Análise de estresse e resiliência durante a pandemia da COVID-19 a partir de dados de redes sociais

Diansley Raphael dos Santos Peres

Universidade Federal de Uberlândia
Faculdade de Computação
Programa de Pós-Graduação em Ciência da Computação

Uberlândia
2023

Diansley Raphael dos Santos Peres

Análise de estresse e resiliência durante a pandemia da COVID-19 a partir de dados de redes sociais

Dissertação de mestrado apresentada ao Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação

Orientador: Profª Drª Maria Camila Nardini Barioni
Coorientador: Profª Drª Elaine Ribeiro de Faria Paiva

Uberlândia
2023

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

P437 Peres, Diansley Raphael dos Santos, 1987-
2023 Análise de estresse e resiliência durante a pandemia da COVID-19 a partir de dados de redes sociais [recurso eletrônico] / Diansley Raphael dos Santos Peres. - 2023.

Orientadora: Maria Camila Nardini Barioni.
Coorientadora: Elaine Ribeiro de Faria Paiva.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Ciência da Computação.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2023.658
Inclui bibliografia.

1. Computação. I. Barioni, Maria Camila Nardini,1978-, (Orient.). II. Paiva, Elaine Ribeiro de Faria,1980-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Ciência da Computação. IV. Título.

CDU: 681.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:

Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Ciência da Computação
Av. João Naves de Ávila, 2121, Bloco 1A, Sala 243 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4470 - www.ppgco.facom.ufu.br - cpgfacom@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Ciência da Computação				
Defesa de:	Dissertação de Mestrado 26/2023, PPGCO				
Data:	24 de novembro de 2023	Hora de início:	08:00	Hora de encerramento:	09:49
Matrícula do Discente:	12112CCP008				
Nome do Discente:	Diansley Raphael dos Santos Peres				
Título do Trabalho:	Análise de estresse e resiliência durante a pandemia da COVID-19 a partir de dados de redes sociais				
Área de concentração:	Ciência da Computação				
Linha de pesquisa:	Ciência de Dados				
Projeto de Pesquisa de vinculação:	-				

Reuniu-se, por videoconferência, a Banca Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Computação, assim composta: Professores Doutores: Elaine Ribeiro de Faria Paiva - FACOM/UFU (Coorientadora), Marcelo Zanchetta do Nascimento - FACOM/UFU, Nádia Félix Felipe da Silva INF/UFG, Maria Camila Nardini Barioni - FACOM/UFU, orientadora do candidato.

Os examinadores participaram desde as seguintes localidades: Nádia Félix Felipe da Silva - Goiânia/GO, os demais membros e o discente participaram da cidade de Uberlândia/MG.

Iniciando os trabalhos a presidente da mesa, Prof.ª Dr.ª Maria Camila Nardini Barioni, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir a senhora presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos,

conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Marcelo Zanchetta do Nascimento**, **Professor(a) do Magistério Superior**, em 24/11/2023, às 10:50, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Maria Camila Nardini Barioni**, **Professor(a) do Magistério Superior**, em 24/11/2023, às 10:51, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **NÁDIA FÉLIX FELIPE DA SILVA**, **Usuário Externo**, em 24/11/2023, às 14:55, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Elaine Ribeiro de Faria Paiva**, **Professor(a) do Magistério Superior**, em 24/11/2023, às 15:52, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4906763** e o código CRC **288C2489**.

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

Agradeço à minha mãe, Nilda Maria Santos, que possibilitou e incentivou os meus estudos. Agradeço ao meu companheiro de vida, Denis dos Reis Oliveira, por todo o incentivo, ajuda e paciência durante a minha formação como mestre. Agradeço a todos os alunos da Faculdade de Computação da Universidade Federal de Uberlândia que me auxiliaram em atividades e possibilitaram os resultados deste trabalho, em especial aos colegas de pesquisa Cássio de Alcântara, Lara Mondini Martins e Gean Fernandes Silva. Agradeço às minhas orientadoras Prof.ª Dr.ª Maria Camila Nardini Barioni e Prof.ª Dr.ª Elaine Ribeiro de Faria Paiva pelos direcionamentos e contribuições valiosas para a minha formação.

"Toda a negatividade é causada pelo acúmulo de tempo psicológico e pela negação do presente. O desconforto, a ansiedade, a tensão, o estresse, a preocupação e todas essas formas de medo são causadas por excesso de futuro e pouca presença. A culpa, o arrependimento, o ressentimento, a injustiça, a tristeza, a amargura, todas as formas de incapacidade de perdão são causadas por excesso do passado e pouca presença."

(Eckhart Tolle em O Poder do Agora)

Resumo

A pandemia do coronavírus (SARS-CoV-2) assolou países de todo o mundo desde 2019, quando surgiu na China. Somente no Brasil, até outubro de 2023, mais de 37 milhões de casos da doença foram confirmados com aproximadamente 706 mil óbitos. Além dos diversos efeitos físicos da doença, efeitos psicológicos puderam ser verificados através do desenvolvimento de doenças como alcoolismo, depressão e desenvolvimento de estresse pós-traumático. Percebe-se que, durante a pandemia, as pessoas utilizaram as redes sociais com diferentes finalidades, de forma que identificar o conteúdo publicado pode contribuir para uma resposta adequada por parte das autoridades em contextos de emergência. Esse trabalho buscou, portanto, investigar o impacto da COVID-19 sobre o estresse percebido, resiliência e o Transtorno de Estresse Pós-Traumático (TEPT) em voluntários da Universidade Federal de Uberlândia usando dados de redes sociais *online*. Foram aplicados ainda questionários auto avaliativos para mensurar os aspectos psicológicos investigados na mineração. Como um dos resultados da pesquisa foi possível validar o algoritmo para detecção de estresse e relaxamento em textos, *TensiStrength*, adaptado para o português. As bases de dados extraídas com características dos voluntários possibilitaram correlacionar dados de formulários com o estresse e relaxamento presentes nos textos e estatísticas de publicações dos usuários nas redes sociais. Apesar do tamanho amostral insuficiente das bases de usuários, foi possível encontrar evidências de que o *X* é uma rede com maior presença de estresse em comparação com o *Instagram*. Em geral, o estresse presente no texto não foi diretamente proporcional ao estresse percebido no usuário. Foram verificadas correlações entre os fatores de resiliência do usuário e estatísticas de publicação extraídas das redes sociais.

Palavras-chave: Estresse, Resiliência, TEPT, TensiStrength, Redes Sociais.

Abstract

The coronavirus pandemic (SARS-CoV-2) has ravaged countries worldwide since 2019, when it emerged in China. In Brazil alone, by October 2023, more than 37 million disease cases had been confirmed with approximately 706 thousand deaths. In addition to the various physical effects of the COVID-19, psychological effects could be seen through the development of diseases such as alcoholism, depression and post-traumatic stress. It is clear that, during the pandemic, people used social networks for different purposes, so identifying published content can contribute to an adequate response by authorities in emergencies. This work therefore sought to investigate the impact of COVID-19 on perceived stress, resilience and Post-Traumatic Stress Disorder (PTSD) in volunteers from the Federal University of Uberlândia using data from social networks. Self-evaluative questionnaires were also applied to measure the psychological aspects investigated in mining. As one of the research results, it was possible to validate the algorithm for detecting stress and relaxation in texts, *TensiStrength*, adapted to portuguese. The databases extracted with the volunteers' characteristics made it possible to correlate form data with the stress and relaxation present in the texts and statistics of users' publications on online social networks. Despite the insufficient sample size of the user bases, it was possible to find evidence that X is a network with a greater presence of stress compared to Instagram. In general, the stress present in the text was not directly proportional to the stress perceived in the user. Correlations were verified between users' resilience factors and publication statistics extracted from social networks.

Keywords: Stress, Resilience, PTSD, TensiStrength, Social Media.

Lista de ilustrações

Lista de tabelas

Lista de siglas

Sumário

Capítulo 1

Introdução

A Organização Pan-Americana da Saúde, OPAS (2023), relembra que no dia 31 de dezembro de 2019, a Organização Mundial da Saúde (OMS) recebeu um alerta a respeito de múltiplos casos de pneumonia na cidade de Wuhan, na província de Hubei, na República Popular da China. Esses casos estavam associados a uma nova variante de coronavírus que até então não tinha sido detectada em seres humanos. Uma semana depois as autoridades chinesas confirmaram a descoberta de uma nova cepa de coronavírus. Os coronavírus estão amplamente distribuídos e ocupam o segundo lugar como causa mais frequente de resfriados, sendo superados apenas pelos rinovírus.

Até as últimas décadas, raramente ocasionavam enfermidades mais sérias em seres humanos, geralmente se limitando a causar sintomas semelhantes aos do resfriado comum. Até o presente momento, foram identificados sete tipos de coronavírus que afetam os seres humanos: HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1, SARS-COV (responsável pela síndrome respiratória aguda grave), MERS-COV (causador da síndrome respiratória do Oriente Médio) e o coronavírus mais recente, inicialmente denominado temporariamente como 2019-nCoV e, em 11 de fevereiro de 2020, oficialmente designado como SARS-CoV-2. Este novo coronavírus é o agente responsável pela doença conhecida como COVID-19 (OPAS, 2023).

É comum durante o enfrentamento de pandemias, que a população demonstre repostas associadas ao desenvolvimento de estresse e ansiedade, tais como comportamentos compulsivos, respostas defensivas, angústia emocional, aversão, preocupações excessivas, pesadelos e pensamentos intrusivos (CULLEN; GULATI; KELLY, 2020; TAYLOR et al., 2020; BROWN et al., 2020).

Pesquisas realizadas em diferentes países forneceram evidências de que houve um aumento da ocorrência de transtornos mentais sobre a população em decorrência da COVID-19. Entre os fatores associados aos transtornos verificam-se dificuldades financeiras, o isolamento social como medida de prevenção, a incidência da doença e óbito de pessoas próximas, a pré-disposição a ocorrência de doenças mentais, exposição excessiva à notícias sobre a COVID-19 e níveis inferiores de escolaridade (BROWN et al., 2020; TAYLOR et

al., 2020; PEDROZO-PUPO; PEDROZO-CORTÉS; CAMPO-ARIAS, 2020; REHMAN et al., 2021; ZHU et al., 2021; CASTELLI et al., 2020; CAMPOS et al., 2020).

Apesar de efeitos adversos, em grande parte negativos, também se verifica durante o enfrentamento de crises o fortalecimento da capacidade dos indivíduos de se transformarem positivamente e superarem adversidades através da resiliência. Essa característica pode ser aprendida, adquirida ou cultivada em comunidade e se constitui um capital social por conferir ao indivíduo um senso de propósito e adaptação significativa com a sobrevivência de uma crise (PECONGA et al., 2020).

As redes sociais atualmente são estudadas por profissionais de diferentes áreas por possibilitarem o entendimento de muitos fenômenos, através da troca intensiva de informação entre as pessoas, característica desse meio (SOUZA; QUANDT, 2008). Além disso, se constituem um ambiente propício ao estudo de diversos temas da computação relacionados à organização e tratamento de dados não estruturados e aplicação de técnicas de mineração de dados (BENEVENUTO; ALMEIDA; SILVA, 2011).

