Centro Federal de Educação Tecnológica de Minas Gerais
Modelagem Matemática Computacional

Análise da correlação entre sentimentos e avaliações de qualidade sobre diálogos entre *chatbots* e seus usuários

Diego Ascânio Santos

Orientador: Prof. Dr. Gustavo Campos Menezes
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Daniel Hasan Dalip
Centro Federal de Educação Tecnológica de Minas Gerais

Belo Horizonte
Novembro de 2022

Diego Ascânio Santos

Análise da correlação entre sentimentos e avaliações de qualidade sobre diálogos entre *chatbots* e seus usuários

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional.

Linha de pesquisa: Sistemas Inteligentes.

Orientador: Prof. Dr. Gustavo Campos Menezes
Centro Federal de Educação Tecnológica de Minas Gerais

Coorientador: Prof. Dr. Daniel Hasan Dalip
Centro Federal de Educação Tecnológica de Minas Gerais

Centro Federal de Educação Tecnológica de Minas Gerais
Modelagem Matemática Computacional
Belo Horizonte
Novembro de 2022

Santos, Diego Ascânio
S237a Análise da correlação entre sentimentos e avaliações de qualidade sobre diálogos entre chatbots e seus usuários / Diego Ascânio Santos. – 2022.
66 f.

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Gustavo Campos Menezes.
Coorientador: Daniel Hasan Dalip.
Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1. Robôs de bate-papo – Teses. 2. Percepção de padrões – Teses. 3. Controle de qualidade – Teses. 4. Correlação (Estatística) – Teses. I. Menezes, Gustavo Campos. II. Dalip, Daniel Hasan. III. Centro Federal de Educação Tecnológica de Minas Gerais. IV. Título.

CDD 006.3

Elaboração da ficha catalográfica pela bibliotecária Jane Marangon Duarte, CRB 6º 1592 / Cefet/MG

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO CURSO DE MESTRADO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"ANÁLISE DA CORRELAÇÃO ENTRE SENTIMENTOS E AVALIAÇÕES DE QUALIDADE SOBRE DIÁLOGOS ENTRE CHATBOTS E SEUS USUÁRIOS"

Dissertação de Mestrado apresentada por **Diego Ascânio dos Santos**, em 30 de novembro de 2022, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Prof. Dr. Gustavo Campos Menezes
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Daniel Hasan Dalip
Centro Federal de Educação Tecnológica de Minas Gerais

Profª. Drª. Michele Amaral Brandão
Instituto Federal de Minas Gerais

Profª. Drª. Livia Maria Dutra
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. Dr. Thiago Magela Rodrigues Dias
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida à impressão,

Profª. Drª. Elizabeth Fialho Wanner
Presidenta do Colegiado do Programa de Pós-Graduação em
Modelagem Matemática e Computacional

Dedico este trabalho a todas as vítimas da pandemia da COVID-19, bem como, a todos(as) os(as) colegas do CEFET-MG que pereceram durante este período, vitimados ou não por esta doença, em especial ao saudoso professor Dr. Paulo Eduardo Maciel de Almeida.

Agradecimentos

Agradeço ao Criador pela graça de desfrutar desta etapa na caminhada de cumprimento da sentença sobre esta terra.

Agradeço aos meus ancestrais e principalmente aos meus velhos que, contrariando todas as estatísticas, derramaram sangue, suor e lágrimas para que eu chegasse até aqui. Não existem palavras e atos que sejam suficientes para retribuir proporcionalmente a paciência, o carinho, a compreensão, o amor e o cuidado que tiveram comigo. Tenho esta dívida impagável para sempre.

Agradeço a meus orientadores Gustavo Campos Menezes e Daniel Hasan Dalip por todo apoio, paciência, compreensão e por não terem desistido de mim, apesar de todas as debilidades, fraquezas e vícios que me afetaram neste período conturbado da pandemia. Minha gratidão eterna a vocês e meu pedido de desculpas por todos os distúrbios que meus transtornos causaram.

Agradeço a meus amigos de caminhada do CEFET-MG, especialmente aos senhores (e senhoras) Marcos Prado Amaral, Luís Fernando Linhares, Joice Ferreira Rodrigues Souza, Ítalo José Sena Costa, Patrícia Rodrigues Gomes Fidalgo, Anderson Rodrigues Gomes e efusivamente a meu camarada Renato Stangherlin Castanheira. Todos vocês contribuíram imensamente para que pudesse alcançar este resultado, jamais teria conseguido sem o apoio de vocês.

Agradeço a meu grupo de amigos de infância — Fodetrões — destacando a figura do meu amigo Thales Nascimento Xavier que durante este mestrado esteve muito próximo a mim, dando um apoio moral (afetivo e psicológico) essencial para chegar até aqui.

Agradeço a minha mãe, meu pai, minha irmã, meu cunhado, meus afilhados e meus avós pelas orações e pelo sustento na fé (e no pragmatismo), pois, vocês moveram o sobrenatural (e o ordinário) para me tirar do fundo do poço.

Minhas palavras são insuficientes para expressar gratidão e carinho pelo CEFET-MG. Amo este lugar e é um privilégio incalculável ter minha história (con)fundida com a desta casa. Catorze anos seguidos de muito tesão, muita alegria, muita garra, dedicação, entrega e de muita loucura para de alguma forma, mesmo que da forma mais heterodoxa de todas, retribuir todo o bem (e todo o mal) que esta casa fez a mim (e eu a ela).

Agradeço imensamente ao Instituto Federal do Espírito Santo (IFES) pelo fomento ao projeto Oficinas 4.0, que possibilitou construir grande parte desta dissertação, através do valoroso apoio de todos os componentes deste projeto: alunos do CEFET-MG, os professores Daniel Hasan Dalip, Lívia Maria Dutra, Gustavo Campos Menezes e Sandro Renato Dias, bem como, a empresa Take Blip e seus funcionários (e meus amigos) Arthur Iperoyg Rodrigues Carvalho, Renan Santos Mendes e Gabriel Oliveira Assunção.

A todos vocês que mencionei acima, quando o abismo parecia inevitável, vocês (e o Criador) me estenderam a mão e não me deixaram cair. Obrigado por várias e várias vezes me salvarem. Por fim, agradeço a quem sacrificou uma noite de sono para me auxiliar a escrever o pré-projeto que possibilitou meu ingresso no PPGMMC.

"Não, não pares. É graça divina começar bem. Graça maior, persistir na caminhada certa, manter o ritmo... mas a graça das graças é não desistir, podendo ou não podendo, caindo, embora, aos pedaços, chegar até o fim. "

(Servo de Deus dom Helder Pessoa Câmara)

Resumo

Chatbots com boa qualidade correspondem às expectativas de seu funcionamento e assim, aumentam a retenção, engajamento e satisfação de seus usuários. A opinião dos usuários — medida através de questionários — é utilizada para mensurar esta qualidade, mas, possui limitações como métrica de qualidade tais quais: ser sucetível a viéses e avaliações displicententes, bem como, não poder ser aferida em tempo real. Para lidar com estas limitações, existem atributos de qualidade que as mitigam. O sentimento textual é um destes atributos, pois, o sentimento contém informações implícitas que mitigam avaliações displicentes, além de poder ser inferido em tempo real. Destarte, a partir de referências que exploram o sentimento textual como apromixador da qualidade de *chatbots* são realizados experimentos de correlação entre estas duas variáveis sobre o corpo de diálogos (entre humanos e *chatbots*) ConvAI, mencionado por uma das referências — cuja correlação entre sentimento e qualidade dos diálogos não foi abordada — atestando por fim que existe uma correlação entre sentimento textual e a qualidade dos chatbots do ConvAI: ρ-spearman = 0,3221, valor-p << 0,001. Outros resultados incluem: uma ferramenta de rotulação construída para realizar reclassificação de qualidade e anotação de sentimentos dos diálogos, bem como, o reforço da prática de reclassificação de diálogos por meio de comitês, prática presente na literatura.

Palavras-chave: *Chatbot*, Sentimento Textual, Qualidade de *Chatbots*, Correlação.

Abstract

Chatbots with good quality are able to maintain service expectations and therefore, increase users' retention, engagement, and satisfaction. Users' opinion — available through surveys — is considered to measure chatbots' quality, but, it has several limitations such as: being subject to biases and careless evaluations from users, as well, the impossibility to being measured in real-time. To address these limitations, there are quality attributes able to mitigate them. Textual sentiment is one of those attributes, since textual sentiment has implicit details which mitigate careless evaluations and can be measured in real-time. Therefore, through literature references that investigate the correlation between textual sentiment and chatbots' quality, this work performs correlation experiments over a specific dialogue corpora mentioned by one of the references (ConvAI) — whose correlation results between sentiment and dialog quality weren't shown — assuring that there is a correlation between textual sentiment and chatbots' quality over this specific dialogue dataset: ρ-spearman = 0,3221, p-value << 0,001. Other results include a labeling tool built to perform quality reclassification and sentiment anotation, as well, the reinforcement of subsequential dialogues evaluation through comitees, as performed in literature.

Keywords: Chatbots, Sentiment Analysis, Sentiment Anotation, Chatbots' Quality, Correlation

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Sumário

1 Introdução

Mensurar a qualidade de *chatbots*[1], vinculada ao desempenho que apresentam, é tarefa essencial, pois, agentes conversacionais que apresentem boa qualidade serão capazes de corresponder demandas apresentadas por usuários e consequentemente, aumentar a retenção, satisfação e engajamento de seus utilizadores (KIM; LEVY; LIU, 2020; FEINE; MORANA; GNEWUCH, 2019). Aplicações na indústria e na academia frequentemente mensuram a qualidade desses agentes através da opinião dos usuários (ou de avaliadores humanos externos) manifestada em respostas a questionários sobre aspectos desejáveis no desempenho de *chatbots* após o fim de atendimentos (LIANG; ZOU; YU, 2020; KIM; LEVY; LIU, 2020; FEINE; MORANA; GNEWUCH, 2019).

Apesar de serem bastante utilizados, esses questionários apresentam limitações para aferir qualidade dos agentes, tais quais: suscetibilidade a viéses, suscetibilidade a respostas displicentes por parte dos usuários e impossibilidade de fornecer informações de qualidade durante os atendimentos — em tempo real (LIANG; ZOU; YU, 2020; KIM; LEVY; LIU, 2020; LOGACHEVA et al., 2018). Ainda assim, a ubiquidade desses questionários para aferir qualidade de *chatbots* (como também qualidade de outros sistemas e serviços em geral) faz com que sejam obrigatórios para a finalidade que se destinam. Dessa forma, é necessário buscar por aspectos complementares de qualidade que possam, junto a esses questionários, mitigar essas limitações elencadas.

O sentimento textual, como verificado na literatura — (TAUSCZIK; PENNEBAKER; NERBONNE; KÜSTER; KAPPAS, 2010, 2014, 2017 apud FEINE; MORANA; GNEWUCH, 2019) — , é um aspecto complementar da qualidade dos agentes conversacionais visto que é correlato à qualidade de atendimentos de *chatbots*, bem como, pode ser inferido em tempo real, mitigando assim: a suscetibilidade a viéses de displicência e a impossibilidade de aferir qualidade durante os atendimentos. Assim, a seguinte pergunta de pesquisa é apresentada: **O sentimento textual de diálogos entre humanos e *chatbots* oferece indicativos da qualidade de desempenho?**

Feine, Morana e Gnewuch (2019), Kim, Levy e Liu (2020) indicam que sim, por meio de experimentos que atestam a correlação entre qualidade atribuída por usuários e sentimento textual contido nas mensagens que enviam a *chatbots*. Para realizar esses experimentos, os autores usaram três corpos de diálogos cujas conversas foram avaliadas por questionários: ConvAI (LOGACHEVA et al., 2018) e ExpCorpus (GNEWUCH et al., 2020), utilizados por Feine, Morana e Gnewuch (2019), e *Amazon Alexa Prize Challenge* (KHATRI et al., 2018), utilizado por Kim, Levy e Liu (2020).

Sumarizando experimentos efetuados pelas referências, Feine, Morana e Gnewuch (2019)

[1] Agentes autônomos concebidos como programas de computador capazes de se comunicar e produzir diálogos com humanos através de linguagem natural (SHAWAR; ATWELL, 2007).

restringiram o escopo da investigação da correlação entre qualidade e sentimento ao corpo de diálogos ExpCorpus, utilizando o corpo ConvAI, que também contém avaliações de qualidade aptas a terem correlações com sentimentos textuais estudadas, apenas para testar algoritmos de análise de sentimento. Kim, Levy e Liu (2020) avaliaram a correlação entre qualidade e sentimento de conversas ocorridas através de mensagens de voz trocadas entre usuários e a assistente virtual *Alexa*. Por conterem dados oriundos de consumidores finais, que possívelmente possuem restrições para compartilhamento, a base de conversas utilizada pelos autores — *Amazon Alexa Prize Challenge* — não está disponível publicamente para realização de experimentos, entretanto, a partir da leitura de Kim, Levy e Liu (2020), é possível extrair boas práticas metodológicas para a condução de experimentos em outras bases de diálogos.

Destarte, esse trabalho tem por proposta verificar se a resposta à pergunta de pesquisa — **O sentimento textual de diálogos entre humanos e *chatbots* oferece indicativos da qualidade de desempenho?**[2] — se sustenta sobre o corpo de diálogos ConvAI que, até o momento, não possui — na literatura — resultados apresentados relativos a esse fato.

Ao cumprir essa proposta, esse trabalho atesta que sentimento e qualidade também são correlacionados no corpo ConvAI, porém, de forma impactada pela presença de viéses e displicências nas avaliações de qualidade efetuadas pelos usuários dos *chatbots* pertencentes a ele.

Esse impacto causado por viéses e avaliações displicentes pode explicar o fato de Feine, Morana e Gnewuch (2019) não terem apresentado os resultados de correlação entre qualidade e sentimento sobre o corpo de diálogos ConvAI em seu trabalho, sendo essa, a principal contribuição dessa dissertação, na forma de uma investigação científica (e reprodução experimental) do trabalho desses autores, apontando melhorias metodológicas que podem ser aplicadas em contextos similares.

1.1 Objetivos

O objetivo desse trabalho — para verificar se o sentimento textual de diálogos entre humanos e *chatbots* oferece indicativos da qualidade de desempenho — consiste em investigar a correlação entre qualidade e sentimento textual presente em diálogos produzidos por humanos e *chatbots* a partir de:

- Estudo dos principais conceitos e de trabalhos que avaliam qualidade de *chatbots*;
- Investigação dos experimentos de correlação entre qualidade e sentimento realizados por Feine, Morana e Gnewuch (2019), Kim, Levy e Liu (2020);
- Realização de experimentos de correlação entre sentimento e qualidade sobre o corpo de diálogos ConvAI (LOGACHEVA et al., 2018) através de:

[2] verificada como positiva por Feine, Morana e Gnewuch (2019), Kim, Levy e Liu (2020) em seus respectivos contextos.

- Filtragem e seleção de diálogos avaliados displicentemente, conforme recomendado por Logacheva et al. (2018) e praticado em contextos similares da literatura;
- Reclassificação de qualidade (LOGACHEVA et al., 2018; KIM; LEVY; LIU, 2020) e anotação manual de sentimentos textuais (KIM; LEVY; LIU, 2020) das conversas;
- Cálculo das correlações entre qualidade e sentimento dos diálogos antes e depois das reclassificações de qualidade;

1.2 Contribuições

As contribuições desse trabalho visam corroborar o que empiricamente foi afirmado por outras referências — de que o sentimento textual e a qualidade são correlacionados — aplicado ao contexto de diálogos presentes no corpo ConvAI, o que até o momento, no melhor dos esforços de verificação empreendidos, não é demonstrado por outras referências, apontando por fim, melhorias em práticas metodológicas que podem ser aplicadas em contextos similares. Além disso, o trabalho também disponibiliza uma ferramenta de rotulação de diálogos que pode ser utilizada para tarefas de reclassificação de conversas de outros corpos textuais, bem como, disponibiliza o corpo de diálogos ConvAI com diálogos revisados (e sentimentos anotados) por especialistas humanos — como recomendado por Logacheva et al. (2018) — com resultados de correlação entre sentimento e qualidade dos diálogos pertencentes ao corpo antes e depois da revisão efetuada.

1.3 Organização do trabalho

Após essa introdução, estão presentes no Capítulo 2 os principais conceitos que embasam o desenvolvimento desse trabalho e no Capítulo 3, trabalhos que utilizam questionários para mensurar qualidade de diálogos produzidos por *chatbots* e humanos, bem como, os trabalhos de Feine, Morana e Gnewuch (2019), Kim, Levy e Liu (2020), principais referências (até o momento) que estudam a correlação entre qualidade e sentimento no contexto de *chatbots*.

O Capítulo 4 apresenta os experimentos e artefatos realizados para estudar a correlação entre sentimento e qualidade de diálogos sobre o corpo ConvAI, cujos resultados, que corroboram a existência de correlação entre qualidade e sentimento, encontram-se relatados no Capítulo 5. Por fim, são apresentadas conclusões e perspectivas de trabalhos futuros no Capítulo 6.

2 Fundamentação Teórica

Estão dispostas neste capítulo as principais formulações téoricas que embasam o teor deste trabalho. A Seção 2.1 apresenta uma definição sobre o que são *Chatbots* e a contextualização histórica do surgimento desses agentes. A Seção 2.1.1 apresenta considerações sobre aspectos de qualidade que permeiam *Chatbots*, a Seção 2.2 apresenta escalas de medida e tipos de variáveis que podem ser utilizados para mensurar qualidade de *chatbots*, sendo dedicada uma Seção específica (2.2.1) para discussões aprofundadas sobre escalas Likert, métricas tradicionalmente associadas na aferição de qualidade de *chatbots*. A Seção 2.3 contém definições e considerações sobre o sentimento, com destaque para o sentimento textual, objeto de estudo do presente trabalho em investigações que buscam inquirir sua associação a qualidade de *chatbots* por meio de ferramentas de correlação apresentadas na Seção 2.4.

2.1 *Chatbots*

Chatbots são agentes autônomos (robôs) concebidos como programas de computadores capazes de se comunicar e produzir diálogos com seres humanos por meio de linguagem natural (SHAWAR; ATWELL, 2007). Esses agentes podem ser disponibilizados em vários canais de comunicação como programas de mensagens instantâneas (*Telegram*, *WhatsApp*, *Signal*), *websites* ou redes sociais (*Facebook*, *Twitter*, *Instagram*), sendo acessíveis por meio de múltiplos dispositivos (*smartphones*, computadores, dentre outros). Por serem autônomos, podem realizar atendimentos de forma ininterrupta (24 horas por dia, 7 dias por semana) sendo vantajosos (sob a perspectiva de custos operacionais) quando comparados a atendentes humanos. Nesse sentido, apresentam um grande potencial econômico para empresas e negócios que precisem manter canais de atendimento/vendas a seus clientes (MCTEAR; CALLEJAS; GRIOL, 2016 apud FEINE; MORANA; GNEWUCH, 2019, p. 5).

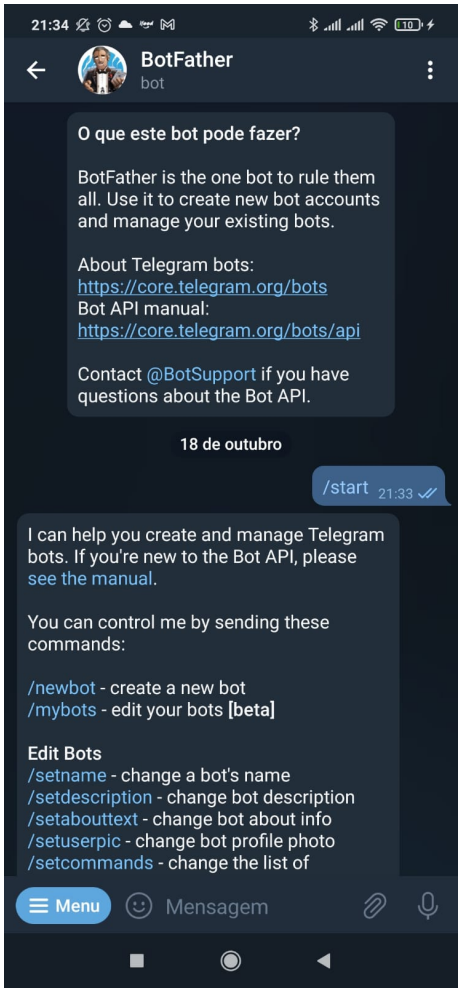
A origem dos *chatbots* remete ao trabalho do célebre matemático Alan Turing. Considerado o pai da ciência da computação, Turing defendeu que a forma de processamento dos computadores evoluiria a um ponto comparável à capacidade cerebral humana e para verificar essa capacidade propôs em seu artigo *Computing Machinery and Intelligence* um teste conhecido como Jogo da Imitação capaz de identificar habilidades de pensamento e de reprodução do comportamento humano por parte de autômatos, em que um autômato poderia ser considerado inteligente quando em uma conversa natural com um avaliador humano, o avaliador não conseguisse distinguir se conversava com uma máquina ou outro humano (TURING, 1950 apud AMORIM, 2014).

