UNIVERSIDADE DE SÃO PAULO

FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

ANA KATARINY DE SOUZA CACHETA

Exploração da mineração de texto em documentos da saúde em diferentes idiomas para acompanhamento médico de pacientes com doenças crônicas

Orientador: Prof. Dr. Antonio Pazin Filho

Ribeirão Preto

2018

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

ANA KATARINY DE SOUZA CACHETA

Exploração da mineração de texto em documentos da saúde em diferentes idiomas para acompanhamento médico de pacientes com doenças crônicas

Dissertação apresentada ao Programa de Mestrado Profissional em Gestão de Organizações de Saúde, da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para a obtenção do título de Mestre em Ciências.

Orientador: Prof. Dr. Antonio Pazin Filho

Ribeirão Preto

2018

FICHA CATALOGRÁFICA

Cacheta, Ana Katariny de Souza

Exploração da mineração de texto em documentos da saúde em diferentes idiomas para acompanhamento médico de pacientes com doenças crônicas. Ribeirão Preto, 2018.

68 f. : il.

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 2018. Área de concentração: Neurologia.

Orientador: Pazin Filho, Antônio.

1. Processamento de linguagem natural. 2. Doenças congênitas. 3. CleverCare.

Nome: Cacheta, Ana Katariny de Souza

Título: Exploração da mineração de texto em documentos da saúde em diferentes idiomas para acompanhamento médico de pacientes com doenças crônicas.

Dissertação apresentada ao Programa de Mestrado Profissional em Gestão de Organizações de Saúde, da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para a obtenção do título de Mestre em Ciências.

Aprovada em:

Banca Examinadora

Prof. Dr. _____

Instituição: _____

Julgamento: _____

Prof. Dr. _____

Instituição: _____

Julgamento: _____

Prof. Dr. _____

Instituição: _____

Julgamento: _____

Agradecimentos

Agradeço aos meus pais, Danilo e Rita de Cassia, por todo o incentivo e amor e por acreditarem na minha capacidade de alcançar meus objetivos, mesmo quando eu mesma duvidava, provendo todo apoio e educação para isso.

Agradeço aos meus irmãos, Daniel e Beatriz, por toda bagunça realizada e todos doces constantemente oferecidos.

Agradeço a todos meus amigos, que de alguma forma contribuíram para meu crescimento e caráter. Entre eles toda a família do Projeto Sopa de Taquaritinga, que tiveram a incrível missão de me animar e manter a alegria em todos os meus sábados.

Agradeço a todo suporte que a Kidopi ofereceu, especialmente aos grandes amigos e mentores que tive a oportunidade de encontrar na empresa, principalmente a Juliana Tarossi Pollettini e o Rafael dos Santos Elias que contribuíram, não apenas com o apoio emocional, mas que também dedicaram tempo e com muito carinho ajudaram no desenvolvimento deste projeto.

Agradeço também a Maria Claudia Propheta Alves, a mais paciente secretária existente, que com muita calma sempre resolveu prontamente todas as minhas dúvidas.

Agradecimento especial ao meu orientador Dr. Antonio Pazin Filho, por toda compreensão, e paciência dedicada ao meu trabalho e especialmente a mim, o que contribuiu muito para meu crescimento pessoal, científico e intelectual.

Agradeço a todos que participaram do experimento e contribuíram de alguma maneira para o desenvolvimento deste projeto

Agradeço à Universidade de São Paulo por permitir que eu me torne uma profissional qualificada e prover meios para obter uma formação acadêmica de alta qualidade.

“Palavras são, na minha não tão humilde opinião, nossa inesgotável fonte de magia. Capazes de ferir e de curar. “

(J.K. Rowling)

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

SUMÁRIO

RESUMO

CACHETA, A. K. S. **Exploração da mineração de texto em documentos da saúde em diferentes idiomas para acompanhamento médico de pacientes com doenças crônicas**. 2018. 68 p. Dissertação (Mestrado Profissional de Gestão de Organizações de Saúde) - Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, [2018].

O CleverCare é um framework para controle, gestão e orientação de pacientes que necessitam de acompanhamento médico contínuo. O sistema possui ferramentas de mineração de textos responsáveis por compreender o conteúdo das mensagens e integrar com serviços de mensagem para envio e recebimento das mesmas, onde inicia diálogos com o paciente para gerenciar atividades rotineiras personalizadas e permite, inclusive, ao paciente fazer perguntas a respeito de uma enfermidade ou condição clínica. Desta forma, a comunicação com o paciente é a base para o sucesso do CleverCare, o qual atualmente possui suporte para o português, atuando por meio de suporte e empoderando o paciente ao cuidado de sua saúde. Compreender as implicações lógicas e adaptações necessárias para a compreensão de textos em diferentes idiomas pode fornecer informações para a aplicação dos mesmos procedimentos a outros idiomas, correlacionando informações e estabelecendo lógicas para traduções e tratamento de termos específicos da área, permitindo atender a uma maior demanda de pacientes que necessitam de tratamento contínuo. Para o desenvolvimento do projeto foram utilizadas abordagens e técnicas visando a escalabilidade e expansão de idiomas de maneira dinâmica. Para isso além das decisões de alterações específicas do sistema foram utilizadas ferramentas como o NLTK para o aperfeiçoamento e realização das adaptações necessárias ao projeto, uma vez que essa ferramenta possui suporte a diversos idiomas e está em constante melhoria. Os resultados, analisados por meio de técnicas de acurácia, precisão e revocação, demonstram que a melhoria observada com as adaptações do sistema para suporte aos idiomas de interesse foram positivas e significativas, com aumento de 13% nos indicadores de revocação e acurácia e manutenção da precisão em 100%. Sendo assim, o CleverCare apresentou um bom desempenho e foi capaz de classificar corretamente as mensagens, permitindo ao sistema reconhecer e classificar corretamente diferentes idiomas. Esta solução permite ao sistema não apenas fazer o processamento de diálogos em português, inglês e espanhol, mas

também ingressar no mercado internacional com a possibilidade de expansão e escalabilidade para outros idiomas.

Palavras-Chave: Doenças Crônicas. Processamento de linguagem natural. CleverCare.

ABSTRACT

CACHETA, A. K. S. **Exploration of text mining in health documents in different languages for medical follow-up of patients with chronic diseases**. 2018. 68 p. Dissertação (Mestrado Profissional de Gestão de Organizações de Saúde) - Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, [2018].

CleverCare is a framework for the control, management, and guidance of patients who need ongoing medical follow-up. The system has text-mining tools responsible for understanding the content of the messages and integrating with message services to send and receive messages, where it initiates dialogues with the patient to manage personalized routine activities and allows the patient to ask questions about them in relation to an illness or clinical condition. In this way, communication with the patient is the basis for the success of CleverCare, which currently has support for Portuguese, acting through support and empowering the patient to take care of their health. Understanding the logical implications and adaptations required for the understanding of texts in different languages can provide information for the application of the same procedures to other languages, correlating information and establishing logics for translations and treatment of specific terms of the area, allowing to supply a greater demand of patients who require continuous treatment. For the development of the project, it was used approaches and techniques aimed at scaling and language expansion in a dynamic way. For this in addition to the system-specific changes decisions tools like NLTK were used, aiming at the improvement and accomplishment of the necessary adaptations to the project, since this tool has support to several languages and is constantly improving. The results, analyzed using accuracy, precision and recall techniques, demonstrate that the improvement observed with the system adaptations to support the languages of interest were positive and significant, with an increase of 13% in recall and accuracy indicators and maintenance of 100% of precision. Thus, CleverCare performed well and was able to classify messages correctly, allowing the system to correctly recognize and classify different languages. This solution allows the system not only to process dialogues in Portuguese, English and Spanish, but also to enter the international market with the possibility of expansion and scalability for other languages.

Keywords: Chronic Diseases. Natural language processing. CleverCare.

1. INTRODUÇÃO E JUSTIFICATIVA

A necessidade de redução de leitos hospitalares remete a um fenômeno conhecido como desospitalização (NETO; MALIK, 2007), o qual consiste em dar alta ao paciente e prover suporte ao tratamento utilizando modelos de cuidado alternativos àqueles prestados em ambiente hospitalar. Dentre os modelos alternativos temos a assistência domiciliar, para realizar cuidados em domicílio quando estes já não são de alta complexidade, mas sim de extrema importância.

No entanto, o processo de desospitalização deve ser realizado com cautela, para evitar rehospitalizações. Utilizando o método de alta-assistida, por exemplo, o hospital Dr. João Machado foi capaz de reduzir o índice de rehospitalizações (BEZERRA; DIMENSTEIN, 2011).

Reduzir taxas de rehospitalização atraiu a atenção dos formuladores de política hospitalar como a maneira de melhorar a qualidade do cuidado e de reduzir custos (JENCKS; WILLIAMS; COLEMAN, 2009). A implantação eficaz de um plano terapêutico bem-sucedido de cuidados para pacientes é dependente de participação do paciente e da conformidade com o regime de tratamento (GRADY et al., 2000).

Além disso, em ensaios clínicos e farmacêuticos, um grande problema é o acompanhamento e a necessidade de garantir a aderência dos sujeitos da pesquisa aos protocolos clínicos. A execução incorreta dos protocolos por parte dos sujeitos pode acarretar em resultados inconsistentes, levando a conclusões errôneas ou invalidando os estudos (MARTIN et al., 2005). O contato em tempo real, que pode ser viabilizado através de mensagens instantâneas, garante que sujeitos de pesquisa sigam corretamente os protocolos de pesquisa e tenham suas dúvidas sanadas de forma rápida e pontual, propiciando uma maior aderência aos ensaios.

O presente projeto vem de encontro a esses problemas, buscando viabilizar e potencializar projetos na área da saúde que necessitem soluções informatizadas para a gestão de seus pacientes com o objetivo de, por exemplo, reduzir o número de doenças e mortes evitáveis, complicações, sequelas e internações desnecessárias, bem como garantir uma correta aderência a planos de tratamento e protocolos de pesquisa, uma vez que a aderência do paciente ao seu plano de tratamento é essencial para o desfecho positivo de sua enfermidade. Este projeto está vinculado a um projeto da fase 1 do programa PIPE da Fapesp (processo número 16/00774-4), sob o título “Aprimoramento do sistema CleverCare por meio do desenvolvimento de

novas funcionalidades, adaptações para internacionalização e desenvolvimento de novos métodos de acesso ao sistema", na qual uma de suas vertentes teve como objetivo a adaptação do sistema e preparação para o processo de internacionalização.

O CleverCare é um *framework, ou seja,* uma aplicação semipronta de estrutura reutilizável que pode ser especializado para fazer aplicações sob encomenda (FAYAD; SCHMIDT, 1997). O CleverCare é utilizado para o controle, gestão e orientação de pacientes que necessitam de acompanhamento médico contínuo. O sistema utiliza mensagens de celular para realizar a comunicação com o paciente de forma personalizada.

Abordagens de mineração de textos permitem que o sistema compreenda a mensagem, responda-a e execute ações de forma automática e personalizada (DEZEMBRO, 2015). Sendo assim, a comunicação e linguagem empregadas são de extrema importância para o CleverCare, o qual foi desenvolvido inicialmente para a língua portuguesa.

A possibilidade de as mesmas técnicas empregadas para o processamento de documentos em português serem também utilizadas como ferramentas para processamento de documentos em outros idiomas pode ajudar a compreender aspectos importantes da mineração de textos, além de correlacionar a eficácia e adaptações necessárias para seu funcionamento de maneira adequada.

O CleverCare apresenta impacto positivo consolidado no acompanhamento de pacientes de diabetes e reeducação alimentar do Hospital Albert Einstein, provendo um aumento no controle glicêmico por meio da diminuição da hemoglobina glicada (AVANSI, A. F. et al, 2016), além disso, vem sendo reconhecido nacional e internacionalmente por seu caráter inovador. Em outubro de 2015, o CleverCare foi eleito como um dos melhores softwares em saúde do mundo pela ONU (http://www.wsis-award.org/winners). O reconhecimento alcançado fez com que a estratégia de mercado do CleverCare fosse remodelada, visando a capacidade de atender ao mercado internacional.

Em 2014, com apoio do Redempreendia o diretor executivo da Kidopi, Mario Sérgio Adolfi Jr., participou de uma missão empresarial para Medellín na Colômbia visando avaliar a potencialidade do uso do software em mercado latino americano (https://www.redemprendia.org/pt/actualidad/noticias/seis-empresas-universitarias-do-brasile-portugal-exploram-oportunidades-de-negocio-em-medellin-gracas-ao-redemprendiatrading).

Ainda com relação ao processo de internacionalização do sistema, acordos de parceria e convênio estão sendo firmados com as universidades colombianas: UDES (Universidad de Santander) e UNAB (Universidad Autónoma de Bucaramanga).

O Ethnologue pode ser descrito como um catálogo abrangente das línguas conhecidas faladas no mundo (PAOLILLO, 2006). Onde é possível verificar que os idiomas Inglês e espanhol estão entre os mais falados do mundo (http://www.ethnologue.com/).

A adaptação do CleverCare para os idiomas inglês e espanhol é também uma estratégia para difusão e adesão de novos usuários, beneficiando uma quantidade maior de pessoas que necessitam de tratamento contínuo, como pacientes diabéticos e hipertensos. Além de permitir compreender a utilização de métodos de mineração de texto e suas adaptações em documentos provenientes da área da saúde em diferentes idiomas, assim como as relações entre termos.

As inovações propostas para o CleverCare permitem potencializar o caráter inovador e prepará-lo para uma fase de expansão nacional e internacional.

2. OBJETIVOS

Objetivo Geral

1. Investigar as adequações necessárias para o funcionamento correto do CleverCare nos idiomas inglês e espanhol;

Objetivos Específicos

1. Avaliar as adaptações e necessidades específicas para cada idioma realizadas por meio de estratégias de mineração de texto para documentos da área da saúde;
2. Realizar as adaptações necessárias às ferramentas para o adequado funcionamento nestes idiomas;
3. Avaliar os resultados obtidos por meio dos testes realizados.

3. REVISÃO DA LITERATURA

3.1 CleverCare

A Kidopi desenvolveu e comercializa o CleverCare (www.clevercare.com.br), um framework voltado para o controle, gestão e orientação de pacientes que necessitam de acompanhamento médico contínuo.

Além de interfaces gráficas para usuários e APIs de acesso, o CleverCare possui uma máquina de estados, sistema de gestão de diálogos, ferramentas de Processamento de Linguagem Natural (PLN) e integração com serviços de mensagens como Telegram, Facebook, aplicativos dedicados e SMS[1], que permitem a comunicação entre usuário e sistema de forma personalizada e ativa, proporcionando um cuidado próximo. O usuário pode iniciar o diálogo com o CleverCare por meio de perguntas que serão tratadas como FAQs (*Frequently Asked Questions*) ou o CleverCare pode iniciar o diálogo com o paciente por meio de roteiros desenvolvidos previamente.

Para o desenvolvimento de suporte aos idiomas inglês e espanhol foram experimentadas ferramentas de mineração de texto. A mineração de texto permite descobrir e extrair conhecimento relevante de documentos não estruturados (KAO et al., 2007). Dados textuais podem ser tratados por meio das estratégias de análise semântica, análise estatística ou, ainda, a combinação de ambas (POLLETTINI, 2008). Onde a análise estatística avalia a frequência com que os termos ocorrem no texto e a análise semântica que utiliza métodos relacionados ao processamento de linguagem natural (DOU; HU, 2012).

A coleta de dados realizada pelo CleverCare se dá por meio de mensagens de celular trocadas com o usuário, que para a compreensão do conteúdo das mensagens possui um framework de recuperação de informação que classifica por similaridade as mensagens recebidas, utilizando a codificação *bag-of-words* para o armazenamento de termos.

[1] O CleverCare não disponibiliza uma integração com o WhatsApp pois, até o momento, não existe uma API (Interface de Programação de Aplicações) oficial disponibilizada para tal integração.

A mineração de textos inclui os processos de coleta de documentos de interesse, pré-processamento de textos, mineração de dados e avaliação de resultados.

Podemos definir três processos fundamentais no CleverCare: uma máquina de estados que controla o fluxo de mensagens entre os usuários, ferramentas de recuperação textual que interpretam as respostas de usuário e os serviços de mensagem que são responsáveis pelo envio e recebimento das mesmas. O diálogo entre o sistema e o usuário é bidirecional ou seja tanto o usuário como o sistema podem iniciar o diálogo (DEZEMBRO, 2015). O fluxo de execução é mostrado na Figura 1.