Percebe-se que, durante a pandemia, as pessoas utilizaram as redes sociais com diferentes finalidades, de forma que identificar o conteúdo publicado pode contribuir para uma resposta adequada por parte das autoridades em contextos de emergência. Analisar o conteúdo postado pelos usuários pode ajudar a identificar demandas relacionadas a medidas de segurança, pedidos de ajuda, combate de rumores e *fake news* e identificação de usuários com diversos fins (LI et al., 2020). Dessa forma, as redes sociais podem atuar como importante meio de gerenciamento de crises, tomadas de ação e meio propagador da consciência situacional de uma crise por parte da população (FREITAS; BORGES; CARVALHO, 2020).

1.1 Motivação

Os trabalhos relacionados à pandemia da COVID-19 estão, em sua grande parte, relacionados à aplicação de mineração de textos para: i) a caracterização da percepção da população sobre a pandemia; ii) o entendimento do uso das redes sociais no gerenciamento de crises e iii) a detecção de sintomas de transtornos psicológicos, como ansiedade e depressão. Não foram encontrados trabalhos que tenham utilizado dados de redes sociais para investigar evidências da ocorrência de estresse percebido, transtorno de estresse pós-traumático (TEPT) e níveis de resiliência durante a pandemia da COVID-19 em associação com o resultado de testes psicológicos de usuários brasileiros.

Foi verificada na literatura a ocorrência de estresse percebido na população de diversos países desde o início da pandemia da COVID-19 (BROWN et al., 2020; TAYLOR et al., 2020; PEDROZO-PUPO; PEDROZO-CORTÉS; CAMPO-ARIAS, 2020; REHMAN et al., 2021; ZHU et al., 2021; CASTELLI et al., 2020; CAMPOS et al., 2020). A literatura também fornece evidências de que a população vivenciou a resiliência como efeito adverso

positivo das dificuldades decorrentes da pandemia (PECONGA et al., 2020; FERREIRA; BUTTELL; CANNON, 2020; KILLGORE et al., 2020).

Diante das evidências da ocorrência de estresse na população durante a pandemia, das consequências desse transtorno já amplamente discutidas na comunidade científica e do desenvolvimento da resiliência como suporte para o enfrentamento das dificuldades, justifica-se a investigação da ocorrência desses fenômenos na população por meio de técnicas de mineração dos textos publicados nas redes sociais.

Esse trabalho buscou, portanto, investigar o impacto da COVID-19 sobre o estresse percebido, resiliência e o Transtorno de Estresse Pós-Traumático (TEPT) nos participantes usando dados de redes sociais. Foi considerada ainda a avaliação de questionários respondidos pelos indivíduos, com a devida ciência e aprovação do projeto de pesquisa pelo comitê de ética da instituição, como referência para os aspectos psicológicos investigados na mineração.

1.2 Objetivos da Pesquisa

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é propor e desenvolver um método computacional, que permita descrever o estresse percebido, a resiliência e o TEPT frente à pandemia da COVID-19 usando dados linguísticos e comportamentais extraídos das redes sociais.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- ❏ Criação de uma base de dados anonimizada a partir da extração de características dos dados brutos de redes sociais e de questionários de estresse e resiliência;
- ❏ Avaliar o algoritmo de classificação de estresse e relaxamento, TensiStrength, como ferramenta para a classificação do nível de estresse de um texto usando bases de dados em português;
- ❏ Criação de um ou mais modelos que consigam descrever o estresse percebido dos usuários por meio de suas postagens em redes sociais;
- ❏ Criação de um ou mais modelos que consigam descrever a resiliência dos usuários por meio de suas postagens em redes sociais;
- ❏ Criação de um ou mais modelos que consigam descrever o TEPT dos usuários por meio de suas postagens em redes sociais;

- ❏ Correlacionar os modelos de estresse percebido e resiliência com o resultado de questionários auto avaliativos.

1.3 Hipótese

Este trabalho possui a seguinte hipótese:

- ❏ H1: O estresse percebido, a resiliência e o TEPT podem ser descritos a partir da análise de dados de redes sociais.

1.4 Contribuições

As contribuições do presente trabalho são:

- ❏ Possibilitar o entendimento do impacto da pandemia da COVID-19 sobre a comunidade acadêmica da Universidade Federal de Uberlândia em termos de estresse, TEPT e resiliência;
- ❏ Compreender a relação entre traços psicológicos, o estresse e relaxamento presentes nos textos publicados pelos indivíduos;
- ❏ Contribuir para o fomento de futuras pesquisas que visem associar traços de sintomas psicológicos investigados e técnicas computacionais para o processamento de linguagem natural.

1.5 Organização da Dissertação

Os demais capítulos desta produção acadêmica estão assim organizados:

- ❏ o Capítulo 2 trata de conceitos básicos relacionados à mineração de textos, redes sociais, coleta de dados de redes sociais, pré-processamento, representação de dados, extração de tópicos, algoritmos de classificação, aplicação do TS e formulários para a avaliação de escalas e testes psicológicos;
- ❏ o Capítulo 3 traz alguns trabalhos relacionados à pesquisa na literatura;
- ❏ o Capítulo 4 mostra uma visão geral da proposta, os detalhes da metodologia e os resultados verificados na validação do TensiStrength adaptado para a avaliação do estresse e relaxamento em textos em português (TSpt) e caraterização das discussões ocorridas durante a pandemia da COVID-19;

- o Capítulo 5 trata do processo de caracterização dos voluntários da UFU participantes da pesquisa com base em formulários para a avaliação de estresse percebido, resiliência e TEPT e estatísticas de publicação no *Instagram* e *X*;
- o Capítulo 6 traz as principais conclusões obtidas, sugestões de trabalhos futuros e contribuições em produção bibliográfica.

Capítulo 2

Fundamentação Teórica

Nas seções que se seguem serão discutidos conceitos e técnicas relacionados à mineração de textos; importância e coleta de dados de redes sociais; etapas de pré-processamento verificas na literatura; representação de documentos; extração automática de tópicos; algoritmos de classificação; aplicação do TS para a mensuração do estresse e relaxamento no texto e aplicação de formulários para avaliação de escalas e testes.

2.1 Mineração de Textos

A mineração de textos engloba uma gama de diferentes abordagens e métodos que possuem em comum o texto como entrada de dados. Trata-se de um campo multidisciplinar com conceitos e técnicas oriundas da Mineração de Dados, Linguística, Estatística e Ciência da Computação. Entre as diversas técnicas existentes estão inclusas a classificação e agrupamento de textos, sumarização de documentos, análise latente de corpus e técnicas de recuperação da informação (FEINERER, 2008).

Na mineração de textos, os padrões são extraídos a partir de textos escritos em linguagem natural (HEARST, 2003). Ao invés de dados armazenados de forma estruturada em tabelas, as bases de dados são não estruturadas, amorfas e difícil de serem manipuladas por um algoritmo. Ao contrário de outras técnicas de mineração, que visam encontrar padrões desconhecidos, na mineração de textos a informação a ser extraída está claramente situada no texto e não oculta de fato. Normalmente são analisados textos informativos ou opinativos com o objetivo de realizar a extração automática da informação sem a necessidade de um intermediário humano (WITTEN, 2004).

Este campo de estudo tem ganhado atenção nos últimos anos devido ao volume de dados em texto disponível nas redes sociais, aplicações web e outras aplicações centradas na informação, segundo Aggarwal e Zhai (2012). Os autores explicam que dados não estruturados são um tipo muito comum devido a possibilidade de serem gerados a partir de qualquer cenário de aplicação. O resultado dessa variedade de fontes é a necessidade de algoritmos e métodos capazes de processar essa diversidade e volume de textos dispo-

níveis. A classificação de textos tem sido amplamente estudada e possui aplicações em diversos domínios, tais como filtragem e organização de textos, recuperação e organização de documentos, mineração de opiniões, classificação de e-mails e filtragem de *spams*. Diversas técnicas podem ser aplicadas, entre as mais comuns existem as árvores de decisão, classificadores baseados em Máquina de Vetor de Suporte ou *Support Vector Machine* (SVM), redes neurais e classificadores Bayesianos.

Nas redes sociais, as fronteiras entre os usuários são apenas virtuais, de modo que as pessoas podem expressar suas opiniões e interagir entre si através de postagens, comentários, mensagens e reações. As redes sociais possibilitam ao usuário compartilhar seus pensamentos, sentimentos e opiniões com outras pessoas de forma fácil e instantânea (ÖZTÜRK; AYVAZ, 2018).

Além disso, as mídias sociais são úteis para diferentes sujeitos, tais como cidadãos, mídia e serviços de emergência, possibilitando a troca de informações úteis no enfrentamento de crises. Não é uma tarefa trivial o processamento de dados das redes sociais, devido à grande complexidade dos dados, característica do *Big Data*. Além do grande volume de dados compartilhados, o processamento da informação em diferentes formatos é um desafio, demandando o desenvolvimento de diferentes técnicas de mineração de texto e métodos de classificação e de aprendizado de máquina (REUTER; STIEGLITZ; IMRAN, 2020).

A forma como as pessoas utilizam as redes para expressar sua opinião sobre os eventos têm levado ao surgimento de técnicas para explorar os sentimentos expressos nas publicações de forma otimizada. Os métodos visam inferir o sentimento presente em um texto e demandam o uso do Processamento de Linguagem Natural (PLN) com diversas aplicações. A presença de múltiplas linguagens, escrita informal, erros de digitação, gramaticais e lógicos nos textos dificultam a análise do sentimento presente nas postagens (CHAKRABORTY; BHATTACHARYYA; BAG, 2020).

A análise de sentimento, ou mineração de opinião, é o estudo computacional com técnicas que visam investigar a opinião das pessoas, avaliações, atitudes e sentimentos sobre organizações, indivíduos, eventos, tópicos e seus atributos. Trata-se de uma tarefa desafiadora e útil, a exemplo de empresas que buscam identificar a opinião de clientes sobre seus produtos e serviços. Enquanto um ser humano poderia ter dificuldade em analisar de forma consistente um grande volume de dados textuais devido a limitações físicas e mentais, a mineração automática de opinião e sistemas de sumarização de textos superam essas limitações através de uma análise objetiva do sentimento presente no texto (AGGARWAL; ZHAI, 2012).

As postagens extraídas de *microblogs* como o *X*, em geral, apresentam uma opinião pessoal do autor da postagem em relação ao assunto mencionado. A análise de sentimento possibilita a mineração da opinião e sentimentos presentes nesses textos, conforme descrito por Chong, Selvaretnam e Soon (2014).