Assim, o teste de Turing motivou a construção de autômatos que conseguissem "ludibriar" julgadores humanos em uma conversa natural, originando assim os primeiros *chatbots*. Passar nesse teste se tornou (inicialmente) um parâmetro de qualidade ao qual os *chatbots* deveriam

satisfazer e os primeiros desenvolvidos buscavam apresentar esse parâmetro, tais quais: *ELIZA*[1], *Parry*[2], *Jabberwacky*[3], dentre outros.

Esse teste foi durante muito tempo a principal (quiçá única) métrica de qualidade dos chatbots, entretanto, Shawar e Atwell (2007) apontam também ser possível analisar o desempenho de um *chatbot* pela sua capacidade de auxiliar uma pessoa a alcançar determinado objetivo (agendar uma reserva de hotel) ou de responder corretamente a um questionamento levantado por ela (a temperatura ambiente, a cotação diária do dólar, etc.). Além das capacidades levantadas pelos autores, outros fatores também podem ser considerados para avaliar a qualidade de um *chatbot* e são abordados na próxima Seção. Por fim a Figura 1 mostra um exemplo de *chatbot* fornecido pelo aplicativo *Telegram*.

Figura 1 – *Chatbot Botfather*, responsável pela criação de demais *chatbots* existentes no *Telegram*



[1] Weizenbaum (1966)
[2] Colby (1999)
[3] Carpenter (2007)

2.1.1 Qualidade de *Chatbots*

Medir a qualidade de um *chatbot* em suas interações com humanos é uma atividade multifacetada, pois, a qualidade é uma variável subjetiva, ou seja, assume diferentes características a partir da perspectiva pela qual é considerada.

Em uma abordagem orientada a dados textuais genéricos Tejay, Dhillon e Chin (2004) afirmam que a qualidade desses dados pode ser organizada em níveis semióticos: sintático, semântico e pragmático.

Por serem *chatbots* entidades produtoras de dados textuais em suas interações com humanos, é razoável considerar tais níveis semióticos para extrair características de qualidade inerentes às suas interações? Antes de discorrer sobre tal consideração, é necessário definir o próprio conceito de semiótica para uma compreensão adequada a respeito do que cada nível mencionado representa e do porquê podem ser utilizados para extrair características de qualidade inerentes aos *chatbots*.

De acordo com Palatini e Zischner (2014) e pela definição do dicionário *LEXICO*[4], a semiótica pode ser definida como o "estudo de signos, símbolos, seus usos e interpretações". Letras são exemplos de signos (e símbolos) e palavras (e frases) são exemplos de usos e interpretações. Portanto, são objetos de estudo da semiótica, assim como os dados produzidos por *chatbots*, por serem constituídos por palavras representadas em linguagem natural.

Demonstrado isso, é atestada a razoabilidade de se aplicar os níveis semióticos para a análise da qualidade de *chatbots*[5], sendo necessário explicar o que significam, para situar suas contribuições em relação à qualidade textual. Recapitulando Tejay, Dhillon e Chin (2004), esses níveis são categorizados em três: sintático, semântico e pragmático e representam respectivamente a como um texto é apresentado, ao seu significado e às intenções que seus interlocutores possuem.

Como essas representações ainda são abstratas, aspectos concretos vinculados a cada nível se fazem necessários. Para o nível sintático, aspectos de corretude (respeito às normas gramaticais vigentes) são dimensões concretas para determinação de qualidade. Para o nível semântico, aspectos de ambiguidade e sentimento (do sentido textual) e para o nível pragmático, aspectos de utilidade (se os interlocutores alcançaram os objetivos esperados na interação) representam dimensões concretas de qualidade sob a perspectiva de cada um desses dois níveis. No contexto dos comunicadores autônomos, aspectos de corretude (sintáticos) podem ser utilizados para verificar se erros ortográficos impactam as respostas fornecidas aos usuários, se o *chatbot* pode responder adequadamente quando um usuário envia mensagens contendo erros ortográficos ou ainda, se o agente pode responder (enviar) mensagens na norma-padrão. Já os aspectos de sentimento (semânticos) fornecem indícios associados a satisfação de um usuário a partir do conteúdo de sua resposta: sentimentos positivos tendem a estar associados a interações satisfatórias e negativos a interações insuficientes. Por fim, o aspecto de utilidade permite verificar

[4] SEMIOTICS (2021)

[5] Bem como a qualquer circunstância que envolva representações textuais, como em Dalip et al. (2017) que propõem métricas de qualidade de documentos colaborativos criados na *Web* 2.0 baseadas nos níveis semióticos

o quão útil a resposta de um *chatbot* é para auxiliar seu usuário a atingir determinado objetivo.

Estabelecidos os níveis semióticos e como podem contribuir para a definição da qualidade de interações de autômatos conversacionais, é necessário estudar outros possíveis indicadores de qualidade vinculados a *chatbots*. A partir disto, RADZIWILL e BENTON (2017) realizaram um *survey* a respeito destes indicadores, tendo identificado três principais: **eficácia**, **eficiência** e **satisfação**. Em relação a cada um desses indicadores, é possível inferir que:

- A **eficácia** representa quando um *chatbot* é tolerante a falhas e é robusto, conseguindo apresentar uma resposta relevante ao usuário mesmo para diferentes entradas e formas de escrita (COHEN; LANE, 2016; THIELTGES; SCHMIDT; HEGELICH, 2016; MORRISSEY; KIRAKOWSKI, 2013).
- A **eficiciência** está relacionada a quando um *chatbot* usa uma linguagem apropriada e gramaticalmente correta além de uma forma de interação parecida com um humano - se for necessária (EEUWEN, 2017; MORRISSEY; KIRAKOWSKI, 2013; WEIZENBAUM, 1966; WALLACE, 2003).
- A **satisfação** está relacionada à capacidade de um *chatbot* apresentar informações úteis e também é relacionada a aspectos de ética e de comportamento que um *chatbot* apresenta (THIELTGES; SCHMIDT; HEGELICH, 2016; NEFF; NAGY, 2016; APPLIN; FISCHER, 2015).

Quando os *chatbots* são orientados a objetivos — responsáveis por realizar atendimentos — a satisfação é essencial na qualidade das interações entre os agentes e seus usuários visto que por meio dela, o usuário expressa se recebeu uma resposta adequada a sua demanda, sendo, portanto, indispensável para o sucesso (a longo prazo) das relações entre clientes e empresas provisoras desses agentes (AU; NGAI; CHENG; OLIVER; JONES; SASSER et al., 2002, 2014, 1995 apud FEINE; MORANA; GNEWUCH, 2019).

Para *chatbots* não orientados a objetivos (conversacionais, por exemplo) a satisfação também é um fator importante, mas, não necessariamente determinante, pois, atributos relativos à eficácia e eficiência possuem igual relevância, como em Logacheva et al. (2018), em que atributos como profundidade do diálogo (respostas relevantes aos usuários — **eficácia**) e engajamento do agente (aspectos de interação — **eficiência**) são considerados para a avaliação da qualidade das interações.

Por fim, é possível e suficiente — para os propósitos desse trabalho — aferir a qualidade dos atendimentos de *chatbots* por meio das respostas de seus próprios usuários a questionários relacionados aos indicadores discutidos, como realizado em Logacheva et al. (2018), Feine, Morana e Gnewuch (2019), Kim, Levy e Liu (2020) (aprofundados nos Trabalhos Relacionados). E, tais respostas são variáveis quantificadas em escalas de medidas abordadas na próxima Seção.

2.2 Escalas de Medida

Uma escala de medida (ou nível de medida) é um sistema de classificação que descreve a natureza de dados associados a variáveis (objetos, assuntos, criaturas, conceitos) permitindo ao menos agrupá-las a partir das características que apresentam (KIRCH, 2008; CAVALCANTE; SOUSA; SOUSA, 2019). Escalas de medida podem ser utilizadas para classificar / estimar comporamentos humanos (com aplicações nos campos das ciências sociais e da psicologia), medir grandezas físicas (como temperatura, distância, tempo e peso) e, em geral, distinguir / quantificar conceitos que permeiam a vida cotidiana (como o dinheiro, a catalogação de endereços, a numeração de residências, etc). Elas podem ser categorizadas em quatro tipos: **nominal**, **ordinal**, **intervalo** e **razão**, servindo para mensurar variáveis qualitativas e quantitativas (STEVENS et al., 1946 apud TREIBLMAIER; FILZMOSER, 2011).

Variáveis qualitativas (ou categóricas) são aquelas cujas propriedades podem ser categorizadas ou ordenadas, mas, não podem ser discriminadas numericamente por não possuírem quantificantes de intensidade formalmente estabelecidos como, por exemplo, classes gramaticais, gênero, espécies de animais, níveis de satisfação, dentre outras, cujas distinções ou ordens podem ser estabelecidas, mas, não mensuradas[6]. Já variáveis quantitativas são variáveis de representações essencialmente numéricas, como peso, idade, altura, temperatura, dentre outras, que podem ser discriminadas numericamente por possuírem quantificantes de intensidade (intervalos) formalmente estabelecidos e igualmente distanciados, apresentando por isso, maior facilidade de entendimento e mensuração (MELO; STEINKE, 2014; SOARES; SIQUEIRA, 2002; STEVENS et al., 1946; MANGIAFICO, 2016).

Escalas nominais mensuram variáveis qualitativas não ordenadas e são as escalas mais simples existentes por disporem somente de operação matemática de igualdade que limita o escopo da comparação dos valores assumidos pelas variáveis a um procedimento de agrupamento em duas ou mais categorias disjuntivamente exclusivas[7] (COOPER; BLUMBERG; SCHINDLER, 2014 apud DALATI, 2018). Por mensurar variáveis não ordenadas, a operação de medida das escalas nominais também é definida como categorização (classificação) e as variáveis medidas por essas escalas são aquelas em que não faz sentido existir qualquer tipo de ordem tais quais: gênero, espécies de animais, classes gramaticais e também, variáveis que usem rótulos numéricos sem qualquer implicação quantitativa como CEP (códigos de endereço postal), números em camisas de jogadores de futebol, índices em catálogos, dentre outras (DALATI, 2018; MANGIAFICO, 2016; STEVENS et al., 1946). Por finalidade explicativa, exemplos de escalas nominais e de variável qualitativa que utiliza rótulo numérico são apontados pelas Figuras 2 e 3 respectivamente.

[6] Tomando como exemplo a variável qualitativa **Avaliação da Administração do Atual Prefeito**, extraída de Melo e Steinke (2014) e que pode assumir valores como **excelente, ótima, boa, regular, ruim, muito ruim e péssima**, é possível captar a diferença de avaliações emitidas entre cidadãos distintos — um avalia como **ótima** e outro como **regular** — mas não é possível definir exatamente qual a diferença da intensidade entre a opinião de ambos, apenas que o primeiro aprova a administração do atual prefeito mais do que o segundo.

[7] Para cada mensuração de uma variável qualitativa não ordenada em uma determinada escala nominal, tal variável pode pertencer a apenas uma única categoria dentre duas ou mais categorias descritas pela escala, pois, tais categorias são disjuntivamente exclusivas.

Figura 2 – Exemplos de escalas nominais que definem os valores que as variáveis qualitativas sexo e cor dos olhos podem assumir.

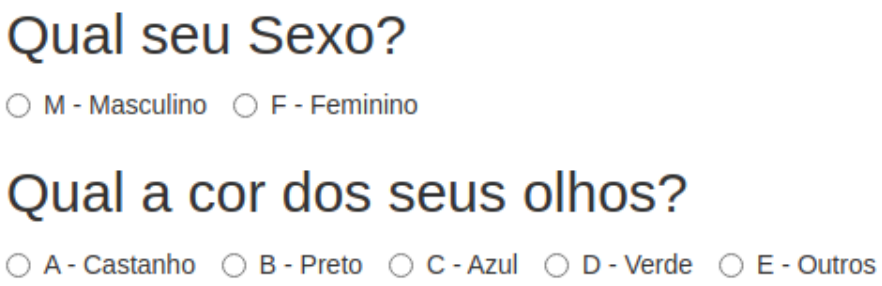


Figura 3 – Exemplo de variável qualitativa representada pelo número da camisa do ex-atleta Wayne Rooney (antigo jogador do Manchester United F.C.)

Fonte: Dalati (2018)

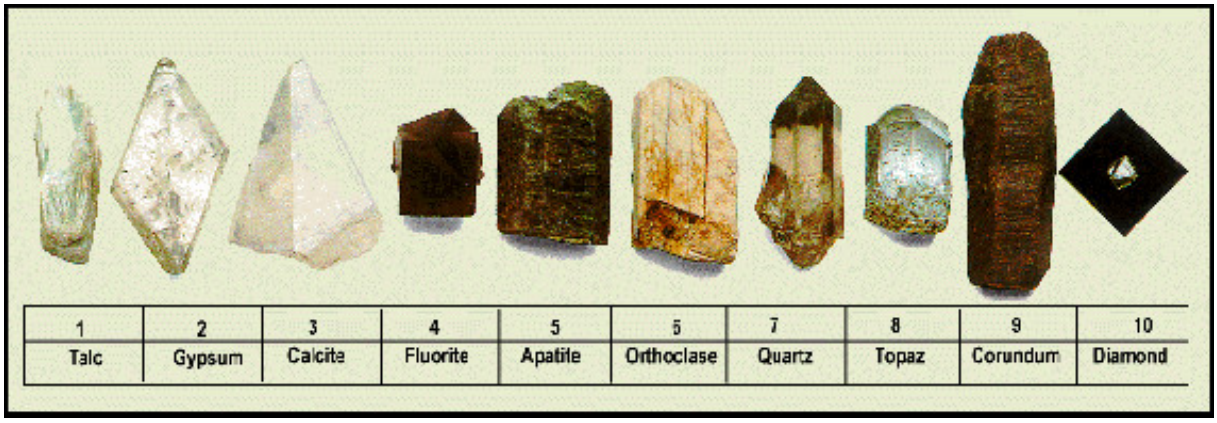
Escalas ordinais são escalas que classificam e ordenam variáveis qualitativas a partir da significância que possuem quando apresentam alguma relação capaz de ser ordenada (ZIKMUND; CARR; GRIFFIN; ZIKMUND, 2013, 1991 apud DALATI, 2018). De acordo com Stevens et al. (1946), Dalati (2018), essas escalas emergem a partir da operação matemática da comparação, que pode determinar se uma variável é maior (ou menor) do que a outra, mas, sem quantificar o quão maior (ou menor) é essa variável.

São exemplos destas escalas, escalas não absolutas tais quais: escala de Mohs de dureza mineral e escalas regionais de tamanhos de calçados — ilustradas respectivamente pelas Figuras 4 e 5 — em que se verifica ser possível ordenar, por meio da comparação, variáveis distintas mensuradas por essas escalas, mas, é impossível quantificar a diferença de magnitude existente entre tais variáveis[8].

Escalas de intervalo mensuram variáveis quantitativas e existem para permitir o cálculo de diferenças (e distâncias) existentes entre essas variáveis, pois, tais escalas possuem intervalos iguais e afixados entre as múltiplas representações que definem. O calendário gregoriano, ilustrado pela Figura 6, é um exemplo de escala de intervalo, pois, o intervalo existente entre dois dias seguidos nesse calendário — 1 dia / 24 horas — permanece invariável para quaisquer pares de dois dias seguidos representados nessa escala. Outros exemplos de escalas de intervalo são

[8] Um calçado nº 8 na escala americana é maior que outro nº 4, mas, a diferença de tamaho entre esses calçados não é quantificável, pois, o calçado nº 8 não apresenta o dobro do tamanho do calçado nº 4 visto que em termos absolutos possuem respectivamente 26cm e 22cm de tamanho.

Figura 4 – Escala de Mohs de Dureza Mineral



Fonte: In (2014)

Figura 5 – Múltiplas escalas regionais de tamanhos de calçados e suas correspondências absolutas em centímetros e polegadas

U.K.	U.S.	E.U.	Japan	CM	IN
2	3	34.5	210	21.0	8 1/4
3	4	35.5	220	22.0	8 5/8
4	5	37	230	22.9	9
5	6	38	240	24	9 1/2
6	7	39.5	250	25	9 7/8
7	8	40.5	260	26	10 1/4
8	9	42	270	27	10 5/8
9	10	43	280	28	11
10	11	44.5	290	29	11 3/8
11	12	45.5	300	30	11 3/4
12	13	47	310	31	12 1/4
13	14	48.5	320	32	12 5/8
14	15	49.5	330	33	13
15	16	50.5	335	33.5	13 1/4

Fonte: Dalati (2018)

as escalas de temperatura *Fahrenheit* (F) e *Celsius* (C) que também dispõem de distâncias invariáveis entre as múltiplas graduações que possuem (STEVENS et al., 1946; DALATI, 2018).

Por disporem de intervalos afixados e medirem variáveis lineares genéricas, da forma $x' = ax + b$, as escalas de intervalo permitem que operações de soma e subtração sejam realizadas entre essas variáveis, bem como, comparações a padrões, médias aritméticas e desvios padrões, operações matemáticas indisponíveis sobre as escalas nominal e ordinal (STEVENS et al., 1946).

Escalas de razão segundo Dalati (2018), Stevens et al. (1946) representam o tipo de escala de medida mais completo dentre os quatro abordados, pois, permitem classificar informações como as escalas nominais, ordenar valores como as escalas ordinais, quantificar / computar diferenças como as escalas de intervalo e por fim, estimar proporções entre magnitudes distintas de variáveis do mesmo tipo por abrangerem o conceito de **zero absoluto**, que indica a total ausência da variável a ser medida. Assim como as escalas de intervalo, também mensuram variáveis quantitativas, mas diferem destas, pelo fato de que o zero, por ser absoluto, apresenta relevância de magnitude (e não apenas de convenção, como nas escalas de intervalo) e por isso,

Figura 6 – Calendário gregoriano

CALENDÁRIO GREGORIANO

Janeiro							Fevereiro							Março							Abril							Maio							Junho						
1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
8	9	10	11	12	13	14	8	9	10	11	12	13	14	8	9	10	11	12	13	14	8	9	10	11	12	13	14	8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	19	20	21	15	16	17	18	19	20	21	15	16	17	18	19	20	21	15	16	17	18	19	20	21	15	16	17	18	19	20	21
22	23	24	25	26	27	28	22	23	24	25	26	27	28	22	23	24	25	26	27	28	22	23	24	25	26	27	28	22	23	24	25	26	27	28	22	23	24	25	26	27	28
29	30	31					29							29	30	31					29	30						29	30	31					29	30					

Julho							Agosto							Setembro							Outubro							Novembro							Dezembro						
1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
8	9	10	11	12	13	14	8	9	10	11	12	13	14	8	9	10	11	12	13	14	8	9	10	11	12	13	14	8	9	10	11	12	13	14	8	9	10	11	12	13	14
15	16	17	18	19	20	21	15	16	17	18	19	20	21	15	16	17	18	19	20	21	15	16	17	18	19	20	21	15	16	17	18	19	20	21	15	16	17	18	19	20	21
22	23	24	25	26	27	28	22	23	24	25	26	27	28	22	23	24	25	26	27	28	22	23	24	25	26	27	28	22	23	24	25	26	27	28	22	23	24	25	26	27	28
29	30	31					29	30	31					29	30						29	30	31					29	30						29	30	31				

Dias: 365 + 1 intercalar.
Em vermelho: dia intercalar de ano bissexto.
Ano bissexto: os múltiplos de 400 (1600, 2000, etc), os múltiplos de 4, exceto os que também forem múltiplos de 100, mas não de 400 (como 1900 e 2100).
Equinócios: 19/20/21 de março e 22/23 de setembro.
Solstícios: 20/21 de junho; 21/22 de dezembro.
Estações: Primavera, Verão, Outono, Inverno.
Anno mundi: 1. Nascimento de Cristo.
Introdução: 15 de Outubro de 1582, pelo Papa Gregório XIII em "Calendarium Gregorianum Perpetuum" (o dia anterior, no Juliano, foi 4 de Outubro de 1582).

Fonte: Serra (2018)

garantem que uma razão (proporção) seja absoluta em todos os casos que existir[9], permitindo assim que operações aritméticas de divisão e multiplicação, bem como médias geométricas e harmônicas sejam calculadas sobre as variáveis mensuradas por essas escalas. Para facilitar a compreensão, um comparativo entre as quatro escalas (extraído diretamente de Stevens et al. (1946)) é representado pelo Quadro 1 em que são resumidas as operações e níveis de medida oferecidos por cada escala.