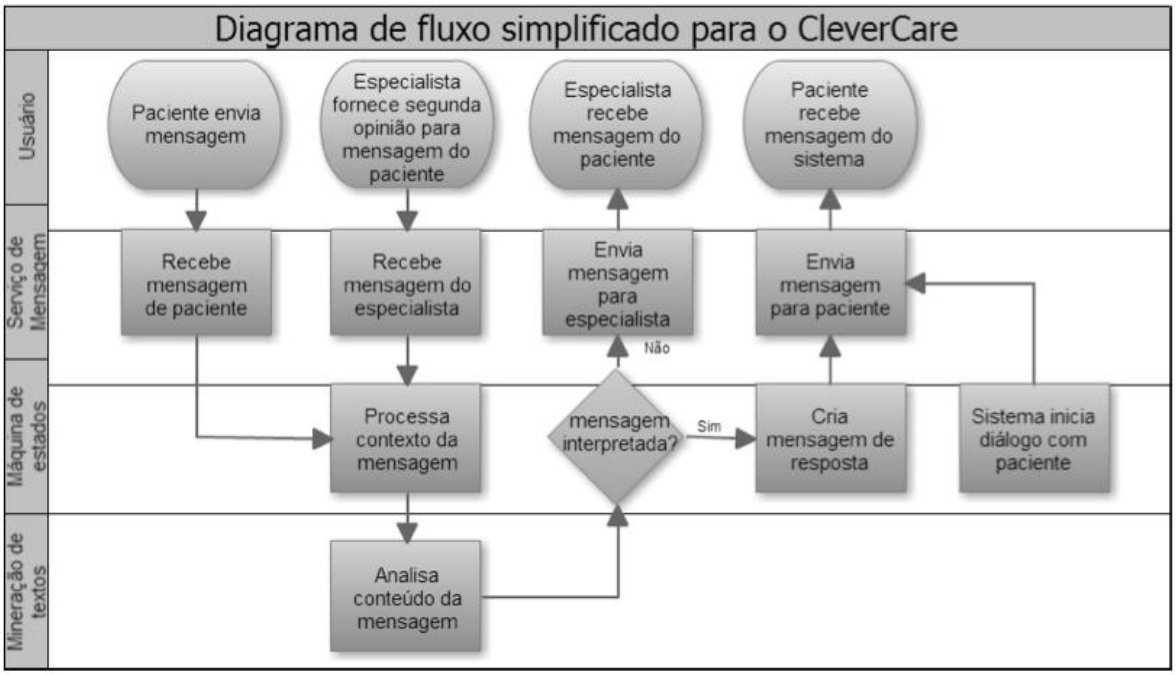

Figura 1 - Diagrama de fluxo simplificado de CleverCare (DEZEMBRO, 2015).

Quando um usuário responde a uma pergunta feita pelo sistema, o mesmo analisa esta resposta com tecnologias de reconhecimento de informação e aprendizado de máquina, verifica a semelhança com outras respostas presentes no banco e retorna uma resposta adequada. Em caso de não existir informações previamente cadastradas este passa a ser respondido pelo especialista. O sistema, por sua vez, pode iniciar uma conversação com o usuário, nesse caso ele envia uma mensagem, que terá sua resposta analisada pelo sistema de reconhecimento textual.

Este esquema de perguntas e respostas foi esquematizado em forma de árvore de decisões. É assim que para cada pergunta temos associado um conjunto de possibilidades de resposta (Figura 2).

O CleverCare utiliza o modelo vetorial para recuperação de informação, onde cada documento é um vetor no espaço vetorial. A ordenação dos documentos com relação ao espaço vetorial é dada pela similaridade entre eles, onde a similaridade é calculada pelo cosseno do ângulo entre os vetores.

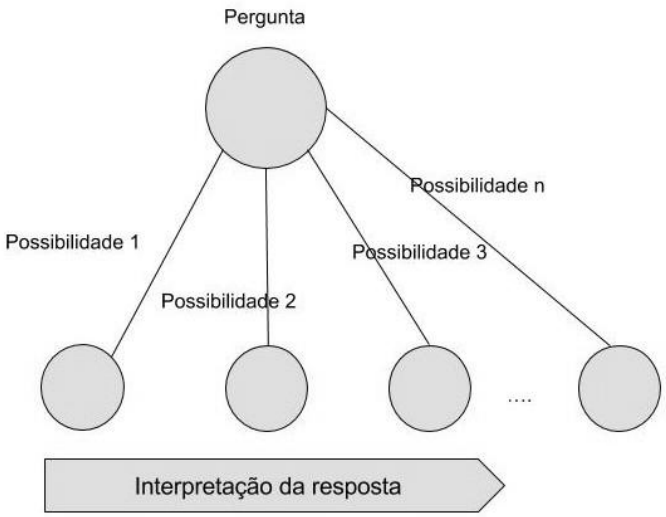

Figura 2 - Componentes do sistema de perguntas e respostas (PESSOTTI, 2017).

As perguntas definem as regras do atendimento e monitoramento do paciente. Estas encontram-se estruturadas em uma árvore, onde os nós representam os diálogos e seus ramos ou arcos representam as possibilidades de resposta (PESSOTTI, 2017). As possibilidades de resposta são definidas pelos especialistas, limitando as possíveis interpretações das respostas do usuário. Com esse mecanismo, o CleverCare garante que toda informação é consistente e curada por um especialista.

Para garantir a qualidade das informações o CleverCare possui um modelo de auditoria e de segunda opinião para controle e acompanhamento dos diálogos do sistema por especialistas, onde a auditoria é quando o sistema é capaz de classificar uma determinada mensagem e o especialista apenas confirma essa classificação, enquanto a segunda opinião é quando o especialista categoriza ou responde mensagens.

A funcionalidade de segunda opinião do CleverCare é importante para casos em que o sistema não consegue tomar uma decisão quanto à classificação de uma mensagem. Sendo assim, é uma solução que, com auxílio humano, realiza a união da “inteligência artificial” e “inteligência humana” de maneira fluida e imperceptível ao paciente. No entanto, o ideal é que essa funcionalidade seja utilizada o mínimo possível, uma vez que demanda trabalho humano.

3.2 Mineração de textos

A sobrecarga de informação é um fenômeno contemporâneo observado a partir do crescimento exponencial na disposição de informações, registrada principalmente após a popularização e a expansão da Internet (SANTOS et al., 2014).

Com o avanço tecnológico as informações passaram a ser digitalizadas, possibilitando a extração de informações com maior facilidade e rapidez quando comparado ao fluxo em papel. Os textos são descritos em linguagem natural, desta forma utilizamos mineração de texto e processamento de linguagem natural para compreender e extrair conceitos relevantes de um documento.

A mineração de texto, também conhecida como mineração de dados de texto ou descoberta de conhecimento a partir de bancos de dados textuais, refere-se ao processo de extração de padrões e conhecimentos interessantes e não triviais de documentos de texto (TAN, 1999).

> Mineração de Textos é um campo multidisciplinar. Para o tratamento de textos e obtenção de conhecimento presente neles, fez-se necessário buscar e empregar avanços, técnicas e conceitos de diversas áreas como Ciência Cognitiva, Processamento de Linguagem Natural, Aprendizado de Máquina, Estatística, Recuperação de Informação e, principalmente, Mineração de Dados, da qual teve seu ponto de partida (CRISTINA; BARION; LAGO, 2008).

3.3 Recuperação de informações

Recuperação de Informação lida com a representação, armazenamento, organização e acesso a itens de informação, e tem por objetivo maior prover ao usuário acesso facilitado à informação de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 1999).

O processo de recuperação de informação consiste em identificar no conjunto de documentos (corpus) de um sistema, que atendem à necessidade de informação do usuário (FERNEDA, 2003).

A área de Recuperação de Informação desenvolveu modelos para a representação de grandes coleções de textos que identificam documentos sobre tópicos específicos. (MORAIS; AMBRÓSIO, 2007).

Um dos modelos mais utilizados é o modelo vetorial, onde cada documento é representado por um vetor de termos e cada termo possui um valor associado que indica o grau de importância desse no documento, também conhecido como peso.

As principais métricas de avaliação utilizadas em Mineração de Textos foram adotadas da área de Recuperação de Informação e são baseadas na noção de relevância. Um documento é considerado relevante quando possui importância para o tópico considerado (AZEVEDO, 2008).

	t_1	t_2	t_3
eBUSCA$_1$	0.2	0.35	0.1

	t_1	t_2	t_3
DOC$_1$	0.3	0.0	0.5
DOC$_2$	0.5	0.4	0.3

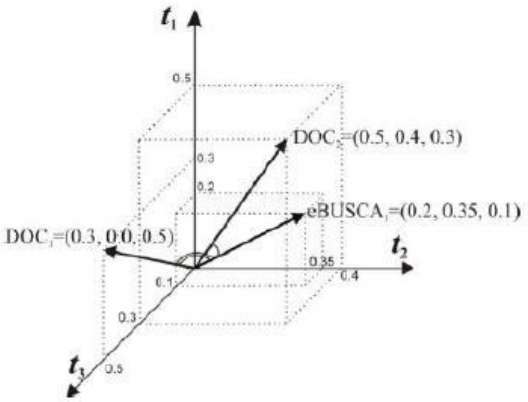

Figura 3 – Representação vetorial de dois documentos e uma expressão de busca. Fonte: (EDBERTO, 2017)

Na Figura 3 temos um exemplo de modelo vetorial utilizado para a recuperação de informação. Neste exemplo temos um espaço vetorial com uma expressão de busca (eBusca) e dois documentos (DOC1 e DOC2).

A utilização de uma mesma representação tanto para os documentos como para as expressões de busca permite calcular o grau de similaridade entre uma determinada busca e cada um dos documentos do corpus (EDBERTO, 2017).

3.4 Processamento de linguagem natural

O processamento de linguagem natural (PLN) surge como uma possível solução aos problemas relacionados à recuperação de informação pela simples observação de que os documentos e as expressões de busca são objetos linguísticos. Sendo assim, o processamento de linguagem natural, é a tentativa de extrair uma representação de significado mais completo do texto livre (KAO et al., 2007).

O processamento de linguagem natural é um conjunto de técnicas computacionais para a análise de textos em um ou mais níveis linguísticos, com o propósito de simular o processamento humano da língua.

> O objetivo final do Processamento de Linguagem Natural é fornecer aos computadores a capacidade de entender e compor textos. Entender um texto significa reconhecer o contexto, fazer análise sintática, semântica, léxica e morfológica, criar resumos, extrair informação, interpretar os sentidos e até aprender conceitos com os textos processados (CRISTINA; BARION; LAGO, 2008).

Na Mineração de texto, as técnicas de processamento de linguagem natural são utilizadas, principalmente, na fase de pré-processamento (SANTOS et al., 2014).

3.5 Bag-of-Words

A metodologia de *bag-of-words (BoW)* foi proposta inicialmente no domínio de recuperação de texto para a análise de documentos de texto (BOSCH; MUÑOZ; FREIXENET, 2007) , com o objetivo de transformar os dados não estruturados em um formato estruturado, representado por uma tabela atributo-valor (MARTINS; MONARD; MATSUBARA, 2003)

> A abordagem mais simples e mais utilizada na estruturação de textos é a representação de *bag-of-words*, onde o documento é representado por um vetor das contagens de palavras que aparecem nele. Dependendo do método de classificação, o vetor pode ser normalizado à unidade e dimensionado de modo que palavras comuns sejam menos importantes do que palavras raras, como na representação tf*idf. (BOULIS; OSTENDORF, 2005)

Essa abordagem ignora informações estruturais, de pontuação e de ordem das palavras, mas armazena o número de aparições de cada palavra. Assim, a codificação

bag-of-words, apesar de não ser suficiente para uma interpretação completa da linguagem é suficiente para processos de clusterização ou de recuperação de informação (POLLETTINI, 2008).

Sendo assim, na codificação *bag-of-words*, os dados textuais podem ser representados por meio de uma tabela atributo-valor, como a Tabela 1, que representa os documentos de textos sendo classificados de acordo com atributos que permitem que o documento seja classificado de alguma maneira. Um atributo que pode ser utilizado é a frequência relativa, como representada na tabela abaixo, por meio de três documentos e a frequência dos termos “Eu”, “sempre”, “doação”, “sangue” nestes documentos.

Tabela 1 - Tabela atributo-valor para representação de documento.

	Eu	**Sempre**	**Doação**	**Sangue**
DOC 1	1	1	1	
DOC 2	1		1	1
DOC 3	2			

Uma representação direta do modelo *bag-of-words* consiste no Modelo de Espaço Vetorial (*Vectorial Space Model – VSM*), o qual utiliza representação geométrica para representar documentos.

3.6 Pré-processamento Textual

3.6.1 Tokenização

Tokenização é o processo de quebrar um fluxo de conteúdo textual em palavras, termos, símbolos ou outros elementos significativos chamados *tokens* (VIJAYARANI; JANANI, 2016).

> Este processo pode não ser uma tarefa trivial. Caracteres especiais, números, pontuações, *emoticons* e separações silábicas são removidos. É importante decidir se letras maiúsculas e minúsculas são diferenciadas, pois uma mesma palavra pode ser representada por *tokens* diferentes caso não haja esta distinção. Estes e outros aspectos devem ser tratados cuidadosamente pois a extração eficiente de *tokens* gera melhores resultados (Picchi Netto, 2009).

A tokenização é benéfica tanto na linguística quanto na ciência da computação, onde faz a parte da análise lexical (PORTER, 1980a).

O processo de tokenização é principalmente complicado para idiomas escritos em '*scriptio continua*' que não revela limites de palavras como o grego antigo, o chinês ou o tailandês (HEMALATHA; VARMA; A.GOVARDHAN, 2012). Um *continuum Scriptio*, também conhecido como *scriptura continua* ou *scripta continua*, é um estilo de escrita sem espaços ou outras marcas entre as palavras ou frases (VIJAYARANI; JANANI, 2016). O principal uso da tokenização é identificar as palavras-chave significativas (VIJAYARANI; ILAMATHI; NITHYA, 2015).

Eu gosto de caminhar no parque					
Eu	gosto	de	caminhar	no	parque

No exemplo acima está representada uma frase que passou pelo processo de tokenização, tendo como resultado seis *tokens* a partir da frase inicial “Eu gosto de caminhar no parque”.

Esta etapa pode não ser trivial, pois uma extração eficiente exige decisões de tratamento sobre a remoção ou manutenção de *emoticons*, caracteres especiais e diferenciação entre maiúsculas e minúsculas. No CleverCare, por exemplo, optou-se por não realizar a remoção de emoticons, uma vez que este pode ser conter um significado importante na mensagem.

3.6.2 Remoção de Stopwords

Segundo (MORAIS; AMBRÓSIO, 2007) esta fase envolve a eliminação de algumas palavras que não devem ser consideradas no documento, conhecidas como *stopwords*, as quais são palavras consideradas não relevantes na análise de textos, justamente por não traduzirem sua essência. Normalmente fazem parte desta lista as preposições, pronomes, artigos, advérbios, e outras classes de palavras auxiliares. Na Figura 4 temos uma frase que passou pelas etapas de tokenização e remoção de *stopwords*, tendo como resultado as palavras “gosto”, “caminhar” e “parque” após a remoção dos *stopwords* “Eu”, “de” e “no”.

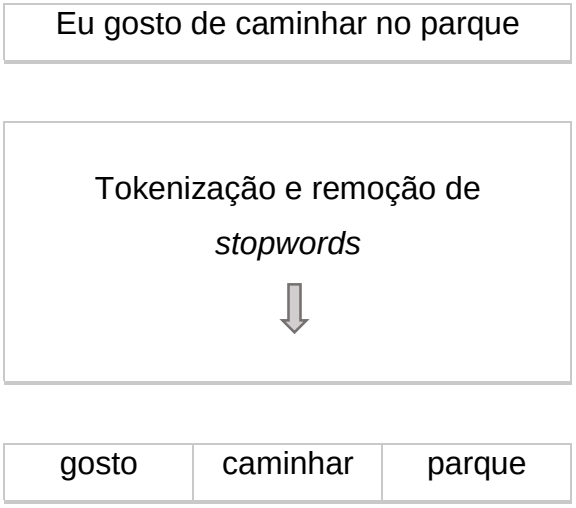

Figura 4 - Exemplo da aplicação de tokenização, seguida da remoção de *stopwords*.

3.6.3 Redução ao radical

Segundo Matsubara, Martins e Monard (2003), algoritmos de *stemming* consistem em uma normalização linguística em que formas variantes de um mesmo termo são reduzidas a uma forma comum denominada *stem* ou radical. Para isso, são removidos prefixos e/ou sufixos de um termo, ou, caso se trate de uma variação de conjugação verbal, é realizada a transformação do verbo para sua forma no infinitivo (POLLETTINI, 2008).

> Durante o processo de indexação, dependendo do caso, torna-se interessante eliminar as variações morfológicas de uma palavra. Elas são eliminadas através da identificação do radical de uma palavra. Os prefixos e os sufixos são retirados e os radicais resultantes são adicionados ao índice. Essa técnica de identificação de radicais é denominada *stemming*, que em inglês significa reduzir uma palavra ao seu radical (MORAIS; AMBRÓSIO, 2007).

Matsubara, Martins e Monard (2003) afirmam ainda que um dos algoritmos de *stemming* mais conhecidos é o de Porter (PORTER, 1980b). Sua implementação original remove sufixos de termo em inglês e tem sido bastante utilizada, motive pelo qual continua sendo adaptado a outros idiomas, uma vez que esta abordagem é extremamente dependente do idioma do documento (POLLETTINI, 2008).

Computador	Algoritmo de radicalização	Comput
Computacionalmente	⇨	
Computar		
Computadores		

Figura 5 - Exemplo de redução de palavras ao seu radical

Qualquer documento possui muitas palavras flexionadas de diversas formas. Um substantivo, por exemplo, pode ter gênero, número e grau. Um verbo possui diversos tempos verbais e conjugações distintas.

Como podemos observar na Figura 5, esta etapa visa reduzir ao radical original as palavras que foram flexionadas, reduzindo a dimensionalidade da mineração de textos.

Embora extremamente dependente do idioma, os algoritmos de *stemming* representam regras básicas da linguística. Em geral, algoritmos de *stemming* retornam uma representação que, em alguns casos, se aproximam do radical latino ou grego (ZWEIGENBAUM; GRABAR, 1999).

3.6.4 Correção ortográfica

Os erros ortográficos são difundidos na escrita informal. Assim as perguntas que as pessoas fazem sobre sua saúde ou a de outra pessoa frequentemente contêm muitos erros ortográficos (ZHANG, 2010).

