Semiótica Computacional para Desenvolvimento de *Chatbots* Cibernéticos para Detecção de Sintomas

Trabalho de Conclusão de Curso

Engenharia de Computação

Ericks da Silva Rodrigues

Orientador: Prof. Fernando Buarque

Coorientador: Rui Nobrega Pontes Filho

Ericks da Silva Rodrigues

Semiótica Computacional para Desenvolvimento de *Chatbots* Cibernéticos para Detecção de Sintomas

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Engenharia de Computação

Escola Politécnica de Pernambuco

Universidade de Pernambuco

Orientador: Prof. Fernando Buarque

Coorientador: Rui Nóbrega de Pontes Filho

Recife - PE, Brasil

Julho de 2018

Ericks da Silva Rodrigues

Semiótica Computacional para Desenvolvimento de *Chatbots* Cibernéticos para Detecção de Sintomas/ Ericks da Silva Rodrigues. – Recife - PE, Brasil, Julho de 2018-

47 p.

Orientador: Prof. Fernando Buarque

Coorientador: Rui Nobrega de Pontes Filho

Trabalho de Conclusão de Curso – Engenharia de Computação

Escola Politécnica de Pernambuco

Universidade de Pernambuco, Julho de 2018.

1. Chatbot. 2. Detecção. 2. Sintoma. 3. Cibernética. 4. Semiótica. 5. Inteligência Artificial. I. Prof. Fernando Buarque. II. Universidade de Pernambuco. III. Escola Politécnica. IV. Semiótica Computacional para Desenvolvimento de *Chatbots* Cibernéticos para Detecção de Sintomas.

Este trabalho é dedicado à minha família

Agradecimentos

Gostaria de agradecer a meus avós, principalmente minha avó Lúcia que mesmo com tantos problemas de saúde, nunca mediu esforços para me tornar uma pessoa melhor, sempre me guiou durante minha infância e me instruiu a ser um bom homem.

Agradeço a meus pais, Ericka e Eurípedes por sempre se sacrificarem por mim, mesmo vindo de origem pobre, nunca me deixaram faltar nada de essencial para que eu pudesse ser quem sou hoje. Nunca passei fome, não sei o que é o real sofrimento por conta dessas duas pessoas maravilhosas.

Gostaria também de agradecer ao meu tio Lúcio por me introduzir no mundo da tecnologia, através de Minigames mostrando a uma criança que nunca tinha visto um computador na vida e alimentando sonhos. Gostaria de agradecer também as duas famílias maravilhosas que tenho, que sempre haja amor entre todos nós.

Agradeço as minhas professoras (Tias) Leila, Helena e Cícera, as quais hoje em dia não tenho mais contato. Elas foram quem disseram a minha avó que me criava que eu tinha um futuro brilhante, disseram para "investir nesse menino", e hoje sou a primeira pessoa da família do meu pai a entrar para a vida universitária, e a segunda pessoa da família da minha mãe. Muito obrigado!

Agradeço também a meus amigos de infância, em especial a um grande amigo, o André Mateus, o qual foi um grande espelho para mim na fase pré-adolescente, o que me impediu de me perder e de perder o foco. Agradeço também aos amigos do ensino médio que me divertiram bastante, mas nunca me levaram para o mal caminho e me levaram a viver uma fase maravilhosa da minha vida.

Agradeço também os meus amigos da universidade, os quais tornaram uma graduação muito dificil algo muito mais agradável e divertida.

Agradeço também a minha namorada Libania, que foi a pessoa que me ensinou abrir minhas asas e conhecer o mundo, e por me suportar durante os momentos mais difíceis da minha graduação.

Agradeço, por fim, a todos os meus educadores que tive durante toda a vida e agradeço a Deus por permitir tantas experiências maravilhosas nesse mundo.

“If I have seen further it is by
standing on the
shoulders of Giants.“
“Isaac Newton“

Resumo

Atualmente, os profissionais de saúde tem buscado apoio em ferramentas que auxiliem seu trabalho. Uma das funções de sistemas de suporte a decisão que mais demandam esse apoio é o diagnóstico médico em especial na anamnese. A anamnese é um procedimento fundamental para o diagnóstico médico, onde o profissional de saúde realiza perguntas com o objetivo de adquirir informações para identificar as prováveis causas do problemas através dos sintomas. Recentemente, os *chatbots* estão ganhando espaço no mundo tecnológico como uma fexível ferramenta de atendimento a pessoas. Os *chatbots* são programas de computador capazes de responder de forma instantânea mensagens enviadas. Assim, eles podem ser ferramentas que podem trazer contribuições muito significativas para a realização de forma eficiente da anamnese. Porém, atualmente, existe uma grande dificuldade para que *chatbots* detectem as diversas variações linguísticas para descrever um mesmo sintoma. Então, o software desenvolvido neste projeto trata-se de um *chatbot* capaz de utilizar os conceitos de semiótica computacional para aprender com as interações dos usuários centralmente se beneficiando dos conceitos de Semiótica Computacional.

Palavras-chave: Chatbot, Detecção, Sintomas, Cibernética, Semiótica.

Abstract

Currently, the professionals of health have been seeking support in tools that help their jobs. One of the functions of decision support systems that requests more help is medical diagnostic, in special, anamnesis. Anaminesis is a fundamental procedure to medical diagnostic, in which the professional of health asks questions with the objective of getting information to identify the probables causes of troubles through the symptoms. Recently, the chatbots are winning a prominent place in the technological world because they end up as like a flexible tool of attendance for people. The chatbots are piece of software able to answer instantaneously messages sent. Therefore, they can bring contributions very significant to the achievement of enficient anamnesis. However, currently, there is a big dificult of chatbots to detect the various linguistic variations that describe the same symptom. So, the software developed in this project is a chatbot able of use the concepts of Computational Semiotic to learn with the interactions of users.

Keywords: Chatbot, Detection, Symptoms, Cybernetic, Semiotic.

Lista de ilustrações

Lista de abreviaturas e siglas

IA	Inteligência Artificial
IDF	*Inverse Document Frequency*
JS	*Javascript*
KDD	*Knowledge Discovery in Databases*
KNN	*K-Nearest Neighbors*
TF	*Term Frequency*

Sumário

1 Introdução

Este trabalho de conclusão de curso tem como objetivo o desenvolvimento de um *chatbot* cibernético que utiliza Semiótica Computacional, que é capaz de adquirir conhecimento através de um especialista para classificar e padronizar os sintomas médicos tratando as variações linguísticas.

A crescente utilização de *chatbots* para prestação de serviços em interfaceamento humano-máquina e os aprendizados antevistos (*i. e.* experiência profissional, técnicas de classificação de texto e etc.) no processo de desenvolvimento do mesmo foram os principais motivadores da escolha do tema.

Este capítulo descreve a motivação ao problema em 3 seções. Na Seção 1.1 são descritos os motivos e é oferecida uma descrição do problema de forma geral.

Nas Seções 1.2 e 1.3 são descritos os objetivos que devem ser atingidos para a solução do problema. Por fim, a Seção 1.4 detalha a organização da monografia.

1.1 Motivação e Caracterização do Problema

Com a crescente quantidade de usuários utilizando a internet para solicitar atendimentos que normalmente seriam feitos por telefone ou presencialmente, os *chatbots* estão se tornando cada dia mais uma alternativa efetiva para diversas organizações. Por oferecerem a possibilidade de atenderem pessoas, ao mesmo tempo, com mais velocidade, e a qualquer hora do dia, eles estão reduzindo bastante a necessidade de atendimento humano[1]. Um exemplo de área que pode ser fortemente auxiliada por um *chatbot* é a de ciências da saúde. Em aplicações voltadas para saúde, *chatbots*, como ferramenta tecnológica de interface com o usuário podem ser muito efetivas para tornar amigáveis os diagnósticos médicos; mais especificamente na detecção de sintomas.

Os sintomas são alterações ao padrão de normalidade que a pessoa percebe, desejavelmente um profissional de saúde, referente ao corpo, sensações ou metabolismo de um paciente [2]. Essas informações são normalmente obtidas através de uma consulta médica, num processo chamado de anamnese. Com um *chatbot* realizando as perguntas sobre os sintomas, o processo poderia ser agilizado, até mesmo à distância, e realizado de forma mais eficiente e eficaz.

Apesar do sucesso, muitos *chatbots* falham ao cobrir uma grande variedade de possíveis perguntas e respostas para os mais diversos tipos de usuários, que utilizam os mais diferentes linguajares para a comunicação. Como exemplo na área de atenção a saúde, um mesmo sintoma pode ser descrito de diversas formas por diferentes pacientes, a saber,

um paciente pode informar que está sentindo uma dor de cabeça e outro pode dizer que está com uma dor no crânio, e as duas declarações estão ligadas ao mesmo sintoma: cefaleia [3]. E essas características linguísticas variam entre diferentes regiões, o que torna muito difícil o processo de um *chatbot* ser capaz de identificar e classificar os mais diversos tipos de sintomas.

Então, este trabalho irá propor uma solução utilizando semiótica computacional, que possui propostas baseadas no processo de semiose para resolução de problemas complexos de forma computacional.

1.2 Objetivo Geral

Este trabalho de conclusão de curso tem como objetivo o desenvolvimento de um *chatbot* cibernético capaz de aprender com a interação de médicos e pacientes, identificar os sintomas descritos pelos pacientes, e que de acordo com o contexto que ele for utilizado via Semiótica Computacional, seja capaz de adequar-se às variações linguísticas e tornar-se mais autônomo à medida que adquira novas informações.

1.3 Objetivos Específicos

- Estudo sobre elementos relevantes para detecção de sintomas;
- Desenvolvimento de um classificador de texto capaz de adaptar-se as variações linguísticas;
- Interface gráfica capaz de atender tanto a pacientes, quanto a médicos;
- Realização de testes para validar o modelo apresentado.