No contexto de *Chatbots*, as escalas **nominal**, **ordinal** e de **intervalo** são suficentes para inquirir aspectos de qualidade junto aos usuários, pois, por meio das escalas nominais é possível inquirir aspectos categóricos e/ou questões polares[10] e por meio de escalas ordinais e de intervalo é possível, para cada tipo de escala respectivamente, categorizar a opinião de usuários em níveis ordinais de concordância e vinculá-las a notas numéricas atribuídas aos aspectos indagados, ações também possíveis de serem realizadas com o uso de escalas **Likert**, abordadas na próxima Seção.

Por fim, é possível questionar o porquê das escalas de razão terem sido discutidas, visto que a priori, não são necessárias para inquirir aspectos de qualidade dos agentes conversacionais autônomos. Entretanto, as escalas de razão possuem o acesso a todas as ferramentas estatístico-matemáticas disponíveis nas demais escalas e, portanto, podem ser utilizadas para inquirir os mesmos aspectos, bem como, por possuírem uma maior gama de operações, podem ser úteis para mensurar o sentimento e/ou outras métricas candidatas a aproximadores de qualidade de chatbots, justificando dessa forma a sua apresentação.

[9] Aprofundando este conceito (para finalidades explicativas) Dalati (2018) menciona que um objeto que apresente temperatura de 80° na escala *Fahrenheit* não é duas vezes mais quente do que outro objeto com as mesmas características físicas que apresente uma temperatura de 40° na mesma escala, pois, o zero desta escala não indica a ausência completa da variável a ser medida.

[10] Questões cujas respostas sejam apenas sim/não

Quadro 1 – Comparativo entre as quatro escalas de medida

Escala	**Operações empíricas básicas**	**Estrutura do grupo matemático**	**Operações estatísticas possíveis**
Nominal	Determinação de igualdade	Grupo de permutação $x' = f(x)$ Onde $f(x)$ implica em qualquer substituição um-para-um	Número de casos Moda Correlação de contingência
Ordinal	Determinação de maior ou menor	Grupo isotônico $x' = f(x)$ Onde $f(x)$ implica em qualquer função monotônica crescente	Mediana Percentis
Intervalo	Determinação de igualdade intervalos ou diferenças	Grupos lineares genéricos da forma $x' = ax + b$	Média Desvio padrão Correlação de ordem de classificação (*Spearman*) Correlação de produto-momento (ρ de *Pearson*)
Razão	Determinação de igualdade de razões	Grupos de similaridade da forma $x' = ax$	Coeficiente de variação

Fonte: Stevens et al. (1946, p. 678)

2.2.1 Escalas Likert

Escalas Likert são escalas ordinais de respostas que podem ser interpretadas (extrapoladas) como escalas de intervalo [11] que mensuram opiniões, atitudes, conhecimentos, percepções, valores e mudanças comportamentais de indivíduos por meio de pontos relacionados a estes aspectos, os quais os indivíduos selecionam para expressar suas respostas às indagações levantadas (VOGT, 1999; MANGIAFICO, 2016; DALATI, 2018).

Esses pontos são apresentados em itens (perguntas) Likert de forma gradual, ordenada — podendo ou não possuir intervalos numéricos equidistantes arbitrados — e são restritos a um escopo de simetria entre os extremos onde se encontram, que sempre expressam idéias opostas entre si Normalmente esses pontos estão associados a pares **<nota: descrição textual> — 1: muito insatisfeito, 2: insatisfeito, 3: indiferente, 4: satisfeito, 5: muito satisfeito** — a símbolos como *smiles* (*emojis*), estrelas (presentes em escalas de avaliação 5-estrelas[12]), a estados de emoção, dentre outros (BURNS; BUSH, 2007; MANGIAFICO, 2016; DALATI, 2018).

Ainda em relação a itens Likert e seus pontos, quando a quantidade de pontos é ímpar,

[11] De acordo com Mangiafico (2016), não existe consenso a respeito das escalas Likert serem ordinais ou de intervalo.

[12] Uma escala de tipo Likert, de acordo com Lee e Yang (2015 apud CHEN, 2017)

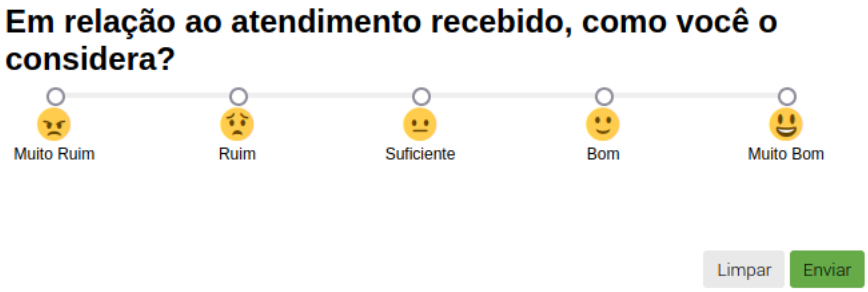
normalmente assume-se a existência de neutralidade/indiferença/incerteza em relação ao item inquirido, sendo permitido ao respondente expressá-la, o que não ocorre quando a quantidade de pontos é par, pois, qualquer ponto escolhido por um respondente de um item apresentará uma tendência de concordância (ou discordância) em relação ao aspecto indagado. As Figuras 7, 8, 9 apresentam:

1. Exemplo de escala 5-estrelas (Fig. 7);
2. Exemplo de escala de *smiles* (Fig. 8);
3. Exemplo de escala Likert de dois itens com 5 e 4 pontos cada (Fig. 9);

Figura 7 – Escala 5-estrelas

Fonte: W3Schools (2022)

Figura 8 – Avaliação de atendimento mensurada em escala Likert com pontos associados a *smiles*

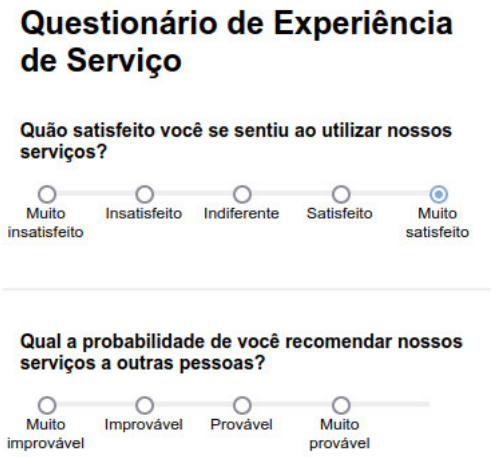


A possibilidade de se arbitrar intervalos numéricos equidistantes para pontos não (necessariamente) numéricos presentes nos itens das escalas Likert é objeto de controvérsia, pois, é uma extrapolação de escalas ordinais que mensuram variáveis qualitativas para escalas numéricas de intervalos que mensuram variáveis quantitativas. Isto, segundo Mangiafico (2016) é indesejável devido ao desconhecimento dos níveis de intervalos existentes entre pontos em escalas ordinais, mas admissível — sob perda de precisão e presença de ruídos e viéses — em itens Likert com um número expressivo de pontos.

Apesar das indesejáveis situações mencionadas acima, aplicações da indústria (e também da academia) utilizam frequentemente escalas Likert para tomar a opinião de usuários de sistemas, produtos e serviços acerca de suas (in)satisfações e experiências junto aos artefatos que utilizam[13]. No contexto de *Chatbots*, é solicitado aos próprios usuários destes (ou a avaliadores externos) que efetuem a qualificação do desempenho dos agentes conversacionais por meio da escolha

[13] Sem necessariamente oferecerem um número expressivo de pontos Likert

Figura 9 – Questionário de experiência de serviço em escala Likert com dois itens que apresentam 5 e 4 pontos respectivamente



de pontos em itens Likert que indagam aspectos de qualidade diversos, como: profunidade do diálogo, qualidade do atendimento, se os objetivos pretendidos foram alcançados, dentre outros (LIANG; ZOU; YU, 2020).

Referendando as já relatadas situações indesejáveis [14] em relação ao uso de escalas Likert para mensurar qualidade de *Chatbots*, Kulikov et al.; Ram et al.; Venkatesh et al. (2018, 2018, 2018 apud LIANG; ZOU; YU, 2020) vão além e criticam também a ausência de fatores motivacionais (econômicos, sociais, dentre outros) para que usuários cooperem diligentemente com o processo de avaliação de qualidade. Entretanto, por serem utilizadas frequentemente, as escalas Likert constituem ao menos um referencial comparativo para possíveis novas métricas de qualidade relacionadas ou não a si.

Por mais trivial que possa parecer, existem indícios quanto a influência do sentimento textual presente nas expressões de usuários e dos próprios *Chatbots* na qualidade atribuída pelos primeiros (por meio de escalas Likert) em relação ao desempenho dos últimos, de que quando a qualidade do desempenho dos *Chatbots* é insuficiente, o sentimento textual das expressões dos diálogos apresentará tendências negativas, valendo também para a situação oposta (TAUSCZIK; PENNEBAKER; NERBONNE; KÜSTER; KAPPAS, 2010, 2014, 2017 apud FEINE; MORANA; GNEWUCH, 2019). Mas, antes de aprofundar nos argumentos que justificam estes indícios (apresentados na Seção 3.2, replicados no Capítulo 4 e referendados no Capítulo 5), é necessário definir o conceito de sentimento, o tema da próxima Seção.

2.3 Sentimento

O sentimento pode ser definido como uma opinião / ponto de vista que uma pessoa sustenta em relação a um tópico — produto, serviço, crença, ideologia, local, costume, indivíduo,

[14] Vinculadas a ruídos, viéses e perdas de precisão

etc. — com tendência positiva, neutra ou negativa. Este sentimento pode ser expresso por meio de signos, gestos, linguagem oral e linguagem escrita e pode ser examinado manualmente por humanos ou automaticamente por computadores (KIM; LEVY; LIU, 2020; YAAKUB; LATIFFI; ZAABAR, 2019; SOUZA; PEREIRA; DALIP, 2017).

Quando esse exame é conduzido por computadores, recebe o nome de **Análise de Sentimento** (ou mineração de opinião) e é realizado sobre mídias textuais[15] por meio de técnicas de processamento de linguagem natural (*NLP — Natural Language Processing*) e/ou aprendizado de máquinas (*ML - Machine Learning*)[16] que podem ser aplicadas para a construção de novos algorítmos de análise de sentimento ou estarem contidas em ferramentas — proprietárias ou de código aberto (*open source*) — já existentes (YAAKUB; LATIFFI; ZAABAR, 2019; FEINE; MORANA; GNEWUCH, 2019; SOUZA; PEREIRA; DALIP, 2017). Os algorítimos de análise de sentimento, segundo Gonçalves (2015 apud SOUZA; PEREIRA; DALIP, 2017), realizam este procedimento a partir da polaridade emocional — graus de positividade, neutralidade ou negatividade — presente em frases e palavras que permite computar *scores* que apontam se o sentimento dos excertos textuais analisados é negativo, neutro ou positivo. Geralmente, os scores são apresentados em escalas de intervalo/razão simétricas, que iniciam em um extremo negativo, passam por um referencial nulo de neutralidade e terminam em um extremo positivo, podendo ou não serem normalizadas[17].

O exame de sentimentos nas ocasiões em que é conduzido por humanos também é conhecido como **Anotação (Rotulação) de Sentimentos** e recebe este nome porquê é realizado para vincular sentimentos a dados textuais (palavras e frases) que ainda não foram processados e que comumente, após este processo de anotação, são utilizados para construir algoritmos de análise de sentimento, como em Kim, Levy e Liu (2020). Vale ressaltar ser possível utilizar humanos para realizar essa tarefa de extração de sentimentos em múltiplos contextos, apesar de não ser recomendável em termos econômicos, sendo preferível utilizar computadores para realizarem essa tarefa, dadas as acurácias que os algoritmos de análise de sentimento possuem. Entretanto, para massas de dados pequenas (ou médias) é viável utilizar a anotação humana.

O que diferencia *scores* computados por algoritmos de análise de sentimento daqueles apontados por humanos em processos de rotulação é a escala de representação desses valores, sendo os primeiros representados por escalas de intervalo/razão e os segundos em escalas ordinais com intervalos afixados entre as graduações [18], pois, para a execução humana, é mais fácil

[15] Quando o sentimento está contido em representações sonoras, como em Kim, Levy e Liu (2020), é benéfico converter tais representações para formato textual, de forma que seja possível utilizar técnicas tradicionais de Análise de Sentimentos. Quando as representações são visuais (gráficas), a detecção de emoções e análise de sentimentos são conduzidas a partir de técnicas de visão computacional (com aprendizado de máquinas) dispensáveis para o escopo deste trabalho. Entretanto, para aqueles que se interessarem, é recomendada a leitura de Gajarla e Gupta (2015) para maiores informações sobre análise de sentimentos em imagens.

[16] Existem outras formas de realizar análise de sentimentos, entretanto, a grande acurácia na predição de sentimentos obtida pelas técnicas de processamento de linguagem natural e de aprendizado de máquinas — até 80% de predições corretas obtidas pelas técnicas NLP e até 96% pelas técnicas ML, segundo Yaakub, Latiffi e Zaabar (2019) — justifica a delimitação do escopo de técnicas abordadas a estas duas mencionadas.

[17] Ver Nielsen (2011), Hutto e Gilbert (2014) e Kim, Levy e Liu (2020) para diferentes apresentações de *scores*.

[18] Escalas Likert extrapoladas para escalas de intervalo.

anotar sentimentos de forma qualitativa (**-3: muito negativo, -2: negativo, -1: ligeiramente negativo, 0: neutro, 1: ligeiramente positivo, 2: positivo, 3: muito positivo**) do que de forma quantitativa (presente em Kim, Levy e Liu (2020)).

Por fim, as representações de *scores* nas escalas mencionadas habilitam a utilização de ferramentas aritmético-estatísticas importantes para análise de sentimentos como a média, bem como correlações, elicitadas na próxima Seção.

2.4 Correlações

Correlações, segundo Anton, Matei e Avram (2019), Reddy (2019) e o dicionário *LEXICO*[19], representam relacionamentos mútuos ou conexões entre duas ou mais entidades. A princípio, esta definição pode confundir mais do que elucidar, mas, o entendimento de correlações pode ser melhor exemplificado[20] como a existência de vínculos que refletem em uma entidade quaisquer alterações que outras entidades que coexistam consigo sofram, ou seja, duas (ou mais) entidades possuem correlação se e somente si, alterações nas estruturas (nos valores) de uma refletirem nos valores de outra(s).

As correlações podem ser verificadas qualitativamente — por meio da percepção de mudanças (ou manutenções) das estruturas de objetos coexistentes quando algum desses objetos é modificado — e quantitativamente pelo cálculo do **coeficiente de correlação** (r ou ρ), que é um número que varia de -1 a 1 e que representa o grau de associação entre duas variáveis X e Y quaisquer em um conjunto de dados. Quão mais próximo esse número for dos extremos do intervalo $[-1, 1]$, mais associadas (correlacionadas) as variáveis avaliadas são e quão mais próximo de 0, menos correlacionadas são [21,22] (ANTON; MATEI; AVRAM, 2019; REDDY, 2019).

Quando o coeficiente se aproxima do valor 1, as variáveis são **positivamente correlacionadas**, ou seja, a modificação no valor de uma implica em uma modificação de mesma natureza na outra[23]. Quando o coeficiente se aproxima do valor -1, as variáveis são **negativamente correlacionadas**, o que implica que modificações em uma variável produzem o efeito oposto na outra variável correlacionada. Em outras palavras, se uma variável aumenta (dimunui) a outra diminui (aumenta)[24] (ANTON; MATEI; AVRAM, 2019).

De acordo com Hinkle, Wiersma e Jurs (2003 apud MUKAKA, 2012) são dois os principais métodos para se calcular coeficientes de correlação: **r-pearson** — coeficiente de

[19] CORRELATION (2021)

[20] Através de definições complementares presentes em Anton, Matei e Avram (2019)

[21] Até o ponto de serem totalmente independentes quando esse número for 0 de fato.

[22] Para categorizar as intensidades (forças) das correlações, este trabalho utiliza a interpretação de correlações de Cohen (1992 apud FEINE; MORANA; GNEWUCH, 2019).

[23] As variáveis **pressão** e **temperatura** pertencentes aos gases ideais são variáveis positivamente correlacionadas, pois, aumentos (diminuições) na temperatura implicam aumentos (diminuições) na pressão e vice-e-versa.

[24] As variáveis **temperatura ambiente** e **altitude em relação ao nível do mar** são exemplos de variáveis negativamente correlacionadas, pois, ao aumentar a altitude diminui-se a temperatura ambiente e vice-e-versa.

correlação produto-momento — e ρ**-spearman** — coeficiente de correlação de postos[25] de Spearman[26]. Estes métodos calculam coeficientes de correlação a partir da razão da covariância[27] das variáveis estudadas sobre o produto dos desvios padrão[28] que apresentam (MILONE, 2009 apud FERREIRA, 2020), como exemplificado pela Equação 1:

$$r = \frac{cov(A,B)}{\sigma_A \cdot \sigma_B} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2} \cdot \sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \tag{1}$$

A, B = Conjunto de variáveis a serem analisadas.
$cov(A,B)$ = Covariância das variáveis A e B com base em seus valores (ou postos).
σ_A (σ_B) = Desvio-padrão da variável A (B) com base em seus valores (ou postos).
a_i (b_i) = Valor (ou posto) da i-ésima amostra da variável A (B).
$\bar{a}$ ($\bar{b}$) = Média dos valores (ou dos postos) das amostras da variável A (B).
n = Quantidade de amostras observadas das variáveis A e B.

A diferença entre **r-pearson** e ρ**-spearman** ocorre nos seguintes aspectos:

1. O cálculo do coeficiente ρ**-spearman** demanda que as amostras das variáveis estudadas sejam colocadas em postos[29].

2. Apenas o coeficiente ρ**-spearman** pode ser utilizado para calcular correlações de variáveis ordinais;

3. Quando as amostras das observações das variáveis não seguem uma distribuição normal[30], os valores da correlação r-pearson não podem ser generalizados para toda a população de observações dessa variável (DEMOS; SALAS, 2017).

O último item da enumeração acima é relacionado ao conceito de relevância estatística, abordado na próxima Subseção, que permite verificar se o valor de correlação calculado para uma amostra de observações de duas variáveis pretendidas representa a correlação dessas variáveis em nível de população — em todas observações existentes, não apenas nas amostras selecionadas.

[25] Ou ordens de classificação.
[26] Existem outras formas de se calcular correlações, mas, segundo Mukaka (2012), são derivadas dos principais métodos (**r-pearson**(e ρ**-spearman**).
[27] Covariância: medida da variabilidade conjunta de duas variáveis (LINDSTEN et al., 2017).
[28] Desvio padrão: “uma medida que expressa o grau de dispersão de um conjunto de dados"(GAMA, 2020).
[29] Colocar em postos consiste em ordenar (em ordem crescente ou decrescente) um conjunto de variáveis qualquer e considerar o valor de cada variável como a posição (o posto) que essa variável ocupa no conjunto ordenado (PONTES, 2010).
[30] Uma função estatística de distribuição de dados cuja concentração da maioria dos valores que representa ocorre em torno do valor médio desses dados (CLAYTON, 2020). Esta função apresenta uma curva em formato de sino.

2.4.1 Relevância Estatística (Valor-P)

Para inferir o coeficiente de correlação para toda a população das variáveis estudadas (não somente para as amostras selecionadas) é necessário apontar a relevância estatística associada ao coeficiente calculado, também conhecida como probabilidade de signifância e/ou **valor-p**.

O valor-p representa a probabilidade de se obter os mesmos resultados de correlação ao acaso, ou seja, quando não existe correlação real entre as variáveis estudadas. Este estado de acaso é determinado a partir do nível de significância α estabelecido para o experimento (usualmente $\alpha = 0.05$) e quando **valor-p** $\geq \alpha$ entende-se que os resultados de correlação obtidos também poderiam ser obtidos ao acaso, não podendo ser rejeitada a hipótese de que não exista correlação entre os conjuntos de variáveis estudados — $\rho = 0$ — mesmo que o coeficiente de correlação calculado seja não nulo (FERREIRA; PATINO, 2015; KOKOSKA; ZWILLINGER, 2000).

Não são apresentadas fórmulas para os cálculos do **valor-p**, pois, é suficiente apenas o que este representa para a admissão ou não de uma correlação. Também corrobora a não apresentação dessas fórmulas o fato de que trabalhos correlatos[31] não explicam todas as fórmulas estatísticas utilizadas. Por fim, referências ao **valor-p** são interpretadas como relevâncias estatísticas das correlações e assim como estas, calculadas por ferramentas computacionais.

Reiterando, quando **valor-p** for menor que o nível de significância α, entende-se que o coeficiente de correlação calculado é pertinente, visto que a probabilidade de não existir correlação entre as variáveis estudadas em toda a população respeita o limite de significância estatística estabelecido.