Erros ortográficos podem não representar uma carga cognitiva significativa para um humano, mas eles podem limitar severamente a eficácia de um sistema automatizado (ZHANG, 2010).

Uma edição pode representar quatro tipos de alterações na palavra, sendo estas: deleção, inserção, transposição ou alteração. Onde a deleção é a remoção de uma letra, uma inserção é a adição de uma letra, uma transposição é a troca de letras adjacentes e uma alteração é trocar de uma letra por outra.

Uma forma de realizar a correção ortográfica é verificar se uma determinada palavra está contida no dicionário. Se estiver, não é necessário realizar nenhuma correção ortográfica, caso contrário, é realizada uma busca por sugestões que é ordenada pela distância máxima, valor este que representa o número de alterações necessárias para transformar a palavra a ser corrigida na correção sugerida. Ou seja,

utiliza-se um cálculo de distância para selecionar a escolha que representa melhor a palavra a ser corrigida (LACHOWICZ; THOMAS, 2018).

Dentre os diferentes algoritmos de verificadores ortográficos temos o modelo probabilístico *n-grams*, o qual sugere palavras e prioriza-as de acordo com o contexto da sentença qual a palavra mal soletrada. Além disso, dois benefícios dos modelos *n-gram* (e algoritmos que os utilizam) são simplicidade e escalabilidade (SUNDBY, 2009).

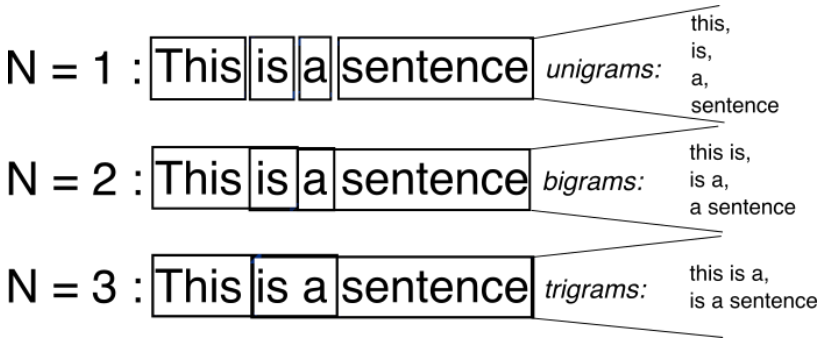

Figura 6 - Representação do modelo probabilístico *n-grams* com o valor de N sendo alterado de 1 a 3. Fonte: (BURSTEIN, 2016).

Na Figura 6 temos a representação da aplicação do modelo n-grams em uma frase para diferentes valores de N. Com esta representação é possível perceber que a frase é tratada de maneira fragmentada, formando *tokens* com N palavras. Ou seja, para o N = 1 temos uma fragmentação onde cada palavra é tratada individualmente, enquanto que para N = 2 temos subgrupos de 2 palavras, e para N = 3 subgrupos de 3 palavras.

Um exemplo deste tipo de aplicação é o Hunspell, um corretor ortográfico e analisador morfológico desenvolvido para idiomas com morfologia rica, palavras compostas e codificação de caracteres complexos, na qual utiliza formatos especiais de dicionários com radicais e afixos validos em cada idioma. Ou seja, cada idioma pode ter arquivos para ortografia, hifenização e um dicionário de sinônimos (OOMS, 2017).

3.7 NLTK

O NLTK é uma biblioteca que contém um conjunto de técnicas de processamento de linguagem natural (“Natural Language Toolkit — NLTK 3.2.5 documentation”, 2017). É um conjunto de ferramentas implementado como uma

coleção de módulos independentes, cada um dos quais define uma estrutura ou tarefa de dados específica (LOPER; BIRD, 2002).

O NLTK define uma infraestrutura que pode ser usada para construir programas de processamento de linguagem natural em Python, por meio de classes básicas para representar dados relevantes, como interfaces padrão para executar tarefas como tokenização, marcação de parte da fala, análise sintática e classificação de texto, além de implementações padrão para cada tarefa que podem ser combinadas para resolver problemas complexos (BIRD; KLEIN; LOPER, 2009).

Desta forma, o NLTK realiza o processamento de linguagem natural simbólica e estatística e está interligado com corpora anotados, podendo ser usado não só como um complexo de treinamento, mas também como uma ferramenta analítica ou base para o desenvolvimento de sistemas de processamento de texto aplicado.

3.8 Aprendizado de máquinas

O aprendizado de Máquinas é um campo focado no desenvolvimento de métodos e técnicas que permitem aos computadores aprenderem tarefas específicas (CARVALHO, 2012). O objetivo principal do aprendizado de máquina é que dado um novo exemplo, de classe desconhecida, seja possível predizer sua classe.

Geralmente um dos fatores mais importantes na determinação é o tipo de retorno disponível para aprendizado no qual o algoritmo se depara (RUSSELL; NORVIG, 1995). No aprendizado de máquinas supervisionado precisamos de um conjunto de treinamento para que o modelo de aprendizagem aprenda com base nele. Ou seja, a partir de um conjunto de dados rotulados que já sabemos qual é a saída correta, tendo a ideia de que existe uma relação entre a entrada e a saída.

3.9 Linguística

O estudo da língua, linguagem e fala permite estabelecer correlações e investigar particularidades no tratamentos de linguagem de comunicação.

Sob o enfoque da Linguística, mesmo um simples documento apresenta abundantes estruturas semânticas e sintáticas, ainda que estas estejam implícitas no texto (CRISTINA; BARION; LAGO, 2008).

Embora as línguas estejam ligadas à cultura, os conceitos expressos por estas línguas são universais (SOERGEL, 1997)

3.10 Validação

O sucesso na recuperação de informação depende igualmente do processo de indexação dos documentos, da linguagem de indexação, da interface usuário-sistema e das estratégias de buscas empregadas (COSTA, 2008).

Os principais critérios e medidas para avaliação de um sistema de recuperação de informação foram definidos por F.W. Lancaster. Para o autor, relevância é uma consideração pessoal, isto é, cada usuário terá uma interpretação diferente para o que é e o que não é sua necessidade de informação o que torna a avaliação um processo complicado (COSTA, 2008).

Dentre as medidas de avaliação, temos a acurácia, a revocação e a precisão. Em geral, um sistema considerado como eficiente é capaz de apresentar um bom equilibro entre precisão e revocação, embora esse pressuposto possa não se aplicar dependendo dos requisitos considerados (MARTINS, 2009).

3.10.1 Acurácia

A acurácia representa como o classificador se saiu de maneira geral, pois mede a quantidade de acertos sobre o todo. Em resumo, é responsável por mensurar o percentual de instâncias classificadas corretamente. A acurácia pode ser representada pela fórmula a seguir, na qual visa-se responder a seguinte pergunta: **No geral, o quão frequente o classificador está correto?** (LEAL, 2017)

$$Acurácia = \frac{Verdadeiros\ Positivos\ (TP) + Verdadeiros\ Negativos\ (VN)}{Total}$$

3.10.2 Revocação

Coeficiente de revocação (recall) de uma busca é a proporção de todos os itens relevantes em uma coleção particular ou banco de dados que a busca é capaz de recuperar (COSTA, 2008). A revocação pode ser representada pela fórmula a seguir,

na qual visa responder a seguinte pergunta: **Quando realmente é de uma determinada classe, o quão frequente você classifica como sendo realmente dessa classe?** (LEAL, 2017)

$$Recall = \frac{Verdadeiros\ Positivos\ (TP)}{Verdadeiros\ Positivos\ (TP) + Falsos\ Negativos\ (FN)}$$

3.10.3 Precisão

É a capacidade do sistema em recuperar somente referências relevantes, eliminando aquelas que não são importantes (lixo). É uma medida muito importante para avaliação quando a busca é realizada por um intermediário. Isto porque na busca realizada pelo próprio usuário ele faz o julgamento de importância no momento que está realizando sua busca (COSTA, 2008). A precisão pode ser representada pela fórmula a seguir, na qual visa responder a seguinte pergunta: **Daqueles que classifiquei como corretos, quantos efetivamente eram?** (LEAL, 2017)

$$Precisão = \frac{Verdadeiros\ Positivos\ (TP)}{Verdadeiros\ Positivos\ (TP) + FalsosPositivos\ (FP)}$$

4. MATERIAIS E MÉTODOS

Para a etapa de revisão bibliográfica foi realizada uma revisão sistemática. Os termos de busca utilizados foram obtidos por meio de consulta aos Descritores em Ciências da Saúde (decs.bvs.br). Esta plataforma foi utilizada para procurar trabalhos indexados através da combinação dos descritores “processamento de linguagem natural” e “doença crônica”. Na pesquisa bibliográfica foram utilizadas as bases SciELO (www.scielo.org), LILACS (bases.bireme.br), MEDLINE (www.ncbi.nlm.nih.gov/pubmed) e Scopus (www.scopus.com/periodicos.capes.gov.br).

A gestão do desenvolvimento do projeto é realizada utilizando uma abordagem de desenvolvimento denominada *Scrum*. Este é um processo de desenvolvimento iterativo e incremental, que foca no desenvolvimento ágil de software (SCHWABER, 2004).

As principais linguagens de programação utilizadas foram PHP versão 5.6.10 (http://php.net), e Python versão 2.7.13 (www.python.org).

O CleverCare realiza o pré-processamento das mensagens utilizando técnicas de mineração de textos. O sistema utiliza, para a recuperação de informação, a abordagem de *bag-of-words*.

As etapas de pré-processamento do CleverCare incluem a remoção de *stopwords*, redução ao radical, correção gramatical e busca fonética. Estas etapas de processamento foram realizadas por meio da ferramenta NLTK (LOPER; BIRD, 2002).

Para a remoção de *stopwords* foram utilizados os corpus no NLTK, sendo este representado por meio das bases de dados do Snowball (http://snowball.tartarus.org) para cada idioma de interesse, onde a lista de palavras recomendadas como *stopwords* foi revisada e alterada conforme a viabilidade para a aplicação no contexto deste trabalho.

Para reduzir as variações de palavras a um radical foi utilizada a técnica de *Stemming,* onde foi aplicado o algoritmo de Porter adaptado às regras gramaticais dos idiomas de interesse já existentes no algoritmo.

Para a correção ortográfica foi utilizado o Hunspell (NÉMETH, 2005) nas respectivas línguas de interesse por meio da biblioteca pyEnchant versão 1.6.11 (KELLY, 2017).

Alguns dos participantes são da região de Ribeirão Preto, principalmente na Universidade de São Paulo - Campus de Ribeirão Preto, sendo estes os alunos de

intercâmbio. O recrutamento foi realizado por meio de contato com o iTeam USP-Ribeirão Preto (https://www.iteamusprp.com.br/) visando recrutar esses alunos participantes fluentes nos idiomas de interesse.

Todos os roteiros do sistema foram construídos nos idiomas de interesse e revisados por pessoas nativas e fluentes nesses idiomas.

Com relação ao tamanho amostral, na mineração de textos, mais do que a quantidade de pessoas participando do projeto temos a quantidade de conteúdo avaliado, ou seja, no caso do CleverCare a quantidade de mensagens trocadas com os usuários e a capacidade do sistema classificar corretamente essas mensagens. A amostra de estudo foi representada por 151 interações, sendo 89 em inglês e 65 em espanhol.

Para a validação dos resultados obtidos foram utilizadas as medidas de acurácia, precisão e revocação, onde esses indicadores foram obtidos por meio de um testador automatizado para o CleverCare desenvolvido pela empresa Kidopi.

O projeto foi aprovado no Comitê de Ética em Pesquisa do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (CAAE: 65303717.2.0000.5440), parecer nº 2.053.036. Antes de realizar os experimentos todos os participantes foram orientados e o prosseguimento do mesmo só prosseguiu após esclarecidas todas as dúvidas. O Termo de Consentimento Livre e Esclarecido encontra-se no APÊNDICE A.

5. RESULTADOS E DISCUSSÕES

5.1 Revisão Sistemática

Realizou-se uma revisão sistemática da literatura mediante busca nas bases de dados SciELO, LILACS, PUBMed e Scopus utilizando os termos “Processamento de linguagem natural e doenças crônicas” (em português, inglês e espanhol). Esta revisão será submetida a um congresso da área e está apresentada na integra no APÊNDICE B.

Dentre os artigos encontrados, catorze estavam presentes em mais de uma plataforma, representando duplicidade. Após a remoção de duplicidade foram eliminados seis artigos, que não permitiam acesso completo, além de catorze devido a avaliação de seu resumo, resultando em uma redução a 20 artigos finais que foram avaliados neste projeto, como representado na Figura 7.

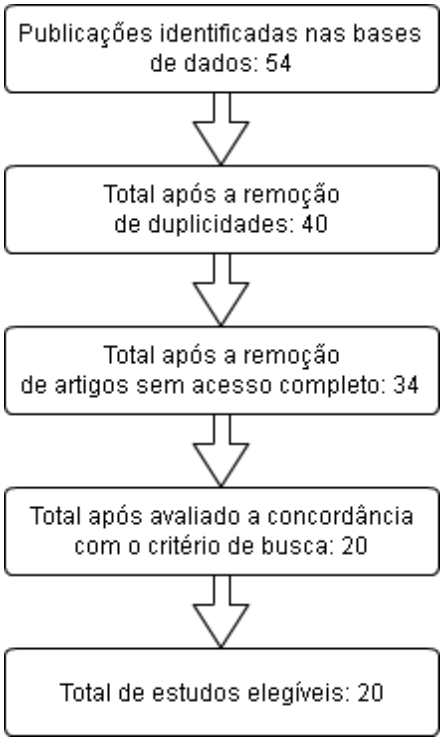

Figura 7 - Número de artigos resultantes em cada etapa do fluxo metodológico.

Entre os 20 artigos resultantes, 9 foram encontrados nas plataformas Scopus e MEDLINE, 1 apenas na Scopus e 10 apenas na MEDLINE. A busca na plataforma SciELO não retornou nenhum resultado, enquanto que a LILACS obteve um artigo, o qual foi eliminado dos estudos finais durante a etapa de avaliação de elegibilidade.

A análise de distribuição dos artigos analisados por ano de publicação demonstra que metade dos estudos publicados estão concentrados nos anos de 2013 a 2015, além de ter o primeiro registro em 1993 (Figura 8).

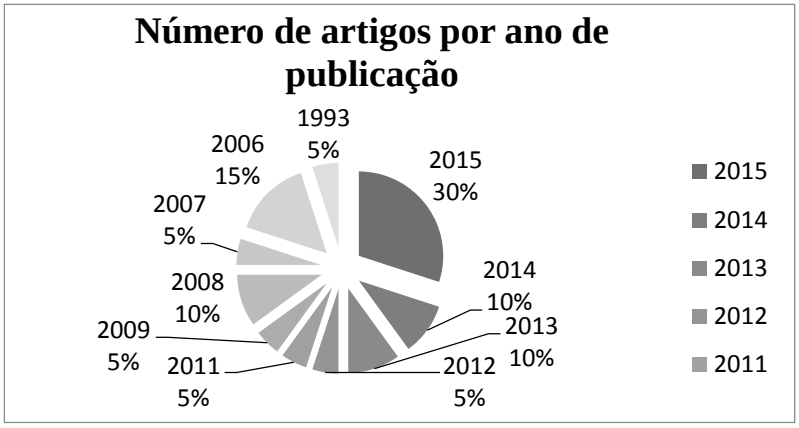

Figura 8 - Porcentagem de artigos publicados por ano.

A análise dos estudos selecionados na presente revisão bibliográfica apontou para a recente pesquisa relacionada a processamento de linguagem natural e doenças crônicas, na qual metade dos artigos resultantes foram publicados nos últimos anos.

Não existe uma associação temporal com relação às técnicas mais frequentemente utilizadas, ou seja, dentre as técnicas mais abordadas no conjunto de artigos estudados não existe um período onde uma determinada técnica foi ou deixou de ser empregada.

Aplicações de processamento de linguagem natural podem empregar abordagens superficiais ou profundas, quando fazem uso de poucas ou muitas informações linguísticas, as quais são normalmente categorizadas em níveis de conhecimento (NÓBREGA, 2013).

A tokenização, segmentação de sentenças e etiquetagem foram as metodologias de destaque entre os artigos estudados. Estas são parte da análise sintática do processamento de linguagem natural.

O primeiro processamento que é efetuado na análise sintática é a identificação das classes das palavras (também conhecidas como classes morfológicas, etiquetas lexicais ou partes da fala). Para proceder esta classificação são implementados *parsers* que identificam nas frases as classes de palavras que as compõe (MULLER, 2003).

A metodologia mais frequente foi a etiquetagem, classificação de palavras, onde as classes de palavras que compõe as frases são identificadas (OLIVEIRA; FREITAS, 2006).

5.2 Análise de requisitos

Em geral, a tarefa de levantamento de requisitos é realizada antes do início do desenvolvimento e, como resultados, têm-se documentos contendo descrições do sistema, modelagens e diagramas que são de grande importância para as etapas de desenvolvimento, testes, implantação e manutenção do sistema.

Foram avaliados os primeiros requisitos para o funcionamento do sistema, que em sua totalidade deverá ser capaz de receber a indicação do idioma de interesse de acordo com as preferências dos usuários.