1.4 Estrutura do Documento

O documento é composto por 4 capítulos:

- Este capítulo que descreve a motivação e os objetivos;
- O capítulo 2, que fornece o embasamento teórico necessário para a compreensão das teorias, conceitos e decisões tomadas durante o desenvolvimento do projeto;
- O capítulo 3, que descreve tanto o *Chatbot* produzido, bem como as decisões tomadas no projeto como também as visualizações de forma geral;
- E por fim o capítulo 5, resume as contribuições e que finaliza listando quais os planos futuros para o projeto.

2 Fundamentação Teórica

Este capítulo fornece o embasamento teórico necessário para a compreensão do desenvolvimento do experimento.

Na seção 2.1, são definidos os conceitos fundamentais da semiótica, bem como os conceitos de semiologia médica e de semiótica computacional. Já na seção 2.2, é esclarecido o conceito de sistemas cibernéticos e as suas funções.

Na seção 2.3 são esclarecidos os conceitos de *chatbots* e também são exemplificados alguns exemplos de *chatbots* em funcionamento atualmente. Por fim, na seção 2.4, estão explicados os conceitos de classificação, e também as definições de algumas técnicas computacionais que são capazes de realizar esse processo.

2.1 Semiótica

Na subseção 2.1.1, são definidos os conceitos gerais da semiótica de um ponto de vista filosófico. Já na seção 2.1.2, são exibidos os conceitos referentes ao processo de entendimento dos sintomas dos pacientes para futuramente um diagnóstico médico. Por fim, na seção 2.1.3, são introduzidos os conceitos de semiótica computacional, que é o processo de significação dos signos através de meios computacionais.

2.1.1 Conceitos Gerais

A semiótica é uma ciência relativamente nova, que está em constante expansão e seu sentido é mutável e difícil de definir com apenas um conceito. Segundo SANTAELLA, a "semiótica é uma ciência que tem por objetivo de investigação todas as linguagens possíveis"[4]. Dada esta definição, é possível entender que a semiótica tem como objetivo estudar, investigar e analisar as formas de comunicação, sejam elas verbais ou não verbais através do significado.

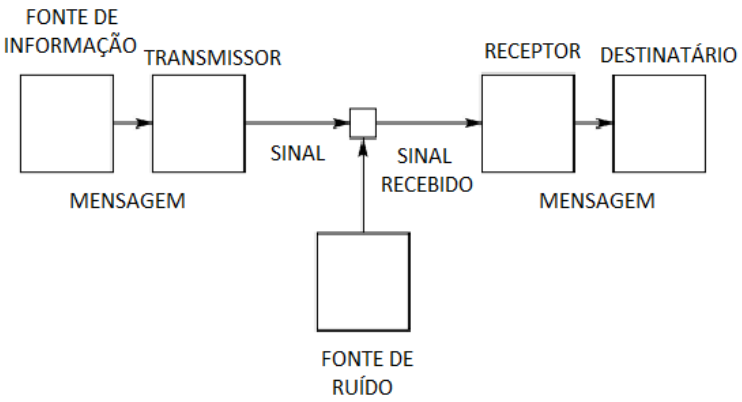
Analisando a formação da palavra, semiótica vem do grego *semeion*, que significa signo. E ainda do grego pode-se derivar o *semeiotiké*, que significa 'a arte dos sinais'[5]. O signo é algo que pode assumir um significado, e não necessariamente ser permanente. Pela definição de semiótica, para que exista uma forma de comunicação, é necessário um meio para que a mensagem seja transmitida. O signo é o meio de comunicação responsável por transportar o significado de quem está emitindo a mensagem para quem a recebe.

Exemplificando de forma que possa ser mais associada com a vida real, uma cor pode ser um signo. A cor vermelha pode trazer o significado de perigo, de sangue, de parar.

Isso depende da situação e de quem está interpretando a mensagem transportada nesse signo.

Abaixo, a Figura 1 descreve como funciona um sistema geral de comunicação.

Figura 1 – Diagrama esquemático de um sistema de comunicação geral



Fonte: A Mathematical Theory of Communication [6]

Como descrito na Figura 1, um sistema de comunicação é composto por uma fonte de informação, que irá se apropriar de um transmisso para enviar uma mensagem codificada em um sinal, que através do canal de comunicação pode sofrer alguma interferência definida como ruído até chegar a um receptor que irá transmitir a mensagem até o destinatário[6]. No entanto, para que o processo de comunicação seja realizado de forma correta, o receptor deve entender o significado da mensagem. Por exemplo, uma cor vermelha em um semáforo para um motorista está significando para o motorista parar. No entanto, para uma pessoa que nunca viu um semáforo e não tem noções das leis de trânsito ela receberá a mensagem porém não terá um significado atribuído ao signo.

Segundo o dicionário Michaelis, semiose é, *"segundo o filósofo e matemático americano Charles Sanders Peirce (1839-1914), o processo de produção de significados; a operação produtora e geradora de signos."*. Ou seja, é o processo em que é construído o entendimento da informação, partindo de um determinado signo. Esse processo é dito como fundamental, pois a partir da percepção dos sinais emitidos pelo meio, que é feito o processo de modelização e compreensão do mundo[7].

Peirce, considerado o fundador da semiótica moderna, foi um filósofo, pedagogo e cientista que trouxe diversas contribuições em diversas áreas de estudo. Ele fundamentou os seus estudos em três conceitos que define como fundamental para o processo semiótico. São eles a *primeiridade*, *secundidade* e *terceiridade*.

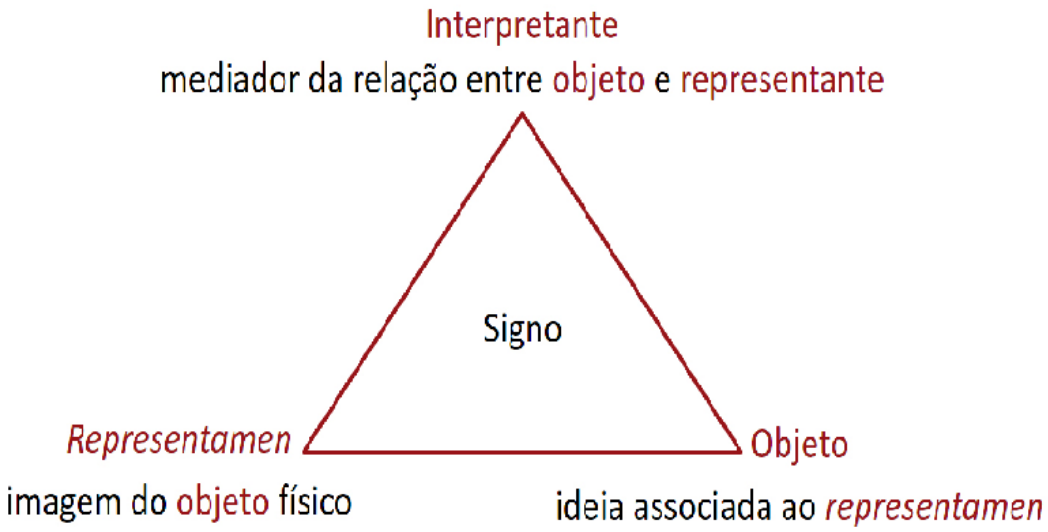
A primeiridade é, como o próprio nome diz, algo que não tem nada antes. É a base de todo o processo de entendimento[8]. É a primeira sensação que se tem ao escutar uma música ou a ver uma cor, porém sem efetuar nenhum processo de inferência. Para exemplificar, considere a visão de um homem e um copo com líquido. Nota-se aqui que houve uma qualificação dos elementos de visão inicial

A secundidade é a reação tida perante o fato, a observação mais aprofundada do fato de acordo com o que existe[8]. Seguindo o exemplo do copo visto anteriormente, nota-se que o homem suado está bebendo o líquido, e que o líquido é transparente e o homem está com roupas de corrida.

Por fim, a terceiridade é a representação que liga os signos ao mundo[8]. Dado o fato que um homem com roupas de corrida está suado e bebendo um líquido transparente, pode-se supor que o líquido é água, e o homem pode ter chegado de uma longa corrida, ou que o homem pode ser um atleta. Percebe-se então diversas possibilidades de ligar os signos a outros signos com outros significados.

Para Peirce, o signo é representado de forma triádica, como ilustrado na Figura 2:

Figura 2 – Relação triádica da semiótica



Fonte: <https://amusearte.hypotheses.org/1075>, Acessado em 03/06/2018

Dada a Figura 2, pode-se definir cada um dos elementos da tríade. O representamen é a parte perceptível do signo. Seguindo o exemplo da água no copo, o representamen seria a água, que trás a lembrança de algo. O objeto é exatamente a coisa, no caso a água. O interpretante é exatamente o significado do que é visto, o que é gerado ao visualizar a água, que pode lembrar um rio, chuva, molhado, dependendo da situação.

2.1.2 Semiologia Médica

A semiologia (*semeion*-signo *logos*-estudo) é definida como o estudo dos signos. Outrora a palavra tinha o mesmo significado que semiótica, porém em meados do século XX houve uma separação de significados e o termo semiótica ficou destinado ao estudo de sistemas de significação[9].

A semiologia médica é o estudo referente a sintomas e sinais das doenças. Tecnicamente falando, é o processo de aquisição de informações em que o profissional, seja ele das áreas de medicina, enfermagem, veterinária, psicologia, etc. adotam para se guiar diante de uma enfermidade, para que a partir dos dados possam efetuar o diagnóstico de forma correta.