[31] Aspectos de qualidade de *chatbots*, análise de sentimento textual, dentre outros.

3 Trabalhos Relacionados

São apresentados nestes trabalhos relacionados as utilizações de escalas Likert para mensurar aspectos de qualidade em *chatbots* e aplicações diversificadas da indústria e da academia — Seção 3.1 — e estudos que consideram apropriado que sentimentos textuais apresentem indícios de qualidade — Subseções 3.2.1 e 3.2.2 da Seção 3.2.

3.1 Utilizações de Escalas Likert para mensurar aspectos de qualidade

Liang, Zou e Yu (2020)[1] afirmam que tanto a indústria quanto a academia dependem da avaliação humana para determinar a qualidade de diálogos produzidos por *chatbots*[2] e essa qualidade é comumente mensurada através de escalas Likert. [3]

Assim, o objetivo desta Seção é apresentar um compêndio de trabalhos e exemplos da indústria e da academia que utilizam estas escalas para mensurar aspectos de qualidade, começando por artefatos de propósito genérico — relatos de experiências de vida, serviços, uso de aplicativos, dentre outros — e terminando em trabalhos que avaliam o desempenho de agentes conversacionais através destas escalas.

Godes e Silva (2012) apresentam uma investigação da evolução de avaliações — mensuradas em escalas tipo Likert de estrelas — de produtos comercializados em meios eletrônicos, especificamente livros *best-sellers* revendidos pela loja Amazon.com. A Figura 10 ilustra avaliações (e a média de todas as avaliações recebidas) do clássico a Arte da Guerra[4], de autoria de Sun Tzu, comercializado pela referida loja.

Apesar de não mencionarem explicitamente a utilização de escalas Likert, Peterson e Wilson (1992) realizam um estudo sobre auto-avaliações de satisfação[5] que são mensuradas por escalas nominais, ordinais e de intervalo que podem (quase em sua totalidade) ser intepretadas como escalas Likert. Os trabalhos abordados pelos autores, representados no Quadro 2, constituem exemplos longevos da utilização de escalas Likert para medir aspectos de qualidade na indústria e na academia.

---

1 Também referenciados na fundamentação teórica (Seção 2.2.1, parágrafos 5 e 6).

2 Principalmente quando estes diálogos são conversacionais, sem restrição de assuntos a serem discutidos ou sem propósitos específicos que fundamentem suas existências, como o alcance de determinados objetivos.

3 Estas escalas são adotadas por serem consagradas para medir qualidade através de pesquisas de opinião sobre assuntos diversos, tais quais: filmes, livros, estabelecimentos, serviços e seus prestadores em geral (LIANG; ZOU; YU, 2020).

4 Tzu (2019).

5 Reportadas por consumidores.

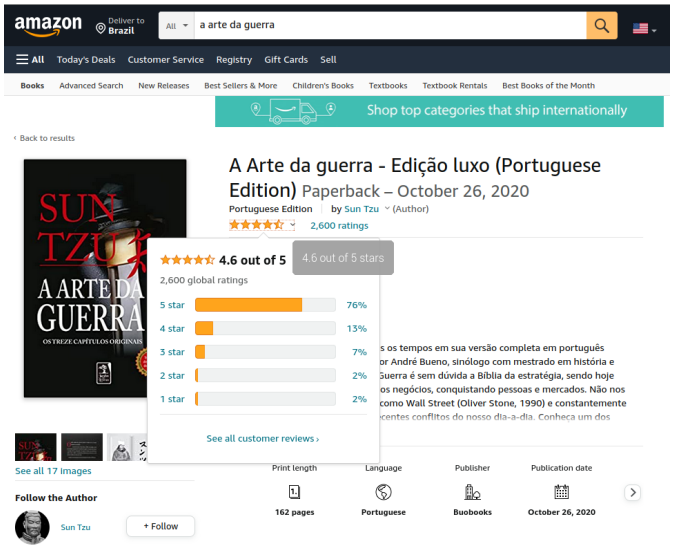
6 O autor realizou uma metanálise de 26 estudos sobre a satisfação de usuários em relação a atendimentos em estabelecimentos psiquiátricos. Para efetuar a leitura dos dados: 12% dos estudos da metanálise indicam que os usuários investigados apresentaram um percentual de satisfação entre 91 e 100% (e desta forma para os demais)

Quadro 2 – Exemplos da literatura e da indústria que usam escalas tipo Likert para mensurar satisfações (relatadas por consumidores) e suas respectivas distribuições

Origem	**Assunto**	**Descrição da População**	**Tamanho da População**	**Distribuição da Satisfação**	
				Escala	**Percentual (%)**
Campbell (1981)	Casamento	Adultos	3700	1 (Completamente Satisfeito)	54
				2	27
				3	9
				4	5
				5	2
				6	2
				7 (Completamente Insatisfeito)	1
				Intervalo de Satisfação (%)	**Percentual (%)**
Lebow (1982)	Estabelecimentos psiquiátricos	Usuários	26[6]	91-100	12
				81-90	38
				71-80	31
				61-70	15
				51-60	4
				Escala	**Percentual (%)**
Protegida por propriedade intelectual	Sistema de Folhas de Pagamento	Pequenos e Micro Empresários	500 +	Muito Satisfeito	76
				Algo Satisfeito	18
				Algo Insatisfeito	3
				Muito Insatisfeito	2
				Não Soube Responder	1
				Escala	**Percentual (%)**
Protegida por propriedade intelectual	Computadores Pessoais	Proprietários de Computadores	4500 +	Muito Satisfeito	64
				Algo Satisfeito	28
				Algo Insatisfeito	6
				Muito Insatisfeito	2
				Escala	**Percentual (%)**
Protegida por propriedade intelectual	Automóveis	Proprietários de Automóveis	38000 +	Totalmente Satisfeito	33
				Muito Satisfeito	44
				Razoávelmente Satisfeito	18
				Algo Insatisfeito	4
				Muito Insatisfeito	1
				Escala	**Percentual (%)**
Opinioni (1989)	Serviço de Telefonia Móvel	Usuários	6553	Muito Satisfeito	70
				Razoávelmente Satisfeito	20
				Não Muito Satisfeito	7
				Não Satisfeito de Forma Alguma	3
				Escala	**Percentual (%)**
Weinstein (1989)	Bancos	Clientes Adultos	1000 +	Muito Satisfeito	70
				Algo Satisfeito	20
				Totalmente Insatisfeito	3

Fonte: Peterson e Wilson (1992, p. 63)

Figura 10 – Avaliações de compradores do livro A Arte da Guerra



Fonte: Amazon (2022a)

O uso de escalas 5-estrelas tipo Likert pelas duas principais lojas de aplicativos móveis do mundo — *App Store* e *Google Play Store*[7] — para avaliar a qualidade de todos os aplicativos disponibilizados em suas plataformas é um exemplo hodierno do uso dessas escalas para mensurar aspectos de qualidade auto-reportados por consumidores, reforçando a ubiquidade delas. Para além disto, esta ubiquidade faz com que os consumidores estejam habituados a estas métricas de avaliação e considerem as informações representadas por elas para a tomada de decisões como verificado em pesquisa realizada por Poschenrieder (2015 apud CHEN et al., 2021) em que 69% dos usuários de *smartphones* inquiridos afirmam considerar as avaliações que um aplicativo recebe como um fator decisivo para a sua instalação e outros 77% dos usuários declaram que não instalariam um aplicativo cuja avaliação média seja inferior a três estrelas.

Para complementar o compêndio de trabalhos anunciado nesta Seção, são apresentados no Quadro 3 trabalhos referentes ao desenvolvimento de *chatbots* e à construção de bases de dados de conversas (entre humanos e inteligências artificiais e entre humanos apenas) que possuam aspectos de qualidade[8] medidos por escalas Likert e avaliados por julgadores humanos[9].

[7] *App Store* e *Google Play Store* são respectivamente as lojas oficiais de aplicativos dos sistemas operacionais móveis *iOS* e *Android* que, de acordo com O´Dea (2021 apud YANKOV, 2021), representam 99% dos sistemas usados em *smartphones* (dados considerados até junho de 2021).

[8] Aspectos relativos à experiência de interação dos usuários, ao engajamento dos agentes conversacionais, à pertinência das respostas dos agentes, e aos outros que possam existir (VENKATESH et al.; BOHUS; HORVITZ; LOWE et al., 2018, 2009, 2017 apud LIANG; ZOU; YU, 2020)

[9] Especialistas externos ou os próprios usuários dos *chatbots* (PÉREZ-ROSAS et al., 2019 apud LIANG; ZOU; YU, 2020).

Quadro 3 – Trabalhos associados a *chatbots* onde aspectos de qualidade são mensurados em escalas Likert

Trabalhos	**Categoria**	**Descrição**	**Observações**
ConvAI (LOGACHEVA et al., 2018), ConvAI2 (DINAN et al., 2019) e ConvAI3 (ALI-ANNEJADI et al., 2020)	Competição de *chatbots* conversacionais	*Chatbots* de escopo aberto, segundo Burtsev e Logacheva (2020), são de difícil construção pelas ausências de dados de treinamento e de métricas automáticas para avaliação de desempenho, sendo necessário confiar na dispendiosa avaliação humana. Visando enfrentar estes problemas, foram propostas as competições ConvAI, ConvAI2 e ConvAI3 para que, por meio da interação (e avaliação) humana junto a *chatbots*, possam ser construídos conjuntos de dados conversacionais úteis para a proposição de (novas) métricas de qualidade e a própria construção de *chatbots*.	ConvAI e ConvAI2 aconteceram em etapa única com diferenças nos diálogos produzidos e nas avaliações de qualidade efetuadas, onde: — Diálogos se pautam sobre tópicos extraídos de artigos da *Wikipedia*[10] no ConvAI enquanto no ConvAI2, (sobre) características de *personas*[11] que os interlocutores têm de interpretar; — Avaliações são mensuradas no ConvAI em escala Likert tridimensional de 5 pontos e no ConvAI2 esta escala possui os mesmos 5 pontos, mas, é unidimensional. Já a competição ConvAI3 aconteceu em duas etapas. Na 1ª etapa usuários faziam solicitações de escopo aberto a robôs como "**O que é a fazenda *Fickle Creek*?**"ou "**Eu quero informações e gravuras sobre dinossauros**"e os robôs deveriam atender as demandas dos usuários ou solicitar maiores explicações para mitigar ambiguidades. Na 2ª etapa julgadores humanos externos avaliavam a qualidade dos diálogos produzidos entre robôs e usuários através de uma escala Likert de 5 pontos. Os conjuntos de dados produzidos nas competições ConvAI, ConvAI2 e na 1ª etapa da ConvAI3 são disponíveis ao público.
Topical Chat: Em direção a conversas de escopo aberto baseadas em conhecimento Gopalakrishnan et al. (2019)	Base de dados conversacionais entre humanos	Assim como Burtsev e Logacheva (2020), Gopalakrishnan et al. (2019) consideram a construção de *chatbots* conversacionais uma tarefa de difícil realização. Por isso propõem o *Topical-Chat*, um conjunto de conversas entre humanos baseadas em 8 assuntos de conhecimentos gerais[12] que pode ser utilizado para o treinamento de *chatbots*	As conversas presentes no conjunto de dados são simuladas por profissionais prestadores de serviço na plataforma *Amazon Mechanical Turk*[13] que são pareados e instruídos a conversar em turnos sobre assuntos dentre os 8 especificados. Em seguida, para cada diálogo, é solicitado a cada um dos participantes que, enquanto esperam seu par enviar uma mensagem, avaliem o sentimento da última mensagem que enviaram em uma escala de 8 pontos, bem como, que avaliem a qualidade da última mensagem enviada pelo seu par em uma escala Likert de 5 pontos. Quando o diálogo chega ao fim é solicitado a cada um dos participantes que avalie a qualidade geral da conversa através de uma escala Likert de 5 pontos.
Alexa Prize — Estado da Arte em Inteligência Artificial Conversacional (KHATRI et al., 2018).	Competição de *chatbots* conversacionais	Para avançar o estado da arte na I.A. conversacional, a empresa *Amazon* oferece uma competição com premiações em dinheiro desafiando equipes universitárias a construir agentes conversacionais (*socialbots*) que possam dialogar de forma coerente e atraente com humanos sobre assuntos de conhecimento geral durante 20 minutos.	A competição é uma oportunidade para que a academia possa realizar pesquisas de grande escala com dados obtidos de interações reais (ao longo de vários meses) de milhões de usuários da assistente virtual Alexa[14] que possuem avaliações em escala de 1 a 5 atribuídas pelos próprios usuários.

Quadro 3 – Trabalhos associados a *chatbots* onde aspectos de qualidade são mensurados em escalas Likert

Trabalhos	**Categoria**	**Descrição**	**Observações**
Além de avaliações reportadas por usuários em escalas Likert: Um modelo comparativo para avaliação automática de diálogos Liang, Zou e Yu (2020).	Proposição de métricas para avaliação de *chatbots* de escopo aberto	Liang, Zou e Yu (2020) abordam outro problema desafiador vinculado a *chatbots* de escopo aberto: a avaliação de seus desempenhos, que segundo os autores, é de difícil realização porque métricas de avaliação automática de desempenho de *chatbots* de escopo fechado não se aplicam corretamente para seus pares de escopo aberto, bem como, porque avaliações reportadas por usuários (em escalas Likert) são passíveis a viéses e apresentam displicências (na maioria dos casos). Para enfrentar este problema, os autores propõem avaliar diálogos de forma comparativa assim como um método para remover viéses de avaliações reportadas por usuários.	Os autores utilizaram dados de diálogos com qualidade avaliada em escala Likert oriundos de 3608 diálogos entre humanos e robôs presentes em *datasets* proprietários pertencentes ao *Amazon Alexa Prize Challenge*. Os viéses mencionados são relativos às distribuições assimétricas das avaliações Likert efetuadas por usuários, que se concentram em valores extremos da escala.
Medindo a satisfação em relação ao serviço prestado por *chatbots* de atendimento ao cliente usando análise de sentimento (FEINE; MORANA; GNEWUCH, 2019)	Desempenho de *chatbots*	Feine, Morana e Gnewuch (2019) abordam as dificuladas para mensurar a qualidade do serviço prestado por *chatbots* de atendimento ao cliente em escalas Likert devido aos vieses presentes nestas escalas, aos obstáculos para inquirir avaliações durante os atendimentos[15] e às displicências dos usuários em relação ao processo avaliativo em si. Neste sentido, entendem ser plausível procurar prever as avaliações dos clientes durante os atendimentos para contornar as limitações existentes nos modelos baseados em escala Likert e o propõem a partir da Análise de Sentimentos, partindo da premissa que avaliações de baixa qualidade podem estar associadas a sentimentos negativos (e vice-e-versa).	A partir da leitura do trabalho, extrapolando a própria divisão que os autores apresentam, é verificado que o trabalho pode ser seccionado em cinco etapas — duas prospectivas e três executórias. As etapas prospectivas são assim nomeadas, pois, nelas são realizadas a prospecção de corpos de diálogos entre humanos e *chatbots* (primeira etapa) e de algorítmos de análise de sentimento (segunda etapa). Já em relação às etapas executórias, que recebem este nome (deduzido da leitura do trabalho) por representarem as etapas onde os experimentos dos autores foram concrtetizados, tem-se que a primeira etapa consistiu em realizar uma análise comparativa e exploratória de algoritmos de análise de sentimento, a segunda consitiu em analisar a correlação (**r-pearson**) entre valores de sentimento e avaliações de satisfação — a nível de diálogo — e a terceira repetiu os passos da segunda etapa — a nível de mensagens (enviadas pelos usuários) — para verificar a hipótese levantada pelos autores de se prever avaliações durante os atendimentos.
Estimativa da satisfação de clientes e análise do sentimento de suas falas durante diálogos com *socialbots* (KIM; LEVY; LIU, 2020).	Desempenho de *chatbots*	Kim, Levy e Liu (2020) assim como Feine, Morana e Gnewuch (2019) relatam limitações das escalas Likert, principalmente as associadas à impertinência de se inquirir durante atendimentos avaliações dos usuários em relação aos serviços recebidos. Visando construir alternativas de avaliação propõem (de forma similiar a Feine, Morana e Gnewuch (2019)) analisar o sentimento presente em interações de usuários junto a assistente virtual Alexa para também investigar a possibilidade de prever qualidade dos atendimentos a partir deste sentimento.	Os autores deste trabalho são vinculados a empresa Amazon e por isso, tem acesso aos conjuntos de dados proprietários do *Amazon Alexa Prize Challenge*, onde primeiramente, investigam a correlação de qualidades de diálogos atribuídas em escalas Likert de 1 a 5 por usuários da Alexa com o sentimento que apresentam ao interagir verbalmente com a assistente. Fazem isso tanto a nível de diálogo quanto a nível de expressões, usando comitês de especialistas humanos externos para anotar as qualidades e sentimentos de amostras dos conjuntos de dados nos níveis de diálogo e de expressões mencionados, bem como, construindo um algoritmo de análise de sentimentos para computar automaticamente os sentimentos expressos pelos usuários da Alexa em suas falas. Em seguida, propõem modelos baseados em aprendizado de máquina que permitem utilizar os valores de sentimento para prever valores de qualidade, apresentando resultados que demonstram que os sentimentos são correlacionados (ρ-**spearman**) à métrica da qualidade e que a partir deles (e dos modelos propostos) é possível prever qualidades dos diálogos.

No Quadro 3 é possível observar que os três últimos trabalhos abordam aspectos indesejáveis das escalas Likert para mensurar qualidade de *chatbots* como existência de vieses nos dados, avaliações displicentes, impossibilidade de mensurar a qualidade durante os atendimentos, dentre outros. E além disto, que estes trabalhos propõem estratégias para lidar com estes aspectos, como estas duas: um método para remoção automática de vieses e ruídos das avaliações em escalas Likert (LIANG; ZOU; YU, 2020) e o uso dos sentimentos dos usuários para inferir avaliações dos diálogos (FEINE; MORANA; GNEWUCH, 2019; KIM; LEVY; LIU, 2020).

Uma terceira estratégia não mencionada, mas, recomendada por Logacheva et al. (2018) para remover ruídos, vieses e avaliações displicentes que existem no *dataset* ConvAI é a estratégia de revisar manualmente os diálogos através de comitês de (ao menos) três especialistas humanos para avaliar adequadamente o desempenho dos agentes[16] e tal estratégia é similar a um dos procedimentos adotados por Kim, Levy e Liu (2020).

Portanto, detalhar Feine, Morana e Gnewuch (2019) e Kim, Levy e Liu (2020) — como realizado na próxima Seção — é suficiente para discorrer sobre o possível uso de sentimento textual para prever qualidade de atendimentos, bem como, (para) discutir potenciais benefícios da remoção manual de ruídos das avaliações no contexto de *chatbots*.

3.2 Estudos sobre a correlação entre avaliações de desempenho em escalas Likert e sentimento textual contido em diálogos entre humanos e *chatbots*

Segundo Tausczik e Pennebaker; Nerbonne; Küster e Kappas (2010, 2014, 2017 apud FEINE; MORANA; GNEWUCH, 2019) é razoável considerar o sentimento textual como um indício da qualidade dos atendimentos de *chatbots*, pois, quando o desempenho dos agentes é insuficiente, acontecerão tendências negativas nos sentimentos expressos pelos humanos partícipes dos diálogos em suas mensagens e do contrário, quando o desempenho dos agentes é satisfatório, os sentimentos manifestados pelos partícipes apresentarão tendências positivas.

Desta forma, esta Seção apresenta dois estudos que consideram plausível sentimentos textuais apontarem indícios de qualidade — Feine, Morana e Gnewuch (2019) (Seção 3.2.1) e Kim, Levy e Liu (2020) (Seção 3.2.2) — expondo para cada estudo o que seus autores fizeram,

[10] Presentes na base de dados SQuAD (RAJPURKAR et al., 2016).
[11] *Personas* são papéis semifictícios baseados em comportamentos e fatos reais (SOUSA, 2018).
[12] Moda, política, livros, esportes, entretenimento de massa, música, ciência, tecnologia e filmes.
[13] *Amazon Mechanical Turk* é uma feira de terceirização que conecta entidades que queiram terceirizar serviços e tarefas a indivíduos dispostos a realizá-los (AMAZON, 2022b).
[14] *Amazon Alexa* é uma assistente virtual — um tipo de agente conversacional — baseada em *interface* de voz (KHATRI et al., 2018).
[15] Avaliações são realizadas ao final dos atendimentos porque interrompê-los é considerado inoportuno (FEINE; MORANA; GNEWUCH, 2019; KIM; LEVY; LIU, 2020).
[16] Estratégia similar também foi adotada por Kim, Levy e Liu (2020) para amostras de diálogos do conjunto de dados do *Alexa Prize Challenge*, como explicitado no Quadro 3.

porque fizeram, como fizeram, resultados que obtiveram e conclusões que alcançaram.

