mining. *American journal of infection control*, 42(3):333–336.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Min, B., Grishman, R., Wan, L., Wang, C., and Gondek, D. (2013). Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, Atlanta, Georgia. Association for Computational Linguistics.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th*

Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pages 1003–1011. Association for Computational Linguistics.

Mitchell, T. M. (1997). Machine learning (mcgraw-hill international editions computer science series).

Monard, M. C., Alves, G. E. d. A. P., Kawamoto, S., and Pugliesi, J. B. (1997). *Uma introdução ao aprendizado simbólico de máquina por exemplos*. ICMSC-USP.

Mooney, R. J. and Bunescu, R. (2005). Mining knowledge from text using information extraction. *ACM SIGKDD explorations newsletter*, 7(1):3–10.

Morlane-Hondère, F., Grouin, C., and Zweigenbaum, P. (2016). Identification of drug-related medical conditions in social media. In *LREC*.

Mota, C. and Santos, D. (2008). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca.

Mota, D. M. (2017). Evolução e resultados do sistema de farmacovigilância do brasil.

Murakawa, C. d. A. A. (2006). *António de Morais Silva: lexicógrafo da língua portuguesa*. Laboratório Editorial da FCL, UNESP.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., and Gonzalez, G. (2015). Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *LREC*.

NLM, N. L. o. M. (2019). This is the portuguese translation of the medical dictionary for regulatory activities (meddra). Disponível em: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MDRPOR/index.html. Data do Acesso: 15 jan. 2019.

Noblat, A. C. B., Noblat, L. A. C. B., de Toledo, L. A. K., de Moura Santos, P., de Oliveira, M. G. G., Tanajura, G. M., Spinola, S. U., and de Almeida, J. R. M. (2011). Prevalência de admissão hospitalar por reação adversa a medicamentos em salvador, ba. *Revista da Associação Médica Brasileira*, 57(1):42–45.

Nunes, A. (2000). Conceitos básicos de farmacovigilância. *Estudos de utilização de medicamentos: noções básicas. Rio de Janeiro: Editora Fiocruz*, pages 106–126.

Ong, M.-S., Magrabi, F., and Coiera, E. (2012). Automated identification of extreme-risk events in clinical incident reports. *Journal of the American Medical Informatics Association*, 19(e1):e110–e118.

Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A. D., and Casillas, A. (2015). On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of biomedical informatics*, 56:318–332.

O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K. L., and Gonzalez, G. (2014). Pharmacovigilance on twitter? mining tweets for adverse drug reactions. In *AMIA annual symposium proceedings*, volume 2014, page 924. American Medical Informatics Association.

Partee, B. B., ter Meulen, A. G., and Wall, R. (2012). *Mathematical methods in linguistics*, volume 30. Springer Science & Business Media.

Patel, H., Bell, D., Molokhia, M., Srishanmuganathan, J., Patel, M., Car, J., and Majeed, A. (2007). Trends in hospital admissions for adverse drug reactions in england: analysis of national hospital episode statistics 1998–2005. *BMC clinical pharmacology*, 7(1):9.

Pedrós, J. L. and Tognoni, G. (1993). *Principios de epidemiología del medicamento*. Ediciones Científicas y Técnicas.

Peng, F. and McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Information processing & management*, 42(4):963–979.

Pereira, J. G. (2002). Reações adversas a medicamentos. *fármaco*, 2(4):6–7.

Pfaffenbach, G., CARVALHO, O., Bergsten-Mendes, G., et al. (2002). Reações adversas a medicamentos como determinantes da admissão hospitalar. *Revista da Associação Médica Brasileira*.

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., Farrar, K., Park, B. K., and Breckenridge, A. M. (2004). Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *Bmj*, 329(7456):15–19.

Pollit, D., Beck, C. T., and Hungler, B. P. (2004). Fundamentos de pesquisa em enfermagem: métodos, avaliação e utilização. *Porto Alegre: Artmed*.

Prybys, K. and Gee, A. (2002). Polypharmacy in the elderly: clinical challenges in emergency practice. *Part 1 Overview, Etiology, and Drug Interactions, Emergency Medicine Reports*, 23(11):145–151.