Os sintomas são manifestações corpóreas que o paciente percebe em seu corpo. É uma informação voltada para o subjetivo. Já os sinais são características objetivas, são dados que são percebidos externamente por outros ou pelo profissional médico através dos seus sentidos ou por exames laboratoriais[3].

2.1.3 Semiótica Computacional

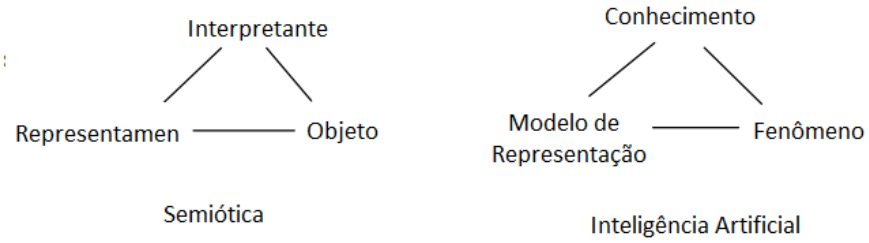
Como falado na subseção 2.1.1, A semiótica é uma ciência que tem como objetivo estudar o processo de significação dos signos. A semiótica computacional é uma área recente a qual tem o objetivo de produzir sistemas computacionais que sejam capazes de realizar o processo de semiose[10].

No entanto, para realizar o processo de semiose automática, é necessário um sistema inteligente. O termo inteligência artificial refere-se a pesquisas científicas referentes a simular de forma artificial inteligência em sistemas computacionais[11]. Inteligência por sua vez é, de acordo com o dicionário do Google, "*capacidade de compreender e resolver novos problemas e conflitos e de adaptar-se a novas situações*".

Assim, pode-se estabelecer uma comparação entre a semiótica e um sistema inteligente da forma descrita na Figura 3.

Pode-se associar o representamen com o modelo representativo do sistema para a IA. Já o interpretante que é a referência de informações para quem faz o processo de semiose do sinal é equivalente a base de conhecimentos do sistema inteligente, e o objeto real o qual é associado o signo e o representante passa a ser o fenômeno, ou a ocorrência que é definida pela IA partindo do modelo representativo do sistema e de sua base de conhecimentos[12].

Figura 3 – Mapeamento entre Inteligencia Artificial e a Semiótica



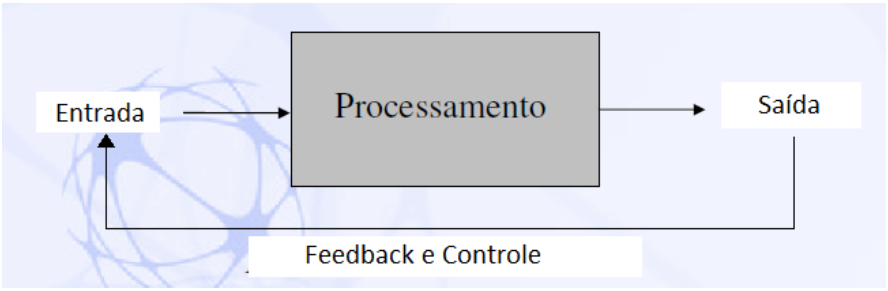
Fonte: Computational Semiotics : An Approach for the Study of Intelligent Systems Part I : Foundations[12]

2.2 Cibernética

A palavra cibernética vem do grego *kibernetiké*, que significa timoneiro. O timoneiro é o responsável por manter o controle do timão, que é o objeto que controla uma embarcação. Diante dessa função, pode-se observar que a função do timoneiro é guiar o barco diante de um fluxo de água, seja ele do mar ou em um rio, e em caso de adversidade, ele deve observar e corrigir a trajetória do barco para que ele siga seu caminho sem grandes problemas.

O termo cibernética foi citado inicialmente pelo cientista estadunidense Norbert Wiener que ficou conhecido como o pai da cibernética com o livro "*Cybernetics: or the Control and Communication in the Animal and the Machine*". Um sistema cibernético é um sistema que é capaz de se adaptar através do processo de retroalimentação, ou *feedback*. Ou seja, partindo da entrada do sistema, são realizadas análises das saídas qualificando os dados de forma que, caso sejam necessários ajustes, esses dados são tratados e enviados para o mecanismo que modela a saída do sistema[13].

Figura 4 – Representação de um sistema com *feedback*



Fonte: <https://nunocamiloblog.wordpress.com/>, acessado em 04/06/2018.

Como na Figura 4 é exibido, o sistema recebe uma entrada, é realizado um processamento, e na saída existe um "arco cibernético", que conecta o valor de saída com

à entrada novamente, para que ajustes sejam realizados. Um bom exemplo de sistema cibernético são os radares dos aviões. Eles recebem os valores de entrada referentes a posição inicial e investigam a área procurando os dados desejáveis, podendo ser um avião inimigo ou aliado, e em tempo real, ele atualiza a posição tanto atual quanto dos outros aviões e informa dados referentes à distância e velocidade.

2.3 *Chatbot*

Nesta seção, é explicado o que é um *chatbot*. Também são exemplificados alguns bots da plataforma Messenger do Facebook.

2.3.1 O que são *Chatbots*?

Os *chatbots* são robôs que são capazes de interagir com humanos através de um *chat* (interações conversacionais). Esses robôs de *software* não são robôs desenvolvidos fisicamente. Segundo o dicionário Aurélio, a robótica é uma "*ciência e técnica da concepção e construção de robôs*". A palavra robô vem da palavra tcheca *robota* que significa trabalho forçado. Então, os *chatbots* são mecanismos pré programados para responderem perguntas e de forma rápida.

Os *chatbots* estão ganhando cada vez mais espaço no mercado. Além de não gerarem custos adicionais, eles podem trabalhar 24 horas por dia, 7 dias por semana, sem pausa e sem descanso e pode atender vários clientes ao mesmo tempo. Além disso, dados mostram que os chatbots são capazes de reduzir os custos de atendimento em até 70% [14].

2.3.2 Exemplos de aplicações atuais

Existem vários exemplos de *chatbots* que estão ativos atualmente. Um deles é o *chatbot* da *Mastercard*, que é uma empresa de cartões de crédito e débito. A partir da página do *Facebook* da Mastercard, pode-se interagir através do Messenger com o bot e ele poderá dar informações referentes aos produtos e promoções disponíveis[15].

Já um *case* de Portugal, existe uma imobiliária chamada ERA, que possui um *chatbot* para onde pode-se obter informações sobre imóveis para alugar e comprar e também é possível enviar sua localização através do *bot* para conseguir localizar os imóveis próximos[16].

Existe também o *chatbot* da Suvinil que é uma marca de tintas. Esse *chatbot* é capaz de fornecer informações sobre tintas e locais de entrega, além de também ser possível o envio da imagem de um cômodo da casa e o chatbot é capaz de indicar as melhores cores para que possa ser pintada[17].

2.4 Classificadores

Esta seção é responsável por esclarecer os conceitos referentes a classificadores e apresentar técnicas computacionais de classificação.

2.4.1 O que são classificadores?

Antes prover a definição de classificador, é necessário entender o que é uma classificação. A classificação é o processo de organizar ou ordenar objetos por determinados rótulos, de acordo com os parâmetros deste objeto. Por exemplo, dentro de uma sala de aula pode haver diversas formas de classificar os discentes. Caso os alunos tenham nota maior que 7 eles são classificados como aprovados, caso contrário são classificados como reprovados.

No entanto, o processo de classificação pode se tornar complicado diante da complexidade das características do objeto. Pode-se citar como exemplo o processo de aprovação de crédito para uma pessoa. Para esse problema de classificação devem ser analisados diversas características da pessoa, tais como renda salarial, renda da família, idade, profissão e vários outros fatores que podem influenciar no processo de escolha do rótulo para a pessoa em questão. Um processo que acaba se tornando muito complexo para um humano realizar.

Então, a partir dai entra o classificador. O classificador é um sistema baseado em técnicas de inteligência computacional que podem adquirir de forma supervisionada com exemplos anteriores a capacidade de classificar objetos de acordo com seus atributos. Essas técnicas são normalmente baseadas em modelos estatísticos, na qual algumas delas serão citadas nas próximas seções.

2.4.2 Naive Bayes

O classificador *Naive Bayes* (*naive* vem do inglês e significa ingênuo) é uma técnica de classificação muito utilizada em diversos problemas de classificação. Apesar de considerar os atributos da situação analisada são independentes (daí a origem do nome *naive*), o algorítmo é muito utilizado por conta da sua simplicidade de desenvolvimento.

Baseado no teorema de Bayes, desenvolvido por Thomas Bayes, o algorítmo se apropria de exemplos anteriores para realizar uma predição de acordo com as informações que possui no momento.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

A equação 2.1 descreve como o Naive Bayes calcula suas probabilidades. Sendo A

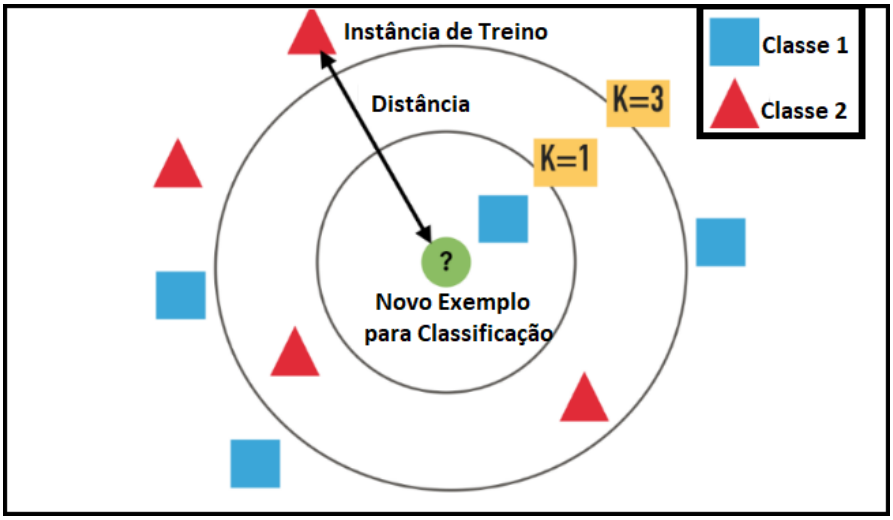
e B eventos, a probabilidade de A ocorrer dado que B ocorreu é dada pela probabilidade de B ocorrer dado que A aconteceu vezes a probabilidade de A acontecer, dividido pela probabilidade de B. Isso permite que possa ser reduzido na equação, partindo de resultados anteriores, que dependendo da distribuição dos eventos, uns podem ser mais significativos que outros no quesito de certeza. Quanto mais relevante é o termo, maior a certeza que ele fornece sobre o resultado.