Segundo Sahayak, Shete e Pathan (2015), as técnicas tradicionais de classificação de sentimentos se baseiam em trechos de textos maiores como *feedbacks* e *reviews*. Porém textos publicados em microblogs se diferenciam de fontes tradicionais em termos de propósito e casualidade. Um *framework* para a análise de sentimento em textos de microblogs foi proposto pelos autores com as seguintes etapas: i) extração de *tweets*; ii) pré-processamento de dados com filtragem e remoção de caracteres e termos seguida da tokenização; iii) construção de n-gramas; iv) treinamento e teste de algoritmos de classificação; v) escoragem e classificação dos sentimentos dos *tweets*.

2.2 Redes Sociais

As redes sociais, entendidas como estruturas dinâmicas e complexas formadas por pessoas com valores e /ou objetivos em comum e interligadas de forma horizontal e predominantemente descentralizadas, têm sido estudadas por profissionais de diversas áreas para explicar uma série de fenômenos caracterizados pela troca intensiva de informação entre as pessoas. Além disso, podem assumir diferentes formatos e níveis de formalidade e surgir em torno de objetivos diversos. Além disso, possuem um alto fluxo de comunicação e não exigem contratos formais reguladores do resultado da interação (SOUZA; QUANDT, 2008).

Confunde-se, em geral, mídias e redes sociais. Apesar de estarem no mesmo universo, as mídias sociais são entendidas como o meio pelo qual determinada rede social utiliza para se comunicar. São muitos os tipos de mídias existentes, com diferentes focos e público-alvo, tais como contatos profissionais, amizades, relacionamentos amorosos, pesquisas, dentre outros. As mídias sociais dispõem de ferramentas que facilitam a comunicação através do conteúdo gerado pelos seus indivíduos por meio de mensagens instantâneas e textos, compartilhamento de vídeos, áudios e imagens. Com o constante crescimento da internet e das mídias sociais, redes sociais bem definidas estão se formando com os mais diversificados perfis expondo opiniões e compartilhando momentos (CIRIBELI; PAIVA, 2011).

Segundo Benevenuto, Almeida e Silva (2011), as redes sociais online são um ambiente propício para o estudo de vários temas da computação, como sistemas distribuídos, padrões de tráfego na Internet, mineração de dados, sistemas multimídia e interação entre ser humano e computador. Além disso, por permitirem ao usuário criarem conteúdo, as redes sociais têm se tornado um tema chave em pesquisas relacionadas à organização e tratamento de grandes quantidades de dados, além de se constituírem um ambiente adequado para a extração de conhecimento e aplicação de técnicas de mineração de dados.

2.3 Coleta de Dados de Redes Sociais

Embora os dados das redes sociais sejam acessíveis através de APIs, devido ao valor comercial dos dados, a maior parte das fontes, tais como *Facebook* e *Google*, têm dificultado o acesso acadêmico aos dados brutos, segundo Batrinca e Treleaven (2015). Os autores explicam que uma minoria das redes sociais atuais oferece opções viáveis de acesso aos pesquisadores. Enquanto diversos serviços cobram uma taxa para o acesso, o *X* disponibiliza módulos de acesso gratuitos aos *tweets* públicos e dados históricos, possibilitando aos pesquisadores obterem *insights* a partir do seus vastos conjuntos de dados. Verifica-se nessa rede uma taxa de publicação superior a 500 milhões de *tweets* por dia.

No *Facebook (2021)*, a coleta ocorre através da *Graph API*, principal forma de inserir e retirar dados da plataforma. A API é baseada em HTTP, de forma que os aplicativos podem usá-la para consultar dados programaticamente, publicar novas histórias, gerenciar anúncios, carregar fotos e realizar uma ampla variedade de tarefas. A API do *Facebook* é baseada em nós, bordas e campos. Em geral, os nós são usados para a obtenção de dados sobre um objeto específico, as bordas são usadas para obter coleções de objetos sobre um objeto único e os campos para obter dados sobre um objeto único ou sobre cada objeto de uma coleção (Facebook for Developers, 2021b).

A API do *Instagram (2021)* é compatível com o mesmo protocolo de versões e ciclo de lançamentos da Graph API do *Facebook*. Novas versões são lançadas a cada 3 meses, aproximadamente. Cada versão fica disponível por cerca de 2 anos até se tornar obsoleta. O aplicativo pode obter códigos de autorização e permissões dos usuários. É possível trocar os códigos de autorização por tokens de acesso do usuário do *Instagram*, que precisam ser incluídos ao consultar a mídia ou o perfil de um usuário do aplicativo. A autorização do acesso aos dados é controlada pelos usuários do aplicativo e feita por meio das permissões de leitura do nó e das mídias do usuário. Os usuários devem conceder essas permissões via janela de autorização para que o aplicativo possa acessar os dados (Facebook for Developers, 2021a).

A API do *Twitter (2021)* oferece ferramentas para que os desenvolvedores possam analisar as conversas online que ocorrem na plataforma. Atualmente a API possui 3 módulos: *Standard*, *Premium* e *Enterprise*. A versão *Standard* é gratuita e, segundo a plataforma, a mais adequada para desenvolvedores iniciantes, possibilitando a testagem, integração e validação de conceitos. Essa versão possibilita a postagem de conteúdo e acesso aos dados publicados no site e no aplicativo móvel. Com a versão *Standard* é possível, entre outras funcionalidades, publicar, interagir e coletar *tweets* realizados num período de até 7 dias anteriores à data de coleta, gerenciar configurações e perfis de contas, criar e interagir com eventos da plataforma, realizar o *upload* de mídias, acessar tendências e obter informações sobre um local. Para um acesso mais escalonável as versões *Premium* e *Enterprise* são as mais indicadas (Twitter API, 2021).

Uma forma de coletar os dados da API do *X* via Python ocorre através da biblioteca

Tweepy (2021), um pacote *open source* composto por um conjunto de classes e métodos que representam os modelos da plataforma e *endpoints* da API. O uso da biblioteca evita a necessidade de lidar com detalhes de baixo nível que demandam tempo e podem resultar em erros, tais como requisições HTTP, serialização e codificação de dados, autenticação, paginação de resultados e limites de taxas (Miguel Garcia, 2021).

2.4 Pré-processamento

Segundo Kannan et al. (2014), o pré-processamento dos textos é uma parte essencial de qualquer sistema de Processamento de Linguagem Natural (PLN), uma vez que caracteres, palavras e sentenças mapeadas nesse estágio são as unidades fundamentais processadas nas etapas posteriores da mineração de textos. O pré-processamento é composto por um conjunto de técnicas necessárias, já que dados textuais podem conter caracteres em formatos especiais como números, datas e palavras comuns que não são úteis para a mineração (artigos, preposições e pronomes podem ser descartados nessa etapa).

O pré-processamento dos textos pode consumir até 80% de todo o esforço envolvido na mineração de textos (MOHBEY; TIWARI, 2011). Os autores argumentam ainda que as etapas necessárias nessa fase dependem do objetivo a ser alcançado. Em geral, busca-se uma homogeneização do texto através da substituição de caracteres especiais e partes estruturais que precisam ser manipuladas separadamente. O pré-processamento pode envolver um certo nível de análise de linguagem natural. A análise morfológica pode ser utilizada para generalizar os dados ao substituir, por exemplo, certas palavras por partes do discurso.

Nessa etapa ocorre o processo de limpeza do texto com a remoção de qualquer parte desnecessária para o processo de mineração, tais como caracteres especiais, links, tabelas, figuras e fórmulas. O propósito da limpeza do texto consiste em simplificar os dados textuais e eliminar ao máximo os fatores dependentes da linguagem. Artigos, por exemplo, são usados na linguagem natural para facilitar o entendimento humano, mas, na mineração de textos, esse tipo de dado não é processado de maneira trivial pelos algoritmos (KUMAR; BHATIA, 2013; TONG; ZHANG, 2016).

Entre as etapas típicas do pré-processamento dos dados para a mineração de textos faz-se necessário, em geral, a limpeza do texto, remoção de *stopwords*, *stemming*, conversão de palavras com letras maiúsculas, identificação de sinônimos, identificação de partes do discurso e *tokenização* (MOHBEY; TIWARI, 2011; SUMATHY; CHIDAMBARAM, 2013; FEINERER, 2008).

2.4.1 Tokenização

Nos estudos envolvendo PLN faz-se necessário a separação do texto em unidades básicas de análise ou *tokens*, processo definido como *tokenização*. A partir dessa etapa os

tokens do texto são identificados e processados nas etapas subsequentes. Com a tokenização, o texto é divido em palavras, termos, símbolos ou algum elemento importante para o processo de mineração. O objetivo da tokenização é a exploração das palavras em um documento. Dados textuais podem ser entendidos, a princípio, como bloco de caracteres, de forma que, para que ocorra a extração da informação do texto, as palavras precisam ser identificadas (WEBSTER; KIT, 1992; VIJAYARANI; JANANI et al., 2016; VERMA; RENU; GAUR, 2014). Um exemplo desse processo pode ser visualizado na Figura 1.

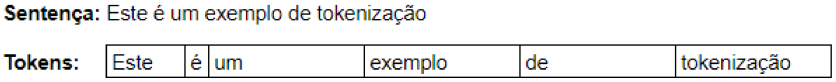

Figura 1 – Exemplo de tokenização - autoria própria.

2.4.2 Remoção de *Stopwords*

Stopwords são palavras que ocorrem com frequência em linguagem natural e não têm uma importância significativa nos processos de PLN como Agrupamento, Sumarização de textos e Recuperação da Informação. Quase todas as aplicações removem essas palavras na fase de pré-processamento, o que aumenta o desempenho da aplicação ao reduzir o tempo de processamento e o tamanho dos documentos a serem analisados. Em geral, tais palavras são categorizadas como conjunções, preposições, advérbios e artigos, conforme exemplo da Figura 2. O agrupamento dessas palavras em uma lista é chamado de "lista de *stopwords*"ou *stopwords corpus* (RAULJI; SAINI, 2016; SARICA; LUO, 2021; KAUR; BUTTAR, 2018).

Dentre várias abordagens para a remoção de stopwords, Raulji e Saini (2016) apresentam um algoritmo aplicado e validado em textos de diferentes domínios com os passos que se seguem: (1) o documento a ser classificado é tokenizado e as palavras são armazenadas individualmente em um vetor; (2) uma palavra da lista de stopwords é lida pelo algoritmo; (3) esta palavra é comparada com cada palavra do documento de texto através de uma pesquisa sequencial; (4) caso a palavra no documento coincida com a stopword, esta é removida e a comparação continua até o fim do arquivo; (5) após a remoção de todas as palavras do arquivo que coincidam com a stopword, outra palavra da lista é carregada, o algoritmo retorna ao segundo passo e o texto é processado de forma contínua até que todas as palavras sejam comparadas; (6) por fim, o texto com as stopwords removidas é exibido e as estatísticas de remoção são apresentadas, tais como número de palavras removidas do documento, número de palavras presentes originalmente, número de palavras após a remoção das stopwords e contagem de cada stopword identificada.

Sentença: Palavras que contribuem pouco para o entendimento do texto devem ser removidas
Remoção de *stopwords*: Palavras contribuem entendimento texto devem removidas

Figura 2 – Exemplo da remoção de *stopwords* - autoria própria.