Pushpa, A. and Kamakshi, S. (2018). Survey on extracting adverse drug reaction using natural language processing. *International Journal of Pure and Applied Mathematics*, 119(Special Issue 7C):2261–2266. cited By 0.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rajapaksha, P. and Weerasinghe, R. (2015). Identifying adverse drug reactions by analyzing twitter messages. In *Advances in ICT for Emerging Regions (ICTer), 2015 Fifteenth International Conference on*, pages 37–42. IEEE.

Ramesh, B. P., Belknap, S. M., Li, Z., Frid, N., West, D. P., and Yu, H. (2014). Automatically recognizing medication and adverse event information from food and drug administration's adverse event reporting system narratives. *JMIR medical informatics*, 2(1).

Rang, R., Ritter, J. M., Flower, R. J., and Henderson, G. (2015). *Rang & Dale Farmacologia*. Elsevier Brasil.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Reshamwala, A., Mishra, D., and Pawar, P. (2013). Review on natural language processing. *IRACST Engineering Science and Technology: An International Journal (ESTIJ)*, 3(1):113–116.

Riloff, Ellen ANDNunca ouvifalar nisso Lehnert, W. (1994). Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems (TOIS)*, 12(3):296–333.

Rocktäschel, T., Weidlich, M., and Leser, U. (2012). Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

Rosen, A. (2017). Twitter testa aumento do limite de caracteres para 280. Disponível em: `https://blog.twitter.com/pt_br/topics/product/2017/Twitter-testa-aumento-do-limite-de-caracteres-para-280.html`. Data do Acesso: 05 jun. 2019.

Ross, M. K., Lin, K.-W., Truong, K., Kumar, A., and Conway, M. (2013). Text categorization of heart, lung, and blood studies in the database of genotypes and phenotypes (dbgap) utilizing n-grams and metadata features. *Biomedical informatics insights*, 6:BII–S11987.

Sandi, G., Supangkat, S. H., and Slamet, C. (2016). Health risk prediction for treatment of hypertension. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6. IEEE.

Santorini, B. (1990). Part-of-speech tagging guidelines for the penn treebank project (3rd revision, 2nd printing). *Ms., Department of Linguistics, UPenn. Philadelphia, PA*.

Santos, D. and Cardoso, N. (2007). Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área.

Santos, D., Seco, N., Cardoso, N., and Vilela, R. (2006). Harem: An advanced ner evaluation contest for portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*.

Sardinha, T. B. (2004). *Lingüística de corpus*. Editora Manole Ltda.

Sarker, A. and Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics*, 53:196–207.

Sato, K. and Sakakibara, Y. (2005). Rna secondary structural alignment with conditional random fields. *Bioinformatics*, 21(suppl_2):ii237–ii242.

Scarinci, R. G. (1997). Ses: Sistema de extração semântica de informações. Master's thesis, Universidade Federal do Rio Grande do Sul.

Segaran, T. (2007). *Programming collective intelligence: building smart web 2.0 applications*. "O'Reilly Media, Inc.".

Segura-Bedmar, I., Martínez, P., Revert, R., and Moreno-Schneider, J. (2015). Exploring spanish health social media for detecting drug effects. In *BMC medical informatics and decision making*, volume 15, page S6. BioMed Central.

Settles, B. (2004). Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 104–107. Association for Computational Linguistics.

Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics.

Singapore, W. A. S. and hootsuite (2018). 2018 q2 global digital statshot. Disponível em: `https://www.slideshare.net/wearesocialsg/2018-q2-global-digital-statshot-94084375`. Data do Acesso: 15 ago. 2018.

Sobhana, N., Mitra, P., and Ghosh, S. (2010). Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications*, 1(3):143–147.

Speck, R. and Ngomo, A.-C. N. (2014). Ensemble learning for named entity recognition. In *International semantic web conference*, pages 519–534. Springer.

Srihari, R. (2000). A hybrid approach for named entity and sub-type tagging. In *Sixth Applied Natural Language Processing Conference*.

Statistics Portal, T. S. P. (2018). Media usage in an internet minute as of june 2018. Disponível em: `https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/`. Data do Acesso: 16 ago. 2018.