2.4.3 K-Nearest Neighbors

O *K-Nearest Neighbors* ou KNN é uma técnica de aprendizado supervisionado relativamente simples. Seu nome vem do inglês e significa K-Vizinhos mais próximos. O K deve ser definido como um valor numérico o qual será um fator importante no processo de classificação. O processo do KNN consiste em investigar os indivíduos de uma determinada amostra e compará-los com o novo indivíduo a ser classificado. Os K-vizinhos mais próximos irão definir o rótulo que classifica esse indivíduo[3]. Normalmente essa distância é determinada pela distância euclidiana.

Para esclarecer o KNN, pode-se supor um conjunto de indivíduos contidos em um espaço, organizado de acordo com suas características; então surge um novo indivíduo necessário para realizar o processo de classificação; supondo que K seja igual a 1, o elemento que possua maior proximidade com esse indivíduo fará com que o novo indivíduo tenha o mesmo rótulo que ele. O objetivo é formar áreas de acordo com a distribuição dos parâmetros para classificar de forma mais aproximada.

Figura 5 – Representação do KNN



Fonte: <https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7>, acessado em 05/06/2018.

De acordo com a Figura 5, pode-se enxergar de forma mais clara como o processo

de classificação é realizado. E de acordo com o valor K (número de elementos levados em consideração na classificação), o valor da classificação também pode mudar.

2.5 Pré-processamento de texto

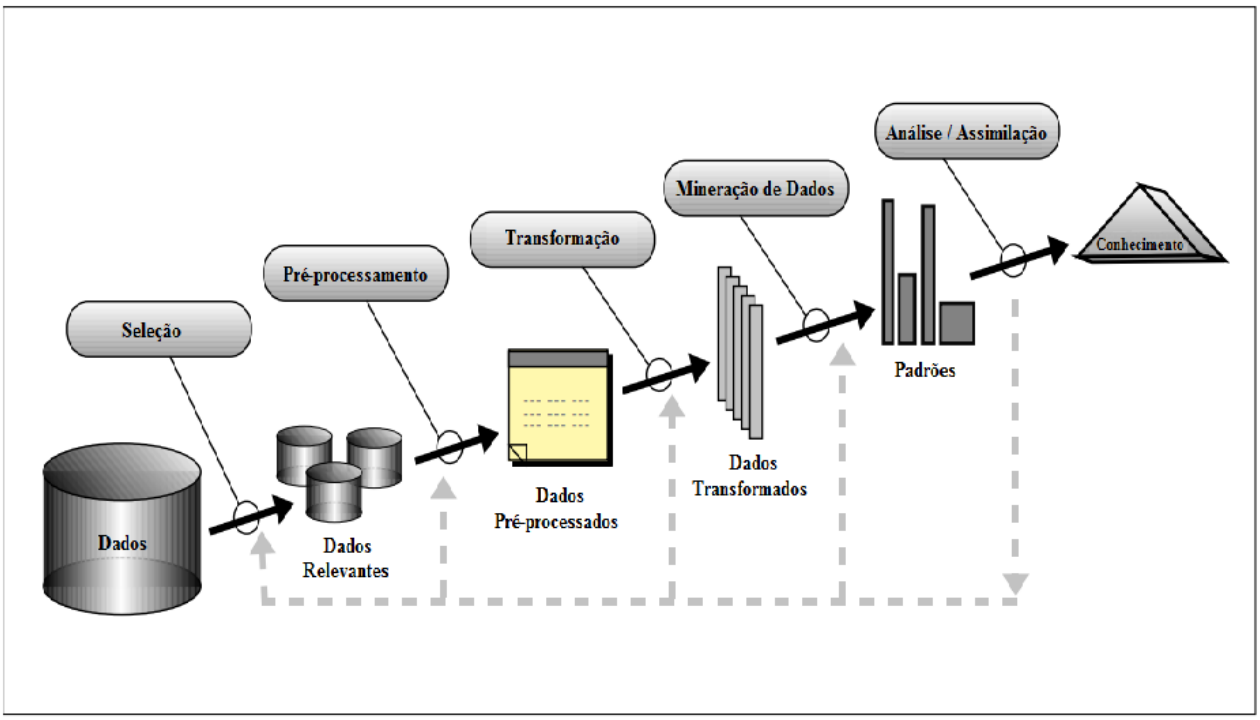
Esta seção estabelece os conceitos referentes ao pré-processamento de texto, e também as técnicas utilizadas no desenvolver do projeto. Nas próximas subseções serão explicadas as técnicas de pré-processamento textual que foram adotadas no projeto.

2.5.1 Pré-processamento de texto

Antes de falar sobre o pré-processamento textual, é necessário elaborar melhor o conceito de pré-processamento de dados.

O pré-processamento de dados é a etapa inicial do processo de mineração de dados (vide Figura 6). O processo de mineração de dados visa extrair conhecimentos relevantes de uma extensa base de dados, para produzir modelos que executem de forma mais satisfatória e precisa análises matemáticas e estatísticas com o objetivo de apoiar decisões de forma rápida e precisa.

Figura 6 – Representação do KDD



Fonte: <https://www.devmedia.com.br/mineracao-de-texto-analise-comparativa-de-algoritmos-revista-sql-magazine-138/34013>, acessado em 05/06/2018.

O pré-processamento de dados é a fase que visa reduzir os ruídos das amostras de dados. Nessa etapa é possível remover dados irrelevantes, informações desnecessárias, como também identificar dados corrompidos e corrigir problemas gerais[18].

Com esses conceitos definidos, entende-se que o processo de mineração de texto é uma subárea da mineração de dados, e a etapa de pré-processamento de texto é a fase de preparação do texto para a remoção de conteudos irrelevantes.

2.5.2 *Stemming* e *Stopwords*

O processo de Stemming visa extrair do texto padrões interessantes e que não são redundantes. O processo de stemming visa a partir da formação da palavra extrair a sua forma mais natural e relevante, para que possam ser comparadas com outros termos sem ter seu sentido alterado por uma alteração verbal[18]. É importante observar que para cada língua, o processo de *stemming* funciona de maneira específica.

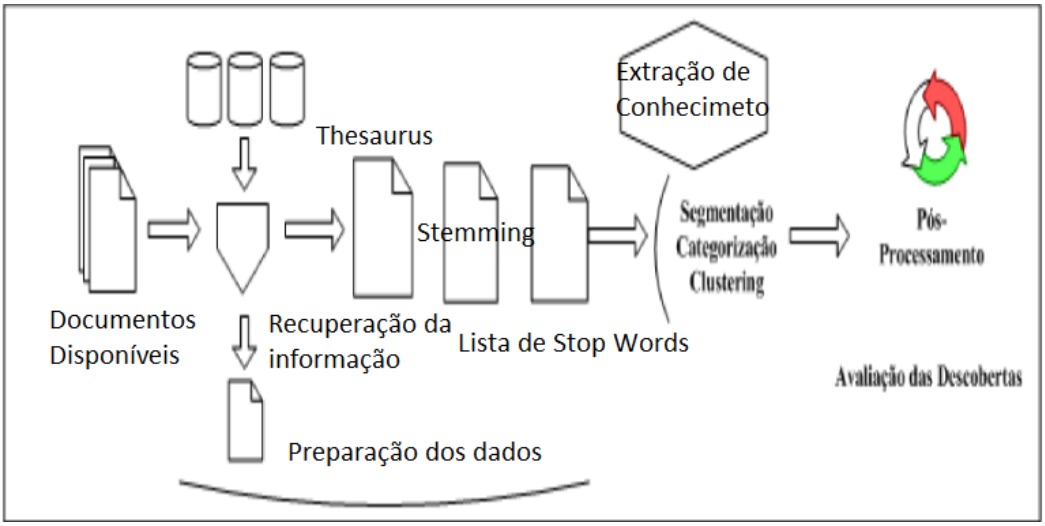
Em linhas gerais, se uma pessoa diz ter "tontura"e outra diz estar "tonta", ambas possuem o mesmo sentido, no entanto, do ponto de vista computacional, se as palavras forem levadas exatamente da forma que foram escritas, elas serão consideradas de formas diferentes por possuírem caracteres diferentes. Com o processo de stemming, as duas palavras se tornaram um único termo, buscando a origem da palavra, que seria "tont". E então os algoritmos de classificação textual, poderiam entender as duas palavras tendo o mesmo sentido.

Já o processo de remoção de stopwords, visa eliminar palavras que não trazem sentido nenhum para o texto como um todo. Essas palavras são normalmente artigos e preposições[18].