2.4.3 *Stemming*

Esta etapa do pré-processamento trata da fusão de variadas formas de uma palavra em uma única representação, i.e. a raiz da palavra. A redução não precisa resultar em uma palavra válida, mas sim capturar o sentido da palavra de origem. Existem inúmeras maneiras de realizar essa etapa, desde métodos manuais e automáticos, dependentes ou independentes do idioma (SHARMA; CSE, 2012).

Esse processo de remoção, em geral, varia conforme a formação das palavras em cada idioma. Por exemplo, o *stemming* reduz as palavras *joga*, *jogar* e *jogando* para a raiz *jog* (ver Figura 3). Uma vez que o termo reduzido à sua raiz representa um conceito mais amplo que o original, o processo de *stemming* possibilita o aumento dos termos recuperados durante buscas (JIVANI et al., 2011; HAROON, 2018). De acordo com Liu et al. (2019), mídias e redes sociais são uma excelente fonte para coleta de dados e se apresentam como uma grande plataforma para a aplicação de algoritmos de *stemming* para a análise de textos.

Sentença: Palavras costumam ser invariavelmente reduzidas nesse processo
***Stemming*:** Palavr costum ser vari reduzi nesse process

Figura 3 – Exemplo de *stemming* - autoria própria.

2.4.4 Lematização

A lematização é um processo similar ao *stemming*, exceto pelo fato de não reduzir as palavras do texto às suas raízes. Na lematização cada palavra é transformada na sua forma inflexiva (lema) ao substituir o sufixo por outro que resulte na forma normalizada dessa palavra, como exemplificado na Figura 4. Em alguns casos o resultado dos processos de *stemming* e da lematização são o mesmo, como no caso das palavras em inglês *working*, *works* e *worked* que resultam em *work* em ambos os casos. Em outros casos os resultados dos dois processos se diferenciam, como no caso das palavras em inglês *computes*, *computing* e *computed* que resultam em *comput* após o *stemming* e *compute* após a lematização (PLISSON et al., 2004).

Sentença: Palavras flexionadas devem ser normalizadas
Lematização: Palavra flexionar dever ser normalizar

Figura 4 – Exemplo de lematização - autoria própria.

2.5 Representação dos Dados

A efetividade da categorização de textos não depende apenas do desempenho de algoritmos de aprendizado, o tipo de representação textual escolhido também é um fator determinante. Documentos de textos não podem ser naturalmente interpretados por classificadores. Antes de realizar a classificação é necessário transformar os documentos em representações adequadas ao algoritmo e à tarefa de classificação. A escolha da representação depende da unidade de texto considerada significativa, bem como das regras de linguagem natural para a combinação dessas unidades. Em geral, cada documento é representado por um vetor de termos ponderados ou as unidades textuais significativas para a categorização (SONG; LIU; YANG, 2005). A seguir serão apresentadas algumas das representações mais utilizadas na literatura.

2.5.1 *Bag of Words* (BOW)

Nessa abordagem, cada documento é representado como um vetor de palavras que ocorrem no documento ou, em representações mais sofisticadas, como frases ou sentenças. Esse tipo de representação é considerado uma das mais simples técnicas que apresentam um bom desempenho. Entre as desvantagens dessa representação destacam-se a alta dimensionalidade e a alta incidência de valores esparsos, uma vez que cada palavra é um possível atributo. Essa abordagem exige ferramentas computacionais específicas que transformem automaticamente os documentos em uma representação estruturada, realizando, ao mesmo tempo, a redução da dimensionalidade (MATSUBARA; MARTINS; MONARD, 2003). A Figura 5 ilustra esse tipo de representação.

	este	é	um	exemplo	outro	terceiro
Este é um exemplo	1	1	1	1	0	0
Este é outro exemplo	1	1	0	1	1	0
Este é um terceiro exemplo	1	1	1	1	0	1

Figura 5 – Exemplo de representação BOW - autoria própria.

2.5.2 N-gramas

N-gramas são sequências contínuas de palavras em um texto. Essa abordagem normalmente resulta em centenas ou milhares de novas variáveis, cada uma representando a frequência que uma dada sequência ocorre no texto. Uma variável indicativa da ocorrência de uma única palavra ou termo é chamada de *unigrama*. Na representação de unigramas cada coluna representa um termo do documento e cada célula fornece a ocorrência de uma palavra no texto. A última coluna *n-token* fornece evidências do número de palavras no documento após o pré-processamento. O grau(n) dos n-gramas indica

a quantidade de palavras representadas na sequência. Dessa forma, uma sequência de 2 palavras é chamada de *bigrama* (ver Figura 6). Para evitar o aumento demasiado de colunas na representação, utiliza-se um limite mínimo para representar uma sequência de termos (SCHONLAU; GUENTHER, 2017).

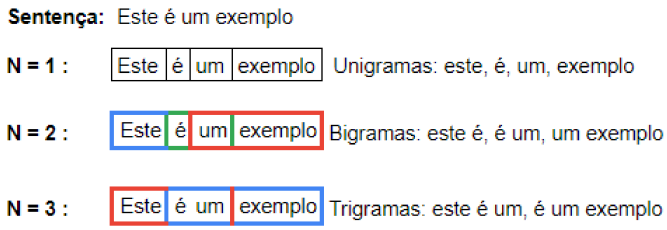

Figura 6 – Exemplo da n-gramas - autoria própria.

2.5.3 TF-IDF

O termo TF-IDF é uma abreviação de *Term Frequency - Inverse Document Frequency*. Nessa abordagem determina-se a frequência relativa dos termos no documento em relação à proporção inversa do termo em todo o conjunto de documentos. Basicamente esse cálculo indica o quão relevante um termo é em um documento. Palavras comuns em um único ou pequeno grupo de documentos tendem a apresentar um valor mais alto de TF-IDF do que palavras comuns em todo o corpus como artigos e preposições (RAMOS et al., 2003).

Segundo Ramos et al. (2003), os procedimentos para a representação dos dados nessa abordagem podem apresentar algumas variações, a depender da aplicação, mas as etapas podem ser resumidas, em geral, como se segue. Dado um conjunto de documentos D, uma palavra w e um documento $d \in D$, calcula-se

$$w_d = f_{w,d} \times log(|D| f_{w,D}), \tag{1}$$

em que $f_{w,d}$ representa a quantidade de vezes que w ocorre em d, $|D|$ o tamanho do corpus e $f_{w,D}$ o número de documentos com a presença de w em D. A Figura 7 apresenta um exemplo para o cálculo TF-IDF para os termos presentes em três sentenças.

	este			é			um			exemplo			outro			terceiro		
	TF	IDF	TF-IDF	TF	IDF	TF-IDF	TF	IDF	TF-IDF	TF	IDF	TF-IDF	TF	IDF	TF-IDF	TF	IDF	TF-IDF
S1	1/4 = 0.25	log(3/3) = 0	(0,25).(0) = **0**	0.25	0	**0**	0,25	0,18	**0,05**	0,25	0	**0**	0	0,47	**0**	0	0,47	**0**
S2	1/4 = 0.25	log(3/3) = 0	(0,25).(0) = **0**	0.25	0	**0**	0	0,18	**0**	0,25	0	**0**	0,25	0,47	**0,12**	0	0,47	**0**
S3	1/5 = 0.20	log(3/3) = 0	(0,20).(0) = **0**	0.20	0	**0**	0,2	0,18	**0,04**	0,2	0	**0**	0	0,47	**0**	0,2	0,47	**0,09**

Figura 7 – Exemplo do cálculo TF-IDF - autoria própria.

Algumas circunstâncias podem ocorrer para um determinado termo de acordo com os valores de $f_{w,d}$, $|D|$ e $f_{w,D}$. Caso $|D| \sim f_{w,D}$, o tamanho do *corpus* é aproximadamente igual à frequência de w sobre D. Caso, $1 < log(|D| f_{w,D}) < c$, para uma constante c

de valor baixo, então w_d apresenta um valor inferior a $f_{w,D}$, porém ainda positivo, o que indica que w é relativamente comum em todo o corpus mas ainda apresenta alguma relevância em D. Por fim, caso $f_{w,d}$ seja grande e $f_{w,D}$ seja pequeno, então $log(|D|f_{w,D})$ será ainda maior, da mesma forma que w_d. Este último caso é, em geral, o de maior interesse, já que representa os casos em que uma palavra w é importante em d, mas não é comum em D. Nesse caso considera-se que w tem um grande poder de discriminação.

2.6 Extração de Tópicos

É comum em tarefas de mineração de textos a partir de coleções de documentos que se queira dividir o *corpus* em grupos naturais que podem ser avaliados separadamente. A modelagem de tópicos é um método de classificação não supervisionada, similar a outras técnicas de agrupamento de dados numéricos, que encontra grupos naturais de itens (Julia Silge and David Robinson, 2022).

2.6.1 Alocação Latente de *Dirichlet*

A Alocação Latente de *Dirichlet* ou *Latent Dirichlet allocation* (LDA) é um método popular de modelagem de tópicos que trata cada documento como um uma mistura de tópicos e cada tópico como uma mistura de palavras. Essa estratégia permite a sobreposição de documentos em termos de conteúdo, comum no uso de linguagem natural, ao invés de formar grupos separados e discretos (Julia Silge and David Robinson, 2022).

De acordo com Blei, Ng e Jordan (2003), o LDA é um modelo probabilístico generativo de um corpus, que consiste na ideia de que os documentos são representados como misturas aleatórias sobre tópicos latentes, de forma que cada tópico é caracterizado por uma distribuição sobre palavras.

Faleiros, Lopes et al. (2016) explicam que a distribuição de *Dirichlet* ($Dir(z, \alpha)$), denotada pela Equação 2, durante o processo generativo aloca palavras de diferentes tópicos e que preencherão os documentos, ou seja, esse modelo se propõe a alocar tópicos latentes distribuídos segundo essa distribuição.

$$Dir(z, \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} z_k^{\alpha_k - 1}, \tag{2}$$

em que $z = (z_1, \ldots, z_K)$ é uma varável K-dimensional, $0 \leq z \leq 1$ e $\sum_{i=1}^{K} z_i = 1$. $\alpha = (\alpha_1, \ldots, \alpha_K)$ são os hiper-parâmetros da distribuição. A função $B(\alpha)$ é denominada *função Beta* e é expressa por meio da função Γ, descrita na Fórmula 3.

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}, \tag{3}$$

O processo gerador do LDA assume que os tópicos, definidos como distribuições de probabilidade sobre um vocabulário fixo de palavras, são especificados antes que qualquer dado seja gerado. Os documentos são conjuntos aleatórios de palavras pertencentes a uma distribuição de tópicos. Para explicar o processo gerador do modelo LDA, deve-se assumir que cada documento d_j é criado da seguinte forma:

1. Criam-se distribuições $\phi_k \sim Dir(\phi_k, \beta)$, para cada tópico k, com $0 \leq k \leq K$;
2. Criam-se distribuições $\theta_j \sim Dir(\theta, \alpha)$ para cada documento d_j;
3. Para cada posição i das palavras em d_j:
   - ❏ Escolhe-se aleatoriamente um tópico $z_{j,i} \sim Multinomial(\theta_j)$;
   - ❏ Escolhe-se aleatoriamente uma palavra $w_{j,i}$ com probabilidade $p(w_{j,i}|\phi_{zj,i})$.