3.2.1 Medição da satisfação em relação ao serviço prestado por *chatbots* de atendimento ao cliente usando análise de sentimento

O trabalho de Feine, Morana e Gnewuch (2019)[17] investiga (e reconhece) a aplicação do sentimento computado por métodos automáticos de análise de sentimento como uma aproximação (*proxy*) para medir a qualidade — satisfação do cliente — em relação a atendimentos prestados por *chatbots* conversacionais.

Os autores fizeram essa investigação porque a maioria das abordagens existentes para avaliação de qualidade de *chatbots* são limitadas a aplicação de questionários Likert após os atendimentos dos agentes, que são raramente respondidos e frequentemente sujeitos a vieses (como relatado na Seção 3.1).

Assim, essa investigação foi realizada através do estudo da correlação (**r-pearson**) entre satisfação[18] atribuída e sentimento textual — considerado tanto na totalidade do diálogo quanto em sequências (de tamanho variável) de mensagens — expresso por 79 pessoas em relação ao serviço prestado por um *chatbot* de atendimento ao cliente de uma companhia de telefonia móvel fictícia[19].

Para explicar este estudo, são apresentadas cinco etapas que fundamentaram sua realização — duas prospectivas (1 e 2) e três executórias (3 a 5). Esta apresentação não segue a ordem apresentada por Feine, Morana e Gnewuch (2019) em seu artigo, mas, contém as principais idéias e técnicas que utilizaram para realizar o trabalho. Eis as etapas:

1. Prospectar algoritmos de análise de sentimento;
2. Prospectar bases de diálogos entre humanos e *chatbots*;
3. Conduzir experimentos de comparação de algoritmos de análise de sentimento;
4. Investigar a correlação entre pontuações de sentimento e satisfação do atendimento em nível de diálogo;
5. Investigar a correlação entre pontuações de sentimento e satisfação do atendimento em nível de sequências de mensagens;

Para concretizarem a primeira etapa, os autores buscaram pesquisas de desempenho de algoritmos de análise de sentimento — Ribeiro et al. (2016), Arbide, Romero e Fernández (2017)

[17] Medindo a satisfação em relação ao serviço prestado por *chatbots* de atendimento ao cliente usando análise de sentimento (FEINE; MORANA; GNEWUCH, 2019).

[18] Mensurada por uma escala tipo Likert de sete pontos e dezessete itens denominada CSES (*Chatbot Service Encounter Satisfaction*) proposta pelos autores com base no trabalho de Verhagen et al. (2014).

[19] Essas pessoas, que interpretam clientes dessa companhia de telefonia móvel, foram municiadas com contas telefônicas (igualmente) fictícias e orientadas a executar a tarefa experimental de solicitar planos telefônicos adequados às restrições impostas pelas contas ao *chatbot* de atendimento da companhia.

— das quais selecionaram os algoritmos AFINN, VADER, *IBM Watson*, *Google Cloud* e *Microsoft Azure* para realizar as tarefas de análise de sentimentos. Estes algoritmos calculam pontuações de sentimentos em intervalos distintos, porém, Feine, Morana e Gnewuch (2019) padronizaram (e normalizaram) as pontuações computadas por cada algritmo de análise de sentimento sobre o intervalo $[-1, 1]$, onde -1 representa a polaridade totalmente negativa, 0 representa a polaridade neutra e 1 representa a polaridade totalmente positiva.

Na segunda etapa, prospectaram dois corpos de diálogos — ExpCorpus (GNEWUCH et al., 2020) e ConvAI (LOGACHEVA et al., 2018) — que contém diálogos entre humanos e *chatbots* com aspectos de qualidade avaliados por escalas do tipo Likert, com o ExpCorpus contendo satisfação dos usuários mensurada pela escala CSES[18] e o ConvAI apresentando atributos de profundidade, engajamento e qualidade geral dos diálogos mensurados por escala tipo Likert tridimensional de cinco pontos[20].

Relatando mais informações sobre os corpos de diálogos, o ExpCorpus é o conjunto de 79 diálogos utilizado pelos autores para estudar a correlação (etapas 4 e 5) entre sentimento e qualidade, sendo utilizado também para comparar os algorítmos de análise de sentimento (etapa 3). Apesar do ConvAI apresentar atributos de qualidade mensurados em escala tipo Likert, os autores não utilizaram essa base de diálogos para estudar correlação entre qualidade e sentimento, pois, restringiram o estudo dessa correlação a corpos mensurados pela métrica CSES. Neste sentido, utilizaram 2180 diálogos entre humanos e *chatbots* extraídos do ConvAI apenas na terceira etapa.

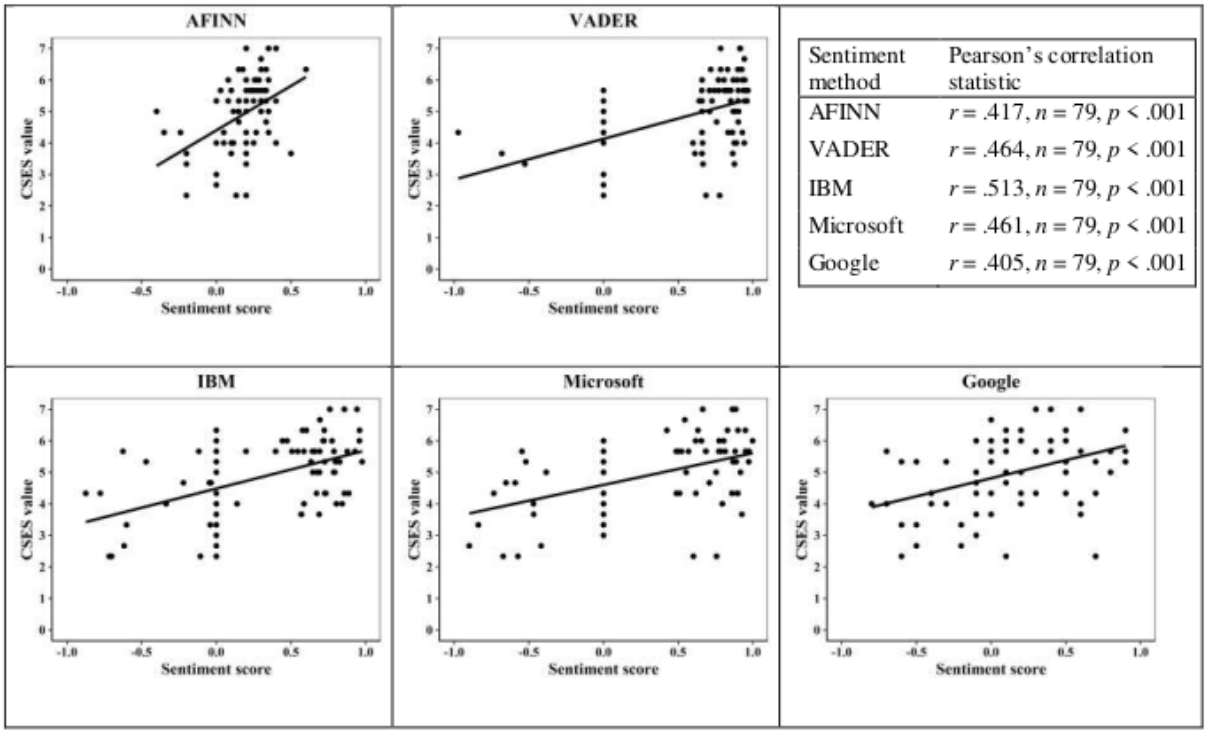
Na terceira etapa — comparação de algoritmos de análise de sentimento — Feine, Morana e Gnewuch (2019) investigaram o quão similares são os valores dos sentimentos computados pelos algoritmos ao analisar os dois corpos prospectados (em nível de diálogos e em nível de mensagens) através do cálculo de correlações **r-pearson** entre os sentimentos calculados pelos cinco algoritmos. Poderia ser questionado o porquê da realização dessa etapa, pois, a princípio, sua execução não apresenta motivo evidente. Entretanto, de acordo com Brooks et al. (2013 apud FEINE; MORANA; GNEWUCH, 2019), algoritmos de análise de sentimento têm um desempenho melhor (e mais convergente) em conjuntos textuais cuidadosamente elaborados, mas, frequentemente enfrentam obstáculos quando analisam sentimentos de textos informais presentes em comunicações *online*. Como as características textuais do ExpCorpus e do ConvAI são similares às que provocam obstáculos, as correlações de pontuações de sentimento calculadas pelos algoritmos averiguam o quão convergentes (ou divergentes) são ao analisar esses corpos e, consequentemente, se esses corpos produzem (ou não) obstáculos para a análise de sentimento.

A quarta etapa apresenta o cálculo das correlações entre os sentimentos computados pelos algoritmos prospectados e as qualidades dos diálogos do ExpCorpus, mensuradas pela métrica CSES. Nessa etapa, Feine, Morana e Gnewuch (2019) atestaram a hipótese levantada — sentimento textual pode ser usado como *proxy* para a qualidade do atendimento prestado por um

[20] Escala tipo Likert tridimensional de cinco pontos consiste em uma escala likert com três itens para mensurar profundidade, engajamento e qualidade geral do diálogo com pontos que variam de 1 (ruim) até 5 (bom).

chatbot — através dos cálculos de correlações realizados, cujos resultados (e dispersões entre sentimentos computados pelos algoritmos e qualidade CSES) são mostrados pela Figura 11.

Figura 11 – Análises de correlação r-pearson (e dispersão) entre sentimento (em nível de diálogo) e valores CSES para o ExpCorpus



Sentiment method	Pearson's correlation statistic
AFINN	$r = .417, n = 79, p < .001$
VADER	$r = .464, n = 79, p < .001$
IBM	$r = .513, n = 79, p < .001$
Microsoft	$r = .461, n = 79, p < .001$
Google	$r = .405, n = 79, p < .001$

Fonte: Feine, Morana e Gnewuch (2019)

Embasando os resultados alcançados nesta quarta etapa, Feine, Morana e Gnewuch (2019) lançaram mão da interpretação de correlações proposta por Cohen (1992) para afirmar que todas as pontuações de sentimentos computadas pelos algoritmos são ao menos moderadamente correlacionadas com a qualidade CSES dos diálogos e que o sentimento computado pelo algoritmo IBM *Watson* possui correlação forte com a qualidade CSES dos diálogos.

Na quinta etapa os autores conduziram experimentos em busca do(s) tamanho(s) mínimo(s) da(s) sequência(s) de mensagens dos diálogos do ExpCorpus que permitisse(m) demonstrar uma correlação entre valores de sentimento e valores de satisfação mensurados pela métrica CSES usando para tal o algoritmo IBM *Watson*, pois, este computou os valores de sentimentos mais correlaciondos à métrica CSES. Em seguida, verificaram os impactos nas correlações causados por distintas combinações de sequência(s) de mensagens do(s) tamanho(s) mínimo(s) encontrado(s) ao longo dos diálogos.

Ao término destas etapas, os autores concluíram que valores de sentimento podem ser utilizados como uma aproximação automática e objetiva de valores de satisfação (CSES) dos usuários em relação a atendimentos *online*, corroborando Küster e Kappas (2017) que afirmam que métodos de análise de sentimentos coincidem surpreendemente bem com respostas sobre

estados emocionais[21] relatadas por indivíduos. Mais ainda, Feine, Morana e Gnewuch (2019) sugeriram novas perspectivas ao projeto de *chatbots*, ao apontar que desenvolvedores de *chatbots* (e demais canais de atendimento virtual) considerem a análise de sentimentos para identificar consumidores insatisfeitos — pelos sentimentos negativos — de forma a mitigar falhas de serviço e/ou evasões de clientes descontentes com os atendimentos prestados através do disparo de gatilhos de recuperação de serviço como envio de excusas, oferecimento automático de benefícios (promoções) e encaminhamento do atendimento para especialistas humanos.

Por fim, Feine, Morana e Gnewuch (2019) afirmam que seu trabalho possui algumas limitações que devem ser consideradas:

1. Divergências de desempenho entre múltiplos algoritmos de análise de sentimento;
2. Existência de viéses em qualidades mensuradas por questionários Likert Liang, Zou e Yu (2020), Podsakoff et al. (2003);
3. Incertezas quanto a replicabilidade do experimento para outras aplicações[22], em outras línguas distintas do inglês e mensurados com outras escalas do tipo Likert para mensurar qualidade;
4. Impossibilidade de se prever valores CSES a partir dos sentimentos pela ausência de fundamentos para relações causa-e-efeito;

E para cada uma dessas limitações oferecem respectivas orientações de mitigação (quando possível) para aqueles que desejam replicar e/ou aprofundar seus experimentos:

1. Testar diferentes algoritmos de análise de sentimentos antes de aplicá-los a computar sentimentos Klein, Moon e Picard (2002);
2. Revisar diálogos e garantir que os respondentes respondam de forma diligente;
3. Não é possível mitigar as incertezas;
4. Considerar os resultados de aproximação com ainda mais cautela por não ser possível estabelecer uma relação de causa e efeito;

Por fim, (re)afirmam que as contribuições que alcançaram contribuem para fundamentar trabalhos de pesquisa na análise (em tempo real) do sentimento do usuário visando inferir fraquezas na qualidade do serviço e disparar gatilhos de reparação de relacionamentos junto a clientes, dentre outros. O trabalho de Kim, Levy e Liu (2020), que menciona Feine, Morana e Gnewuch (2019), é discutido na próxima Subseção em aparente continuidade a estas contribuições.

[21] A satisfação, por exemplo.
[22] Venda de passagens aéreas, por exemplo.

3.2.2 Estimativa da satisfação de clientes e análise do sentimento de suas falas durante diálogos com *socialbots*

Kim, Levy e Liu (2020)[23] apresentam uma nova abordagem para estimar automaticamente a satisfação de clientes (CSAT[24] — *Custommer Satisfaction*) através do sentimento (codificado em informações textuais e sonoras) presente na fala de usuários da assistente virtual *Alexa*.

Os autores apresentam esta abordagem por entenderem que medir (e se adaptar) à satisfação dos clientes é essencial para o sucesso de qualquer *chatbot* — orientado a tarefas ou conversacional — pois, mensurar tal satisfação possibilita entender a percepção dos usuários acerca do funcionamento dos agentes e consequentemente, implantar medidas que aumentem o engajamento e a retenção dos usuários.

Entretanto, segundo os autores, os métodos tradicionais para se mensurar qualidade — questionários Likert respondidos por clientes — possuem incovenientes como a impossibilidade de se capturar continuamente as mudanças na satisfação dos clientes, bem como, potenciais distrações que esses questionários causam nos usuários da *Alexa* e por isso, Kim, Levy e Liu (2020) buscam prever avaliações CSAT através do sentimento que esses usuários expressam em suas falas.

Desta forma, os autores inicialmente investigaram a correlação entre avaliações CSAT submetidas por usuários da *Amazon Alexa* em relação às quantificações de sentimento presentes em suas falas durante as interações, em nível de diálogo e de mensagens, através da formulação (e posterior confirmação) da seguinte hipótese **O sentimento do usuário é correlacionado à métrica de qualidade CSAT?** Em seguida investigaram modelos baseados em aprendizado de máquinas para atestar a segunda (e principal) hipótese que levantaram, **A satisfação CSAT em nível de diálogo pode ser predita a partir do sentimento de um usuário?**

Apesar de corresponder à principal contribuição de Kim, Levy e Liu (2020), a segunda hipótese não é tema de estudo da análise principal desta dissertação. Portanto, o desenvolvimento (e verificação) dessa hipótese não é abordado. Realizada esta consideração, as etapas que os autores executaram para investigar a hipótese H_1 são apresentadas:

1. Selecionar conversas da Alexa e seus usuários;
2. Anotar sentimentos em amostras das conversas selecionadas;
3. Construir algoritmo de análise de sentimentos;
4. Verificar correlações de postos de Spearman entre sentimentos das conversas e as qualidades CSAT atribuídas pelos usuários;

[23] Estimativa da satisfação de clientes e análise do sentimento de suas falas durante diálogos com *socialbots* (KIM; LEVY; LIU, 2020).

[24] CSAT: uma aproximação de qualidade usada por Kim, Levy e Liu (2020) que consiste em uma escala 5-estrelas do tipo Likert que inquire os usuários sobre como eles se sentem, se existem chances (em um intervalo de 1 a 5) em conversar com o *socialbot* outra vez.

Por terem possuído vínculos com a empresa *Amazon*, Kim, Levy e Liu (2020) tiveram acesso a diálogos entre usuários da *Alexa* e *chatbots* construídos por participantes da competição *Amazon Alexa Prize* (KHATRI et al., 2018) e assim, selecionaram aleatoriamente 6.308 diálogos (com satisfação CSAT avaliada pelos usuários) dessa competição. Esta seleção de diálogos não possui referências (verdade fundamental — *ground truth*) de sentimentos anotadas por seres humanos e por isso, compõem uma base de dados nomeada pelos autores como **diálogos não rotulados do *Alexa Prize***. Para construir um conjunto referência de sentimentos, Kim, Levy e Liu (2020) selecionaram aleatoriamente outras 952 conversas da competição *Alexa Prize*, cujos aspectos de sentimento foram anotados por especialistas humanos, sendo esse conjunto de conversas nomeado **diálogos rotulados do *Alexa Prize***. Adicionalmente, ocultaram de ambos os conjuntos de diálogos as mensagens de avaliações CSAT reportadas pelos usuários para garantir que os modelos preditivos (não abordados, construídos para verificar H_2) não tenham quaisquer informações sobre as referências (verdades fundamentais) das avaliações CSAT.

Para anotar sentimentos das 952 conversas do conjunto **diálogos rotulados do *Alexa Prize***, os autores instruíram especialistas humanos a rotular os sentimentos de cada mensagem do conjunto em três dimensões — Ativação, Valência e Satisfação — considerando aspectos do tom da fala e das palavras usadas pelos usuários da *Alexa*. Cada mensagem foi rotulada por ao menos três especialistas que tiveram de atribuir valores de -3 a $+3$ (com o valor 0 indicando neutralidade) para cada dimensão de sentimento considerada.

Ativação (empolgado / calmo) e valência (positivo / negativo) são aspectos de sentimento frequentemente usados em trabalhos de reconhecimento de sentimentos da fala (WU; PARSONS; NARAYANAN; GHOSH et al.; EYBEN et al., 2010, 2016, 2010 apud KIM; LEVY; LIU, 2020). Já a satisfação pode ser considerada similar à valência, mas, segundo Kim, Levy e Liu (2020), deve ser interpretada exclusivamente no âmbito das aplicações de assistentes por voz. Por isso, orientaram os especialistas a considerar a satisfação como um atributo da fala que provê informações relativas ao nível de satisfação/prazer (satisfação positiva) ou insatisfação/descontentamento (satisfação negativa) presente em trechos das falas dos usuários da *Alexa*. Portanto, os autores indicam que a satisfação captura o grau de prazer ou descontentamento de um usuário ao utilizar a assistente virtual ao passo que a valência captura a positividade do usuário como um todo.

Ao terem os sentimentos das 952 conversas anotados por especialistas humanos, Kim, Levy e Liu (2020) usaram as médias desses sentimentos como verdade fundamental treinar um algorítmo de análise de sentimento baseado em técnicas de aprendizado de máquinas.

Em seguida, computaram os sentimentos das 6.308 conversas do conjunto de **diálogos não rotulados** e calcularam as correlações de postos Spearman dos sentimentos computados pelo algoritmo com os valores CSAT dessas 6.308 conversas — relatadas pela Tabela 1 — como também, calcularam as correlações das médias de sentimentos anotados pelos especialistas em relação às avaliações CSAT dos 952 diálogos que as originaram — correlações relatadas pela Tabela 2.