Como exemplo, será utilizado a frase "Estou com tontura". Analisando a frase, ela tem um verbo de ligação (estou) e uma preposição (com), que não agregam muito sentido no geral do texto (apenas o tempo ser presente), então eles são removidos do texto sobrando apenas a palavra tontura, que após ter a remoção de stopwords efetuada, passa pelo processo de stemming relatado acima.

Como observado na Figura 7, o processo de remoção de stopwords deve ser realizado antes do processo de stemming, pois o processo de stemming pode prejudicar a detecção de stopwords por alterar a forma das palavras.

Figura 7 – Etapas do Processo de Mineração de Texto



Fonte: Investigação do Processo de Stemming na Língua Portuguesa[18].

2.5.3 TF-IDF

O TF-IDF é uma forma de quantificar o quão forte é a relação entre o termo e um texto e também em uma relação de documentos[19].

$$TF = vezesQueTermoAparece/quantidadeTermos \tag{2.2}$$

$$IDF = log_e(totalTermos/qtdDocumentosTermoAparece) \tag{2.3}$$

$$TFIDF = TFxIDF \tag{2.4}$$

Logo, com essas definições, pode-se observar que o quanto mais um termo aparece em um documento, maior o valor TF-IDF vai ficar. No entanto, caso algum termo apareça em vários documentos, o peso será reduzido por conta do valor do IDF.

3 Desenvolvimento do *Chatbot*

Este capítulo expõe a forma que o sistema foi desenvolvido, desde as escolhas das ferramentas, técnicas de classificação e a criação do arco cibernético, para a retroalimentação utilizados na ferramenta produzida.

Na seção 3.1, é exibida uma descrição geral do projeto do *chatbot*, explicando de forma geral as etapas do sistema.

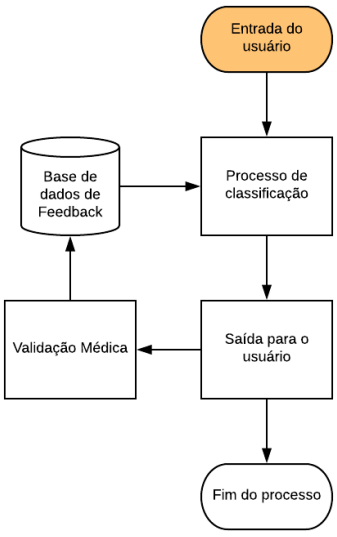
Na seção 3.2, pode-se obter informações das ferramentas, linguagem de programação e frameworks utilizados no desenvolvimento do *Chatbot*.

3.1 O *Chatbot* Cibernético

O *chatbot* cibernético foi desenvolvido baseado nos conceitos de cibernética, expostos na seção 2.2. Em seu funcionamento, o *chatbot* recebe uma retroalimentação de acordo com as respostas produzidas anteriormente.

De forma geral, o chatbot comporta-se de acordo com a Figura 8:

Figura 8 – Fluxograma da arquitetura geral do *Chatbot*



Fonte: o autor.

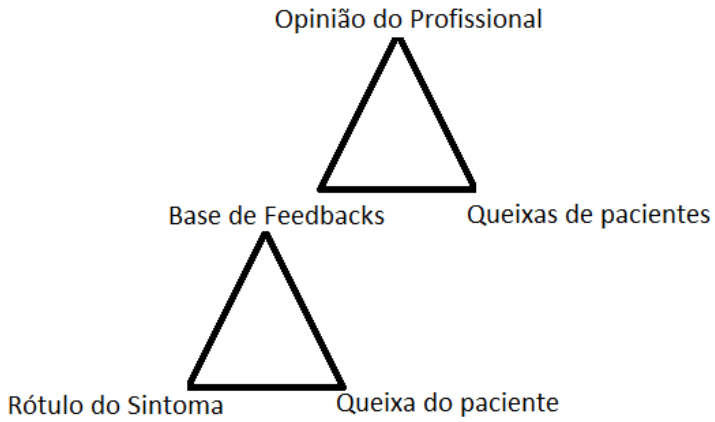
Como exibido na Figura, o *chatbot* aguarda uma entrada de texto do usuário. Essa entrada de texto será exatamente o sintoma que o paciente esteja sentindo, de forma semelhante a que ele responderia durante o processo de anamnese.

No processo de classificação, são efetuadas as etapas de pré-processamento textual, utilizando o *Stemming*, TF-IDF e *Stop Words* para remover termos que possivelmente não agregarão valor lógico para o classificador. Também é feita uma coleta da base dados, a qual é de grande importância, pois ela será responsável pelo processo de armazenagem dos feedbacks recebidos dos textos informados pelos pacientes.

Levando em consideração a os conceitos de semiótica, descrito na subseção 2.1.1, pode-se fazer um processo de relacionamento entre a arquitetura geral do *bot* da Figura 8 e os pilares que nortearam a Semiótica Peirciana. Na etapa em que o usuário envia a mensagem para o bot, as palavras são o signo, onde ao recebê-lo, ocorre a primeiridade, ou seja, a primeira impressão que irá desencadear os demais processos do sistema. Após isso acontecer, os elementos internos, que carregam o sentido lógico são adquiridos, ou seja, a secundidade ocorre. Então, para finalizar, são utilizados conceitos externos, vindos da base de feedback que são alimentadas pelos especialistas, onde o processo de classificação expõe algo novo partindo das informações vindas. Logo o processo em que são definidos os rótulos pode ser relacionado com a terceiridade.

No entanto, é necessário considerar que o profissional de saúde também realiza semioses ao receber o sintoma do paciente. Com isso em mente, pode-se estabelecer um modelo semiótico triádico de Peirce, conforme a Figura 9.

Figura 9 – Triângulo Semiótico em cadeia do *chatbot* cibernético



Fonte: o autor.

Após ser feito o processo de classificação, é necessário informar se o *bot* foi capaz de classificar o sintoma de forma padrão. Caso o *bot* não tenha "certeza"da decisão tomada ou não foi capaz de determinar qual sintoma foi descrito pelo paciente de acordo com os rótulos estabelecidos pelo médico, o texto é guardado para um futuro feedback profissional, ou mesmo consultas de segunda opinião médica.

Já no processo de feedback, médico seleciona um termo de uma lista e define o

rótulo que caracteriza o texto escrito. Ao fazer isso, esse texto é preprocessado e guardado na base de feedbacks, que receberá um novo componente aumentando por conseguinte o conhecimento.

3.2 Ferramentas e Frameworks

Para desenvolvimento do sistema, foi utilizado um Notebook com as seguintes configurações:

- Processador Intel Core i5-7200U;
- Memória RAM de 8GB;
- HD de 1TB;
- Placa de vídeo GeForce 940MX 2GB.

Como linguagem de programação, foi adotado a linguagem *JavaScript* (JS), que é uma linguagem que possui uma tipagem dinâmica, fraca e implícita. Ela foi desenvolvida inicialmente para navegadores, porém por conta da sua simplicidade e capacidade tem se tornado a cada dia mais popular, com diversos *frameworks*.

Para reforçar ainda mais a capacidade do JS, será utilizado também o Node.JS, que é um Runtime Environment, ou seja, uma ferramenta de execução em tempo real que o torna capaz de executar códigos JavaScript dentro de um servidor, e não mais preso em um navegador Web.

Como gerenciador de pacotes do Node.JS, foi utilizado o NPM, que é responsável por gerenciar as dependências de módulos que possuem funções. Os módulos mais importantes foram o *Express*, como *framework* para desenvolver um servidor Web, o *axios*, responsável por fazer as requisições HTTP, o *Stopword*, que executa o processo de remoção de *stopwords*, e o *natural* que é um módulo que possui diversas ferramentas para processamento de linguagem natural, onde foi utilizado a função de stemmer.

Como *framework* para o *front-end*, foi utilizado o React.JS. O React.JS é um *framework* que tem como objetivo fornecer interfaces interativas com o usuário, onde as interações do usuário com as interfaces geram renderização em tempo real da página que o usuário acessa.

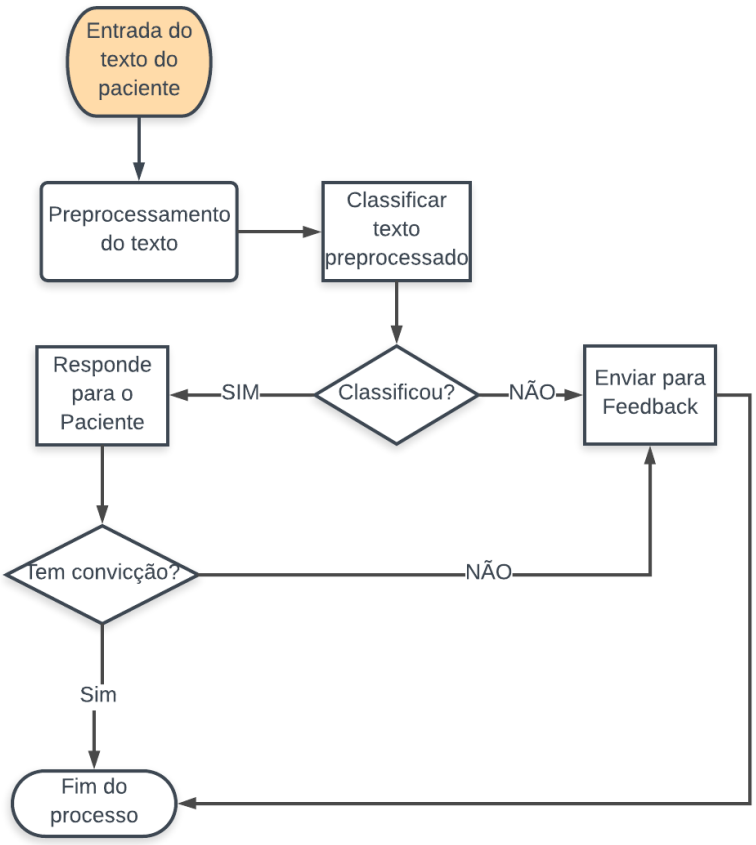
Para finalizar, como base de dados, foi utilizado uma aplicação disponibilizada pelo *Firebase*, que é uma plataforma adquirida pelo Google que possui diversas funcionalidades para aplicações dos mais variados tipos. Como base de dados ela possui uma opção preparada para consultas NoSQL, onde é bastante simples realizar consultas e monitorar a base de dados.