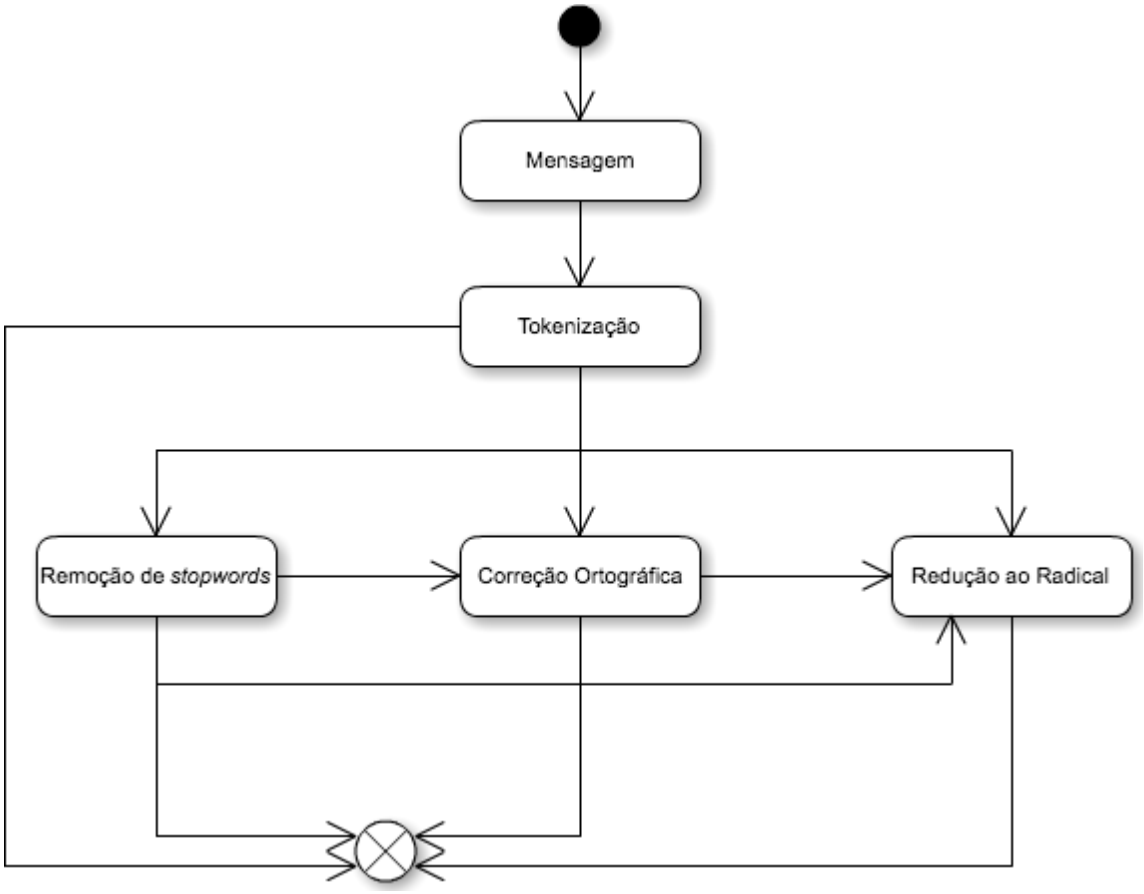

Figura 9 - Diagrama de atividades relacionadas à etapa de pré-processamento do CleverCare.

Como representado na Figura 9, as mensagens trocadas com o CleverCare, após a tokenização, podem não sofrer outros pré-processamentos textuais, assim

como apresentar os processos de remoção de *stopwords*, correção ortográfica e redução ao radical. Estes processos são modulares e podem ser realizados individualmente, ou em combinação entre os diferentes processos. Para a combinação de dois ou mais módulos, a ordem de execução dos algoritmos deve ser respectivamente a remoção de *stopwords*, correção ortográfica, e redução ao radical.

A ordem de correção ortográfica após a remoção de *stopwords* foi definida visando custo-efetividade. Ela melhora o desempenho (por não desperdiçar recursos para corrigir palavras que serão descartadas) sem prejudicar o processo, uma vez que a correção existente no momento é probabilística e não dependente de contexto (PESSOTTI; POLLETTINI, 2017).

Para a etapa de pré-processamento são necessários, portanto, para cada idioma de interesse: um *corpus* de treinamento visando a correção gramatical, além de uma lista de *stopwords* e de algoritmos de *stemming* específicos.

Além do pré-processamento tornaram-se necessárias alterações com o objetivo de permitir configurar as preferências de idiomas de cada usuário. Sendo assim, ao enviar uma mensagem, o sistema espera a resposta do usuário no mesmo idioma da mensagem enviada.

Para o desenvolvimento do projeto foram utilizadas linguagens de programação considerando o contexto de *framework* e ferramentas de auxílio pré-existentes, para isso foram utilizadas principalmente duas linguagens diferentes, sendo estas PHP e Python.

PHP

PHP (Hypertext Preprocessor) é uma linguagem de criação de scripts embutida em HTML no servidor. O interpretador PHP funciona de duas maneiras, uma para uso com sites e uma que você executa a partir da linha de comando, independente da Web (VALADE, [s.d.]). O CleverCare possui grande parte do seu desenvolvimento realizado em PHP.

Python

Python é uma linguagem de propósito geral de alto nível, multi paradigma (suporta o paradigma orientado a objetos, imperativo, funcional e procedural) e altamente modular (LABAKI, [s.d.]). Possui tipagem dinâmica, a qual reduz a quantidade de tempo de planejamento prévio, sendo um mecanismo importante para garantir flexibilidade e simplicidade das funções. Devido às suas características, ela é principalmente utilizada para processamento de textos e dados científicos.

5.3 Adaptações na estrutura de banco de dados e do módulo de processamento de linguagem natural

Com o objetivo de atender os requisitos e minimizar a necessidade pela funcionalidade de segunda opinião, diversas alterações foram realizadas no sistema para a inclusão de múltiplos idiomas e ligação de cada mensagem e cada documento a seus respectivos idiomas. Com essas alterações, é permitido que o usuário possua múltiplos idiomas, além de os idiomas poderem ser definidos por meio do ambiente/área, que o usuário está associado, ou por meio de diálogos que podem ser realizados em diversos idiomas em um mesmo ambiente. Ao enviar uma mensagem, o sistema espera a resposta do usuário no mesmo idioma da mensagem enviada.

Os ajustes realizados incluem modificações no banco de dados do módulo de processamento de linguagem natural do CleverCare, assim como modificações no banco de dados principal da plataforma para armazenar a informação de idioma e nível de proficiência do usuário. Dentre as adaptações realizadas, temos também a criação de uma tabela para representação dos diferentes idiomas e relacionamento da mesma com tabelas importantes para o funcionamento do sistema, como *Document*, *Collection*, *Query* e *Stopword*.

Adaptações relacionadas ao módulo de processamento textual incluem: alterações no cadastro e verificação de *stopwords*; ajustes nos métodos para verificação de termos já cadastrados no banco de dados; alterações no cadastro de informações essenciais ao funcionamento do CleverCare (classe que realiza configurações iniciais fundamentais para o funcionamento do sistema); alterações na classe *Term* (esta classe representa um termo presente em um documento/coleção), assim como alterações nas classes *Document* e *Collection* que representam os documentos e coleções; criação de coleções genéricas de exemplos positivos e negativos para os dois idiomas; adaptações em interface de gerenciamento de coleções e documentos do CleverCare.

Os algoritmos específicos para pré-processamento textual também devem permitir a flexibilidade de idiomas. Para o processamento de mensagens em diferentes idiomas é importante que os algoritmos sejam flexíveis para a adaptação e suporte aos respectivos novos idiomas. Para isso, a abordagem escolhida foi por meio da utilização da ferramenta NLTK (*Natural Language Toolkit*).

Sendo assim, adaptações foram realizadas no CleverCare para a utilização do NLTK nas etapas de pré-processamento. Com a utilização do NLTK as adaptações

foram realizadas no sentido de utilizar *corpora* de treinamento específicos de cada idioma, assim como lista de *stopwords* e algoritmo de Porter referentes às regras de cada idioma.

5.3.1 Stopwords

O algoritmo de remoção de *stopwords* tem seu funcionamento previsto para todos os idiomas por se tratar de um mecanismo simples de remoção de palavras contidas em uma lista, sendo assim este algoritmo não precisou de nenhuma alteração. Porém, a base de dados utilizada para cada idioma, a lista de termos a serem removidos, precisou passar por uma análise e ser adaptada, uma vez que continham palavras que para o nosso contexto não seria prudente eliminar, como exemplo as palavras “onde” e “porque”. Com esta medida permitimos que o sistema reconheça a diferença entre “Onde aplicar insulina?“ e “Porque aplicar insulina?”.
A Tabela 2 apresenta a relação de todas as palavras que foram removidas da lista de *stopwords* para cada idioma.

Tabela 2 - Relação de palavras removidas das listas de *stopwords* utilizadas neste projeto.

Português	Não, Sim, Quando, Quem, Qual, Como
Inglês	What, Which, Who, Whom, When’s, Where’s, Why’s, How’s, Below, Above, When, Where, Why, Over, Under, No, Not
Espanhol	No, Sí, Cuándo, Porque, Cuál, Cómo, Dónde

5.3.2 Redução ao radical

O algoritmo de *stemming* de Porter já tem seu mecanismo desenvolvido para diversos idiomas, tratando as suas especificidades e regras gramaticais, dentre esses idiomas já existem os algoritmos para português, inglês e espanhol. Sendo assim, não foi necessário realizar adaptações deste algoritmo, apenas utilizamos o idioma para determinar as regras específicas necessárias.

5.3.3 Correção ortográfica

Não foram necessárias adaptações no algoritmo utilizado pra a correção ortográfica, apenas houve a seleção do corretor de interesse dentre os disponíveis pela biblioteca pyEnchant. No caso, foi escolhido o Hunspell, que além de estar em desenvolvimento e atualização é utilizado por softwares e aplicativos reconhecidos, como o Chrome, LibreOffice, Firefox, entre outros (KHANNA, 2016).

5.4 Adaptações de interface do sistema

Na Figura 10, Figura 11, e Figura 12, são apresentados exemplos das modificações de interface gráfica realizadas. Os diálogos no CleverCare são organizados como grafos (ou árvores), em que cada nó representa uma possibilidade de interação com o paciente/usuário (PESSOTTI; POLLETTINI, 2017). As modificações foram projetadas para permitir duas abordagens diferentes:

1. Nós com múltiplos idiomas em diferentes linhas de diálogos: Nesta abordagem, em uma mesma árvore, cada nó pode ter textos (possíveis linhas de diálogo) nas diferentes línguas, onde cada linha de diálogo tem associado seu idioma;

2. Idiomas definidos por meio de árvores com idioma fixo: Nesta abordagem pré-definimos um idioma específico para cada árvore ou ambiente de trabalho.

Na Figura 10 pode ser visto um exemplo de criação de conteúdo para um nó de diálogo disponível nas línguas portuguesa, espanhola e inglesa, representando a abordagem definida por nós com múltiplos idiomas. Para facilitar a visualização, quando exibido o grafo completo de um plano de diálogo, os nós são exibidos apenas no idioma preferencial do usuário, no entanto, existe a opção de exibir o grafo em cada idioma no qual está disponível, conforme pode ser observado no destaque da Figura 11.

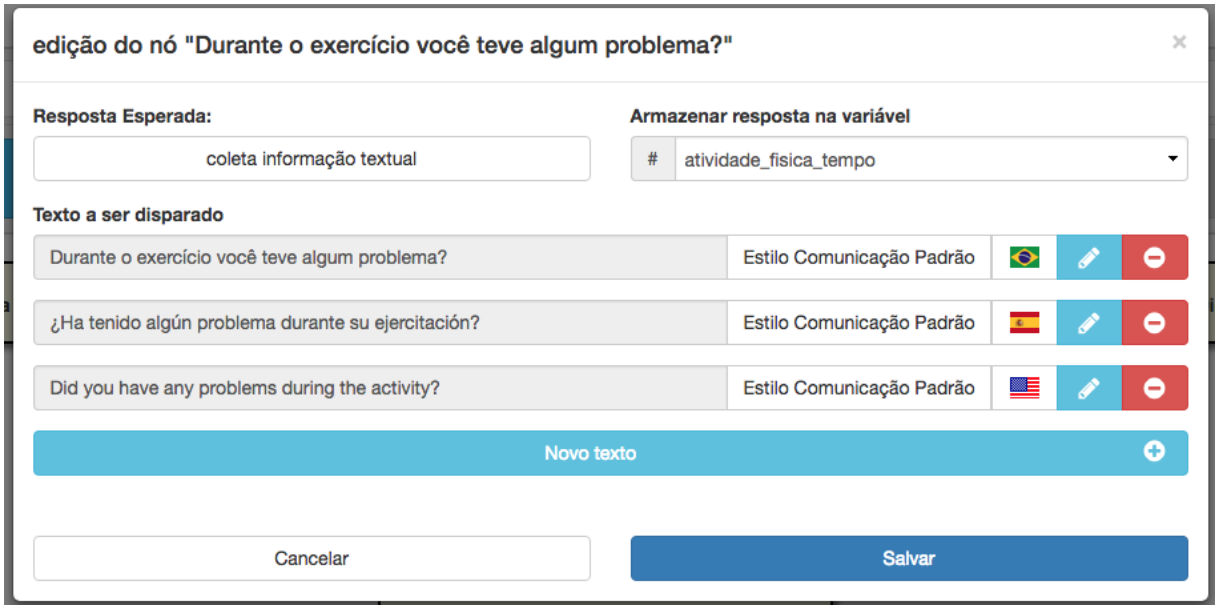

Figura 10 - Exemplo de criação de conteúdo para um nó de diálogo disponível nas línguas inglesa, espanhola e inglesa.

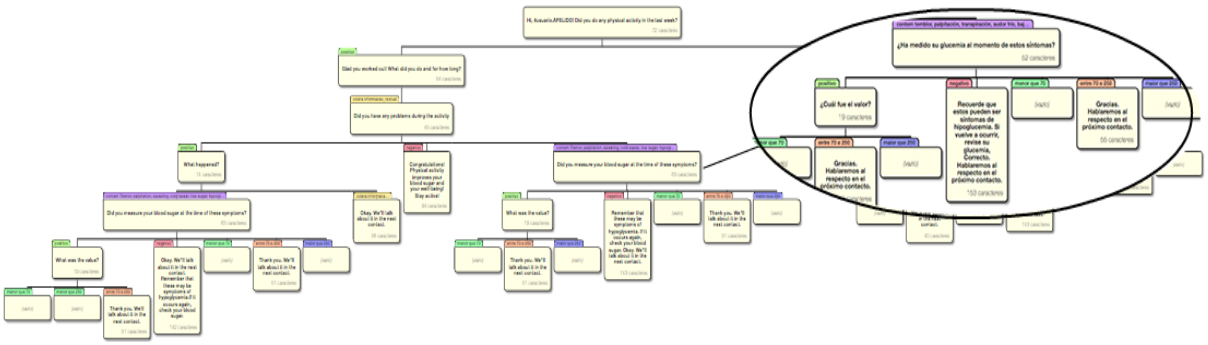

Figura 11 - Exemplo de criação de conteúdo para um diálogo (grafo) disponível simultaneamente nas línguas inglesa, espanhola e inglesa.

Na Figura 12 pode ser vista a interface de gerenciamento de usuários do sistema, incluindo os idiomas conhecidos pelo usuário e seu respectivo nível de fluência em cada idioma. Com estas informações é permitido, não apenas que o sistema reconheça que um usuário pode se comunicar por um determinado idioma, mas também que ele escolha sempre o idioma de maior fluência disponível para o usuário. Sendo assim, os idiomas do paciente são priorizados de acordo com seu nível se for fluente, avançados, intermediário ou básico, e quando um determinado roteiro vai ser agendado as possibilidades de idiomas daquele roteiro serão analisadas e a de maior habilidade do paciente será enviada ao mesmo.