A variável ϕ é n-dimensional e composta por n palavras do vocabulário. θ é uma variável K-dimensional, composta por K tópicos. Ambas são geradas pela distribuição de *Dirichlet* com seus respectivos hiper parâmetros β e α e descrevem distribuições de probabilidade, com $\sum_j^n \phi_j = 1$, $\phi_i > 0$, $\sum_i^K \theta_i = 1$ e $\theta_i > 0$.

Assume-se que cada documento possui sua própria distribuição de tópicos θ_j, de forma que um mesmo documento pode estar relacionado com vários tópicos em diferentes proporções de relevância.

Todo o processo generativo pode ser representado por meio de uma rede Bayesiana com três níveis. O primeiro representa a distribuição de tópicos em toda a coleção de documentos, o segundo a distribuição de tópicos para cada documento e, no último nível, repete-se a distribuição dos tópicos internamente para as palavras em um documento.

A interpretação dos hiper parâmetros permite observar que um valor alto de α indica que cada documento provavelmente contém uma maior mistura de tópicos, enquanto um valor baixo de α indica uma maior concentração de poucos tópicos nos documentos. De forma análoga, um valor alto de β indica uma maior probabilidade de cada tópico conter misturas de várias palavras e um valor baixo desse hiper parâmetro indica que cada tópico é composto por poucas palavras.

Transcrevendo as probabilidades de todas as variáveis latentes do modelo, segundo as informações *a priori*, tem- a seguinte distribuição conjunta detalhada pela Fórmula 4:

$$p(z, w, \phi, \theta|\alpha, \beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \left(\prod_{i=1}^{V} p(z_{j,i}|\overrightarrow{\theta}_j) p(w_{i,j}|z_{i,j}, \phi_{z_{j,i}}) \right). \quad (4)$$

O grande problema computacional do LDA é inferir $p(z, \phi, \theta, |w, \alpha, \beta)$, em que w representa todas as palavras observadas na coleção de documentos. Pelo teorema de Bayes, é possível formular a probabilidade $p(z, \phi, \theta, |w, \alpha, \beta)$ como o cálculo da *a posteriori* do LDA, deifinido pela Equação 5, em que o numerador é a probabilidade marginal dos dados observados expressa pela Equação 4.

$$p(z, \phi, \theta, |w, \alpha, \beta) = \frac{p(z, w, \phi, \theta|\alpha, \beta)}{p(w)}. \quad (5)$$

Dessa forma, o problema computacional central pode ser resolvido inferindo a probabilidade *a posteriori*, descrita pela Equação 5, calculada, teoricamente pela soma da distribuição conjunta de todos os valores possíveis atribuídas às variáveis não observadas. O número de atribuições possíveis é exponencialmente grande, mas existem vários métodos disponíveis na literatura para aproximar essa distribuição.

2.6.2 NMF

O algoritmo baseado em matrizes não negativas de fatoração (do inglês *Non-negative Matrix Factorization* ou NMF), conforme explica Egger e Yu (2022), é um método decomposicional, não probabilístico que pertence ao grupo dos algoritmos algébricos-lineares. O NMF utiliza os dados representados por meio do TF-IDF e gera 2 matrizes de *rankings* baixos, considerando a característica dessa representação de avaliar a importância de uma palavra em uma coleção de documentos.

Conforme pode ser observado na Figura 8, o NMF decompõe a base de dados textuais A em um produto da matriz de termos-tópicos W, que contém os vetores bases, e da matriz de termos documentos H com os pesos correspondentes. Os valores de W e H são modificados iterativamente, sendo necessário que W e H sejam não negativas, uma vez que seria difícil a interpretação de tópicos com valores negativos.

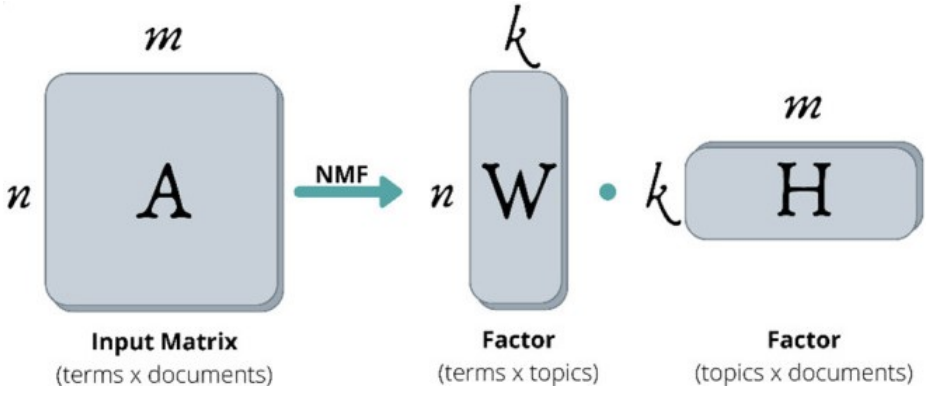

Figura 8 – Ilustração do NMF - autoria Egger e Yu (2022).

Essa técnica exige que os dados sejam pré-processados por meio de um *pipeline* clássico de PLN com a transformação do texto para caixa baixa, remoção de palavras de parada, lematização, *stemming*, bem como a remoção de números e sinais de pontuação.

2.6.3 BERTopic

De acordo com Grootendorst (2022), o BERTopic gera a representação de tópicos por meio de 3 etapas: na primeira o documento é convertido em uma camada de incorporação

embedding por meio de um modelo de linguagem pré-treinado, em seguida a dimensionalidade da camada *embedding* é reduzida e, por fim, é realizado o agrupamento das camadas de incorporação. Dos grupos de documentos, representações de tópicos são extraídas através de uma variação da representação TF-IDF.

Camada de Incorporação

Os documentos de textos são incorporados no BERTopic a fim de criar camadas de representação em um espaço vetorial que possam ser comparadas semanticamente. Considera-se que documentos que contenham o mesmo tópico são semanticamente similares. O *framework* utilizado pelo algoritmo, *Sentence-BERT* (SBERT), permite que o usuário converta sentenças e parágrafos para vetores densos de representação por meio de modelos de linguagem pré-treinados. As camadas de incorporação são utilizadas pelo algoritmo com o objetivo de realizar o agrupamento de documentos similares e não são utilizadas diretamente na extração de tópicos.

Agrupamento de Documentos

Em espaços altamente dimensionais, como os utilizados para representação de documentos textuais, o conceito de localização espacial torna-se mal definido e medidas de distâncias precisam ser mais bem adequadas. Uma abordagem mais adequada se dá pela redução de dimensionalidade das camadas de incorporação, antes do agrupamento. No BERTopic a redução da dimensionalidade ocorre por meio da *Aproximação e Projeção de Manifold Uniforme para Redução de Dimensões* ou *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) (MCINNES; HEALY; MELVILLE, 2018), devido ao poder dessa técnica de preservar características locais e globais de dados altamente dimensionais em dimensões menores projetadas. Além disso, a redução UMAP pode ser usada em diferentes modelos de linguagem com espaços dimensionais distintos.

Após a redução de dimensionalidade, as camadas de incorporação são agrupadas através do algoritmo de *Agrupamento Baseado em Densidade Hierárquica* ou *Hierarchical Density Based Clustering* (HDBSCAN) (MCINNES; HEALY; ASTELS, 2017), que gera grupos de forma que ruídos sejam modelados como *outliers*. Conforme explica Grootendorst (2022), a literatura evidencia a melhora na performance do agrupamento em termos de tempo de acurácia de agrupamento após a redução UMAP.

Representação de Tópicos

A representação de tópicos é modelada alocando todos os documentos de um mesmo grupo em um mesmo tópico. Para cada tópico investiga-se o que o torna diferente de outros tópicos baseado na distribuição de palavras do seu grupo em relação ao demais

grupos. Para isso uma modificação na representação TF-IDF é realizada, concatenando os documentos de um grupo. A representação TF-IDF é extraída da seguinte forma:

$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{tf_t}), \tag{6}$$

em que a frequência de um termo t é modelada em uma classe c de documentos concatenados para cada grupo. Em seguida a frequência de documento inversa é substituída pela frequência de classe inversa com o objetivo de mensurar quanta informação um termo fornece para uma classe. A importância de um termo para uma classe é calculada através do logaritmo do número médio de palavras por classe A dividido pela frequência do termo t em todas as classes (c-TF-IDF).

Dessa forma, a representação c-TF-IDF baseada em classes produz modelos de importância de palavras em grupos ao invés de documentos. Por fim, ao fundir iterativamente as representações do tópico menos comum com o tópico mais similar a este, a quantidade de tópicos pode ser reduzida a um número especificado pelo usuário.

2.6.4 Coerência NPMI

Uma forma de quantificar a coerência de tópicos extraídos por meio de algoritmos de extração de tópicos se dá através do indicador de *Informação Mútua Pontual Normalizada* ou *Normalized Pointwise Mutual Information* (NPMI), proposto por Bouma (2009).

De acordo com o autor, quando duas palavras ocorrem simultaneamente, a chance de encontrar uma delas é a mesma de encontrar a segunda palavra, que é a mesma chance de encontrar ambas as palavras juntas. Nesse caso, a *Informação Mútua Pontual* ou *Pointwise Mutual Information* (PMI) é dada por:

$$i(x, y) = -\ln p(x) = -\ln p(y) = -\ln p(x, y), \tag{7}$$

quando X e Y são perfeitamente correlacionadas e $p(x, y) > 0$.

A normalização pode ser feita utilizando combinações de $-\ln p(x)$ e $-\ln p(y)$ ou ainda por $-\ln p(x, y)$. A última opção se mostra mais eficaz devida à propriedade os limites inferior e superior do indicador NPMI, definido como:

$$i_n(x, y) = \frac{\ln \frac{p(x,y)}{p(x)p(y)}}{-\ln p(x, y)}. \tag{8}$$

Bouma (2009) destaca algumas propriedades importantes dessa métrica de coerência:

1. Quando duas palavras ocorrem juntas, $i_n(x, y) = 1$;

2. Quando duas palavras são distribuídas de forma independente, $i_n(x, y) = 0$, uma vez que o denominador é 0;

3. Quando existe a ocorrência das duas palavras, mas elas não aparecem juntas, $i_n(x, y) = -1$, já que $p(x, y)$ se aproxima de 0 e $p(x)$ e $p(y)$ são fixos.

Ou ainda, conforme Campagnolo, Duarte e Bianco (2022), pode-se dizer que a métrica NPMI é uma medida normalizada no intervalo $[-1, 1]$, em que os valores -1, 0 e 1 indicam, respectivamente, nenhuma coocorrência, independência e total coocorrência entre termos.