3.3 Detalhamento do Sistema

Esta seção foi criada para um exclarecimento referente as tomadas de decisões do *chatbot*, como o processo é executado, desde a entrada do usuário até o processo de feedback.

Esta seção apresenta de forma detalhada como o sistema foi desenvolvido. Para o detalhamento geral da ferramenta, visualizar a Figura 10:

Figura 10 – Fluxograma da entrada do usuário



Fonte: o autor.

3.3.1 O Pré-Processador Textual

O pré-processador de texto desenvolvido foi baseado nas ideias da seção 2.5. Esta etapa tem como objetivo realizar uma limpeza das entradas, eliminando valores ruidosos para um melhor processo de classificação.

Como pode ser observado na Figura 10, o processo se inicia com a entrada de texto do paciente. Ao receber o texto, o *bot* irá iniciar o passo de pré processamento textual, supondo que o paciente esteja informando: "estou com vontade de vomitar", primeiro

são removidas as *stopwords*, que são os termos que possivelmente não agregarão valor ao contexto da frase, o que resultará na frase "vontade vomitar". Após isso, o *stemmer* é executado, removendo os sufixos que podem prejudicar na comparação, o que resultará nos termos "vontad, vomit".

Após isso, os dois termos restantes serão enviados para o processo de classificação, o qual irá determinar em qual rótulo as sentenças se encaixam, isso se anteriormente um feedback tenha sido atribuído. Como descrito na Figura 8, a etapa de classificação recebe dados do banco de *feedbacks*, o qual recebe um texto pré processado. Supondo que anteriormente um médico tenha dado um único *feedback* que resultou no termo que está no banco "vomit".

3.3.2 O Classificador

O classificador foi desenvolvido baseando-se em características das duas técnicas de classificação, as quais foram explanadas na seção 2.4. Também em técnicas de pré-processamento de texto, citados na seção 2.5.

A primeira técnica que serviu de inspiração foi o KNN. O objetivo do classificador é localizar, a partir de uma base de dados, do termo de entrada com um termo já classificado anteriormente pelo médico (que está na base de feedback). No entanto, para determinar o quão termos são próximos um do outro, é necessário uma representação numérica.

Para realizar uma indicação de distância com os termos já contidos na base de feedback, é necessário realizar uma conversão numérica das palavras, e o TF-IDF serve muito bem para a distinção entre os termos. Quanto maior o valor do TF-IDF, maior a semelhança entre as duas amostras, ou menor a distância do KNN. E com a presença do IDF, eu tenho um penalizador no valor do peso, pois quanto mais um termo se repete menos ele é relevante, o que leva a segunda técnica, o Naive Bayes, que busca os rótulos e o ganho de informação de acordo com a relevância das entradas.

Quando o valor de TF é igual a zero, indica que as frases de entrada e a sendo analisada não possuem termos em comum, logo elas possuem uma "distância infinita". Caso todas as amostras da base de feedbacks não calculem um valor de TF maior que zero, logo a frase é absolutamente nova, nenhum dos termos foi visualizado anteriormente na base de feedbacks, então o classificador não é capaz de indicar o rótulo e a frase é enviada diretamente para a base de feedbacks.

Caso algum termo seja encontrado em alguma das amostras da base de feedback, porém o termo não é totalmente igual a frase, o valor de TF irá variar entre 0 e 1, e quanto maior, mais próximo a frase em questão. A frase da base de feedbacks que obtiver maior valor TFxIDF servirá para rotular a frase de entrada, porém o classificador também informará que não está certo da resposta.

No entanto, caso o valor do TF seja igual a 1, indica que toda a frase de entrada está contida na base de feedbacks em questão. Logo a base é a mais parecida, ou mais próxima possível uma da outra. Caso isso ocorra, o classificador irá determinar certeza na resposta contida.

Após calcular os valores para cada valor do feedback, ou atingir o grau de certeza, o classificador define qual o mais próximo de acordo com o valor na maior função de ativação. O grau de certeza é determinado diretamente ao valor TF. Caso o valor atinja o valor máximo, os termos de entrada podem ser considerados como totalmente semelhantes ao valor da base de feedback.

3.3.3 O Arco Cibernético

Após o processo de classificação, caso o indicador de certeza não tenha sido ativado, então a frase é enviada para uma base de dados que a manterá até que um especialista indique um rótulo para a frase. Após a frase receber um rótulo, então ela é preprocessada e enviada para a base de feedback, e então outras entradas poderão ser comparadas com elas.

Para exemplificar, será levado em consideração um suposto paciente que entraria com um texto "Estou com dor nas costas". O *chatbot* de início não conseguirá identificar uma resposta para o texto. Então esse texto será encaminhado para o supervisor (médico) e ele definirá um rótulo (dor nas costas) para o sintoma. Após o rótulo ser definido, o texto será enviado para a base de *feedbacks* e após isso, caso o usuário entre com um texto semelhante ("Minhas costas doem", por exemplo), o *chatbot* será capaz de classificar com certeza ou com dúvidas que o paciente afirmou ter dor nas costas, mesmo se a grafia esteja diferente. Esse processo pode ser visualizado de forma clara na Figura 8.

4 Testes e Resultados

Para iniciar o processo de testes, foi criado um formulário do Google com o objetivo de adquirir uma base de dados informal, onde as pessoas pudessem declarar que estão com algum sintoma. Entretanto, perguntar como as pessoas diriam a um médico que estão com um sintoma é um processo tendencioso (caso se pergunte como uma pessoa diz ao médico que ela está com dor de cabeça, ela vai dizer que ela está com dor de cabeça por conta das palavras usadas). Então, para incentivar as pessoas a responderem sem enviesar o texto, foi utilizado uma inspiração do conceito de semiótica: apenas um signo para cada sintoma sem informar palavras foi descrito, e pedido para que o voluntário respondesse como iriam se reportar a um profissinal de saúde caso sentisse o sintoma da imagem.

Então, foram convidadas diversas pessoas para responder o questionário online, cada uma com seu ponto de vista referente aos signos expostos. No total, foram coletados textos de 15 pessoas diferentes, onde cada uma respondeu as 5 respostas, para as imagens, e cada imagem recebeu um rótulo. Os rótulos foram cefaleia, dor nas costas, frio, tosse e tontura, onde as imagens de cada estão disponíveis para visualização no ANEXO IV.

As respostas foram repassadas para o bot de forma manual, e o bot tentava realizar o processo de significação. Os dados de cada pessoa foram passados em ordem, e quando o processo de inserir as mensagens de uma pessoa era finalizado, então o autor assumiu o processo de correção com o objetivo de realizar testes para comprovar o processo de aprendizado do bot.

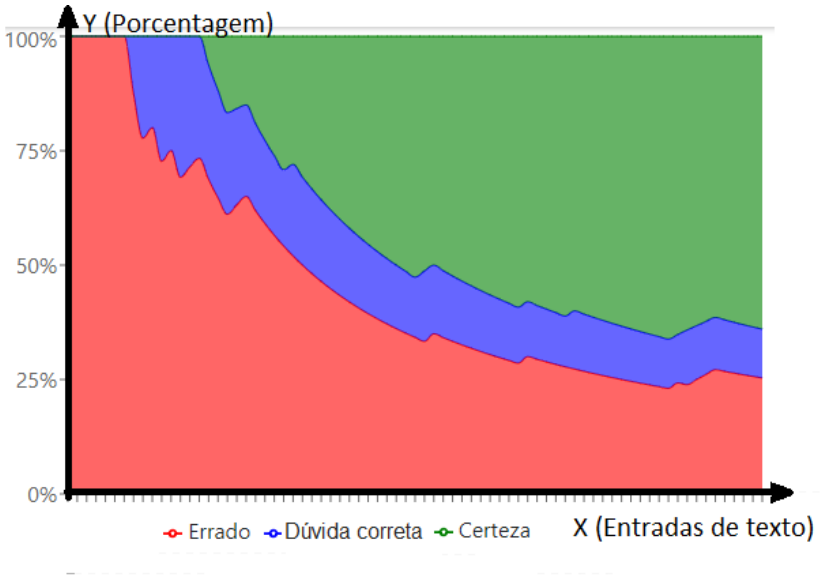
Como resultado, aproximadamente 71% das amostras foram classificadas de forma correta. O que aparentemente é um número muito baixo. Porém, é necessário destacar que o bot não possuia nenhum conhecimento prévio, ele começou com a base de feedback vazia. Com os 10 primeiros textos (os textos de duas pessoas diferentes) apenas 10% dos resultados foram de acertos. Caso os 10 primeiros itens fossem removidos teria uma acurácia de 75% e se fossem considerados apenas os 20 termos mais recentes, a acurácia atingiria os 80%. Isso mostra que o processo de feedback realmente está fazendo com que o bot aprenda com eventos passados e através de eventos ocorridos em tempo real.

Abaixo segue uma figura do gráfico que representa a distribuição de Erros, Dúvidas e Certezas durante o processo de treinamento do bot:

A coleta da série de dados é feita de forma automática, quando o bot tem certeza de que a entrada do usuário é um determinado rótulo, o valor referente a certeza é incrementado. Onde certeza, dúvida e erro são:

- Certeza: O *chatbot* tem convicção no resultado;

Figura 11 – Gráfico de distribuição de porcentagem entre certeza, dúvidas acertadas e erros.