Stephens, M. (2011). *Stephens' detection and evaluation of adverse drug reactions: principles and practice*. John Wiley & Sons.

Sullivan, D. (2000). The need for text mining in business intelligence. *DM REVIEW*, 10:12–16.

Sumathy, K. and Chidambaram, M. (2013). Text mining: concepts, applications, tools and issues-an overview. *International Journal of Computer Applications*, 80(4).

Sutton, C. and McCallum, A. (2006). *An introduction to conditional random fields for relational learning*, volume 2. Introduction to statistical relational learning. MIT Press.

Sutton, C., McCallum, A., et al. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Tan, A.-H. et al. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, volume 8, pages 65–70. sn.

Taskar, B., Abbeel, P., and Koller, D. (2002). Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 485–492. Morgan Kaufmann Publishers Inc.

Tkachenko, M. and Simanovsky, A. (2012). Named entity recognition: Exploring features. In *KONVENS*, pages 118–127.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Travaglia, L. C. (2006). *Gramática e interação: uma proposta para o ensino de gramática.* Cortez.

Turland, M. (2010). php| architec'ts guide to web scraping. *Victoria (Canadá): Nanobooks*, 1.

Twitter, I. (2006). Twitter. Disponível em: `https://about.twitter.com/pt.html`. Data do Acesso: 15 jul. 2018.

Twitter, I. (2018). Gentileza gera gentileza. Disponível em: `https://twitter.com/TwitterBrasil/status/1059862844672483329`. Data do Acesso: 05 jun. 2019.

UNESCO, organização das nações unidas para a educação, a. c. e. a. c. (2019). "40ª sessão da conferência geral". Disponível em: `http://en.unesco.org/generalconference/40/apx`. "Data do Acesso: 01 dez. 2019".

Van, D. H. C. S., Sturkenboom, M. C., Van Grootheest, K., Kingma, H. J., and Stricker, B. H. C. (2006). Adverse drug reaction-related hospitalisations. *Drug safety*, 29(2):161–168.

Vasani, K. (2014). Content evocation using web scraping and semantic illustration. *IOSR J. Comput. Eng.(IOSR-JCE)*, 16(3):54–60.

Vazquez, M., Krallinger, M., Leitner, F., and Valencia, A. (2011). Text mining for drugs and chemical compounds: methods, tools and applications. *Molecular Informatics*, 30(6-7):506–519.

Venulet, J. and Ten, H. M. (1996). Methods for monitoring and documenting adverse drug reactions. *International journal of clinical pharmacology AND therapeutics*, 34(3):112.

Walker, D. E. and Amsler, R. A. (1986). The use of machine-readable dictionaries in sublanguage analysis. *Analyzing language in restricted domains: Sublanguage description and processing*, pages 69–83.

Wang, X., Hripcsak, G., Markatou, M., and Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337.

Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. (2010). *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media.

WHO, W. H. O. (2002). The importance of pharmacovigilance. *Essential medicines and health products*.

WHO, W. H. O. (2005). Segurança dos medicamentos: um guia para detectar e notificar reações adversas a medicamentos. Disponível no portal da ANVISA em: `http://www.anvisa.gov.br/farmacovigilancia/trabalhos/seguranca_medicamento.pdf-` Acessado em: 15 jul. 2018.

Xu, R. and Wang, Q. (2014). Large-scale combining signals from both biomedical literature and the fda adverse event reporting system (faers) to improve post-marketing drug safety signal detection. *BMC bioinformatics*, 15(1):17.

Yeleswarapu, S., Rao, A., Joseph, T., Saipradeep, V. G., and Srinivasan, R. (2014). A pipeline to extract drug-adverse event pairs from multiple data sources. *BMC medical informatics and decision making*, 14(1):13.

Zolezzi, M. and Parsotam, N. (2005). Adverse drug reaction reporting in new zealand: implications for pharmacists. *Therapeutics and clinical risk management*, 1(3):181.

Zopf, Y., Rabe, C., Neubert, A., Gassmann, K., Rascher, W., Hahn, E., Brune, K., and Dormann, H. (2008). Women encounter adrs more often than do men. *European journal of clinical pharmacology*, 64(10):999.