2.7 Algoritmos de classificação

A classificação tem sido amplamente estudada em diversas aplicações, tais como banco de dados, mineração de dados e recuperação da informação. O problema da classificação, segundo Aggarwal e Zhai (2012), pode ser assim definido: dado um conjunto de treinamento $D = X_1, ..., X_n$, em que cada registro é rotulado com um valor de classe dentre k valores discretos possíveis indexados por $1, ..., k$, têm-se que o conjunto de treinamento D é usado na construção de um *modelo de classificação*, que relaciona as características de D a um dos k rótulos de classe.

Nos últimos tempos houve um crescimento exponencial na quantidade de documentos complexos e textos que requerem um entendimento profundo de técnicas de aprendizado de máquina que sejam capazes de realizar a classificação de textos de forma precisa. As variadas abordagens têm demonstrado resultados cada vez mais promissores no processamento de linguagem natural. A eficácia de cada técnica depende da sua capacidade de entender modelos complexos e relações não lineares presentes nos dados. O desafio para os pesquisadores consiste em aplicar as estruturas, arquiteturas e técnicas mais adequadas para a classificação de textos. O passo mais importante no *framework* de classificação de textos é escolher o classificador mais adequado, que só pode ser determinado a partir do entendimento conceitual de cada algoritmo (KOWSARI et al., 2019).

2.7.1 Classificador *Naive Bayes*

O classificador *Naive Bayes* talvez seja o mais simples e amplamente utilizado. Este classificador modela a distribuição de documentos em cada classe segundo uma função probabilística em que termos diferentes são distribuídos independentemente um do outro. Mesmo em contextos reais em que essa suposição é claramente falsa, este classificador costuma apresentar um desempenho satisfatório (ALLAHYARI et al., 2017).

Dentre os diversos trabalhos encontrados na literatura que utilizaram esse algoritmo para realizar a classificação de sentimentos de textos encontram-se: i) Wongkar e Angdresey (2019) utilizaram o algoritmo *Naive Bayes* para analisar o sentimento da comunidade acerca dos candidatos à eleição presidencial em 2019 na República da Indonésia a partir de publicações do X; ii) Laksono et al. (2019) classificaram o nível de satisfação de clientes que visitaram um restaurante utilizando avaliações publicadas em um site de viagens; iii)

Novendri et al. (2020) analisaram o sentimento presente em comentários publicados no Youtube sobre uma série produzida pela Netflix.

De acordo com McCallum, Nigam et al. (1998), diferentes aplicações utilizam modelos probabilísticos de primeira ordem com a mesma suposição *bayesiana*, o *Modelo Multivaridao de Bernoulli* e o *Modelo Multinomial*. Ambos os casos assumem que os documentos de texto são gerados por um modelo de mistura parametrizado por θ. O referido modelo de mistura é composto por componentes $c_j \in C = c_1, ..., c_{|C|}$. Cada componente é parametrizada por um subconjunto disjunto de θ. Assim, um documento d_i é criado a partir da Equação 9, selecionando um componente de acordo com a probabilidade *a priori* $P(c_j|\theta)$, então a Equação 10 fornece a componente de mistura, gera um documento com seus próprios parâmetros e distribuição $P(d_i|c_j;\theta)$. A probabilidade de ocorrência de um documento pode ser definida segundo a soma da probabilidade de todos os componentes de mistura (Equação 9).

$$P(d_i|\theta) = \sum_{j=1}^{|C|} P(c_j|\theta)P(d_i|c_j\theta). \tag{9}$$

$$P(d_i|c_j;\theta) = \prod_{t=1}^{|V|}(B_{it}P(w_t|cj;\theta)) + (1 - B_{it})(1 - P(w_t|c_j;\theta)). \tag{10}$$

Cada documento tem um rótulo de classe, de forma que não existe uma relação entre as classes dos componentes de mistura, possibilitando representar c_j como o j-ésimo componente de mistura e a j-ésima classe.

Na Equação 10, V representa um conjunto de termos ou vocabulário, em que cada componente $t, t \in 1, ..., |V|$, corresponde a uma palavra w_t desse conjunto. A dimensão t do vetor de documentos d_i é indicada por B_{it}, com valores 0 ou 1, a depender da ocorrência da palavra w_t no documento. Nessa representação segue-se a suposição *bayesiana* de que a probabilidade de ocorrência de uma dada palavra no documento é independente da ocorrência de todas as outras palavras no documento. Assim, a probabilidade de ocorrência de um documento d_i, tal que $d_i = w_1, w_2, ..., w_{n_i}$, pode ser calculada por:

$$P(w_1, w_2, ..., w_{n_i}|c_j;\hat{\theta}_j) = \prod_{i=1}^{n_i} P(w_i|c_j;\hat{\theta}), \tag{11}$$

.

em que, $\hat{\theta}$ é a estimativa do parâmetro de classificação (ALLAHYARI et al., 2017).

2.7.2 Classificador KNN

Trata-se de um classificador baseado em proximidade que usa medidas baseadas em distância para realizar a classificação. Neste algoritmo, documentos com a mesma classificação tendem a serem mais similares ou próximos entre si, com base em alguma medida de proximidade (ALLAHYARI et al., 2017).

Wongkar e Angdresey (2019) utilizaram o algoritmo KNN em comparação ao *Naive Bayes* para avaliar a impressão dos usuários do *X* sobre a eleição presidencial em 2019 na República da Indonésia. Shamrat et al. (2021) avaliaram a opinião de usuários do *X* através de técnicas de NLP e classificação de sentimentos sobre a segurança e efetividade de vacinas contra a COVID-19 através do classificador KNN. Huq, Ali e Rahman (2017) realizaram a classificação de sentimentos presentes em *tweets* utilizando esse mesmo algoritmo.

Para classificar um documento desconhecido d_0, o classificador KNN (*K-Nearest Neighbors* ou K Vizinhos Mais Próximos) ranqueia os documentos vizinhos na base de treinamento e usa o rótulo de classe dos k vizinhos mais similares para prever a classe do documento de entrada (TAN, 2006). Para medir a similaridade de forma eficiente, utiliza-se a distância de cosseno dada por

$$Sim(\vec{d_1}, \vec{d_2}) = \frac{\vec{d_1} \cdot \vec{d_2}}{||\vec{d_1}|| \cdot ||\vec{d_2}||} = \frac{\sum_{l=1}^{V} d_{1l} \times d_{2l}}{\sqrt{\sum_{l=1}^{V} d_{1l}^2}\sqrt{\sum_{l=1}^{V} d_{2l}^2}}, \tag{12}$$

em que V representa a dimensão do vetor de documentos $\vec{d_i}$.

As classes dos vizinhos são ponderadas segundo a similaridade de cada vizinho em relação a d_0, de acordo com a Equação 13

$$score(\vec{d_0}, C_i) = \sum_{\vec{d_j} \in KNN_{\vec{d_0}}} Sim(\vec{d_0}, \vec{d_j})\delta(\vec{d_j}, C_i), \tag{13}$$

em que $KNN(\vec{d})$ denota os K vizinhos mais próximos do documento $\vec{d_0}$ e $\delta(\vec{d_j}, C_i)$ a classificação do documento $\vec{d_j}$ em relação a C_i, dada pela Equação 14.

$$\delta(\vec{d_j}, C_i) = \begin{cases} 1 & \vec{d_j} \in C_i \\ 0 & \vec{d_j} \notin C_i \end{cases}. \tag{14}$$

A regra de decisão do classificador KNN é dada pela Equação 15.

$$C = argmax_{c_i}(score(\vec{d_0}, C_i)) = argmax_{c_i}\left(\sum_{\vec{d_j} \in KNN(\vec{d_0})} Sim(\vec{d_0}, \vec{d_j})\delta(\vec{d_j}, C_i)\right). \tag{15}$$

2.7.3 Árvores de Decisão

Também conhecidos como métodos de classificação *ensemble*, os algoritmos baseados em árvores de decisão constroem um conjunto de classificadores base e rotulam novos casos a partir do voto de suas predições e se constituem uma das principais abordagens na comunidade de aprendizado de máquina (SHI et al., 2010).

Fitri, Andreswari e Hasibuan (2019) utilizaram árvores de decisão e outras duas técnicas para mensurar o sentimento presente em *tweets* com reações à uma campanha anti-LGBT realizada na Indonésia. Bayhaqy et al. (2018) compararam a classificação de

sentimentos a partir de textos publicados por clientes de comércios eletrônicos no *X* por meio de árvores de decisão e dos algoritos KNN e *Naive Bayes*. Rathi et al. (2018) classificaram o sentimento de *tweets* por meio de um classificador baseado em árvores de decisão combinadas com Máquinas Vetores de Suporte, discutidas na Subseção 2.7.4.

As árvores de decisão particionam o conjunto de treinamento recursivamente em subdivisões menores baseadas em um conjunto de testes definidos em cada nó ou ramo da árvore. Cada nó da árvore é um teste de alguns atributos da instância de treinamento e, cada ramo descendente do nó representa um dos valores desse atributo. Uma instância é classificada a partir do nó raiz, percorrendo os ramos da árvore de acordo com os valores dos atributos em uma instância (ALLAHYARI et al., 2017).

Árvores mais simples utilizam a estratégia de conquistar e dividir em que um conjunto de treinamento M, no qual a palavra t_i que forneça maior ganho de informação é selecionada. Então, M é particionado em dois subconjuntos, M_i^+ contento a palavra t_i e M_i^- sem a presença da palavra t_i. O procedimento é repetido em M_i^+ e M_i^- até que todos os documentos em um subconjunto pertençam a uma classe L_c. Por fim, uma árvore de regras é gerada para a determinação da classe de uma determinada folha (HOTHO; NÜRNBERGER; PAASS, 2005).

2.7.4 SVM

Máquinas de Vetores de Suporte consistem em encontrar um hiperplano ótimo de separação entre os casos positivos e negativos para a característica classificada. Este hiperplano é o que apresenta uma margem máxima de separação entre os exemplos de treinamento mais próximos ao hiperplano (vetores de suporte). Uma vez que o hiperplano é definido, novos casos podem ser classificados com base na posição que ocupam no hiperplano (SILVA; RIBEIRO, 2007).

Ahmad, Aftab e Ali (2017) realizaram a classificação da polaridade de sentimentos por meio desse algoritmo a partir de bases de dados compostas por *tweets*. Os autores utilizaram as métricas de precisão, revocação e *F1-escore*, apresentadas na Subseção 2.7.6, para a validação de resultados. Kumari, Sharma e Soni (2017) analisaram o sentimento presente em textos avaliativos sobre um *smartphone* publicados no *X*. Prastyo et al. (2020) utilizaram classificadores SVM para entender tendências públicas de opinião sobre a COVID-19 na Indonésia sobre uma perspecitiva geral e econômica em textos do *X*.