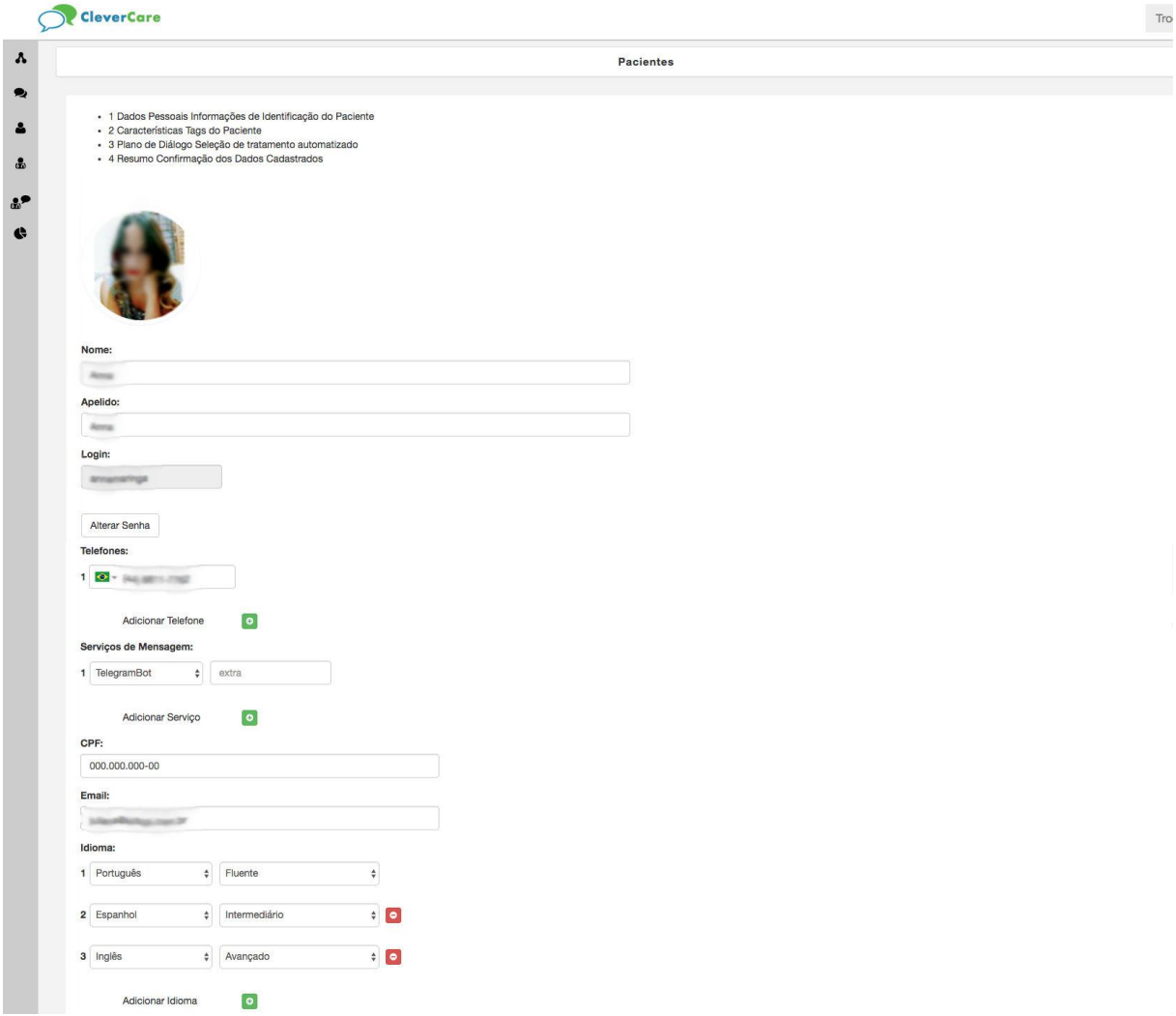

Figura 12 - Gerenciamento de usuários, idiomas conhecidos e respectivas fluências.

5.5 Desenvolvimento das bases de treinamento

Foram desenvolvidos diálogos e FAQs relacionados a prática de atividades físicas. Com relação aos FAQs foram criadas bases de dados pensadas para o desenvolvimento do experimento, como textos contendo explicações sobre o projeto, mensagens de agradecimento e finalização, além de mensagens explicando sobre a alteração de glicemia, referência ao contexto de doenças crônicas (diabetes).

5.6 Experimentação

Para os experimentos construímos a base do teste relacionada à atividade física, onde podemos comparar o fluxo do sistema para espanhol, inglês em um

contexto de monitoramento de atividade física já existente no conteúdo de doenças crônicas em português. Diálogos associados à pratica de atividade física são comuns e de grande importância para pacientes que possuem doenças crônicas, como o caso da diabetes, a qual foi utilizada como base e de onde este roteiro foi adaptado para os demais idiomas.

Uma das maiores dificuldades encontradas para o desenvolvimento do projeto foi, na etapa experimental, encontrar e conseguir a adesão de pessoas que fossem fluentes ou nativas nos idiomas de interesse para os experimentos, por isso os experimentos foram realizados com pessoas fluentes/nativas nos idiomas e não necessariamente pacientes com doenças crônicas, ainda que baseado em diálogos aplicados neste contexto.

Desta forma, utilizamos a estratégia de abordar um tema associado à saúde em um contexto geral, como a prática de atividade física, possibilitando que pessoas fluentes participassem do experimento mesmo sem possuir uma doença crônica.

Em ambos os idiomas o CleverCare foi capaz de classificar as mensagens corretamente, algo que teve por respaldo e suporte a modalidade de segunda opinião do sistema (Figura 13).

A funcionalidade de segunda opinião existe naturalmente no sistema para casos em que o sistema não consegue tomar sozinho uma decisão quanto à classificação de uma mensagem. Todavia, é ideal que esta funcionalidade seja utilizada o mínimo possível, uma vez que demanda trabalho humano. As etapas de pré-processamento textual auxiliam nessa redução de demanda por segunda opinião humana.

As comparações realizadas na etapa de validação foram focadas, com auxílio do testador automatizado, nas classificações do CleverCare antes de realizar as alterações relacionadas ao idioma e após esse processo, incluindo o processamento de linguagem natural, controle e armazenamento das informações relacionadas ao idioma específico.

Devido a uma base de dados previamente construída para a classificação de respostas esperadas no idioma português, além de pré-processamentos e coleções de documentos genéricos criados especificamente para este idioma durante todo o seu tempo de vida, não foi possível realizar experimentos sem esses recursos. Por este motivo não foi apresentada uma comparação retratando o desempenho neste

idioma em dois estágios do sistema, sendo estes antes e após as modificações aplicadas para o suporte a múltiplos idiomas.

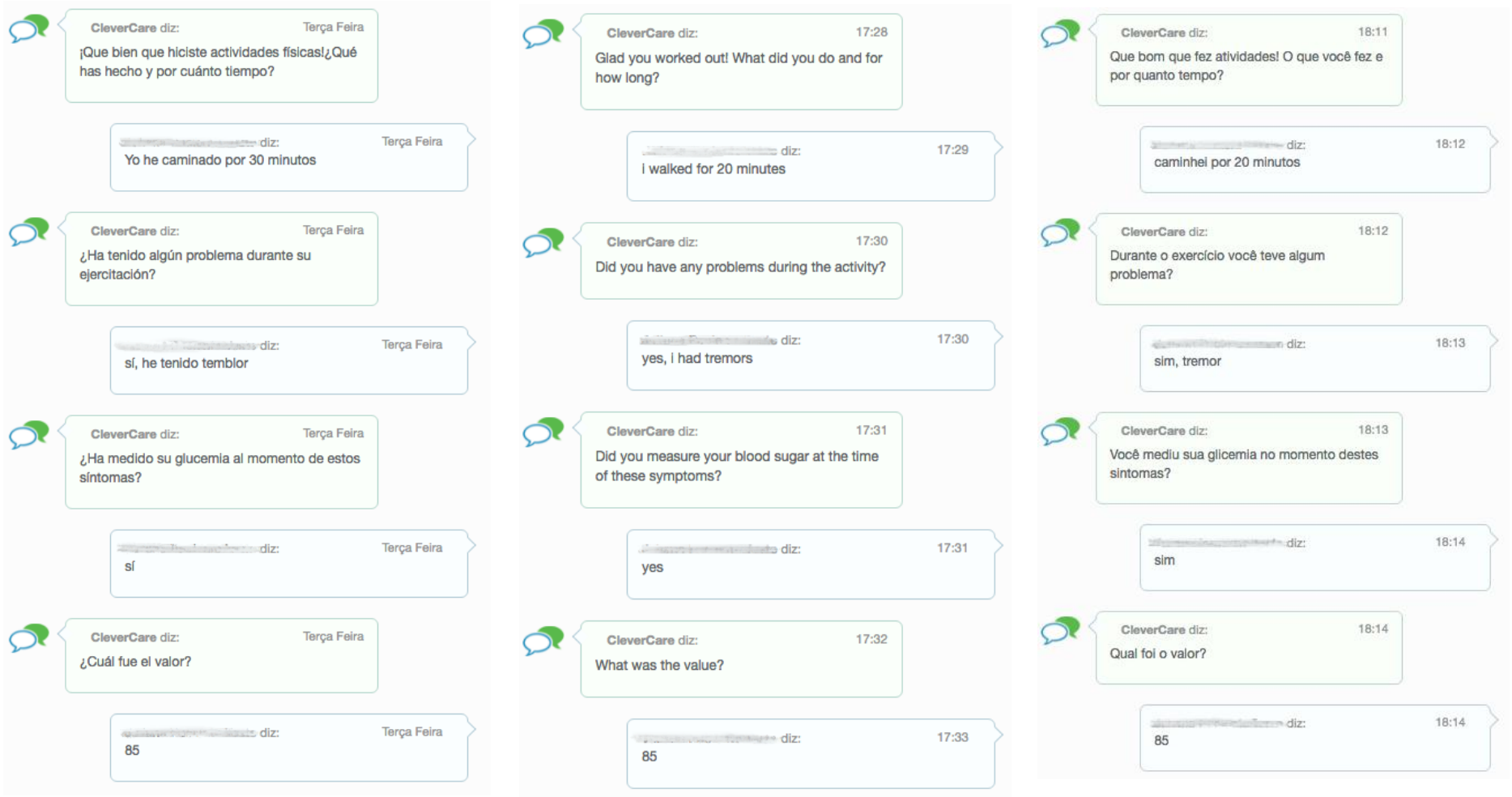

Figura 13 – Histórico parcial de diálogo relacionado à prática de atividade física realizado com o CleverCare em espanhol, inglês e português, respectivamente.

5.7 Desenvolvimento do testador

Com o objetivo de facilitar a experimentação das alterações realizadas no CleverCare e permitir comparações de efetividade entre as diferentes versões do sistema, a Kidopi desenvolveu um testador automatizado para o CleverCare.

O testador utiliza dados de contextos de interesse para simular esses contextos como se estivessem acontecendo novamente. Ao realizar a comparação entre o resultado obtido no passado e o resultado atual (com as configurações atuais vigentes no CleverCare), o testador permite analisar se uma determinada modificação na estrutura do CleverCare possibilitou uma melhoria nos resultados, se manteve os resultados anteriores, ou, até mesmo, se eventualmente ocasionou uma piora nos resultados (o que demandaria uma reversão ao estado anterior do sistema).

O testador permite analisar resultados quantitativos (relativos a precisão, revocação e acurácia), além de permitir a visualização gráfica de contextos finalizados corretamente ou não, ou seja, indica se o diálogo foi completamente percorrido ou mensagens deixaram de serem respondidas. Com isso é possível realizar por meio desta ferramenta a comparação entre o resultado do contexto original e do contexto simulado pelo testador. Exemplos dessa visualização gráfica dos resultados podem ser vistos nas Figura 14 e Figura 15.

Para os nossos testes foram utilizados os contextos nos idiomas de interesse em duas etapas, sendo a primeira considerando o sistema no período anterior às modificações referentes ao idioma e a segunda com o sistema já alterado para o suporte e processamento desses idiomas.

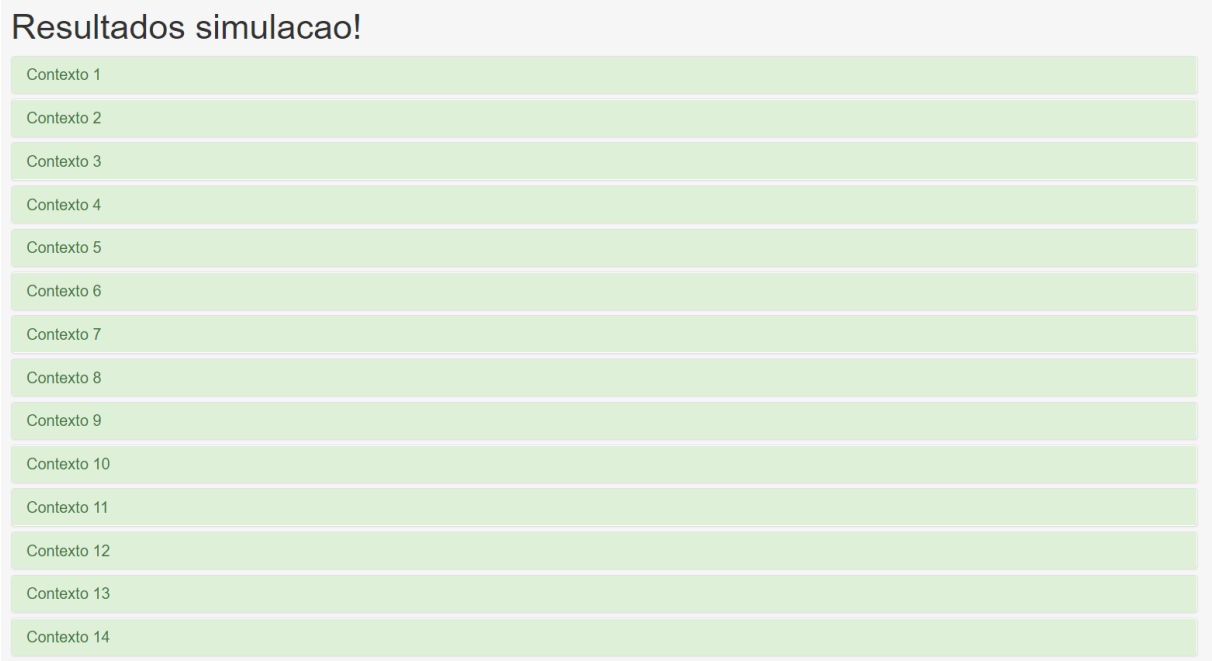

Figura 14 - Resumo de resultados para contextos simulados no testador automatizado.

Resultados

MsgHistorica	MsgSimulada
¡Hola, [redacted] ¿Ha realizado alguna actividad física durante la semana pasada?	¡Hola, [redacted] ¿Ha realizado alguna actividad física durante la semana pasada?
Si.	Si.
Si. -> positivo? Mensagem "Si." de "[redacted]" como resposta para "¡Hola, Vero! ¿Ha realizado alguna actividad física durante la semana pasada?" é positiva. As outras respostas possíveis são: negativo. Estou certo? (A: confirmar, B: negar)	Si. -> positivo? Mensagem "Si." de "[redacted]" como resposta para "¡Hola, Vero! ¿Ha realizado alguna actividad física durante la semana pasada?" é positiva. As outras respostas possíveis são: negativo. Estou certo? (A: confirmar, B: negar)
A	A
¡Que bien que hiciste actividades físicas!¿Qué has hecho y por cuánto tiempo?	¡Que bien que hiciste actividades físicas!¿Qué has hecho y por cuánto tiempo?
Gimnasia con aparatos. 1 hora y media.	Gimnasia con aparatos. 1 hora y media.
Gimnasia con aparatos. 1 hora y media. -> coleta_informacao_textual? Mensagem "Gimnasia con aparatos. 1 hora y media." de "[redacted]" como resposta para "¡Que bien que hiciste actividades físicas!¿Qué has hecho y por cuánto tiempo?" é coleta informação textual. Não há outras respostas possíveis. Estou certo? (A: confirmar, B: negar)	Gimnasia con aparatos. 1 hora y media. -> coleta_informacao_textual? Mensagem "Gimnasia con aparatos. 1 hora y media." de "[redacted]" como resposta para "¡Que bien que hiciste actividades físicas!¿Qué has hecho y por cuánto tiempo?" é coleta informação textual. Não há outras respostas possíveis. Estou certo? (A: confirmar, B: negar)
a	a
¿Ha tenido algún problema durante su ejercitación?	¿Ha tenido algún problema durante su ejercitación?
No.	No.
No. -> negativo? Mensagem "No." de "[redacted]" como resposta para "¿Ha tenido algún problema durante su ejercitación?" é negativa. As outras respostas possíveis são: positivo, temblor, palpitación, transpiración, sudor frío, baja de azúcar, hipoglucemia, alteración de glucemia. Estou certo? (A: confirmar, B: negar)	No. -> negativo? Mensagem "No." de "[redacted]" como resposta para "¿Ha tenido algún problema durante su ejercitación?" é negativa. As outras respostas possíveis são: positivo, temblor, palpitación, transpiración, sudor frío, baja de azúcar, hipoglucemia, alteración de glucemia. Estou certo? (A: confirmar, B: negar)
a	a
¡Enhorabuena! ¡La actividad física mejora su glucemia y bienestar!Manténgase activo!	¡Enhorabuena! ¡La actividad física mejora su glucemia y bienestar!Manténgase activo!

Figura 15 - Comparação visual entre contexto histórico e contexto simulado pelo testador automatizado

5.8 Validação

Para validação das adaptações realizadas ao CleverCare direcionadas à internacionalização, foi realizado um experimento com usuários respondendo a perguntas de um mesmo contexto nas diferentes línguas de interesse: inglês e espanhol. A análise de desempenho foi feita considerando o sistema em seu estado original e após as modificações realizadas pelo projeto de internacionalização.

Os três principais indicadores de desempenho chave em que nos concentramos para avaliação foram:

1. Precisão (Precision) = TP / (TP + FP)
2. Revocação (Recall) = TP / (TP + FN)
3. Acurácia (Accuracy) = (TP + TN) / (TP + TN + FP + FN)

Esses indicadores são compostos por:

TP (True Positive / verdadeiro positivo) = número de mensagens automaticamente classificadas pelo software e confirmadas como corretas por auditoria humana

FP (False Positive / falso positivo) = Número de mensagens automaticamente classificadas pelo software, mas não confirmadas como corretas pela auditoria humana

FN (False Negative / falso negativo) = número de mensagens não classificadas automaticamente, mas confirmadas por segunda opinião humana como parte de nossa base de dados, sendo estas referentes a FAQ ou diálogos existentes que não foram reconhecidos.

TN (True Negative / verdadeiro negativo) = número de mensagens não automaticamente classificadas pelo software e confirmadas por segunda opinião humana como não fazendo parte do nosso banco de dados, ou seja, mensagem a ser arquivada ou necessitando de resposta humana específica por não existir nada correspondente a ela previamente cadastrado no sistema.