Fonte: o autor.

- Dúvida correta: O *chatbot* conseguiu efetuar o processo de classificação, no entanto nao tinha convicção do resultado e o enviou para uma validação médica, que comprovou que o *chatbot* estava correto;
- Erro: O *chatbot* não conseguiu identificar nenhum rótulo referente as entradas, ou ele classificou de forma incorreta ao ser validado pelo supervisor (médico).

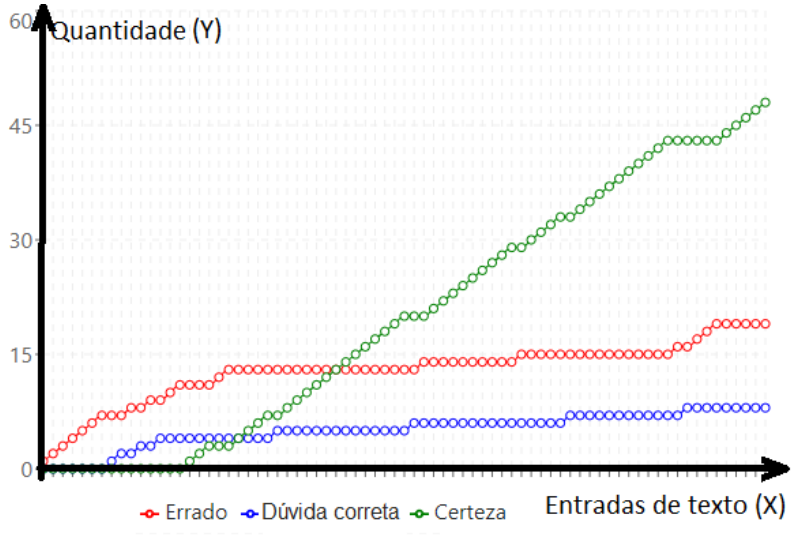
Como avaliação do gráfico da distribuição de porcentagem, é possível notar que de início o bot não consegue identificar nenhum dos rótulos, por não ter nenhum item em sua base de feedback que possa fornecer a informação de forma correta. No entanto, após algumas interações com usuários, é possivel notar um breve processo de apenas dúvidas, e então o bot começou a ter convicções de alguns resultados, que logo se tornaram a maioria e é possivel ver na faixa em azul a distribuição de uma faixa de incerteza, porém que estava correta.

É importante focar também que os erros são a base dessa proposta de classificação, quando existem erros ou termos incertos, indica que o bot não tinha informações suficientes para validar o seu dado, então foi solicitado a ajuda de um supervisor para que possam validar os dados e os adicionar a sua forma de conhecimento.

Lembrando também que não é necessário que o bot comece sem nenhuma base de feedback. Antes do lançamento também pode ser pensado em adicionar ou remover palavras da lista de stop words ou deixar uma base de feedback pré-validada de acordo com o contexto do problema para que ela possa ser aprimorada com outras interações.

Para que se possa perceber de forma mais clara, abaixo será exibido o gráfico que exibe o comportamento em quantidade de erros, incertezas e acertos:

Figura 12 – Gráfico de quantidade de erros, dúvidas e acertos durante o processo de interação com o bot.



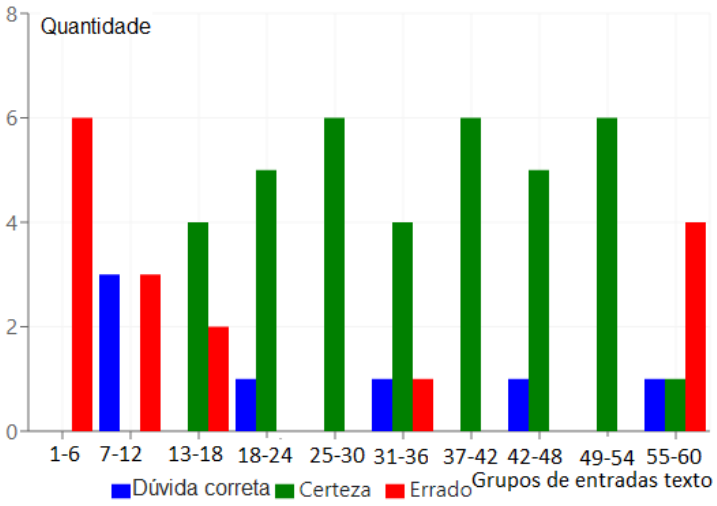
Fonte: o autor.

Como se pode ver na Figura 12, os erros são reduzidos de forma abrupta após algumas interações e pouco depois o processo de acerto por convicção é muito notável para os problemas apresentados.

Por fim, será apresentado um gráfico de barras que agrupou os 60 primeiros casos em 10 grupos de interações, para que o comportamento do bot pudesse ser visualizado de forma particionada sem olhar para os erros do passado e apenas na faixa de observação:

Na Figura 13, pode-se observar que os erros de forma específica só aparecem quando existe uma grande variação linguística inserida no decorrer do processo. No começo, existem muitos erros, que vão se transformando em dúvida, e depois existem períodos que quase não há erro, e os que vão surgindo são as variações linguísticas que o bot não possuia e que com o feedback passará a entender.

Figura 13 – Gráfico de barras dos Erros, Dúvidas e Certezas do bot de forma particionada



Fonte: o autor.

5 Conclusões

Neste trabalho de conclusão de curso foi considerado um ponto de vista diferenciado para o desenvolvimento de *chatbots*. Com a inclusão dos conceitos de Semiótica Computacional e de Cibernética, foi possível estabelecer um modelo que pode ser importado para diversos *chatbots* e que pode contribuir com algumas das limitações atuais nessas ferramentas, tais como tratamento de sinônimos e inclusão facilitada de saberes de especialistas.

Também foi apresentado a possibilidade de que um sistema computacional é capaz de atender um paciente, e realizar o processo de anamnese, podendo auxiliar profissionais de saúde removendo as limitações de horário de trabalho, distância e desgaste.

Também propôs uma tecnica de classificação simples, porém útil e capaz de apresentar bons resultados. Essa técnica utilizou conhecimentos de mineração textual e bases de outros classificadores. É de suma importância também citar que o sistema produzido requer pouquíssimos parâmetros do ponto de vista de usuário final, onde para o paciente é apenas uma tela de diálogo e para o médico uma lista de entradas que não foram rotuladas com certeza pelo *bot* e que eles devem inserir apenas o rótulo, ou criar um novo rótulo caso desejem.

Em geral, o sistema apresentou resultados satisfatórios, no entanto acredita-se que por conta dos poucos elementos da base não se pode observar 100% do potencial do sistema para se adaptar a variações linguísticas.

Ao padronizar entradas textuais, utilizando como base definições definidas por médicos, o *chatbot* também pode se associar a outros sistemas, tais como sistemas de diagnósticos ou outros modelos preditivos, fornecendo parâmetros sólidos para que sistemas possam se interligar e realizar um bom processo de classificação.

Por fim, foi de grande importância a aplicação dos conceitos de semiótica para modelagem geral do projeto. Partindo da tríade de Peirce, foi possível determinar o que inicia o processo de semiose (a primeiridade) como a entrada do texto do usuário, uma qualificação dessa entrada (secundidade) através do pré-processamento, e por fim uma classificação através do conhecimento adquirido, sua conexão com o mundo (terceiridade).

6 Trabalhos Futuros

Existem inúmeras vertentes para trabalhos futuros nos quais este projeto pode ser desenvolvido e implementado. O processo de semiose desenvolvido por ele pode ser aplicado em outras áreas além da semiologia médica, por exemplo. Tendo em vista que esse *chatbot* apenas possui dependência dos rótulos, caso os rótulos sejam criados com um outro objetivo (um site de vendas de imóveis, ou tirar dúvidas referentes a um produto específico). Enfim, a contextualização do *chatbot* pode ser customizada de acordo com a necessidade.

Como pesquisas futuras, podem ser adotados os seguintes caminhos:

- Realizar estudos mais aprofundados com interação direta entre o usuário e um profissional da área de saúde de forma que este o valide de uma maneira mais formal. Além também de ser possível validar o mesmo modelo para outras áreas profissionais e também obter uma base de dados mais concreta.

- Aprimorar o processo de classificação: A técnica utilizada para realizar o procedimento de desenvolvimento do bot é bastante simples, apesar de interessante. Podem ser elaborados sistemas de classificação mais complexos visando uma maior precisão e também uma maior capacidade de identificação das palavras, e o *chatbot* também pode ser melhorado para realizar classificações multiobjetivas, sendo capaz de detectar diversos sintomas em uma mesma frase.

- Integração com outras plataformas de *chatbots*: Esse *chatbot* possui uma divisão modularizada entre o classificador e a parte de interface com o usuário. É possível facilmente acoplá-lo em um módulo *Node.JS* e utilizá-lo em *frameworks* bastante utilizados atualmente, como O *Messenger* do *Facebook* e o *Bot Framework* da *Microsoft*.

- Integração com sistemas de mineração de dados: O *Chatbot* pode ser adaptado para que o processo de classificação seja definido por um sistema mais completo e que seja capaz de tratar grandes massas de dados. Um exemplo seria uma integração entre o bot e o Pentaho.

- Integração com sistemas de diagnósticos: O *chatbot* pode servir também de parâmetro para sistemas que visam através de sintomas encontrar a doenças.

- Estudo e associação com o *Active Learning*, para se apropriar das vantagens da técnica e identificar possíveis futuros problemas.

- Integrar o *chatbot* com outras formas de pré-processamento textual, como por exemplo o *word2vec* e *GloVe*.