Seja o conjunto de treinamento 16

$$(X_i, y_i); i = 1, \ldots, n; x \in R^d; y \in +1, -1, \tag{16}$$

com o rótulo de classificação, utilizado para resolver o problema quadrático de programação, representado pela Equação 17 e sujeito à Equação 18.

$$min\phi(\omega) = \frac{1}{2}||\omega^2|| = \frac{1}{2}(\omega \cdot \omega). \tag{17}$$

$$y_i(\omega \cdot x_i + b) \geq 1; i = 1, 2, \ldots, n. \tag{18}$$

A superfície ótima de classificação, representada pela Equação 19, é derivada. A otimização de Lagrange é utilizada para resolver o problema convertido para um caso dual, através do teorema de Kuhn-Tucker.

$$g(x) = \omega \cdot x + b = 0. \tag{19}$$

A função de classificação ótima 20 é obtida e, caso não seja possível uma separação linear, um relaxamento $\xi_i \geq 0$ pode ser adicionado (JU; WANG; ZHU, 2011).

$$f(x) = sign\left\{\sum_{i=1}^{1} a_i^* y_i (x_i \cdot x) b^*\right\}. \tag{20}$$

2.7.5 Regressão Logística

Trata-se de um modelo discriminativo utilizado para calcular $P(c|x)$ (Equação 21), discriminando os possíveis valores da classe y de acordo com a entrada x.

$$P(c|x) = \sum_{i=1}^{N} w_i \cdot f_i. \tag{21}$$

O valor de $P(c|x)$ não pode ser calculado diretamente através da Equação 21 por resultar em valores de $-\infty$ a ∞. Para que sejam gerados valores no intervalo de 0 a 1, a função exponencial 22 é utilizada.

$$P(c|x) = \frac{1}{z} \exp \sum_{i} w_i \cdot f_i. \tag{22}$$

Com o objetivo de substituir o fator de normalização Z pelo número de características N, a transformação ilustrada pela Equação 23 é aplicada.

$$P(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i \cdot f_i)}{\sum_C \exp(\sum_{i=1}^{N} w_i \cdot f_i(c', x)}. \tag{23}$$

É comum em processamento de textos utilizar valores com características binárias, de forma que não somente a observação x seja considerada, mas também a classe candidata de saída c. Assim, ao invés de f_i ou $f_i(x)$, $f_i(c, x)$ é calculada em que a característica i da classe c é designada como um dado de entrada referente à observação x. Portanto, a Equação 24 calcula a probabilidade de y, uma classe de c, dado x.

$$P(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i \cdot f_i(c, x))}{\sum_{c' \in C} \exp(\sum i = 1^N w_i \cdot f_i(c', x))}. \tag{24}$$

O método da máxima verossimilhança condicional é utilizado pela regressão logística como um estimador de pesos w ao escolher o peso que maximiza a probabilidade de ocorrência da classe y, a partir da observação x, de acordo com as Equações 25 e 26 (INDRA; WIKARSA; TURANG, 2016).

$$\hat{w} = argmax_w \sum_j \log P(y_j|x_j). \tag{25}$$

$$\begin{aligned} L(w) = \sum_j \log P(y_j|x_j) = \log \sum_j \exp\left(\sum_{i=1}^{N} w_i f_i(y^{(j)}, x^{(i=j)})\right) \\ - \log \sum_j \sum_{y' \in Y} \exp\left(\sum_{i=1}^{N} w_i f_i(y^{(j)}, x^{(i=j)})\right) \end{aligned}. \tag{26}$$

Ramadhan, Novianty e Setianingsih (2017) coletaram *tweets* e utilizaram a Regressão Logística Multinominal aplicada sobre vetores binários e vetores transformados por meio do método TF-IDF para realizar a análise de sentimento dos textos coletados. Tyagi e Sharma (2018) classificaram *tweets* em positivos, negativos ou neutros através de classificadores baseados em regressão logística aplicada sobre unigramas. Majumder, Aich e Das (2021) avaliaram, por meio do classificador SVM e regressão logística, o sentimento sobre medidas de combate à COVID-19 em *tweets* coletados.

2.7.6 Medidas de Validação

Existem várias medidas utilizadas para avaliar a efetividade dos modelos implementados na literatura, embora os mais aplicados sejam a precisão, revocação e acurácia. Para um correto entendimento dessas métricas é necessário avaliar se a classificação do documento resulta em um resultado verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN) ou falso negativo (FN), conforme as definições da Tabela 1 (IKONOMAKIS; KOTSIANTIS; TAMPAKAS, 2005).

Tabela 1 – Possíveis Resultados da Classificação de um Documento

Resultado	Definição
VP	O documento é corretamente classificado como pertencente a uma categoria de interesse
FP	O documento é incorretamente classificado como pertencente a uma categoria de interesse
FN	O documento é incorretamente classificado como não pertencente a uma categoria de interesse
VN	O documento é corretamente classificado como não pertencente a uma categoria de interesse

A precisão π_i (Equação 27) é definida como a probabilidade condicional de que um documento aleatório d seja corretamente classificado sob uma categoria c_i.

$$\pi_i = \frac{VP_i}{VP_i + FP_i}. \tag{27}$$

A revocação ou *recall* ρ_i (Equação 28) representa a probabilidade condicional de que um documento que pertença a uma categoria c_i seja corretamente classificado.

$$\rho_i = \frac{TP_i}{TP_i + FN_i}. \tag{28}$$

A acurácia (Equação 29) é uma medida comum utilizada em técnicas de categorização e indica o acerto geral do classificador dentre todas as classes, embora essa medida seja menos sensível à variação na quantidade de decisões corretas em comparação à revocação ou à precisão. Nos casos em que existam poucos documentos sob a classe de interesse, o acerto da classe negativa na composição dessa medida pode mascarar o acerto da classe positiva e causar problemas durante a avaliação do desempenho dos classificadores.

$$A_i = \frac{VP_i + VN_i}{VP_i + VN_i + FP_i + FN_i}. \tag{29}$$

A precisão e a revocação podem ainda ser combinadas de forma a fornecer uma medida mais precisa da perfomance do classificador, de acordo com a Equação 30

$$F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho}, \tag{30}$$

em que π e ρ representam a precisão e a revocação, respectivamente. β é um parâmetro positivo que indica o objetivo da tarefa de avaliação. Caso a precisão seja considerada mais importante que a revocação, o valor de β converge para 0. Caso a revocação seja considerada mais importante, β converge para um valor infinito. Em geral, F_β é calculado com $\beta = 1$, de forma que a precisão e a revocação tenham igual importância.

Na etapa de validação e seleção dos classificadores é comum separar a base de dados em conjuntos de treinamento e teste. Uma abordagem muito utilizada é denominada *K-fold Cross Validation*, na qual a base de dados é dividida em k subconjuntos e então, iterativamente, alguns desses subconjuntos são utilizados na etapa de treinamento, enquanto os demais medem o desempenho dos classificadores, a partir de alguma medida de validação (ANGUITA et al., 2012).

2.8 TensiStregth

Proposto por Thelwall (2017), o *TensiStregth* (TS) é uma adaptação do *software* de detecção da intensidade de sentimento *SentiStrength* (THELWALL et al., 2010), com o objetivo de detectar níveis de estresse e relaxamento no texto.

O TS utiliza uma abordagem léxica baseada em listas de termos associados à estresse e relaxamento. Não apenas sinônimos de estresse, ansiedade e frustração são considerados

na avaliação do estresse, mas também termos relacionados a raiva e emoções negativas, uma vez que o estresse pode ser uma resposta a eventos negativos e podem causar emoções negativas. A Figura 9 ilustra o método de classificação de estresse e relaxamento por meio do TS. Para mensurar o relaxamento como estado oposto ao estresse o algoritmo utiliza uma lista de termos descritivos ou associados a situações ou estados de relaxamento (THELWALL, 2017).

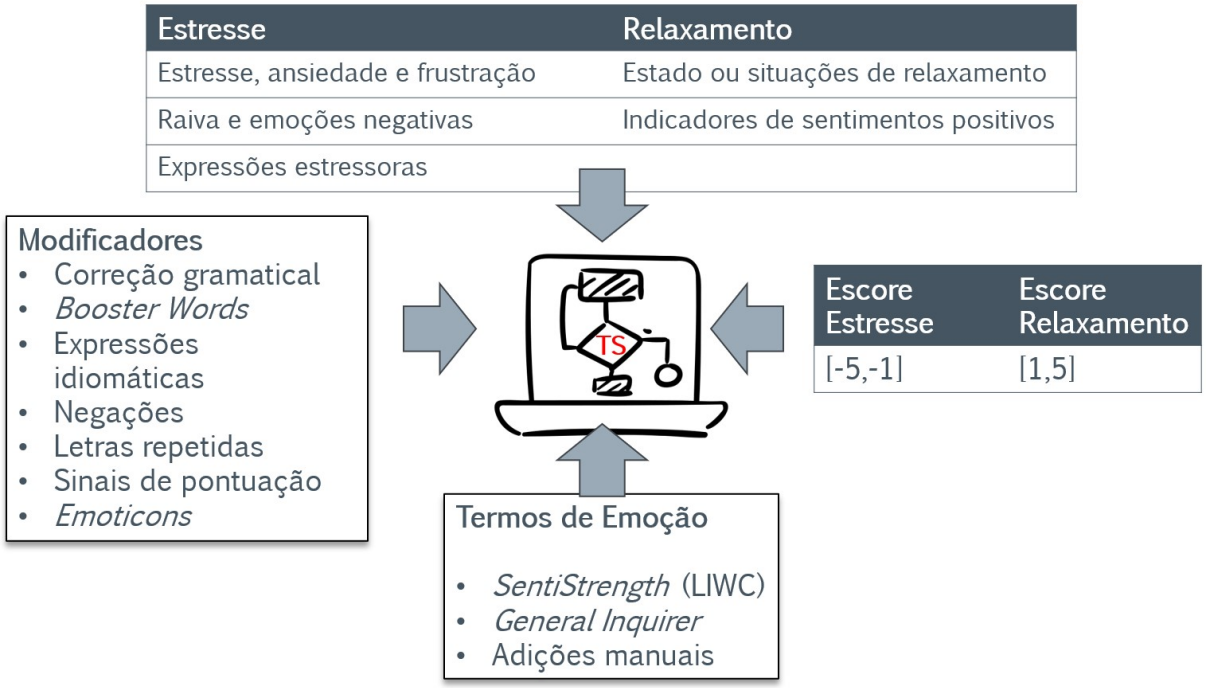

Figura 9 – Ilustração TensiStrength
Autoria própria

Os dicionário em inglês utilizado pelo TS foi obtido por meio do *SentiStrength* e de adições manuais de termos de estresse, indicadores de estressores e situações estressantes obtidos por meio de fontes acadêmicas e não acadêmicas que descrevem estresse em geral ou estressores associados a viagens. O dicionário de termos associados ao relaxamento pelo TS também é utilizado para detectar expressões indiretas de estresse por meio de negações como em "*não relaxado*". Cada termo em ambas as listas de estresse e relaxamento possui um escore que varia de 1 (nenhum) a 5 (muito forte), com sinal negativo para as pontuações negativas para os termos de estresse. As pontuações foram inseridas inicialmente através de uma abordagem não supervisionada e refinadas posteriormente em uma abordagem supervisionada (THELWALL, 2017).