Tabela 1 – Correlação de postos ρ de Spearman entre as médias dos valores de sentimento (Val: valência, Atv: ativação, Sat: satisfação) e avaliação CSAT dos diálogos em nível de conversas e em nível de mensagens para as conversas do conjunto **diálogos não rotulados do *Alexa Prize***

	ρ (Val, CSAT)	ρ (Atv, CSAT)	ρ (Sat, CSAT)
Nível de conversas	**0,1919**	0,046	0,1677
Nível de mensagens	0,0734	0,0427	0,0686

Fonte: Kim, Levy e Liu (2020)

Tabela 2 – Correlação de postos ρ de Spearman entre as médias dos valores de sentimento (Val: valência, Atv: ativação, Sat: satisfação) e avaliação CSAT dos diálogos em nível de conversas e em nível de mensagens para as conversas do conjunto **diálogos rotulados do *Alexa Prize***

	ρ (Val, CSAT)	ρ (Atv, CSAT)	ρ (Sat, CSAT)
Nível de conversas	**0,2584**	0,044	**0,2847**
Nível de mensagens	0,0987	0,0236	0,1133

Fonte: Kim, Levy e Liu (2020)

Os resultados obtidos por Kim, Levy e Liu (2020) possívelmente validam a hipótese H_1 para as dimensões de valência e satisfação do sentimento em nível de diálogos. Ressalta-se que o grifo de possívelmente não é manifestado por Kim, Levy e Liu (2020), mas sim, pela ausência do **valor-p** vinculado às correlações calculadas, pois, em análises de correlação, é essencial conhecer o **valor-p** para saber se as correlações obtidas correspondem à realidade (**valor-p** < 0.05) ou se poderiam ser obtidas ao acaso, ou seja, quando não existe correlação entre as variáveis estudadas. Sem o **valor-p**, é impossível extrapolar as inferências para além das amostras analisadas — para todo o conjunto de dados — e portanto, como Kim, Levy e Liu (2020) não o apresentaram, só é possível interpretar a existência das correlações para as amostras de diálogos que selecionaram.

Sintetizando oportunidades de contribuições à literatura a partir da leitura de Feine, Morana e Gnewuch (2019) e Kim, Levy e Liu (2020) verifica-se a possibilidade de aplicar a revisão manual de diálogos recomendada por Logacheva et al. (2018) [25] sobre o corpo de diálogos ConvAI, bem como, a necessidade de se apresentar a relevância estatística (o **valor-p**) da correlação entre sentimento e qualidade dos diálogos, para generalizar inferências para todo um conjunto de diálogos.

Assim, a revisão manual de diálogos[26], a análise das correlações entre sentimentos e qualidade dos diálogos [27], o impacto da revisão de diálogos sobre a correlação entre sentimento e qualidade, bem como, a relevância estatística das correlações calculadas são contribuições do presente trabalho relatadas no Capítulo 4 e verificadas no Capítulo 5.

[25] Reproduzida de forma similar por Kim, Levy e Liu (2020) sobre o corpo de diálogos *Amazon Alexa Prize*.
[26] E construção de ferramenta para sua realização.
[27] Antes e depois de serem revisados.

4 Metodologia

Conversas com qualidades já atribuídas pelos usuários dos *chatbots* (presentes no ConvAI) são objetos favoráveis para a verificação da correlação entre qualidades e sentimentos textuais, pois, apresentam as duas características necessárias para tal: qualidades anotadas pelos usuários e sentimentos expressos nas mensagens que enviaram aos agentes conversacionais.

Feine, Morana e Gnewuch (2019) estudaram o ConvAI, utilizando-o para verificar o desempenho de algoritmos de análise de sentimento que prospectaram, mas, não apresentaram resultados da correlação entre sentimentos dos diálogos e as qualidades atribuídas pelos seus partícipes, não tendo considerado a existência de diálogos avaliados displicentemente[1] no ConvAI. Desta forma, é necessário investigar o impacto de avaliações displicentes de qualidades dos diálogos na correlação entre sentimentos e respectivas qualidades por meio da realização das seguintes etapas da metodologia:

1. Filtragem e seleção de diálogos do ConvAI avaliados displicentemente (Seção 4.1);
2. Reclassificação de qualidade e anotação manual de sentimentos textuais das conversas por revisores humanos (Seção 4.2);
3. Cálculo das correlações entre qualidade e sentimento dos diálogos presentes no ConvAI, antes e depois das reclassificações de qualidade (Seção 4.3);

Cada uma dessas etapas é relatada por sua respectiva seção neste capítulo, iniciando, a seguir, pela primeira Seção.

4.1 Filtragem de diálogos do ConvAI avaliados displicentemente

Como exemplificado por um diálogo representado no Quadro 4, reproduzido de Logacheva et al. (2018), existem diálogos avaliados de maneira displicente pelos usuários dos *chatbots* registrados no ConvAI.

Por isso, ao efetuarem recomendações àqueles que pretendem utilizar os diálogos presentes no ConvAI para treinar novos *chatbots* [2], Logacheva et al. (2018) preconizam que reclassificações nas qualidades dos diálogos [3] sejam realizadas por comitês de avaliadores humanos e que ao menos os diálogos cujas qualidades tenham sido avaliadas com pontuação máxima sejam revisados [4].

---

[1] Como relatado por Logacheva et al. (2018).
[2] Ou propor novas métricas de qualidade destes agentes.
[3] Para mitigar efeitos das avaliações displicentes dos usuários.
[4] Pois, os autores identificaram que as avaliações máximas dos usuários dos *chatbots* do ConvAI foram as que tiveram maior displicência.

Quadro 4 – Exemplo de diálogo mal sucedido entre um humano e um robô com qualidade máxima atribuída de forma displicente presente na base ConvAI

Contexto	
`In 550 BC, Cyrus the Great, son of Mandane and Cambyses I, took over the Median Empire, and founded the Achaemenid Empire by unifying other city states. The conquest of Media was a result of what is called the Persian Revolt. The brouhaha was initially triggered by the actions of the Median ruler Astyages, and was quickly spread to other provinces, as they allied with the Persians. Later conquests under Cyrus and his successors expanded the empire to include Lydia, Babylon, Egypt, parts of the Balkans and Eastern Europe proper, as well as the lands to the west of the Indus and Oxus rivers`	
Interlocutor	**Mensagem**
Humano	`I will learn that the conquest of Media was a result of what is called the Persian Revolt. Who told you that?`
Humano	`Hello!`
Robô	`a result of`
Humano	`How old are you?`
Robô	`spread`
Humano	`Spread what?`
Robô	`The`

Fonte: Logacheva et al. (2018, p. 6)

Como as conversas do corpo de diálogos foram avaliadas em escala Likert tridimensional (que abrange dimensões de engajamento, profundidade do diálogo e qualidade geral da conversa), então, a princípio, todas essas dimensões deveriam ser consideradas para seleção dos diálogos com maior avaliação. Entretanto, ao avaliar as correlações entre essas dimensões, Logacheva et al. (2018) constataram que todas são fortemente correlacionadas[5] e que teria sido melhor, para a construção do conjunto de diálogos, que apenas uma dimensão (exemplo: qualidade geral da conversa) fosse considerada para avaliar os diálogos[6].

Assim, seguindo essas recomendações, foi escolhida a qualidade geral dos diálogos como dimensão de qualidade a ser considerada e foram selecionadas todas as conversas cujas qualidades gerais obtiveram pontuação máxima — 5 — junto aos usuários, o que deu um total de 112 conversas selecionadas em um total de 2282 diálogos entre humanos e robôs presentes no *dataset*[7]. Admitindo-se a possibilidade de que possam existir conversas avaliadas de forma

[5] **r-pearson** entre 0.86 e 0.87 entre quaisquer dessas dimensões (LOGACHEVA et al., 2018, p. 5).

[6] Logacheva et al. (2018) recomendaram que em versões posteriores do ConvAI, como de fato ocorre no ConvAI2, uma escala Likert unidimensional de 5 pontos fosse utilizada para avaliar qualidade.

[7] (FEINE; MORANA; GNEWUCH, 2019, p. 8).

displicente nos outros quatro pontos de intensidade Likert da qualidade geral, foram selecionadas amostras aleatórias de 112 conversas para cada uma das outras pontuações de qualidade geral (1 a 4)[8] totalizando 560 conversas com qualidades avaliadas.

Em seguida, como Feine, Morana e Gnewuch (2019) encontraram (e desconsideraram) diálogos vazios no ConvAI ao analisar o sentimento dos diálogos, foi realizada uma busca por diálogos vazios neste conjunto de 560 conversas, dos quais foram encontrados e removidos 101 diálogos, pois, não faz sentido submetê-los ao escrutínio da verificação de correlação de sentimento. Dessa forma, o conjunto de diálogos a ter qualidades reclassificadas e sentimentos textuais anotados (procedimentos relatados na próxima Seção) é composto por 459 conversas.

4.2 Reclassificação de qualidade e anotação manual de sentimentos textuais das conversas por revisores humanos

Como pré-requisito para reclassificação de qualidade e anotação manual de sentimentos foi necessário construir uma ferramenta de rotulação de diálogos para que os membros dos comitês — componentes do projeto Oficinas 4.0[9] — pudessem executar, de forma eficiente, estas tarefas. Nesse sentido, foram elicitados requisitos para tornar o funcionamento da ferramenta o mais amigável possível — considerando a (possível) heterogeneidade dos componentes dos comitês — para seus operadores, que foram:

1. Funcionar em múltiplas plataformas e múltiplos dispositivos (com responsividade);
2. Ser capaz de manipular arquivos JSON em operações de leitura e escrita;
3. Permitir rotulação da qualidade do(s) diálogo(s) em muito ruim, ruim, razoável, bom e muito bom;
4. Permitir rotulação do sentimento do(s) diálogo(s) em negativo, neutro, positivo e desconhecido (não sei);
5. Permitir rotulação do sentimento de cada expressão do(s) diálogo(s) em negativo, neutro, positivo e desconhecido (não sei);
6. Ser de fácil utilização;

Em complementação, o segundo requisito da ferramenta é díspar em relação às necessidades dos usuários, pois é vinculado exclusivamente ao ConvAI, disponibilizado por Logacheva et al. (2018) em formato JSON, como mostrado pela Figura 12.

[8] 112 conversas aleatórias com qualidade geral 1, 112 conversas aleatórias com qualidade geral 2 e assim por diante.

[9] As oficinas 4.0 são um projeto de oficinas extracurriculares orientadas à elaboração de soluções de demandas reais originárias do setor produtivo fomentadas pela Secretaria de Educação Profissional e Tecnológica do Ministério da Educação (Setec/MEC) no edital nº 67/2021 do Ministério da Educação (MATOS, 2021).

Figura 12 – Diálogo presente no *ConvAI* em sua representação no formato JSON


```
{'context': 'The decline of the city reached its nadir with the War of Spanish '
            'Succession (1702-1709) that marked the end of the political and '
            'legal independence of the Kingdom of Valencia. During the War of '
            'the Spanish Succession, Valencia sided with Charles of Austria. '
            'On 24 January 1706, Charles Mordaunt, 3rd Earl of Peterborough, '
            '1st Earl of Monmouth, led a handful of English cavalrymen into '
            'the city after riding south from Barcelona, capturing the nearby '
            'fortress at Sagunt, and bluffing the Spanish Bourbon army into '
            'withdrawal.',
 'dialogId': -315877751,
 'evaluation': [{'breadth': 2, 'engagement': 1, 'quality': 4, 'userId': 'Bob'},
                {'breadth': 1,
                 'engagement': 1,
                 'quality': 1,
                 'userId': 'Alice'}],
 'id': -315877751,
 'qualidade_avaliada': False,
 'sentiment': None,
 'sentimento_avaliado': False,
 'sentimento_expressoes_avaliado': False,
 'thread': [{'evaluation': 0,
             'text': 'Hi! As for me, I thought Spain got united much earlier',
             'time': -1,
             'userId': 'Bob'}],
 'users': [{'id': 'Alice', 'userType': 'Human'},
           {'id': 'Bob', 'userType': 'Human'}]}
```


Fonte: Logacheva et al. (2018)

Dessa forma, se construiu a ferramenta de rotulação (com o auxílio de membros do projeto oficinas 4.0 para estilização da ferramenta e funcionamento em dispositivos móveis) que permite a reclassificação dos diálogos. As Figuras 13, 14 e 15 mostram, respectivamente, o fluxograma (simplificado) de funcionamento da ferramenta, sua execução em ambiente *desktop* e sua execução em dispositivo móvel.

Figura 13 – Fluxograma (simplificado) do rotulador de diálogos

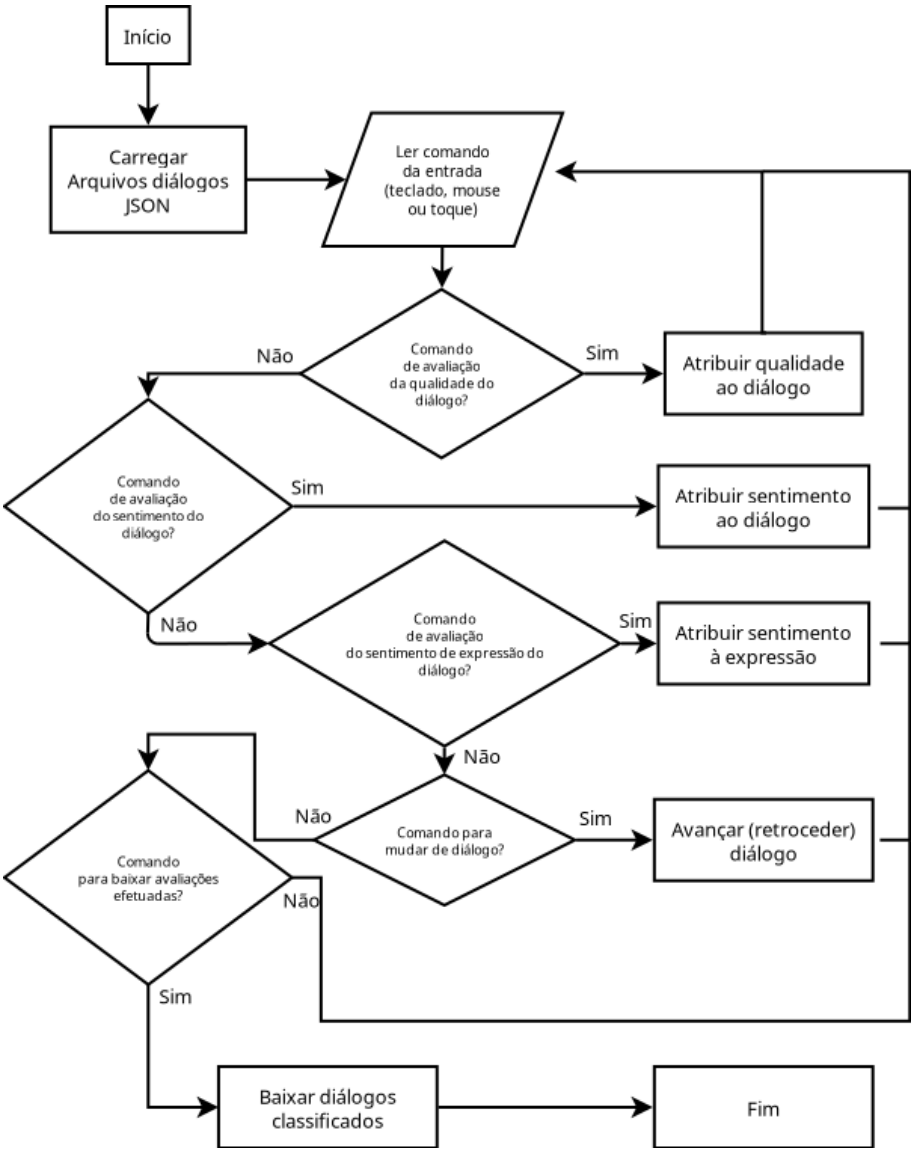


Figura 14 – Rotulador sendo executado em uma plataforma *Desktop*

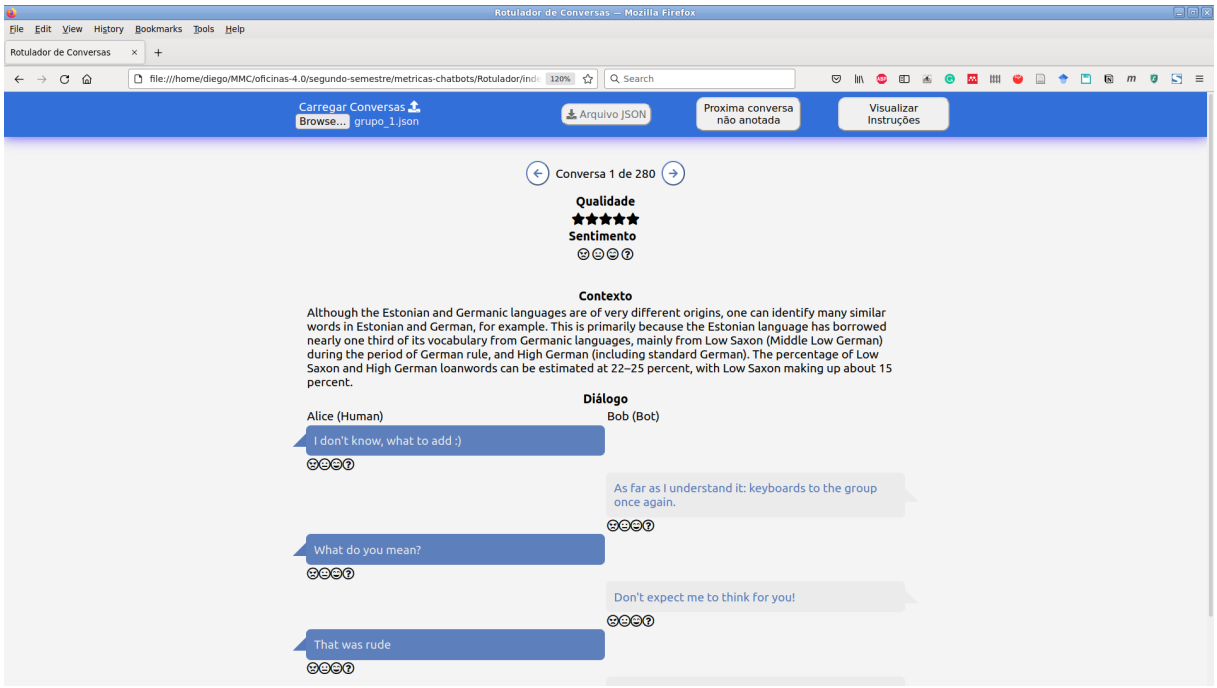
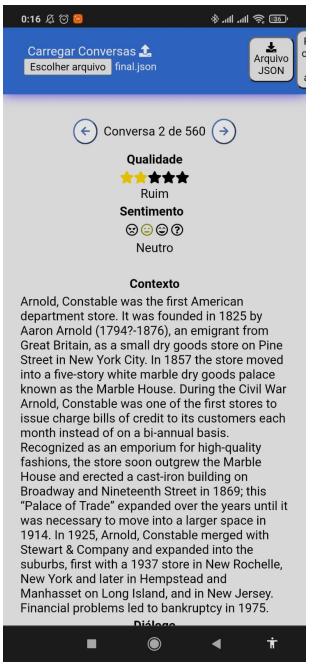


Figura 15 – Rotulador sendo executado em uma plataforma *mobile*



Com a existência da ferramenta de rotulação de diálogos é possível que componentes dos comitês de reclassificação da qualidade dos diálogos realizem de forma eficiente a sua tarefa, como também, a anotação dos sentimentos textuais das conversas e das mensagens presentes nelas, pois, desta forma, convergem-se esforços para a produção dos dados necessários para os estudos de correlação entre sentimento e qualidade.

Assim, para reclassificar os diálogos do ConvAI, foram divididos os recursos humanos disponibilizados pelo projeto oficinas 4.0[10] em dois comitês de três pessoas responsáveis por rotular porções distintas dos 459 diálogos filtrados do ConvAI[11] (comitê 1 — 230 diálogos; comitê 2 — 229 diálogos) em que cada membro de cada comitê foi orientado a rotular cada um dos diálogos designados para seu grupo em relação aos atributos de qualidade geral, sentimento do diálogo e sentimento das mensagens (de cada mensagem presente nas conversas), sendo explicitamente instruídos[12] das seguintes instruções:

- Anotar a qualidade geral dos diálogos entre humanos e *chatbots* atribuindo a ela os seguintes valores:

    1. Muito Ruim: diálogos com termos inaceitáveis (ex. xingamentos) e/ou nenhuma discussão a respeito do tópico proposto para o diálogo (por parte do *chatbot*);
    2. Ruim: diálogos que começaram bem (discutindo o tópico proposto) mas que por qualquer razão degradaram ao longo da conversa, com expressões sem sentido e/ou termos inaceitáveis proferidos pelo *chatbot*;
    3. Razoável: diálogos em que houveram erros, mas, que no geral apresentaram uma discussão adequada a respeito do tópico proposto (por parte do *chatbot*);
    4. Bom: diálogos que apresentaram discussões adequadas a respeito do tópico proposto com pouquíssimos erros (por parte do *chatbot*);
    5. Muito Bom: diálogos que apresentaram discussões perfeitas (sem nenhum erro) a respeito do tópico proposto (por parte do *chatbot*)

- Anotar o sentimento dominante dos diálogos (e de cada uma de suas expressões) entre humanos e *chatbots* atribuindo a ele os seguintes valores:

    – (-1) Negativo: diálogos (expressões) em que é possível inferir que uma (ou ambas) das partes expressa raiva, hostilidades, xingamentos, irritações, críticas, angústia, tristeza, rejeição, terror, medo dentre outros sentimentos negativos e que esses sentimentos predominam no diálogo (expressão);

    – (0) Neutro: diálogos (expressões) em que não existe tanta polaridade (satisfação / insatisfação, positivo / negativo) expressada pelos participantes. Expressões factuais (ex.