Na Tabela 3 é apresentada a Matriz de Confusão utilizada para avaliação dos resultados. Os resultados foram avaliados conforme o resultado do sistema (Condição Predita) com relação ao fato de o sistema ter encontrado ou não uma resposta satisfatória em comparação com os resultados analisados por um avaliador humano especialista (PESSOTTI; POLLETTINI, 2017).

Tabela 3 - Matriz de confusão. Fonte: (PESSOTTI; POLLETTINI, 2017).

Matriz de Confusão		**Condição Predita**	
		Predição positiva	**Predição negativa**
Condição verdadeira (avaliador humano especialista)	**Condição positiva**	Auditoria correta - (TP)	Segunda opinião incorreta - (FN)
	Condição negativa	Auditoria incorreta - (FP)	Segunda opinião correta - (TN)

A Figura 16 representa os indicadores dos resultados obtidos dos experimentos realizados, onde para cada indicador temos o resultado de aplicação de ambos os idiomas, inglês e espanhol, aplicados antes e após as modificações no sistema.

Os valores obtidos demonstram que a precisão foi preservada, enquanto a revocação e acurácia tiveram uma melhoria de 13% em ambos os indicadores após as adaptações aplicadas ao sistema. Essa melhoria indica uma redução no trabalho dependente de humanos na tarefa de segunda opinião no sistema.

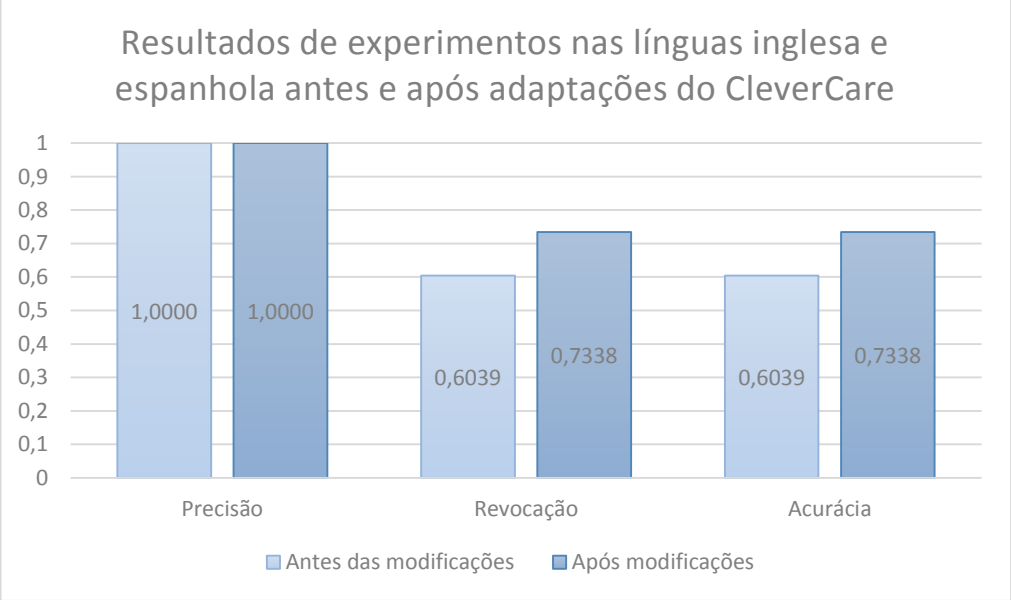

Figura 16 – Gráfico de resultados de experimentos nas línguas inglesa e espanhola antes e após adaptações em algoritmos, artefatos linguísticos e banco de dados.

5.9 Desenvolvimento de parcerias para aplicação do CleverCare em mercados latino-americanos

Em paralelo ao desenvolvimento das atividades, a empresa manteve relações com universidades, hospitais e grupos de pesquisa colombianos. Sendo assim, acordos de parceria e convênio estão sendo firmados com as universidades UDES (Universidad de Santander) e UNAB (Universidad Autónoma de Bucaramanga), ambas localizadas em Bucaramanga, na Colômbia, cidade-irmã de Ribeirão Preto[2].

Na semana de 9 a 14 de outubro, o diretor executivo da Kidopi, Dr. Mário Sergio Adolfi Junior, ministrou uma palestra no II Simpósio Internacional sobre Tecnologias da UDES. Durante este período, participou de diversas reuniões com grupos de pesquisa locais para prospectar oportunidades de parceria para a aplicação do CleverCare e firmar possíveis convênios de cooperação. Os acordos agora estão sendo oficializados e o CleverCare está sendo analisado para a aplicação no acompanhamento de pacientes crônicos de baixo risco ligados à UNAB (Universidad Autónoma de Bucaramanga).

Além disso, em junho de 2018 o diretor executivo da Kidopi, Dr. Mário Sergio Adolfi Junior, participou do evento EmTech digital do MIT junto a IPADE realizado na Cidade do México (https://www.emtechmexico.com/), onde foi convidado a apresentar o CleverCare como uma tecnologia inovadora. Esse contato foi importante para prospectar possíveis cenários de atuação.

5.10 Considerações finais e perspectivas futuras

Este trabalho apresenta a abordagem de técnicas e conceitos de aprendizado de máquinas e processamento de linguagem natural para diferentes idiomas. Inicialmente foi realizada uma revisão sistemática sobre os conceitos de processamento de linguagem natural e doenças crônicas, onde foram avaliados 20 estudos primários selecionados através de critérios de inclusão e exclusão dentre 54 artigos resultantes da busca do termo de interesse nas bases propostas.

Como um dos maiores destaques, temos a quantidade de artigos publicados anualmente, onde podemos observar que além de apresentar um aumento ao passar dos anos temos a primeira publicação realizada em 1993, ou seja, é um tema de

[2] De maneira diplomática, municípios estabelecem acordos de irmandade (acordos de cooperação feitos entre regiões com o objetivo de promover laços culturais e de amizade) em busca de convênios e parcerias.

abordagem com pouco mais de 20 anos. É importante destacar que a aplicação de processamento de linguagem natural associado a doenças crônicas tem maior incidência com relação a extração automática de conhecimento em estudos empíricos da área, bem como em prontuários eletrônicos. Sendo assim, a aplicação deste processamento em um diálogo a tempo real com o paciente é algo ainda mais escassa.

Para os algoritmos e abordagens escolhidas o foco foi a portabilidade da solução entre diferentes idiomas, ou seja, abordagens que podem facilmente ser adaptadas e prover uma expansão do sistema com relação a este tema.

Em ambos os idiomas o CleverCare obteve aproximadamente o mesmo desempenho e foi capaz de classificar as mensagens corretamente, demonstrando o seu potencial e escalabilidade em diferentes idiomas. As adaptações realizadas aos artefatos linguísticos, algoritmos e banco de dados apresentou indicadores que demostraram uma melhora nos índices de precisão e revocação, ou seja, auxiliaram consideravelmente na redução de demanda por segunda opinião humana. Vale notar também que a precisão do sistema, que já era alta devido abordagem mais conservadora do CleverCare na classificação de mensagens não foi prejudicada pelas adaptações realizadas.

Em paralelo ao desenvolvimento das atividades para a adaptação do CleverCare para funcionamento nas línguas inglesa e espanhola, a empresa manteve relações com universidades, hospitais e grupos de pesquisa colombianos.

Com o resultado positivo apresentado pelo CleverCare torna-se possível explorar o mercado internacional. Além disso, as adaptações foram feitas de forma mais genérica possível, visando a possível expansão para outros idiomas além dos atuais e, com isso, inserir o CleverCare em países que possuem diferentes idiomas.

6. Conclusões

6.1 Objetivo Geral

1. Este trabalho permitiu investigar as adaptações necessárias para o funcionamento do CleverCare em inglês e espanhol. Essa investigação foi realizada e as adaptações necessárias envolveram alterações em algoritmos, artefatos linguísticos e banco de dados. Os algoritmos e abordagens escolhidas tiveram como foco abordagens que podem facilmente ser adaptadas a outros idiomas, permitindo a expansão do sistema a cenários com idiomas distintos aos atuais.

6.2 Objetivos Específicos

1. Foi possível avaliar as adaptações e necessidades específicas para cada idioma, avaliada na análise de requisitos para o desenvolvimento do projeto, as quais envolveram a análise das estratégias de mineração de texto para documentos da área da saúde aplicadas pelo CleverCare. Com isso, foram analisadas adaptações em ferramentas de processamento de linguagem natural e no framework do CleverCare que deveriam ser realizadas para o seu adequado funcionamento nos idiomas de interesse.
2. Foi possível realizar as adaptações necessárias às ferramentas de interesse para o adequado funcionamento do CleverCare nestes idiomas. Para isso foram aplicadas modificações no processamento de linguagem natural do CleverCare, utilizando o NLTK e pyEnchant como ferramentas de apoio às etapas de pré-processamento textual. Além disso, foram realizadas adaptações ao CleverCare de forma geral, visando permitir a comunicação com o módulo de processamento de linguagem natural e controle das informações de idioma pelo sistema, incluindo a manipulação dessa informação por meio de sua interface.

3. Foi possível avaliar os resultados obtidos por meio dos testes realizados utilizando os indicadores de acurácia, precisão e revocação. Esses resultados demonstram que obtivemos uma melhoria de aproximadamente 13% nos indicadores de precisão e revocação após as adaptações aplicadas neste projeto.

7. REFERÊNCIAS BIBLIOGRÁFICAS

AVANSI, A. F. ; RIBEIRO, RS ; DAHER, G ; BARBOSA RGV ; MACEDO AISM ; MATTOS GD ; JUNIOR, MARIO SERGIO ADOLFI ; ALVES DO ; POLLETTINI, J. T. **IMPACTO DO MODELO INTEGRADO DE ATENÇÃO PRIMÁRIA E ENFERMEIRA NAVEGADORA APOIADO POR TECNOLOGIAS DE INFORMAÇÃO NO CONTROLE GLICÊMICO**. 12° Congresso Paulista de Diabetes e Metabolismo. **Anais**...São Paulo: 2016

AZEVEDO, F. S. **Metodologia de Mineração de Textos**. [s.l.] Pontífica Universidade católica, 2008.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. [s.l.] Addison-Wesley, 1999.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python**. [s.l: s.n.].

BOSCH, A.; MUÑOZ, X.; FREIXENET, J. Segmentation and description of natural outdoor scenes. **Image and Vision Computing**, v. 25, n. 5, p. 727–740, 2007.

BOULIS, C.; OSTENDORF, M. Text classification by augmenting the bag-of-words representation with redundancy compensated bigrams. **Workshop on Feature Selection in Data Mining**, p. 9–16, 2005.

BURSTEIN, A. **Nasty Fast N-Grams (Part 1): Character-Level Unigrams**. Disponível em: <http://www.sqlservercentral.com/articles/Tally+Table/142316/>. Acesso em: 21 maio. 2018.

CARVALHO, W. S. **Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina**. São Paulo: Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo, fev. 2012.

COSTA, B. I. R. **Coeficiente de revocação (recall) e precisão (Precision) do sistema de recuperação de informação da biblioteca do ICEX da UFMG através da amostra do acervo de teses e dissertações**. [s.l.] Universidade Federal de Minas Gerais, 2008.

CRISTINA, E.; BARION, N.; LAGO, D. **Mineração de textos**Rio de JaneiroPUC-Rio, , 2008.

DEZEMBRO, D. G. **Desenvolvimento e Análise de Métodos de Mineração de Textos para Aprimoramento do Sistema CleverCare Desenvolvimento e Análise de Métodos de Mineração de Textos para Aprimoramento do Sistema CleverCare.** [s.l.] Universidade de São Paulo, 2015.

EDBERTO. OntoSmart: um modelo de recuperação de informação baseado em ontologia. v. 22, n. 2, p. 170–187, 2017.

FAYAD, M.; SCHMIDT, D. C. Frameworks de aplicações orientado a objetos. v. 40, 1997.

FERNEDA, E. **Recuperação de Informação: análise sobre a contibuição da ciência da computação para a ciência da informação**. [s.l.] Universidade de São Paulo, 2003.

HEMALATHA, I.; VARMA, D. G. P. S.; A.GOVARDHAN, D. Preprocessing The Informal Data for Efficient Sentiment Analysis. **International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)**, v. 1, n. 2, p. 58, 2012.

KAO, A. et al. **Natural Language Processing and Text Mining**. [s.l.] Springer, 2007.

KELLY, R. **pyenchant - Python bindings for the Enchant spellchecking system**, 2017. Disponível em: <https://pypi.org/project/pyenchant/1.6.11/#description>

KHANNA, S. **Aspell and Hunspell: A Tale of Two Spell Checkers**. Disponível em: <https://penguindreams.org/blog/aspell-and-hunspell-a-tale-of-two-spell-checkers/>. Acesso em: 13 maio. 2018.

LABAKI, J. **Introdução a Python**, [s.d.].

LACHOWICZ, D.; THOMAS, R. **Enchant**. Disponível em: <https://abiword.github.io/enchant/>. Acesso em: 8 maio. 2018.

LEAL, R. DOS S. **Métricas Comuns em Machine Learning: como analisar a qualidade de chat bots inteligentes — métricas (3 de 4)**. Disponível em: <https://medium.com/as-máquinas-que-pensam/métricas-comuns-em-machine-learning-como-analisar-a-qualidade-de-chat-bots-inteligentes-métricas-1ba580d7cc96>. Acesso em: 30 abr. 2018.

LOPER, E.; BIRD, S. NLTK: The Natural Language Toolkit. 2002.

MARTINS, C. A.; MONARD, M. C.; MATSUBARA, E. T. **PreTexT : uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-wordsInstituto de Ciências Matemáticas e de Computação**. São Carlos: [s.n.].

MARTINS, D. S. Uma abordagem para recuperação de informações sensível ao contexto usando retroalimentação implícita de relevância. p. 108, 2009.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de Textos. p. 29, 2007.

MULLER, D. N. **Processamento de linguagem natural**. Rio Grande do Sul: [s.n.].

Natural Language Toolkit — NLTK 3.2.5 documentation. Disponível em: <https://www.nltk.org/>. Acesso em: 9 mar. 2018.

NÉMETH, L. **Hunspell**. Disponível em: <http://hunspell.github.io/>.

NÓBREGA, F. A. A. Word Sense Disambiguation for portuguese through multilingual mono and multi-document. p. 126, 2013.

OLIVEIRA, C.; FREITAS, M. Classes de palavras e etiquetagem na Lingüística Computacional. **Calidoscópio**, v. 4, n. 3, p. 179–188, 2006.

OOMS, J. **Package “ hunspell ”**. [s.l.] CRAN, 2017. Disponível em: <https://hunspell.github.io>.

PESSOTTI, H. C. **Aprimoramento do sistema CleverCare por meio do desenvolvimento de novas funcionalidades , adaptações para internacionalização e desenvolvimento de novos métodos de acesso ao sistema**. Ribeirão Preto, São Paulo: [s.n.].

PESSOTTI, H. C.; POLLETTINI, J. T. **Aprimoramento do sistema CleverCare por meio do desenvolvimento de novas funcionalidades , adaptações para internacionalização e desenvolvimento de novos métodos de acesso ao sistema**. Ribeirão Preto, São Paulo: [s.n.].

POLLETTINI, J. **Definição automática de medidas que identificam pessoas requerendo diferentes graus de vigilância para atendimento em atenção básica à saúde: uma abordagem utilizando Relevance Feedback e a Classificação Internacional de Doenças**. [s.l.] Universidade de São Paulo, 2008.

PORTER, M. F. Effective Pre-Processing Activities in Text Mining using Improved Porter’s Stemming Algorithm. **International Journal of Advanced Research in Computer and Communication Engineering**, v. 2, n. 12, p. 4536–4538, 1980a.

PORTER, M. F. An algorithm for suffix stripping. **Program**, v. 14, n. 3, p. 130–137, 1980b.

RUSSELL, S.; NORVIG, P. A modern, agent-oriented approach to introductory artificial intelligence. **ACM SIGART Bulletin**, v. 6, n. 2, p. 24–26, abr. 1995.

SANTOS, R. E. S. et al. Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático. **Revista de Sistemas e Computação**, v. 4, p. 116–125, 2014.

SOERGEL, D. Multilingual thesauri in cross-language text and speech retrieval. **AAAI Spring Symposium on Cross-Language Text and Speech Retrieval**, n. August, p. 1–16, 1997.

SUNDBY, D. **Spelling correction using N-grams**. Sweden: [s.n.].

TAN, A.-H. Text Mining: The state of the art and the challenges. **Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases**, v. 8, p. 65–70, 1999.

VALADE, J. PHP 5 for Dummies. [s.d.].

VIJAYARANI, S.; ILAMATHI, J.; NITHYA, M. Preprocessing Techniques for Text Mining - An Overview. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2015.

VIJAYARANI, S.; JANANI, R. Text Mining: open Source Tokenization Tools – An Analysis. **Advanced Computational Intelligence: An International Journal (ACII)**, v. 3, n. 1, p. 37–47, 2016.

ZHANG, Y. Contextualizing consumer health information searching: an analysis of questions in a social Q&A community. **Proceedings of the 1st ACM International Health**, p. 210–219, 2010.

ZWEIGENBAUM, P.; GRABAR, N. A contribution of medical terminology to medical language processing resources: Experiments in morphological knowledge acquisition from thesauri. **Service d'Informatique Médicale: Assistance Publique – Paris Hospitals & Département de Biomathématiques**, p. 16–19, 1999.