Referências

[1] DINO. *Chatbot é aposta para redução de custos de atendimento*. 2018. Exame. Disponível em: <https://exame.abril.com.br/negocios/dino/chatbot-e-aposta-para-reducao-de-custos-de-atendimento/>. Acesso em: 10 abr. 2018. Citado na página 12.

[2] WIKIPéDIA. 2017. Disponível em: <https://pt.wikipedia.org/wiki/Sintoma>. Acesso em: 10 abr. 2018. Citado na página 12.

[3] FILHO, R. N. P. *Computational Semiotics Applied to Medical Semiology*. Dissertação (Mestrado) — Universidade de Pernambuco, 2018. Citado 3 vezes nas páginas 13, 17 e 21.

[4] SANTAELLA, L. *O que é Semiótica*. [S.l.]: Editora Brasiliense, 2002. Citado na página 14.

[5] SANTANA, A. L. Disponível em: <https://www.infoescola.com/filosofia/semiotica/>. Acesso em: 03 jun. 2018. Citado na página 14.

[6] SHANNON, C. E. A mathematical theory of communication. 1948. Citado na página 15.

[7] MACHADO, V. R. I. Semiótica da comunicação: da semiose da natureza à cultura. *Revista FAMECOS*, 2010. Citado na página 15.

[8] MARTINS, W. A. Semiótica de charles peirce: O ícone e a primeiridade. *Revista Contemplação*, 2015. Citado na página 16.

[9] SIGNIFICADOS. Disponível em: <https://www.significados.com.br/semiologia/>. Acesso em: 03 jun. 2018. Citado na página 17.

[10] GUDWIN, R. R. From semiotics to computational semiotics. 2002. Citado na página 17.

[11] SANTOS, M. A. da S. Disponível em: <https://brasilescola.uol.com.br/informatica/inteligencia-artificial.htm>. Acesso em: 04 jun. 2018. Citado na página 17.

[12] GUDWIN, F. G. R. Computational semiotics : An approach for the study of intelligent systems part i : Foundations. 2002. Citado 2 vezes nas páginas 17 e 18.

[13] CHAVES, V. H. C. A revolução cibernética: a nova cultura. 2015. Citado na página 18.

[14] LIMA, B. V. de. Disponível em: <https://medium.com/botsbrasil/chatbots-e-o-reflexo-no-atendimento-ao-seu-cliente-875d3d1a6b9e>. Acesso em: 04 jun. 2018. Citado na página 19.

[15] MASTERCARD. Disponível em: <https://www.facebook.com/MastercardBrasil>. Acesso em: 14 jul. 2018. Citado na página 19.

[16] ERA. Disponível em: <https://www.facebook.com/eraimobiliaria>. Acesso em: 14 jul. 2018. Citado na página 19.

[17] SUVINIL. Disponível em: <https://www.facebook.com/TintasSuvinil>. Acesso em: 14 jul. 2018. Citado na página 19.

[18] VIANA, R. *Investigação do Processo de Stemming na Língua Portuguesa*. Dissertação (Mestrado) — Universidade Federal Fluminense, 2005. Citado 2 vezes nas páginas 23 e 24.

[19] RAMOS, J. Using tf-idf to determine word relevance in document queries. 2003. Citado na página 24.

ANEXO I

Código Fonte TF-IDF

```
const tfIdf = (doc, docs,term) => {
   const tf = 1.0 / doc.length;
   let idf = 1;
   let doclength = 0;
   let counter = 0;

   console.log("TF IDF DOCS", docs)
   for (doc in docs) {
       doclength++
       if (docs[doc].preprocessText.indexOf(term) > -1) {
           counter++;
       }
   }
   if (counter === 0 || counter === 1) {
       return tf
   }
   console.log(counter, doclength, tf)
   idf = Math.log(doclength / counter);

   console.log(idf)
   return tf * idf;

}

const tf = (doc) =>{
   return 1.0 / doc.length;
}

module.exports = {tf, tfIdf};
```

ANEXO II

Código Fonte Classificador

```
const
   stopword = require('./features/stopwords/stopwords'),
   stemmer = require('./features/stemmer/stemmer'),
   tfIdf = require('./features/tf-idf/tf-idf'),
   axios = require('axios'),
   Q = require('q'),
   removeAccents = require('remover-acentos');

const classifier = (text) => {
   const originalText = text;
   text = text.toLocaleLowerCase();
   text = removeAccents(text);
   const stopwordedText = stopword(text);
   const deferred = Q.defer();
   console.log('stopwordedText', stopwordedText);
   const stemmedText = stemmer(stopwordedText);
   axios.get(/*rota de obtenção de dados da base de feedback*/, { validate:
true })
       .then((results) => {
           const ObjectText = {
               text: originalText,
               preprocessText: stemmedText,
               category: ['Não definida'],
               validate: false,
               validationAuthor: 'BOT'
           };
           const docs = {};
           if (results.data) {
               Object.keys(results.data).forEach(key => {
                   if (results.data[key].validate) {
                       docs[key] = { ...results.data[key] }
                   }
               })
           }
           let bestLabel = 'Não definido';
           let bestValue = 0;
```

```
            for (doc in docs) {
                console.log(docs[doc])
                let currentValue = 0;
                let sumTf = 0;
                stemmedText.forEach(term => {
                    if (docs[doc].preprocessText.indexOf(term) > -1) {
                        console.log("TERM", term)
                        sumTf += tfIdf.tf(docs[doc].preprocessText);
                        currentValue += currentValue + tfIdf.tfIdf(stemmedText,
docs, term);
                    };

                });

                if (sumTf > 0.8) {
                    //caso a soma das frequencias dos termos seja maior que 80%
entao é quase certo que o termo já tenha sido validado anteriormente
                    //Então o termo é enviado como validado para o banco de
mensagens
                    axios.get('#adquirir dados do contador, apenas para
gráfico)
                        .then((results) => {
                            const counterValue = results.data
                            let updatedCounter = { ...counterValue }
                            updatedCounter.sure++;
                            axios.put('#rota atualizar contador',
updatedCounter);
                            axios.post('#rota atualizar série de dados',
updatedCounter);
                        });

                    bestLabel = docs[doc].category + " " + "certeza";
                    ObjectText.validate = true;
                    break;
                }


                if (currentValue > bestValue) {

                    bestValue = currentValue;
```

```
                    bestLabel = docs[doc].category;
                }
            }

            ObjectText.category = [bestLabel];
            sendMessageToDataBase(ObjectText);
            deferred.resolve(bestLabel)
        });
    return deferred.promise;
}

const sendMessageToDataBase = (obj) => {
    return axios.post('#rota para postar dados na base de dados', obj);
}

module.exports = classifier;
```

ANEXO III

Dados do formulário em sequência

Minha cabeça dói	Dói bem no final das costas	Estou com uma tosse seca	Estou sentindo muito frio	Sinto como se tivesse levado uma pancada na cabeça
Estou com enxaqueca	Estou com a lombar dolorida	Estou com tosse forte	estou com frio	estou tonto
Dor de Cabeça	Dor na Coluna	Tosse	Febre	Tontura
Dor de cabeça	Dor na lombar	Tosse	Resfriado	tontura
Estou sentindo uma dor na cabeça	Minhas costas estão doendo.	Estou com tosse	Estou com o corpo frio	Estou sentindo tonturas
Dor na cabeça	Dor na lombar	Estou tossindo bastante	Está muito frio	Estou tonto
Sinto dor de cabeça	Sinto dor nas costas	Estou tossindo	Acho que tenho febre	Minha cabeça está girando muito
Estou com dores muito fortes na cabeça	Estou sentindo dores na lombar	Estou tossindo bastante.	Sinto muito frio	Estou me sentindo tonta
minha cabeça dói	dói aqui atrás, nas costas, quando eu sento	sempre que tusso, dói na barriga	estou sentindo muito frio	estou tonto
Tô com uma dor de cabeça	Tem uma dor nas costas me incomodando	Eu to com essa tosse chata	Acho q to com febre	Essa tontura ta me matando
Estou com dor de cabeça	Estou com dor na lombar	Estou com tosse	Estou com calafrios	Estou com tontura
enxaqueca	Dores na região lombar	Tosse seca	Frio	Dor de cabeça
dor no quengo	dor na pleura	dor de barriga	frio	ta despombalizado
Dor de cabeça	Dos na lombar	Tosse	Frio	Tontura
Dor de cabeça	Dor nas costas	Dor de garganta	Febre	Tontura

ANEXO IV

Imagens utilizadas no formulário:

www.shutterstock.com · 753881815

www.shutterstock.com · 520189465

www.shutterstock.com · 503566702

Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação
Projeto de Final de Curso

ESCOLA POLITÉCNICA
DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 19 de julho de 2018, às 11:00 horas, reuniu-se para deliberar a defesa da monografia de conclusão de curso do discente **ERICKS DA SILVA RODRIGUES**, orientado pelo professor **Fernando Buarque de Lima Neto**, sob título **Semiótica Computacional para Desenvolvimento de Chatbots Cibernéticos para Detecção de Sintomas**, a banca composta pelos professores:

Fernando Buarque de Lima Neto
Marcelo Gomes Pereira de Lacerda

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

☒ Aprovada ☐ Aprovada com Restrições* ☐ Reprovada

e foi-lhe atribuída nota: 9,5 (Nove e meio)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O discente terá 7 dias para entrega da versão final da monografia a contar da data deste documento.

F. B. de Lima Neto

FERNANDO BUARQUE DE LIMA NETO

Marcelo Gomes P. de Lacerda

MARCELO GOMES PEREIRA DE LACERDA

* Este documento deverá ser encadernado juntamente com a monografia em versão final.