[10] Seis estudantes de nível técnico e de mestrado do CEFET-MG
[11] Seção 4.1
[12] Por meio da feramenta de rotulação de diálogos.

O Brasil foi descoberto no ano de 1500), expressões curtas e perguntas predominam nesses diálogos (expressões);

- (+1) Positivo: diálogos (expressões) em que é possível inferir que uma (ou ambas) das partes expressa alegria, entusiasmo, elogios, confiança, admiração, otimismo, tranquilidade, orgulho, dentre outros sentimentos positivos e que esses sentimentos predominam no diálogo (expressão);
- (?/0) Não Sei: diálogos (expressões) em que não é possível inferir com clareza a existência de polaridade (positiva / negativa) ou a própria polaridade em si: se o diálogo (a expressão) assume uma conotação negativa ou positiva.

Após a reclassificação dos diálogos, os conjuntos de dados produzidos pelos comitês estão aptos a processamento por um sistema de consenso capaz de computar os *scores* consolidados de qualidade e sentimento dos diálogos. Este sistema de consenso é baseado nas médias das avaliações de qualidade e sentimento (em nível de diálogo e em nível de conversas) atribuídas por cada um dos componentes dos comitês de avaliadores, para que as médias das avaliações de cada comitê se tornem o referencial (*ground truth*) de qualidades e sentimentos computados por avalidores humanos sobre o ConvAI, permitindo assim, o estudo de correlação entre qualidade e sentimento, abordado na próxima Seção.

4.3 Cálculo das correlações entre qualidade e sentimento dos diálogos presentes no ConvAI, antes e depois das reclassificações de qualidade

Conforme relatado na Seção 2.4, a correlação é uma ferramenta estatística que estuda o quão dependentes duas variáveis são entre si, ou seja, o quanto a variação dos valores de uma interfere na variação dos valores de outra. Assim, para verificar a hipótese proposta pelo trabalho, de que o sentimento pode ser utilizado como uma medida de aproximação para a qualidade das interações entre humanos e chatbots, a correlação é uma ferramenta potencialmente apropriada para isto, visto que, seus resultados fornecem respostas concretas para a hipótese levantada, pois, se houver correlação entre sentimento e qualidade, então, o sentimento contém aproximações (*proxies*) da qualidade.

Dessa forma, em posse dos dados referenciais [13] de qualidade e sentimento dos diálogos do ConvAI — antes e depois das reavaliações de diálogos — é possível calcular correlações[14] **r-pearson** e correlações de postos ρ**-spearman** para verificar se o sentimento textual pode ser usado como aproximação da qualidade, a caracterização da correlação entre sentimento e qualidade [15] antes e depois das reclassificações dos diálogos displicentes e o impacto que a

[13] Também conhecido como *ground truth*.
[14] Neste momento, apenas para qualidades e sentimentos em nível de diálogo.
[15] Sob a interpretação proposta por Cohen (1992).

filtragem (e ajuste) das avaliações displicentes produz sobre a correlação entre sentimento e qualidade.

Para realizar tal cálculo, os conjuntos de dados foram submetidos a análise de ferramentas estatísticas disponilizadas por bibliotecas da linguagem de programação *Python*[16] capazes de calcular correlações **r-pearson** e ρ**-spearman**[17] com **valores-p** que apontam a relevância estatística dos dados computados. Os valores de correlação calculados, as relevâncias estatísticas, os gráficos de dispersão entre sentimento e qualidade, as distribuições de frequências de qualidade e de sentimento e por fim, as discussões que fundamentam a correlação entre essas variáveis no ConvAI, existindo ou não avaliações displicentes, são abordados no próximo Capítulo.

[16] Uma linguagem de programação de alto nível concebida para propósitos genéricos PYTHON (2021).

[17] Outras correlações de postos como a τ de kendall não foram utilizadas devido a dificuldade do pleno entendimento dos fundamentos sob os quais se baseiam.

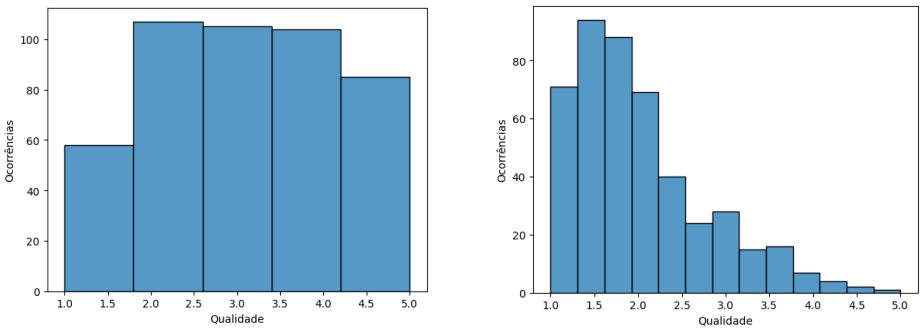
5 Resultados e Discussões

Este Capítulo de Resultados e Discussões apresenta na Seção 5.1 comparativos das qualidades dos diálogos do ConvAI bem como a relação entre as qualidades desses diálogos e os sentimentos neles anotados. Na Seção 5.2 são apresentados resultados das correlações **r-pearson** e ρ**-spearman** entre qualidade e sentimento dos diálogos antes e depois das reclassificações dos diálogos, sendo tecidas na Seção 5.3 considerações acerca dos impactos causados por diálogos avaliados displicentemente.

5.1 Discussões sobre a reclassificação de qualidade do ConvAI e anotação dos sentimentos dos diálogos

A reclassificação da qualidade dos 459 diálogos extraídos do ConvAI reafirma a existência de diálogos displicentemente avaliados, como demonstrado pelas distribuições das frequências das qualidades da amostra de diálogos do ConvAI antes e depois da reclassificação, representadas na Figura 16, em que é possível verificar as constatações de Logacheva et al. (2018), que os diálogos avaliados com nota máxima apresentaram maior displicência e mais além, que os diálogos avaliados com notas 3 e 4 também estão passíveis a terem sido avaliados de forma displicente.

Figura 16 – Distribuições das frequências (histograma) das qualidades de amostras de conversas (n = 459) do ConvAI antes e depois da reclassificação



(a) Distribuição das frequências das qualidades de 459 conversas antes da reclassificação

(b) Distribuição das frequências das qualidades de 459 conversas depois da reclassificação

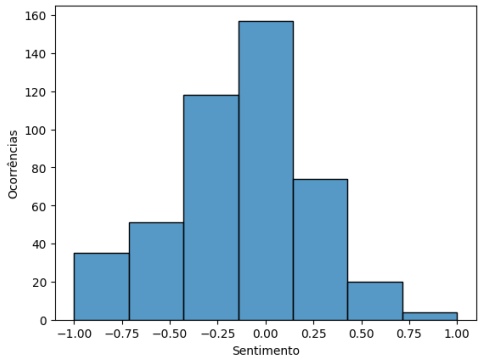
A displicência das avaliações efetuadas pelos usuários indubitávelmente causa prejuízo à real quantificação da qualidade dos agentes conversacionais. E ao prospectar o sentimento como métrica de aproximação da qualidade de *chatbots*, se a qualidade aferida não corresponder a realidade — for negligente — logo, o sentimento não será uma boa aproximação da qualidade.

Portanto, é precipitado desconsiderar o sentimento como aproximação de qualidade se esta estiver sujeita a viéses de displicência. Como no ConvAI houve grande quantidade de

diálogos avaliados de forma desatenta, destarte, a correlação entre sentimento e qualidade tende a ser fraca, o que corrobora a não demonstração dos resultados de sentimento e qualidade do corpo ConvAI no trabalho de Feine, Morana e Gnewuch (2019), visto que não demonstraram a revisão dos diálogos desse corpo em busca de conversas classificadas displicentemente.

No tocante aos sentimentos anotados durante a reclassificação dos 459 diálogos, esses são majoritariamente neutros e possuem mais sentimentos negativos do que positivos, como representado pela Figura 17. Mesmo assim, por serem concentrados em torno do valor médio — neutro — dos sentimentos, podem ser aproximados por uma curva gaussiana (em forma de sino).

Figura 17 – Histograma dos sentimentos dos diálogos anotados durante a reclassificação de qualidade



Se a distribuição das qualidades dos diálogos após a reclassificação seguisse uma distribuição similar à dos sentimentos[1], seria possível utilizar a correlação **r-pearson** para investigar a correlação entre sentimento e qualidade não somente para as amostras de diálogos estudadas como também para toda a população do ConvAI.

Como isso não é possível[2], a correlação ρ**-spearman** é usada para fazer as inferências de correlação para toda a população de diálogos do ConvAI ao passo que a correlação **r-pearson** é restrita à amostra selecionada de diálogos do ConvAI. Os resultados dessas correlações, mostrados na Seção 5.2, demonstram o prejuízo das avaliações displicentes e a melhoria que a remoção dos viéses por meio da avaliação de comitês proporciona na correlação entre sentimento e qualidade.

5.2 Considerações em relação à literatura e resultados das correlações r-pearson e ρ-spearman entre sentimento e qualidade antes e após a reclassificação dos diálogos

Nesta subseção o presente trabalho é situado dentre as referências da literatura mais relevantes para sua existência — Feine, Morana e Gnewuch (2019), Kim, Levy e Liu (2020),

[1] Que pudesse ser aproximada por uma curva normal (gaussiana)
[2] Ver Seção 2.4.

Logacheva et al. (2018) — nos aspectos que possui em comum a elas para, em sequência, serem apresentados:

- Os resultados de correlação encontrados sobre o ConvAI (presentes na Tabela 3);
- As visualizações gráficas das dispersões entre sentimento e qualidade (presentes na Figura 18) que ajudam a interpretar os resultados obtidos.

Situando o trabalho junto às referências da literatura, em relação a Feine, Morana e Gnewuch (2019) ocorre uma contribuição à literatura ao elencar os resultados de correlação **ρ-spearman**[3] e **r-pearson**[4] do corpo de diálogos ConvAI que não foram apresentados por estes autores que também estudaram o ConvAI. Já em relação a Kim, Levy e Liu (2020) ocorre uma contribuição à literatura ao se apresentarem valores de relevância estatística (valor-p) para correlações **ρ-spearman** entre qualidade e sentimento textual antes e depois da reclassificação dos diálogos.

Apesar de serem distintos os corpos de diálogos abordados pelo presente trabalho e por Kim, Levy e Liu (2020), ainda assim, a apresentação do valor-p é uma contribuição que contribui à literatura pelo fato de que a correlação encontrada no presente trabalho pode ser generalizada para todos os diálogos do corpo estudado e o isto, não pode ser realizado em Kim, Levy e Liu (2020) porque eles não apresentaram os resultados de relevância estatística das correlações entre sentimento e qualidade no corpo de diálogos que abordaram.

Por fim, um fato da literatura — a existência de diálogos displicentes no ConvAI Logacheva et al. (2018) — é reafirmado, sendo um fator que prejudica diretamente a correlação entre sentimento e qualidade — como pode ser observado na Tabela 3 — e que justifica a prática adotada na literatura de se realizar a revisão das avaliações de qualidade por comitês de avaliadores humanos:

Tabela 3 – Correlações entre sentimento e qualidade de 459 diálogos do ConvAI antes e depois de revisões de qualidade efetuadas por comitês de avaliadores humanos

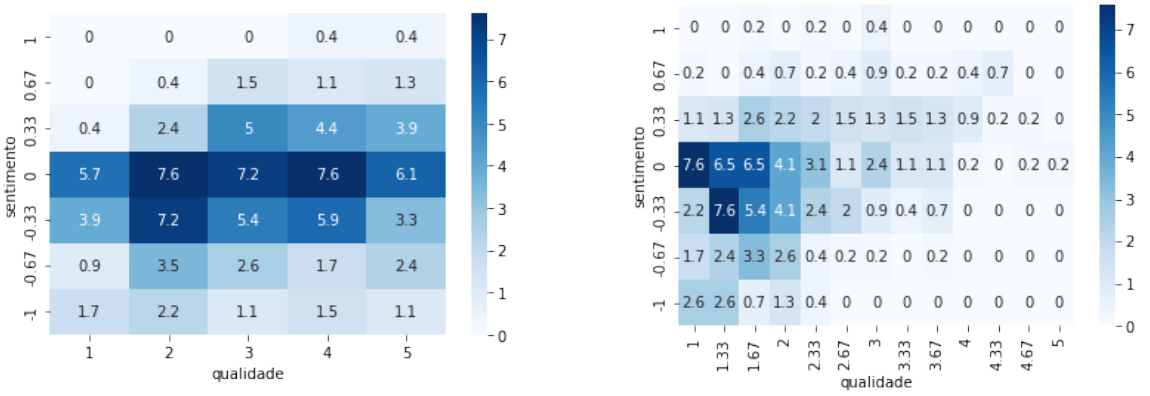
	ρ-spearman	**r-pearson**
Antes da revisão de qualidade	0,1789, valor-p < 0,001	0,1820
Depois da revisão de qualidade	0,3221, valor-p << 0,001	0,3708

Os resultados presentes na tabela 3 realçam o prejuízo que as avaliações displicentes causam na correlação entre qualidade e sentimento. Comparando-se os valores da correlação **ρ-spearman** — generalizados para todos os diálogos do ConvAI devido ao valor-p < 0.001 — observa-se uma melhoria de 80% nessa correlação após a revisão da qualidade dos diálogos displicentes. Esta melhoria verificada pelo aumento da correlação ainda pode ser visualizada nos mapas de calor presentes na Figura 18, que oferecem um panorama em termos percentuais

[3] Que permite fazer inferências para toda a população de diálogos do ConvAI.
[4] Restrito apenas à amostra de 459 diálogos estudadas.

das amostras dos diálogos do ConvAI em relação aos valores de qualidade e sentimento que possuem.

Figura 18 – Mapas de calor das qualidades e dos sentimentos de amostras de conversas (n = 459) do ConvAI antes e depois da reclassificação



(a) Mapa de calor das qualidades e dos sentimentos em termos percentuais do total de conversas antes da reclassificação

(b) Mapa de calor das qualidades e dos sentimentos em termos percentuais do total de conversas depois da reclassificação

A partir da interpretação dos mapas de calor dos diálogos, observa-se que a reclassificação dos diálogos aumentou a monotonicidade — a preservação da ordem — da relação qualidade-sentimento, o que faz com que a correlação de postos (ρ**-spearman**) seja mais forte, como é possível observar pela dispersão muito maior dos valores ao longo das células antes da reclassificação e após esta, uma concentração em torno de células específicas. Por último, os mapas de calor também ajudam a identificar a efetividade das reclassificações de diálogos, pois, demonstram que a maioria dos diálogos são de qualidade ruim e que os diálogos de boa qualidade (qualidade ≥ 4) não apresentam sentimentos negativos.

5.3 Discussões Finais

Os resultados de correlação obtidos, segundo interpretação de Cohen (1992 apud FEINE; MORANA; GNEWUCH, 2019), demonstram que antes das reclassificações das amostras dos diálogos — ρ**-spearman** = 0,1789, valor-p < 0,001 e **r-pearson** = 0,1820 — existe uma correlação fraca entre sentimento e qualidade e que após a reclassificação das avaliações de qualidade displicentes — ρ**-spearman** = 0,3221, valor-p << 0,001 e **r-pearson** = 0,3708 — se torna uma correlação moderada. Por fim, a demonstração dos valores-p das correlações ρ**-spearman** atesta a validade dos valores obtidos não só para as amostras populacionais, como também, para toda a população de diálogos do ConvAI a ser submetida a experimentos similares, visto que, a probabilidade de que se alcancem resultados similares ao acaso — quando não existir correlação entre sentimento e qualidade na população de todos os diálogos do ConvAI — é inferior a 0,1%.

Os resultados, portanto, apoiam as evidências da literatura — Tausczik e Pennebaker; Nerbonne; Küster e Kappas (2010, 2014, 2017 apud FEINE; MORANA; GNEWUCH, 2019)

e Kim, Levy e Liu (2020) — que apontam que a qualidade dos *chatbots* está relacionada ao sentimento expresso pelos seus usuários e corroboram para a prática de submeter diálogos a avaliações de comitês de julgadores externos — como adotado pela indústria e pela academia (LIANG; ZOU; YU, 2020) — para a remoção de viéses de displicência, devido a melhoria que tais práticas promovem na correlação entre sentimento e qualidade.

Destarte, em desfecho dos artefatos expressos, o sentimento textual é indicador de qualidade dos *chatbots* do ConvAI, como também, a submissão de conversas a avaliações de julgadores designados, deve ser considerada (ao menos em nível amostral) para mitigação de displicências e/ou tendências, alcançando-se assim os objetivos deste trabalho. Considerações ulteriores e oportunidades de experimentos futuros são tecidas no Capítulo de Conclusão (6) que vem em sequência.

6 Conclusão

Retomando a pergunta de pesquisa: **O sentimento textual de diálogos entre humanos e *chatbots* oferece indicativos da qualidade de desempenho?** Os experimentos realizados atestam que sim — restritos ao corpo de diálogos ConvAI — e mais ainda, que mesmo com a presença de avaliações displicentes, ainda assim, existe uma correlação entre sentimento e qualidade que não é desprezível, apesar de ser fraca (COHEN, 1992). Portanto, ao menos em contextos similares ao ConvAI, a análise de sentimentos dos diálogos oferece perspectivas que não podem ser descartadas como indicativos de qualidade.

Os experimentos realizados atestam também a importância da prática de submeter avaliações de diálogos a revisões que removam viéses e corrijam avaliações displicentes — como realizado na literatura, por Kim, Levy e Liu (2020), Liang, Zou e Yu (2020), Logacheva et al. (2018) — pois, estas revisões tornam as avaliações de qualidade mais condizentes com a realidade, refletindo melhor os desempenhos dos agentes conversacionais responsáveis pelas interações junto aos usuários. Mais ainda, estas revisões aprimoram o sentimento textual como um indicativo de qualidade de *Chatbots*, pois, após estas revisões, a correlação entre sentimento e qualidade aumentou, passando de fraca a moderada (COHEN, 1992).

Por fim, além de alcançar o objetivo principal — demonstrar que o sentimento textual de diálogos entre humanos e *chatbots* oferece indicativos da qualidade do desempenho destes agentes — foi possível construir artefatos auxiliares como:

- Uma ferramenta de reclassificação (e anotação de sentimentos) de diálogos;
- Uma versão de 459 diálogos do ConvAI com qualidade revisada e sentimentos anotados;
- Objetos estatísticos como histogramas e cálculos de correlações sobre o ConvAI;

6.1 Trabalhos Futuros

São objetos de estudo para trabalhos futuros eventualmente derivados desta dissertação:

- Disponibilizar os artefatos auxiliares em repositórios públicos;
- Verificar a correlação entre sentimento e qualidade em nível de mensagens — sequências de mensagens;
- Utilizar algoritmos de análise de sentimento para calcular o sentimento dos diálogos e de sequências de mensagens;
- Replicar os experimentos sobre outro(s) corpo(s) de diálogos;

Referências

ALIANNEJADI, M. et al. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). **arXiv preprint arXiv:2009.11352**, 2020. Citado na página 35.

AMAZON. **A Arte da guerra - Edição luxo (Portuguese Edition) Tzu, Sun 9788560018000 Amazon.com Books**. 2022. Disponível em: <https://www.amazon.com/Arte-Guerra-Capitulos-Originais-Portugues/dp/856001800X>. Acesso em: 7 de janeiro de 2022. Citado na página 34.

AMAZON. **Amazon Mechanical Turk**. 2022. Disponível em: <https://www.mturk.com/>. Acesso em: 27 de janeiro de 2022. Citado na página 37.

AMORIM, P. F. P. d. A crítica de john searle à inteligência artificial: uma abordagem em filosofia da mente. Universidade Federal da Paraíba, 2014. Citado na página 17.

ANTON, C. A.; MATEI, O.; AVRAM, A. D. Survey of collaborative data mining. **Carpathian Journal of Electrical Engineering**, v. 12, n. 1, 2019. Citado na página 29.

APPLIN, S. A.; FISCHER, M. D. New technologies and mixed-use convergence: How humans and algorithms are adapting to each other. In: IEEE. **2015 IEEE international symposium on technology and society (ISTAS)**. [S.l.], 2015. p. 1–6. Citado na página 20.

ARBIDE, A. C.; ROMERO, M.; FERNÁNDEZ, J. M. Affective computing for smart operations: a survey and comparative analysis of the available tools, libraries and web services. **International Journal of Innovative and Applied Research**, v. 5, n. 9, p. 12–35, 2017. Citado na página 38.