APÊNDICE A - Termo de Consentimento Livre e Esclarecido

Nome da Pesquisa: Exploração da mineração de texto em documentos da saúde em diferentes idiomas para acompanhamento médico de pacientes com doenças crônicas
Pesquisadores Responsáveis: Ana Katariny de Souza Cacheta, Juliana Tarossi Pollettini, Prof. Dr. Antônio Pazin Filho

Gostaríamos de convidá-lo(a) a participar de um estudo que irá coletar informações com relação à sua saúde.
Neste estudo, queremos testar se um programa eletrônico que envia mensagens no seu telefone celular, O CleverCare (www.clevercare.com), é capaz de se comunicar com falantes de idiomas diferentes do português. Este sistema vai enviar mensagens de texto em seu aparelho celular com perguntas com relação à sua saúde, atividades físicas desenvolvidas, entre outras. Você também poderá escrever para o sistema, fazendo perguntas em casos de dúvidas.

Quais são os objetivos deste estudo?
Nosso objetivo será o de avaliar como o uso do sistema CleverCare em inglês e espanhol.

Quais são suas responsabilidades?
Caso concorde em participar desta pesquisa, o Sr.(a) deverá ter um aparelho de celular que possibilite o uso da tecnologia em teste, ou seja, acesso à internet com o aplicativo Telegram (www.telegram.org) ou WhatsApp (www.whatsapp.com/).
O Sr.(a) vai responder a questionários de qualidade de vida e os pesquisadores irão registrar as informações obtidas por meio do sistema.

Suas informações são confidenciais?
Toda a informação será sigilosa e codificada com uma sequência de caracteres que só os investigadores terão acesso. Em nenhum momento seu nome ou qualquer informação sobre a sua saúde será fornecida para qualquer pessoa que não seja um dos investigadores. A informação será utilizada somente para fins de pesquisa. Auditores, monitores, Comitê de Ética em Pesquisa e autoridades regulatórias eventualmente poderão ter acesso direto aos registros médicos originais, com o objetivo de checagem dos dados da pesquisa, respeitando a sua privacidade.

Quais são os riscos e benefícios deste estudo?
No caso deste estudo, o(a) Sr. (a) será submetido(a) a perguntas e diálogos que avaliará uma série de informações, como informações relacionadas à sua saúde e qualidade de vida. A sua participação neste estudo não determina nenhum risco adicional ou dano à sua saúde e é isenta de remuneração ou ônus.
Garantimos que todas suas informações serão mantidas em confidencialidade e sigilo. O(a) senhor(a) será tratado anonimamente, através de uma sequência de caracteres, durante toda a condução do estudo e os resultados desta pesquisa serão revelados também de forma anônima. Os resultados deste trabalho poderão ajudar pacientes de doenças cônicas a melhorar o controle da qualidade de vida e a conhecer mais sobre a doença.

Serei compensado por danos relacionados ao estudo?
Não. A proposta do estudo é coletar dados textuais de perguntas e respostas (por exemplo, sobre a sua saúde), com o intuito de manutenção de uma conversa fluida entre o sistema e você. No entanto, caso ocorra algum dano ao senhor(a), decorrente de sua participação, o senhor (a) será devidamente indenizado, conforme determina a Resolução CNS nº 466 de 2012.

Serei reembolsado por algum custo decorrente da minha participação?
A participação é isenta de remuneração ou ônus, pois os dados serão coletados por meio de conversas de celular em um aplicativo gratuito.

Contatos dos pesquisadores responsáveis pelo estudo
Pesquisadora: Ana Katariny de Souza Cacheta
E-mail: ana@kidopi.com.br
Telefone: (16) 3315-9936

Orientador: Prof. Dr. Antônio Pazin Filho
E-mail: apazin@fmrp.usp.br
Telefone: (16) 3602-2180

Coordenadora: Juliana Tarossi Pollettini
E-mail: juliana@kidopi.com.br
Telefone: (16) 3315-9936

Informações sobre o Comitê de Ética em Pesquisa e contato
O CEP (Comitê de Ética em Pesquisa) é um órgão que tem por objetivo proteger o bem-estar dos indivíduos pesquisados. É responsável pela avaliação e acompanhamento dos aspectos éticos de todas as pesquisas envolvendo seres humanos, visando assegurar a dignidade, os direitos, a segurança e o bem-estar do sujeito da pesquisa.
Se você tiver dúvidas e/ou perguntas sobre direitos como participante deste estudo, poderá entrar em contato com o CEP HCFMRP-USP através do telefone (16) 3602-2228 ou pelo email cep@hcrp.usp.br de segunda a sexta-feira das 08:00 às 17:00.

Eu, _____ fui informado(a) dos objetivos e da justificativa da pesquisa de forma clara e detalhada. Recebi informações sobre os procedimentos desta pesquisa. Também me foi garantido pelo pesquisador, o sigilo e a privacidade dos dados obtidos na pesquisa. Li e compreendi os objetivos do estudo, todos os procedimentos que serão realizados, e em caso de qualquer dúvida, poderei entrar em contato com a equipe do estudo.

_____
Nome do Participante/Representante Legal

_____ _____/_____/_____
Assinatura do Participante/Representante legal Data

_____ _____/_____/_____
Ana Katariny de Souza Cacheta Data

APÊNDICE B - Revisão bibliográfica do trabalho a ser submetida para publicação

REVISÃO BIBLIOGRÁFICA SISTEMÁTICA: APLICAÇÃO DE PROCESSAMENTO DE LINGUAGEM NATURAL NA ÁREA DA SAÚDE PARA ACOMPANHAMENTO DE PACIENTES PORTADORES DE DOENÇAS CRÔNICAS

Ana Katariny de Souza Cacheta[1,2] (anakatariny@gmail.com), Juliana Tarossi Pollettini[2] (juliana@kidopi.com.br), Hugo Cesar Pessotti[2] (hugo@kidopi.com.br), Mario Sergio Adolfi Junior[2] (mario@kidopi.com.br), Rafael dos Santos Elias[2] (rafa.elias@gmail.com), Antonio Pazin Filho[1] (apazin@fmrp.usp.br)

1- Universidade de São Paulo – Ribeirão Preto, SP - Brasil

2- Kidopi Soluções em Informática Ltda – Ribeirão Preto, SP - Brasil

Resumo

Objetivo. Este estudo tem como objetivo identificar e avaliar a aplicação de técnicas de processamento de linguagem natural para o tratamento de doenças crônicas.

Métodos. Realizou-se uma revisão sistemática da literatura mediante busca nas bases de dados SciELO, LILACS, PUBMed e Scopus utilizando os termos “Processamento de linguagem natural e doenças crônicas (em português, inglês e espanhol).

Resultados. Foram encontrados 54 artigos de todas as bases de dados, dos quais 20 foram incluídos na pesquisa após passarem pelos critérios de inclusão definidos pela revisão sistemática. As técnicas mais utilizadas nos artigos foram etiquetagem, segmentação de sentenças e tokenização.

Conclusões. Os estudos analisados apontam que a aplicação de técnicas de processamento de linguagem natural para documentos da saúde associados a doenças crônicas é recente, e suas técnicas são pouco abordadas nos artigos analisados, dificultando a análise e comparação entre metodologias. As metodologias mais frequentes estão associadas à análise sintática no processamento de linguagem natural.

Descritores

Português: Processamento de linguagem natural, doenças crônicas.

Inglês: Natural language processing, chronic disease.

Introdução

A necessidade de redução de leitos hospitalares remete a um fenômeno conhecido como desospitalização (NETO; MALIK, 2007), o qual consiste em dar alta ao paciente e fornecer suporte ao tratamento utilizando modelos de cuidado alternativos àqueles prestados em ambiente hospitalar. Dentre os modelos alternativos temos a assistência domiciliar, para realizar cuidados em domicílio quando estes já não são de alta complexidade, mas sim de extrema importância.

No entanto, o processo de desospitalização deve ser realizado com cautela, para evitar rehospitalizações. Utilizando o método de alta-assistida, por exemplo, o hospital Dr. João Machado foi capaz de reduzir o índice de rehospitalizações (BEZERRA; DIMENSTEIN, 2011).

Reduzir taxas de rehospitalização atraiu a atenção dos responsáveis como a maneira de melhorar a qualidade do cuidado e de reduzir custos (JENCKS; WILLIAMS; COLEMAN, 2009). A implantação eficaz de um plano terapêutico bem sucedido de cuidados para pacientes é dependente de participação do paciente e da conformidade com o regime de tratamento (GRADY et al., 2000).

Além disso, em ensaios clínicos e farmacêuticos, um grande problema é o acompanhamento e a necessidade de garantir a aderência dos sujeitos da pesquisa aos protocolos clínicos. A execução incorreta dos protocolos por parte dos sujeitos pode acarretar em resultados inconsistentes, levando a conclusões errôneas ou invalidando os estudos (MARTIN et al., 2005). O contato em tempo real, que pode ser viabilizado através de mensagens instantâneas, garante que sujeitos de pesquisa sigam corretamente os protocolos de pesquisa e tenham suas dúvidas sanadas de forma rápida e pontual, propiciando uma maior aderência aos ensaios.

O CleverCare é um framework para o controle, gestão e orientação de pacientes que necessitam de acompanhamento médico contínuo. O sistema utiliza mensagens de celular para realizar a comunicação com o paciente de forma personalizada.

Abordagens de mineração de textos permitem que o sistema compreenda a mensagem, responda-a e execute ações de forma automática e personalizada (DEZEMBRO, 2015).

Ao utilizar processamento de linguagem natural em documentos da área da saúde, buscamos viabilizar e potencializar projetos que necessitem soluções informatizadas para a gestão de seus pacientes com o objetivo de, por exemplo, reduzir o número de doenças e mortes evitáveis, complicações, sequelas e internações desnecessárias, bem como garantir uma correta aderência a planos de tratamento e protocolos de pesquisa, uma vez que a aderência do paciente ao seu plano de tratamento é essencial para o desfecho positivo de sua enfermidade.

A Revisão Bibliográfica Sistemática (RBS) é um instrumento para mapear trabalhos

publicados no tema de pesquisa específico para que o pesquisador seja capaz de elaborar uma síntese do conhecimento existente sobre o assunto (TRAVASSOS et al., 2007).

Uma forma de obter maior rigor e melhores níveis de confiabilidade em uma revisão bibliográfica é adotar uma abordagem sistemática. Isso significa, definir uma estratégia e um método sistemático para realizar buscas e analisar resultados, que permita a repetição por meio de ciclos contínuos até que os objetivos da revisão sejam alcançados (CONFORTO; AMARAL; SILVA, 2011).

Sendo assim, o objetivo deste estudo foi realizar uma revisão sistemática da literatura sobre a utilização de processamento de linguagem natural em relação a doenças crônicas.

Justificativa

A sobrecarga de informação é um fenômeno contemporâneo observado a partir do crescimento exponencial na disposição de informações, registrada principalmente após a popularização e a expansão da Internet (SANTOS et al., 2014)

Com o avanço tecnológico, essas informações passaram a ser digitalizadas, possibilitando a extração de informações com maior facilidade e rapidez quando comparado ao fluxo em papel. Os textos são descritos em linguagem natural, desta forma utilizamos mineração de texto e o processamento de linguagem natural para compreender e extrair conceitos relevantes de um documento.

Este trabalho está baseado no desenvolvimento de uma revisão sistemática com o objetivo de identificar, analisar e interpretar dados que relatem o uso das técnicas de processamento de linguagem natural no contexto da saúde, identificando a sua frequência e aplicação associadas a doenças crônicas, bem como a incidência de publicações associadas ao tema nos últimos anos.

Metodologia

Os termos de busca utilizados foram obtidos através de consulta aos Descritores em Ciências da Saúde (decs.bvs.br), onde foi utilizada para procurar trabalhos através da combinação dos descritores “processamento de linguagem natural” e “doença crônica”. Na pesquisa bibliográfica foram utilizadas as bases SciELO (www.scielo.org), LILACS (bases. bireme.br), MEDLINE (www.ncbi.nlm.nih.gov/pubmed) e Scopus (www.scopus.com/periodicos.capes.gov.br/home.url).

Após consultar todas às bases de dados com os termos de interesse, foram identificados

estudos que apresentavam duplicidade entre as bases, ou seja, estudos que apareceram em mais de uma base.

Como critérios de inclusão e exclusão foram excluídos trabalhos que tratavam de resultados referentes a apenas um dos temas de estudo, ou seja, somente processamento de linguagem natural ou somente doenças crônicas. Também foram excluídos estudos primários que faziam somente referência e citações aos temas.

Todos os artigos foram avaliados com base em seu resumo, sendo assim foi utilizado como critério de inclusão seu idioma (em português, inglês ou espanhol), além do livre acesso ou acesso completo do artigo através da rede da Universidade de São Paulo.

Os artigos selecionados foram avaliados e um resumo com informações de interesse foi criado, o qual apresenta as seguintes informações: autor, ano de publicação, objetivo, metodologia e ferramenta de processamento de linguagem natural.

Em muitos estudos a metodologia de processamento de linguagem natural não foi descrita ou foi apresentada de maneira superficial. Nesses casos, novas pesquisas foram realizadas com o objetivo de coletar informações com relação a este fator.

Os estudos resultantes foram sumarizados, onde as informações foram organizadas através de estruturas na forma de tabelas e gráficos para facilitar a compreensão dos dados.

Resultados

A partir das plataformas estabelecidas para revisão bibliográfica, foram encontrados ao todo 54 artigos, onde o número de artigos resultantes da busca em cada uma das plataformas está representado na Figura B-1.

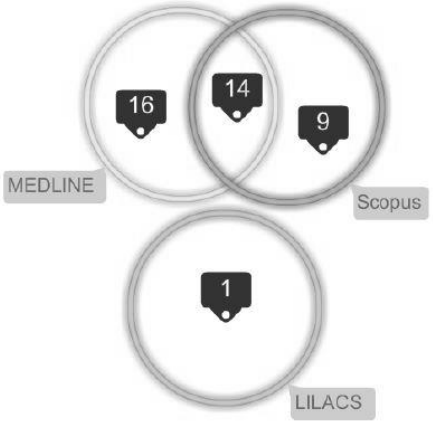

Figura B-1: Diagrama de *Venn* com o resultado por busca em cada base de dados.

Dentre os artigos encontrados catorze estavam presentes em mais de uma plataforma,

representando a duplicidade. Após a remoção de duplicidade foram eliminados seis artigos, que não permitiam acesso completo, além de catorze devido a avaliação de seu resumo, resultando em uma redução a 20 artigos finais que foram avaliados neste projeto, como representado na Figura B-2.

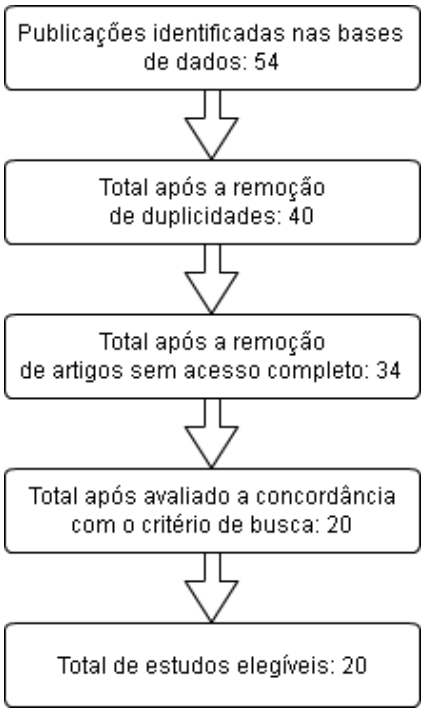

Figura B-2: Número de artigos resultantes em cada etapa do fluxo metodológico.

Dos 20 artigos resultantes, 9 foram encontrados nas plataformas Scopus e MEDLINE, 1 apenas na Scopus e 10 apenas na MEDLINE. A busca por meio da plataforma SciELO não retornou nenhum resultado, enquanto que a LILACS obteve um artigo, no entanto, o mesmo foi eliminado dos estudos finais durante a etapa de avaliação de elegibilidade.

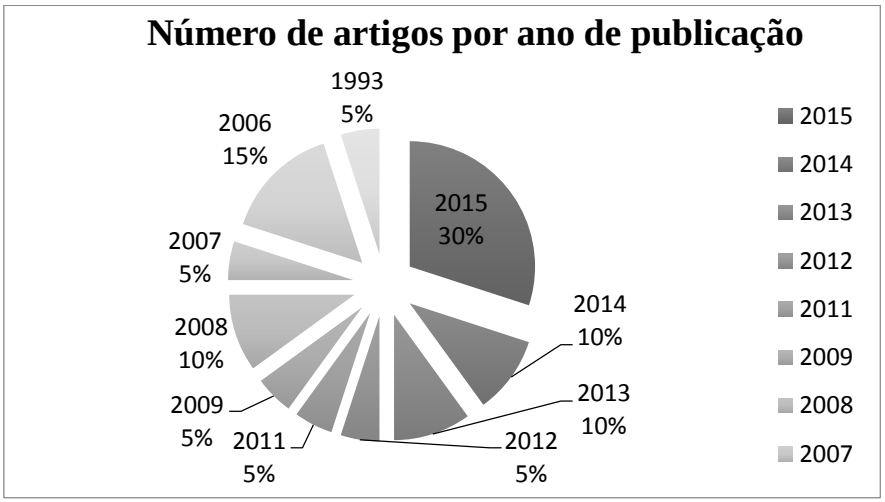

Figura B-3: Porcentagem de artigos publicados por ano.

A análise de distribuição dos artigos analisados por ano de publicação demonstra que metade dos estudos publicados estão concentrados nos anos de 2015 a 2013, além de ter o primeiro registro em 1993 (Figura B-3).

Para o estudo foi criada uma tabela de análise de artigos elegíveis contendo as informações de interesse (**Erro! Fonte de referência não encontrada.**). Por meio da análise da metodologia apresentada por cada artigo podemos analisar as técnicas e algoritmos de processamento de linguagem natural. No entanto, não são todos os artigos que descrevem e detalham as técnicas, muitas vezes ocultando informações com relação a técnica na qual foi avaliada e escolhida para o estudo ou ao algoritmo utilizado.

O algoritmo de processamento de linguagem natural mais frequente foi o HITex, presente em quatro dos artigos estudados, seguido do cTAKES em 3 artigos. As demais ferramentas citadas nos artigos estudados apresentam apenas uma ocorrência dentre os artigos selecionados.

Com relação à metodologia, quatro artigos não tiveram sua metodologia de processamento de linguagem natural descrita e estas também não puderam ser encontradas em buscas adicionais.

O número de técnicas e algoritmos utilizados foi grande, ou seja, apresentou grande variabilidade. Destes alguns obtiveram destaque por sua frequência, como o caso da etiquetagem, segmentação de sentenças e tokenização, apresentando 9, 8 e 6 ocorrências respectivamente.

Tabela B-1: Sumarização de todos os artigos selecionados.

Autor	Ano	Objetivo	Metodologia aplicada	Algoritmo	Plataforma
(KERR et al., 2015)	2015	Medindo a adesão médica com indicadores de qualidade de gota: um papel para o processamento de linguagem natural	Leo - Tokenização, segmentação de sentença, segmentação de seção, anotadores de expressão regular, anotador lógico baseado em regra. UIMA-AS – analisa e realiza etiquetagem de dados (baseado em regras)	Leo Service e UIMA-AS services	Pubmed
(JAMES et al., 2015)	2015	Validando as estimativas de prevalência de doenças não transmissíveis com base em pesquisas domiciliares: o estudo diagnóstico sintomático	Tokenização Mapeamento de palavras sinônimas	Não descrito	Pubmed e Scopus
(CARRELL et al., 2015)	2015	Usando processamento de linguagem natural para identificar problemas na prescrição de opióides.	Métodos PLN baseados em regra. UIMA OpenNLP - Tokenização, segmentação de sentença, etiquetador gramatical, extração de entidade nomeada, segmentação, analise e resolução da correferência	cTAKES (Utiliza UIMA e OpenNLP), adaptação do NegEx	Pubmed
(PALMER et al., 2014)	2015	A prevalência de consumo problemático de opiáceos em pacientes recebendo terapia opióide crônica: avaliação assistida por computador de notas clínicas de registos de eletrônicos de saúde	Métodos PLN baseados em regra. UIMA OpenNLP - Tokenização, segmentação de sentença, etiquetador gramatical, extração de entidade nomeada, segmentação, analise e resolução da correferência	cTAKES(Utiliza UIMA e OpenNLP), adaptação do NegEx	Pubmed
(LIAO et al., 2015)	2015	Métodos para desenvolver um algoritmo de fenótipo de prontuário médico eletrônico para comparar o risco de doença arterial coronariana em torno de 3 coortes de doenças crônicas	Métodos PLN baseados em regra. HITex - utiliza um conjunto de módulos contendo *token*, um gazetteer, segmentador de sentença, etiquetador gramatical, um transdutor de entidades nomeadas e um etiquetador de correferência.	HITex	Pubmed e Scopus
(TRICHE et al., 2014)	2014	Abordagem de bioinformática para a Genética do pré-eclâmpsia	SciMiner - Etiquetagem	SciMiner	Pubmed

(MORRISON et al., 2014)	2014	Razões para descontinuação de medicamentos hipolipemiantes em pacientes com doença renal crônica	TextMiner – baseado em um modelo de linguagem que consiste em conjuntos de classes de palavras e conjuntos de regras de estrutura de frase	TextMiner	Pubmed
(NIGWEKAR et al., 2014)	2014	Quantificar uma doença rara em dados administrativos: O Exemplo de calcifilaxia	Não descrito	Não descrito	Pubmed
(GREENBERG et al., 2013)	2013	Medição significativa: o desenvolvimento de um sistema de medição para melhorar o controle da pressão arterial em pacientes com doença renal crónica	Não descrito	Não descrito	Pubmed e Scopus
(LIYANAGE et al., 2013)	2013	Ontologias para melhorar estudos de pesquisa e qualidade de gerenciamento de doenças crônicas - uma estrutura conceitual.	Não descrito	DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering)	Pubmed e Scopus
(PIVOVAROV; ELHADAD, 2012)	2012	Um conhecimento híbrido e uma abordagem de dados orientado a identificar conceitos semanticamente semelhantes.	Identifica a estrutura do documento, realiza análise sintática (etiquetador gramatical e superficial), e faz o reconhecimento de nome-entidade de conceitos UMLS	Não descrito	Pubmed e Scopus
(KANDULA et al., 2011)	2011	Uso de tópicos de modelagem para recomendar material educativo relevante para pacientes diabéticos.	MALLET - Naïve Bayes, máxima entropia, árvores de decisão, e etiquetador de sequencias	MALLET (Machine Learning for Language Toolkit)	Pubmed e Scopus
(HIMES et al., 2009)	2009	Predição de doença pulmonar obstrutiva crônica em pacientes com asma usando registros médicos eletrônicos.	Métodos PLN baseados em regra. HITex - utiliza um conjunto de módulos contendo *token*, um gazetteer, segmentador de sentença, etiquetador gramatical, um transdutor de entidades nomeadas e um etiquetador de correferência.	Health Information Text Extraction (HITEx)	Pubmed
(HIMES et al., 2008)	2008	Caracterização dos pacientes que sofrem de ataques de asma utilizando dados extraídos de registros médicos eletrônicos	Métodos PLN baseados em regra. HITex - utiliza um conjunto de módulos contendo *token*, um gazetteer, segmentador de sentença, etiquetador gramatical, um transdutor de entidades nomeadas e um etiquetador de correferência.	Health Information Text Extraction (HITEx)	Pubmed e s Scopus

(GUDIVADA et al., 2008)	2008	Identificar genes de doenças-causal usando representação baseada em Web Semântica do conhecimento genômico e fenômico integrado.	Não descrito	MetaMap	Pubmed e Scopus
(PAKHOMOV et al., 2007)	2007	Estudo para avaliar a hipótese de que o PLN do prontuário médico eletrônico melhora a determinação sobre o diagnóstico de angina pectoris.	Métodos PLN baseados em regra. UIMA OpenNLP - Tokenização, segmentação de sentença, etiquetador gramatical, extração de entidade nomeada, segmentar, analisar e resolução da correferência	cTAKES(Utiliza UIMA e OpenNLP)	Pubmed
(CHU; DOWLING; CHAPMAN, 2006)	2006	Avaliar a eficácia de quatro características contextuais na classificação anotada de condições clínicas nos relatórios de departamento de emergência	SySTR - PLN de texto não estruturado	SySTR (Syndromic Surveillance from Textual Records)	Pubmed e Scopus
(CHU; DOWLING; CHAPMAN, 2006)	2006	Extraindo diagnósticos principais, co-morbidade e tabagismo para a investigação da asma: avaliação de um sistema de processamento de linguagem natural.	Métodos PLN baseados em regra. HITex - utiliza um conjunto de módulos contendo *token*, um gazetteer, segmentador de sentença, etiquetador gramatical, um transdutor de entidades nomeadas e um etiquetador de correferência.	Health Information Text Extraction (HITEx)	Pubmed
(LACSON; LONG, 2006)	2006	Processamento de linguagem natural de registros de dieta falados (DSE)	GUTime – etiqueta e normaliza expressões temporais	GUTime	Scopus
(LENERT; TOVAR, 1993)	1993	Ligação automática de descrições de texto livre de pacientes com uma guia de prática médica.	CAPIS - Métodos PLN baseados em conceitos.	Canonical Phrase Identification System (CAPIS)	Pubmed

Discussão

Dos 54 estudos resultantes da busca nas diferentes bases de dados seis trabalhos potencialmente importantes para a revisão sistemática não permitiam acesso completo a seu conteúdo. Desta forma, foram removidos na etapa de inclusão e exclusão de estudos.

A análise dos estudos selecionados na presente revisão bibliográfica aponta para a recente pesquisa relacionada a processamento de linguagem natural e doenças crônicas, na qual metade dos artigos resultantes foram publicados de 2013 a 2015.

A maioria dos artigos não apresentavam com clareza a ferramenta ou metodologia de processamento de linguagem natural utilizada, sendo necessário procurar outros artigos que pudessem esclarecer sobre a abordagem utilizada.

Não existe uma associação temporal com relação às técnicas mais frequentemente utilizadas, ou seja, dentre as técnicas mais abordadas no conjunto de artigos estudados não existe um período onde uma determinada técnica foi ou deixou de ser empregada.

Aplicações de processamento de linguagem natural podem empregar abordagens superficiais ou profundas, quando fazem uso de poucas ou muitas informações linguísticas, as quais são normalmente categorizadas em níveis de conhecimento (NÓBREGA, 2013).

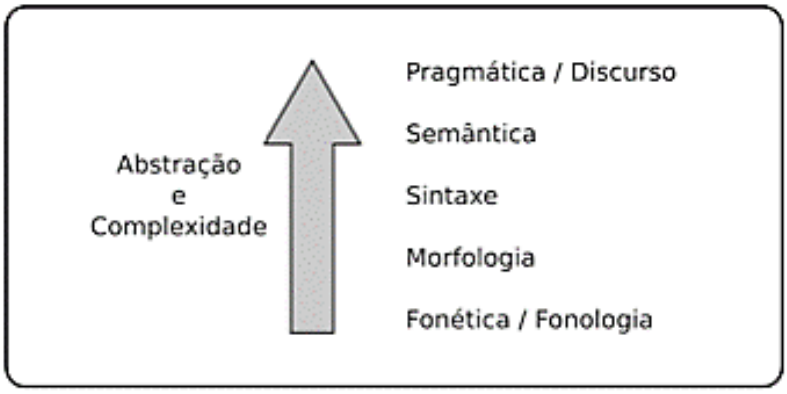

Figura B-4: Níveis de complexidade no processamento de linguagem Natural. Fonte: (NÓBREGA, 2013).

A Figura B-4 representa os níveis de conhecimento de acordo com seu grau de abstração de informação, onde os níveis mais altos são mais subjetivos. Em geral, métodos de processamento de linguagem natural baseados em abordagens superficiais demandam menos tempo para serem desenvolvidos, no entanto, características inerentes à língua natural podem dificultar o processo "superficial" da informação textual (NÓBREGA, 2013).

A tokenização, segmentação de sentenças e etiquetagem foram as metodologias de destaque entre os artigos estudados, estas são parte da análise sintática do processamento de

linguagem natural.

O primeiro processamento que é efetuado na análise sintática é a identificação das classes das palavras (também conhecidas como classes morfológicas, etiquetas lexicais ou partes da fala), para proceder esta classificação são implementados *parsers* que identificam nas frases as classes de palavras que as compõe (MULLER, 2003).

A metodologia mais frequente foi a etiquetagem, a qual é a classificação de palavras, onde as classes de palavras que compõe as frases são identificadas (OLIVEIRA; FREITAS, 2006).

A tokenização está baseada no processo de identificação e separação dos componentes significativos da sentença, assim como palavras e símbolos. A criação de *tokens* de um texto baseada em seus delimitadores é uma estratégia simples e que apresenta bons resultados (SILVA; SOUZA, 2014).

Conclusões

A revisão sistemática foi realizada a partir de 20 estudos primários selecionados através de critérios de inclusão e exclusão dentre 54 artigos resultantes da procura da palavra de busca nas bases de interesse.

É importante destacar que a aplicação de processamento de linguagem natural associado a doenças crônicas tem maior incidência com relação a extração automática de conhecimento em estudos empíricos da área, bem como em prontuários eletrônicos.

Uma limitação recorrente nesta pesquisa foi o conteúdo apresentado nos estudos, os quais muitas vezes não possuíam as informações técnicas ou as mesmas estavam incompletas. Mesmo com pesquisas mais aprofundadas, o levantamento de dados escasso afetou a elaboração de resultados e discussões.

Dentre as metodologias de processamento de linguagem natural mais utilizadas de acordo com a análise de artigos temos destaque para abordagens que são representadas pela análise sintática.

Referências Bibliográficas

BEZERRA, C. G.; DIMENSTEIN, M. O fenômeno da reinternação : um desafio à Reforma Psiquiátrica. p. 417–441, 2011.

CARRELL, D. S. et al. Using natural language processing to identify problem usage of prescription opioids. **International Journal of Medical Informatics**, v. 84, n. 12, p. 1057–1064, 2015.

CHU, D.; DOWLING, J. N.; CHAPMAN, W. W. Evaluating the effectiveness of four contextual features in classifying annotated clinical conditions in emergency department reports. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 141–145, 2006.

CONFORTO, E. C.; AMARAL, D. C.; SILVA, S. L. DA. 1 Roteiro para revisão bibliográfica sistemática: aplicação no desenvolvimento de produtos e gerenciamento de projetos. **Congresso brasileiro de gestão e desenvolvimento de produto**, 2011.

GRADY, K. L. et al. Team Management of Patients With Heart Failure. **Circulation**, v. 6083, p. 2443–2456, 2000.

GREENBERG, J. O. et al. Meaningful measurement: developing a measurement system to improve blood pressure control in patients with chronic kidney disease. **Journal of the American Medical Informatics Association : JAMIA**, v. 20, n. e1, p. e97–e101, jun. 2013.

GUDIVADA, R. C. et al. Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge. **Journal of biomedical informatics**, v. 41, n. 5, p. 717–29, out. 2008.

HIMES, B. E. et al. Characterization of patients who suffer asthma exacerbations using data extracted from electronic medical records. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 308–312, 2008.

HIMES, B. E. et al. Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients Using Electronic Medical Records. **Journal of the American Medical Informatics Association**, v. 16, n. 3, p. 371–379, 2009.

JAMES, S. L. et al. Validating estimates of prevalence of non-communicable diseases based on household surveys: the symptomatic diagnosis study. **BMC medicine**, v. 13, n. 1, p. 15, jan. 2015.

JENCKS, S. F.; WILLIAMS, M. V; COLEMAN, E. A. Rehospitalizations among patients in the Medicare fee-for-service program. **The New England journal of medicine**, v. 360, n. 14, p. 1418–28, 2009.

KANDULA, S. et al. Use of topic modeling for recommending relevant education material to diabetic patients. **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, v. 2011, p. 674–682, 2011.

KERR, G. S. et al. Measuring physician adherence with gout quality indicators: A role for natural language processing. **Arthritis Care and Research**, v. 67, n. 2, p. 273–279, 2015.

LACSON, R.; LONG, W. Natural language processing of spoken diet records (SDRs). **AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium**, p. 454–458, 2006.

LENERT, L. A; TOVAR, M. Automated linkage of free-text descriptions of patients with a practice guideline. **Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care**, p. 274–8, 1993.

LIAO, K. P. et al. Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts. **PLoS ONE**, v. 10, n. 8, p. 1–11, 2015.

LIYANAGE, H. et al. **Ontologies to improve chronic disease management research and quality improvement studies - A conceptual framework**. Studies in

Health Technology and Informatics. **Anais**...2013.Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84894373032&partnerID=tZOtx3y1>

MARTIN, L. R. et al. The challenge of patient adherence. **Therapeutics and clinical risk management**, v. 1, n. 3, p. 189–99, set. 2005.

MORRISON, F. J. R. et al. Reasons for Discontinuation of Lipid-Lowering Medications in Patients with Chronic Kidney Disease. **Cardiorenal Medicine**, v. 4, n. 3–4, p. 225–233, 2014.

MULLER, D. N. **Processamento de linguagem natural**. Rio Grande do Sul: [s.n.].

NETO, G. V.; MALIK, A. M. Tendências na assistência hospitalar. **Ciência & saúde Coletiva**, p. 825–839, 2007.

NIGWEKAR, S. U. et al. Quantifying a rare disease in administrative data: The example of calciphylaxis. **Journal of General Internal Medicine**, v. 29, n. SUPPL. 3, p. 924–931, 2014.

NÓBREGA, F. A. A. Word Sense Disambiguation for portuguese through multilingual mono and multi-document. p. 126, 2013.

OLIVEIRA, C.; FREITAS, M. Classes de palavras e etiquetagem na Lingüística Computacional. **Calidoscópio**, v. 4, n. 3, p. 179–188, 2006.

PAKHOMOV, S. S. et al. Epidemiology of angina pectoris: role of natural language processing of the medical record. **American Heart Journal**, v. 153, n. 4, p. 666–673, 2007.

PALMER, R. et al. The prevalence and characteristics of patients with indicators of opioid abuse within an integrated group practice. **The Journal of Pain**, v. 15, n. 4, p. S25, 2014.

PIVOVAROV, R.; ELHADAD, N. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. **Journal of biomedical informatics**, v. 45, n. 3, p. 471–81, jun. 2012.

SANTOS, R. E. S. et al. Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático. **Revista de Sistemas e Computação**, v. 4, p. 116–125, 2014.

SILVA, E. M. DA; SOUZA, R. R. Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 19, n. 40, p. 1, 2014.

TRAVASSOS, G. H. et al. Scientific research ontology to support systematic review in software engineering. **Advanced Engineering Informatics**, v. 21, n. 2, p. 133–51, 2007.

TRICHE, E. W. et al. Bioinformatic approach to the genetics of preeclampsia. **Obstetrics and gynecology**, v. 123, n. 6, p. 1155–61, 2014.