AU, N.; NGAI, E. W.; CHENG, T. E. A critical review of end-user information system satisfaction research and a new research framework. **Omega**, Elsevier, v. 30, n. 6, p. 451–478, 2002. Citado na página 20.

BOHUS, D.; HORVITZ, E. Learning to predict engagement with a spoken dialog system in open-world settings. In: **Proceedings of the SIGDIAL 2009 Conference**. [S.l.: s.n.], 2009. p. 244–252. Citado na página 34.

BROOKS, M. et al. Statistical affect detection in collaborative chat. In: **Proceedings of the 2013 Conference on Computer Supported Cooperative Work**. New York, NY, USA: Association for Computing Machinery, 2013. (CSCW '13), p. 317–328. ISBN 9781450313315. Disponível em: <https://doi.org/10.1145/2441776.2441813>. Citado na página 39.

BURNS, A. C.; BUSH, R. F. **Basic marketing research using Microsoft Excel data analysis**. [S.l.]: Prentice Hall Press, 2007. Citado na página 25.

BURTSEV, M.; LOGACHEVA, V. Conversational intelligence challenge: Accelerating research with crowd science and open source. **AI Magazine**, v. 41, n. 3, p. 18–27, 2020. Citado na página 35.

CAMPBELL, A. The sense of well-being in america: Recent patterns and trends. 1981. Citado na página 33.

CARPENTER, R. **Jabberwacky**. 2007. Citado na página 18.

CAVALCANTE, R. N. B.; SOUSA, M. H. R.; SOUSA, J. P. R. de. Localização e distâncias no mapa da cidade: Uma conexão entre matemática e geografia. **Ensino De Matemática**, Clube de Autores, p. 181, 2019. Citado na página 21.

CHEN, C.-W. Five-star or thumbs-up? the influence of rating system types on users' perceptions of information quality, cognitive effort, enjoyment and continuance intention. **Internet Research**, Emerald Publishing Limited, 2017. Citado na página 25.

CHEN, Q. et al. How should i improve the ui of my app? a study of user reviews of popular apps in the google play. **ACM Trans. Softw. Eng. Methodol.**, Association for Computing Machinery, New York, NY, USA, v. 30, n. 3, apr 2021. ISSN 1049-331X. Disponível em: <https://doi.org/10.1145/3447808>. Citado na página 34.

CLAYTON, R. Quantifying and reducing uncertainty in tidal energy yield assessments. University of Exeter, 2020. Citado na página 30.

COHEN, D.; LANE, I. An oral exam for measuring a dialog system's capabilities. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2016. v. 30, n. 1. Citado na página 20.

COHEN, J. A power primer. **Psychological bulletin**, American Psychological Association, v. 112, n. 1, p. 155, 1992. Citado 5 vezes nas páginas 29, 40, 52, 57 e 59.

COLBY, K. M. Human-computer conversation in a cognitive therapy program. Springer, p. 9–19, 1999. Citado na página 18.

COOPER, D.; BLUMBERG, B.; SCHINDLER, P. **EBOOK: Business Research Methods**. [S.l.]: McGraw Hill, 2014. Citado na página 21.

CORRELATION. In: CORRELATION | Meaning & Definition for UK English | Lexico.com. Oxford University, 2021. Disponível em: <https://www.lexico.com/definition/correlation>. Acesso em: 28 de outubro de 2021. Citado na página 29.

DALATI, S. Measurement and measurement scales. In: **Modernizing the Academic Teaching and Research Environment**. [S.l.]: Springer, 2018. p. 79–96. Citado 5 vezes nas páginas 21, 22, 23, 24 e 25.

DALIP, D. H. et al. A general multiview framework for assessing the quality of collaboratively created content on web 2.0. **Journal of the Association for Information Science and Technology**, Wiley Online Library, v. 68, n. 2, p. 286–308, 2017. Citado na página 19.

DEMOS, A.; SALAS, C. **A language not a letter: learning statistics in R**. [S.l.: s.n.], 2017. Citado na página 30.

DINAN, E. et al. The second conversational intelligence challenge (convai2). **arXiv preprint arXiv:1902.00098**, 2019. Citado na página 35.

EEUWEN, M. v. **Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers**. Dissertação (Mestrado) — University of Twente, 2017. Citado na página 20.

EYBEN, F. et al. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. **Journal on Multimodal User Interfaces**, Springer, v. 3, n. 1, p. 7–19, 2010. Citado na página 43.

FEINE, J.; MORANA, S.; GNEWUCH, U. Measuring Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis. **Wirtschaftsinformatik 2019 Proceedings**, feb 2019. Disponível em: <https://aisel.aisnet.org/wi2019/track10/papers/2>. Acesso em: 28 de outubro de 2021. Citado 21 vezes nas páginas 14, 15, 16, 17, 20, 27, 28, 29, 36, 37, 38, 39, 40, 41, 44, 45, 46, 47, 55, 56 e 57.

FERREIRA, G. B. A. **Avaliação da radioatividade natural em águas subterrâneas do município de Campos do Jordão, SP**. Dissertação (Mestrado) — Universidade de São Paulo, 2020. Citado na página 30.

FERREIRA, J. C.; PATINO, C. M. O que realmente significa o valor-p? **Jornal Brasileiro de Pneumologia**, SciELO Brasil, v. 41, p. 485–485, 2015. Citado na página 31.

GAJARLA, V.; GUPTA, A. Emotion detection and sentiment analysis of images. **Georgia Institute of Technology**, p. 1–4, 2015. Citado na página 28.

GAMA, B. C. Avaliação comparativa entre medidas de redes complexas para a classificação de dados. Universidade Federal de Uberlândia, 2020. Citado na página 30.

GHOSH, S. et al. Representation learning for speech emotion recognition. In: . [S.l.: s.n.], 2016. Citado na página 43.

GNEWUCH, U. et al. **ExpBot - A dataset of 79 dialogs with an experimental customer service chatbot**. 2020. Citado 2 vezes nas páginas 14 e 39.

GODES, D.; SILVA, J. C. Sequential and temporal dynamics of online opinion. **Marketing Science**, INFORMS, v. 31, n. 3, p. 448–473, 2012. Citado na página 32.

GONÇALVES, P. d. O. Um benchmark para comparação de métodos para análise de sentimentos. Universidade Federal de Minas Gerais, 2015. Citado na página 28.

GOPALAKRISHNAN, K. et al. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In: **Proc. Interspeech 2019**. [s.n.], 2019. p. 1891–1895. Disponível em: <http://dx.doi.org/10.21437/Interspeech.2019-3079>. Citado na página 35.

HINKLE, D. E.; WIERSMA, W.; JURS, S. G. **Applied statistics for the behavioral sciences**. [S.l.]: Houghton Mifflin College Division, 2003. v. 663. Citado na página 29.

HUTTO, C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [S.l.: s.n.], 2014. v. 8, n. 1. Citado na página 28.

IN, G. **How to Study and Identify the Minerals Samples?** 2014. Disponível em: <https://www.geologyin.com/2014/09/how-to-study-and-identify-minerals.html>. Acesso em: 28 de outubro de 2021. Citado na página 23.

JONES, T. O.; SASSER, W. E. et al. Why satisfied customers defect. **Harvard business review**, SUBSCRIBER SERVICE, PO BOX 52623, BOULDER, CO 80322-2623, v. 73, n. 6, p. 88, 1995. Citado na página 20.

KHATRI, C. et al. Alexa prize—state of the art in conversational ai. **AI Magazine**, v. 39, n. 3, p. 40–55, 2018. Citado 4 vezes nas páginas 14, 35, 37 e 43.

KIM, Y.; LEVY, J.; LIU, Y. Speech Sentiment and Customer Satisfaction Estimation in Socialbot Conversations. In: **Proc. Interspeech 2020**. [S.l.: s.n.], 2020. p. 1833–1837. Citado 16 vezes nas páginas 14, 15, 16, 20, 28, 29, 36, 37, 41, 42, 43, 44, 55, 56, 58 e 59.

Level of measurementlevel of measurement. In: KIRCH, W. (Ed.). **Encyclopedia of Public Health**. Dordrecht: Springer Netherlands, 2008. p. 851–852. ISBN 978-1-4020-5614-7. Disponível em: <https://doi.org/10.1007/978-1-4020-5614-7_1971>. Citado na página 21.

KLEIN, J.; MOON, Y.; PICARD, R. W. This computer responds to user frustration: Theory, design, and results. **Interacting with computers**, Oxford University Press Oxford, UK, v. 14, n. 2, p. 119–140, 2002. Citado na página 41.

KOKOSKA, S.; ZWILLINGER, D. **CRC standard probability and statistics tables and formulae**. [S.l.]: Crc Press, 2000. Citado na página 31.

KULIKOV, I. et al. Importance of a search strategy in neural dialogue modelling. **arXiv preprint arXiv:1811.00907**, Nov, 2018. Citado na página 27.

KÜSTER, D.; KAPPAS, A. Measuring emotions online: Expression and physiology. In: **Cyberemotions**. [S.l.]: Springer, 2017. p. 71–93. Citado 5 vezes nas páginas 14, 27, 37, 40 e 57.

LEBOW, J. Consumer satisfaction with mental health treatment. **Psychological bulletin**, American Psychological Association, v. 91, n. 2, p. 244, 1982. Citado na página 33.

LEE, K. Y.; YANG, S.-B. The role of online product reviews on information adoption of new product development professionals. **Internet Research**, Emerald Group Publishing Limited, 2015. Citado na página 25.

LIANG, W.; ZOU, J.; YU, Z. Beyond user self-reported likert scale ratings: A comparison model for automatic dialog evaluation. arxiv preprint arxiv: 200510716. 2020. Citado 9 vezes nas páginas 14, 27, 32, 34, 36, 37, 41, 58 e 59.

LINDSTEN, F. et al. Probabilistic modeling–linear regression & gaussian processes. **Uppsala: Uppsala University**, v. 7, 2017. Citado na página 30.

LOGACHEVA, V. et al. Convai dataset of topic-oriented human-to-chatbot dialogues. In: **The NIPS'17 Competition: Building Intelligent Systems**. [S.l.]: Springer, 2018. p. 47–57. Citado 16 vezes nas páginas 14, 15, 16, 20, 35, 37, 39, 44, 45, 46, 47, 48, 54, 55, 56 e 59.

LOWE, R. et al. Towards an automatic turing test: Learning to evaluate dialogue responses. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.: s.n.], 2017. p. 1116–1126. Citado na página 34.

MANGIAFICO, S. S. Summary and analysis of extension program evaluation in r. **Rutgers Cooperative Extension: New Brunswick, NJ, USA**, v. 125, p. 16–22, 2016. Citado 3 vezes nas páginas 21, 25 e 26.

MATOS, M. **Inscrições abertas para seleção de projetos voltados à implementação das Oficinas 4.0 — Instituto Federal do Tocantins**. 2021. Disponível em: <http://www.ifto.edu.br/noticias/inscricoes-abertas-para-selecao-de-projetos-voltados-a-implementacao-das-oficinas-4.0>. Acesso em: 19 de fevereiro de 2022. Citado na página 47.

MCTEAR, M.; CALLEJAS, Z.; GRIOL, D. The conversational interface: Talking to smart devices: Springer international publishing. **Doi: https://doi. org/10.1007/978-3-319-32967-3**, 2016. Citado na página 17.

MELO, L. A. M. P.; STEINKE, E. T. Um ensaio argumentativo a favor do uso de quantificação em geografia. **Caderno Prudentino de Geografia**, v. 3, n. 36, p. 161–181, 2014. Citado na página 21.

MILONE, G. **Estatística: geral e aplicada**. [S.l.]: Pioneira Thomson Learning, 2009. Citado na página 30.

MORRISSEY, K.; KIRAKOWSKI, J. 'realness' in chatbots: establishing quantifiable criteria. In: SPRINGER. **International conference on human-computer interaction**. [S.l.], 2013. p. 87–96. Citado na página 20.

MUKAKA, M. M. A guide to appropriate use of correlation coefficient in medical research. **Malawi medical journal**, v. 24, n. 3, p. 69–71, 2012. Citado 2 vezes nas páginas 29 e 30.

NEFF, G.; NAGY, P. Automation, algorithms, and politics| talking to bots: Symbiotic agency and the case of tay. **International Journal of Communication**, v. 10, p. 17, 2016. Citado na página 20.

NERBONNE, J. The secret life of pronouns. what our words say about us. **Literary and Linguistic Computing**, Oxford University Press, v. 29, n. 1, p. 139–142, 2014. Citado 4 vezes nas páginas 14, 27, 37 e 57.

NIELSEN, F. Å. A new anew: Evaluation of a word list for sentiment analysis in microblogs. **arXiv preprint arXiv:1103.2903**, 2011. Citado na página 28.

OLIVER, R. L. **Satisfaction: A behavioral perspective on the consumer: A behavioral perspective on the consumer**. [S.l.]: Routledge, 2014. Citado na página 20.

OPINIONI, S. S. **Ricerea SuUa Qualita' Percepita Dogli Utenti Del Servizio Autotelefono**. 1989. Citado na página 33.

O´DEA, S. **Mobile operating systems' market share worldwide from January 2012 to June 2021**. 2021. Disponível em: <https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/>. Acesso em: 22 de janeiro de 2022. Citado na página 34.

PALATINI, K.; ZISCHNER, K. The art of interpretation-chances and risks on interpretation in the field of mobile testing. **Vedecké Práce Materiálovotechnologickej Fakulty Slovenskej Technickej Univerzity v Bratislave so Sídlom v Trnave**, De Gruyter Poland, v. 22, p. 17, 2014. Citado na página 19.

PÉREZ-ROSAS, V. et al. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2019. p. 926–935. Citado na página 34.

PETERSON, R. A.; WILSON, W. R. Measuring customer satisfaction: fact and artifact. **Journal of the academy of marketing science**, Springer, v. 20, n. 1, p. 61–71, 1992. Citado 2 vezes nas páginas 32 e 33.

PODSAKOFF, P. M. et al. Common method biases in behavioral research: a critical review of the literature and recommended remedies. **Journal of applied psychology**, American Psychological Association, v. 88, n. 5, p. 879, 2003. Citado na página 41.

PONTES, A. C. F. Ensino da correlação de postos no ensino médio. **Simpósio Nacional de Probabilidade e Estatıstica (SINAPE)**, v. 19, p. 26–30, 2010. Citado na página 30.

POSCHENRIEDER, M. **77% will not download a Retail app rated lower than 3 stars**. 2015. Disponível em: <https://web.archive.org/web/20151024035208/http://blog.testmunk.com/77-will-not-download-a-retail-app-rated-lower-than-3-stars/>. Acesso em: 23 de janeiro de 2022. Citado na página 34.

PYTHON. In: PYTHON English Definition and Meaning | Lexico.com. Oxford University, 2021. Disponível em: <https://www.lexico.com/en/definition/python>. Acesso em: 28 de outubro de 2021. Citado na página 53.

RADZIWILL, N.; BENTON, M. Evaluating quality of chatbots and intelligent conversational agents. **Software Quality Professional**, v. 19, n. 3, 2017. Citado na página 20.

RAJPURKAR, P. et al. Squad: 100,000+ questions for machine comprehension of text. In: **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2016. p. 2383–2392. Citado na página 37.

RAM, A. et al. Conversational ai: The science behind the alexa prize. **arXiv preprint arXiv:1801.03604**, 2018. Citado na página 27.

REDDY, V. V. The relation between the work motivation and job satisfaction of secondary school teachers. **Age**, v. 30, p. 30yrs, 2019. Citado na página 29.

RIBEIRO, F. N. et al. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. **EPJ Data Science**, Springer, v. 5, n. 1, p. 1–29, 2016. Citado na página 38.

SEMIOTICS. In: SEMIOTICS | Meaning & Definition for UK English | Lexico.com. Oxford University, 2021. Disponível em: <https://www.lexico.com/definition/semiotics>. Acesso em: 28 de outubro de 2021. Citado na página 19.

SERRA, M. **Calendário Gregoriano**. Panmythica, 2018. Disponível em: <https://www.panmythica.com/2018/05/calendario-gregoriano.html>. Acesso em: 28 de outubro de 2021. Citado na página 24.

SHAWAR, B. A.; ATWELL, E. Different measurement metrics to evaluate a chatbot system. In: **Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies**. [S.l.: s.n.], 2007. p. 89–96. Citado 3 vezes nas páginas 14, 17 e 18.

SOARES, J. F.; SIQUEIRA, A. L. Introdução à estatística médica. In: **Introdução à estatística médica**. [S.l.: s.n.], 2002. p. 300–300. Citado na página 21.

SOUSA, I. L. **PLANO DE MARKETING DIGITAL PARA UMA STARTUP COM MODELO DE NEGÓCIO EM SAAS.** 68 f. Monografia (Bacharelado em Administração) — Universidade Federal do Maranhão, São Luís, 2018. Citado na página 37.

SOUZA, K. F. de; PEREIRA, M. H. R.; DALIP, D. H. Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro. **Abakós**, v. 5, n. 2, p. 79–96, 2017. Citado na página 28.

STEVENS, S. S. et al. On the theory of scales of measurement. Bobbs-Merrill, College Division, 1946. Citado 5 vezes nas páginas 21, 22, 23, 24 e 25.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. **Journal of language and social psychology**, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 24–54, 2010. Citado 4 vezes nas páginas 14, 27, 37 e 57.

TEJAY, G.; DHILLON, G.; CHIN, A. G. Data quality dimensions for information systems security: A theoretical exposition. In: SPRINGER. **Working Conference on Integrity and Internal Control in Information Systems**. [S.l.], 2004. p. 21–39. Citado na página 19.

THIELTGES, A.; SCHMIDT, F.; HEGELICH, S. The devil's triangle: Ethical considerations on developing bot detection methods. In: **2016 AAAI Spring Symposium Series**. [S.l.: s.n.], 2016. Citado na página 20.

TREIBLMAIER, H.; FILZMOSER, P. Benefits from using continuous rating scales in online survey research. 2011. Citado na página 21.

TURING, A. M. I.—computing machinery and intelligence. **Mind**, LIX, n. 236, p. 433–460, 10 1950. ISSN 0026-4423. Disponível em: <https://doi.org/10.1093/mind/LIX.236.433>. Citado na página 17.

TZU, S. **A arte da guerra**. [S.l.]: Editora Schwarcz-Companhia das Letras, 2019. Citado na página 32.

VENKATESH, A. et al. On evaluating and comparing conversational agents. **arXiv preprint arXiv:1801.03625**, v. 4, p. 60–68, 2018. Citado 2 vezes nas páginas 27 e 34.

VERHAGEN, T. et al. Virtual Customer Service Agents: Using Social Presence and Personalization to Shape Online Service Encounters*. **Journal of Computer-Mediated Communication**, v. 19, n. 3, p. 529–545, 04 2014. ISSN 1083-6101. Disponível em: <https://doi.org/10.1111/jcc4.12066>. Citado na página 38.

VOGT, W. P. Dictionary for statistics & methodology: A ontechnical guide for the social sciences sage publications. **Inc, Thousand Oaks, CA**, 1999. Citado na página 25.

W3SCHOOLS. **How To Create a Simple Star Rating with CSS**. 2022. Disponível em: <https://www.w3schools.com/howto/howto_css_star_rating.asp>. Acesso em: 22 de janeiro de 2022. Citado na página 26.

WALLACE, R. The elements of aiml style. **Alice AI Foundation**, v. 139, 2003. Citado na página 20.

WEINSTEIN, M. Consumers still like service, but their enthusiasm erodes. **American Banker Consumer Survey**, v. 6, p. 18–20, 1989. Citado na página 33.

WEIZENBAUM, J. Eliza—a computer program for the study of natural language communication between man and machine. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 9, n. 1, p. 36–45, jan. 1966. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/365153.365168>. Citado 2 vezes nas páginas 18 e 20.

WU, D.; PARSONS, T. D.; NARAYANAN, S. S. Acoustic feature analysis in speech emotion primitives estimation. In: **Eleventh Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2010. Citado na página 43.

YAAKUB, M. R.; LATIFFI, M. I. A.; ZAABAR, L. S. A review on sentiment analysis techniques and applications. In: IOP PUBLISHING. **IOP Conference Series: Materials Science and Engineering**. [S.l.], 2019. v. 551, n. 1, p. 012070. Citado na página 28.

YANKOV, M. **A Game for Learning how Epidemics Spread**. 35 f. Monografia (Bacharelado em Ciência da Computação) — Instituto de Informática. Universidade Luís Maximiliano de Munique, Munique, 2021. Citado na página 34.

ZIKMUND, G. **Business Research Methods, 3rd**. [S.l.]: Fort Worth, TX: Harcourt Brace Jovanovich College Press, 1991. Citado na página 22.

ZIKMUND, W. G.; CARR, J. C.; GRIFFIN, M. **Business Research Methods (Book Only)**. [S.l.]: Cengage Learning, 2013. Citado na página 22.