Na classificação do estresse e relaxamento, o TS assinala para cada sentença as pontuações do termo com maior nível de estresse e do termo com maior nível de relaxamento identificados. No caso de textos com múltiplas sentenças são assinaladas as maiores pon-

tuações das sentenças constituintes. Os seguintes modificadores são considerados pelo algoritmo:

- ❏ Correções gramaticais: eliminação de letras repetidas em palavras reconhecidas.
- ❏ Palavras de reforço: palavras que reforçam ou enfraquecem as palavras subsequentes.
- ❏ Expressões idiomáticas: utiliza-se a pontuação das palavras equivalentes.
- ❏ Negação de palavras de relaxamento: consideradas termos de estresse.
- ❏ Ao menos duas letras repetidas: adiciona-se 1 ponto aos termos de estresse ou relaxamento identificados. Nesse caso, *preocupadooo* é considerado com maior carga de estresse do que em *preocupado*.
- ❏ *Emotions*: são interpretados como neutros ou indicativos de estresse ou relaxamento, de acordo com o sentimento correspondente.
- ❏ Pontos de exclamação: acrescenta-se 1 ponto à pontuação de estresse ou relaxamento assinalada.
- ❏ Sinais de pontuação repetidos: no caso de sinais de exclamação repetidos a pontuação assinalada recebe o acréscimo de 1 ponto.

Para identificar o nível de cada termo, bem como termos faltantes e refinar a pontuação de cada termo de sentimento os autores utilizaram uma coleção de 3 mil *tweets* classificados manualmente em termos de estresse e relaxamento na escala de 1 a 5 de pontuação. Os *tweets* foram coletados por meio de um conjunto de palavras-chave associadas a estresse e relaxamento. O sistema de avaliação do TS se dá como nos exemplos a seguir:

- ❏ *Quase em casa e o trem está atrasado*: marcação de 1 ponto para relaxamento (pontuação mínima por não haver nenhuma evidência de relaxamento) e -3 para estresse (a pontuação de estresse identificada no dicionário para *atrasado*).
- ❏ *Dormi e bagunçou o meu cabelo*: o TS assinala uma pontuação de 4 para relaxamento por conta do termo dormir e -1 para estresse (nenhum estresse identificado).
- ❏ *Nunca confie em um homem com uma cozinha imunda*: 1 ponto de relaxamento e -2 de estresse, pois o termo confiar é considerado moderado de relaxamento no dicionário (pontuação 2), mas é negado pelo termo *Nunca*, resultado em uma pontuação de -2 para estresse.

Durante a fase supervisionada o sistema que avalia a força dos termos utiliza uma abordagem de escalada ao avaliar se a alteração da pontuação assinalada em um termo pode alterar a acurácia geral no conjunto de desenvolvimento. Nesse processo o TS

seleciona aleatoriamente um termo do dicionário, incrementa a sua pontuação e aceita a alteração somente quando a soma dos escores positivos e negativos aumenta em pelo menos 2 pontos. No caso de alterações rejeitadas o processo se repete ao remover 1 ponto do peso do termo. O processo é repetido para cada termo do dicionário de forma aleatória. Após a checagem de cada termo, se nenhuma alteração tiver ocorrido, o processo é encerrado.

2.9 Formulários para a Avaliação de Escalas e Testes

A adoção de ferramentas de medição, como escalas e testes que possuam validação tanto em nível nacional como internacional, é incentivada devido à sua capacidade de possibilitar a coleta sistemática de dados e a avaliação quantitativa de fenômenos. Além disso, essa abordagem facilita a análise da correlação entre variáveis por meio de técnicas estatísticas (FEITOSA et al., 2014). Diversas escalas para a avaliação de fenômenos em seres humanos estão disponíveis e validadas na literatura, como as descritas a seguir.

2.9.1 Escala de Estresse Percebido

Luft et al. (2007) traduziram e testaram a escala, proposta inicialmente por (COHEN; KAMARCK; MERMELSTEIN, 1983), na versão completa e reduzida com 14 e 10 questões, respectivamente. A tradução realizada pelos autores mostrou-se clara e confiável para mensurar o estresse percebido em idosos brasileiros, apresentando qualidades psicométricas adequadas.

Essa escala possui 14 questões com opções de resposta que variam de zero a quatro (0=nunca; 1=quase nunca; 2=às vezes; 3=quase sempre 4=sempre). Das 14 questões, metade têm conotação positiva e a outra metade têm conotação negativa. As questões com conotação positiva são somadas de forma invertida (0=4, 1=3, 2=2, 3=1 e 4=0) e as questões negativas são somadas de forma direta. São de conotação positiva as questões 4, 5, 6, 7, 9, 10 e 13. O valor total da escala de estresse percebido é a soma das 14 questões, que pode variar de 0 a 56. Não existe um valor limite para a detecção do estresse percebido nessa escala, quanto maior a pontuação obtida, maior o estresse percebido no indivíduo.

As questões incluem a autoavaliação do estado de saúde, percepção da situação econômica, da memória, satisfação com a vida e acontecimento de eventos negativos. Essas perguntas foram empregadas para analisar as médias de estresse percebido em relação às variáveis mencionadas na literatura, o que proporcionou uma visão sobre a utilização da escala em pesquisas diversas (LUFT et al., 2007). A Figura 10 mostra as questões que compõe a escala.

Neste último mês, com que freqüencia...						
1	Você tem ficado triste por causa de algo que aconteceu inesperadamente?	0	1	2	3	4
2	Você tem se sentido incapaz de controlar as coisas importantes em sua vida?	0	1	2	3	4
3	Você tem se sentido nervoso e "estressado"?	0	1	2	3	4
4	Você tem tratado com sucesso dos problemas difíceis da vida?	0	1	2	3	4
5	Você tem sentido que está lidando bem as mudanças importantes que estão ocorrendo em sua vida?	0	1	2	3	4
6	Você tem se sentido confiante na sua habilidade de resolver problemas pessoais?	0	1	2	3	4
7	Você tem sentido que as coisas estão acontecendo de acordo com a sua vontade?	0	1	2	3	4
8	Você tem achado que não conseguiria lidar com todas as coisas que você tem que fazer?	0	1	2	3	4
9	Você tem conseguido controlar as irritações em sua vida?	0	1	2	3	4
10	Você tem sentido que as coisas estão sob o seu controle?	0	1	2	3	4
11	Você tem ficado irritado porque as coisas que acontecem estão fora do seu controle?	0	1	2	3	4
12	Você tem se encontrado pensando sobre as coisas que deve fazer?	0	1	2	3	4
13	Você tem conseguido controlar a maneira como gasta seu tempo?	0	1	2	3	4
14	Você tem sentido que as dificuldades se acumulam a ponto de você acreditar que não pode superá-las?	0	1	2	3	4

Figura 10 – Formulário: Estresse Percebido.
Autoria: Luft et al. (2007)

2.9.2 Escala de Resiliência para Adultos

Carvalho, Teodoro e Borges (2014) explicam que esta escala tem demonstrado potencial para explicar e intervir em fenômenos psicossociais relacionados ao trabalho e investigaram algumas propriedades do instrumento para servidores públicos. Os resultados da investigação mostraram que a escala tem propriedades de validade e fidedignidade de aplicação, o que a torna um instrumento útil para subsidiar decisões de gestão organizacional.

O instrumento é composto por 33 itens em que os participantes respondem cada item em uma escala de 7 pontos em formato de diferencial semântico. Cada item é organizado como um continuum, cujos opostos apresentam alternativas de resposta com conteúdo positivo e negativo. 6 fatores teóricos compõem a escala: percepção de si mesmo (6 questões), futuro planejado (4 questões), competência social (6 questões), estilo estruturado (4 questões), coesão familiar (6 questões) e recursos sociais (6 questões). Cada fator é composto por questões com escala direta e inversa. Assim como na escala de estresse percebido, não existe um valor limite, quanto maior a pontuação obtida, maior é o escore de resiliência do indivíduo. As Figuras 11 e 12 exibem as questões que compõem a escala.

Escala de Resiliência para Adultos (RSA)			
Instruções: Por favor, leia cuidadosamente as afirmações abaixo e indique o quanto você geralmente, ou no último mês, tem sentido e pensado em relação a você mesmo e em relação a pessoas que são importantes para você. Coloque um X no espaço correspondente que melhor descreve como você se sente.			
1. Quando algo imprevisto acontece	eu geralmente me sinto desnorteado	☐☐☐☐☐☐☐	eu sempre encontro uma solução
2. Os meus planos para o futuro são	difíceis de concretizar	☐☐☐☐☐☐☐	concretizáveis
3. Eu gosto de estar	com outras pessoas	☐☐☐☐☐☐☐	sozinho
4. Na minha família, a concepção do que é importante na vida é	bastante diferente	☐☐☐☐☐☐☐	a mesma
5. Assuntos pessoais	eu não posso discutir com ninguém	☐☐☐☐☐☐☐	eu posso discutir com amigos e familiares
6. Eu funciono melhor quando	eu tenho um objetivo a alcançar	☐☐☐☐☐☐☐	eu vivo um dia de cada vez
7. Os meus problemas pessoais	eu sei como solucioná-los	☐☐☐☐☐☐☐	são impossíveis de solucionar
8. Eu sinto que o meu futuro	é promissor	☐☐☐☐☐☐☐	é incerto
9. Poder ser flexível em relações sociais	é algo que eu não me importo com	☐☐☐☐☐☐☐	é importante para mim
10. Eu me sinto	muito bem com a minha família	☐☐☐☐☐☐☐	não me sinto bem com a minha família
11. Aqueles que me encorajam	são amigos e familiares	☐☐☐☐☐☐☐	ninguém me encoraja
12. Quando vou fazer algo	me atiro direto nas coisas sem planejar	☐☐☐☐☐☐☐	prefiro ter um plano
13. Nos meus julgamentos e decisões	tenho frequentemente incertezas	☐☐☐☐☐☐☐	acredito firmemente
14. Os meus objetivos	eu sei como atingi-los	☐☐☐☐☐☐☐	eu estou incerto sobre como atingi-los
15. Novas amizades	tenho facilidade em me vincular	☐☐☐☐☐☐☐	tenho dificuldades em me vincular
16. A minha família caracteriza-se por	desunião	☐☐☐☐☐☐☐	boa união

Figura 11 – Formulário: Resiliência.
Autoria: Carvalho, Teodoro e Borges (2014)

Escala de Resiliência para Adultos (RSA)			
Instruções: Por favor, leia cuidadosamente as afirmações abaixo e indique o quanto você geralmente, ou no último mês, tem sentido e pensado em relação a você mesmo e em relação a pessoas que são importantes para você. Coloque um X no espaço correspondente que melhor descreve como você se sente.			
17. A solidariedade entre meus amigos	é ruim	☐☐☐☐☐☐☐☐	é boa
18. Eu tenho facilidade para	organizar o meu tempo	☐☐☐☐☐☐☐☐	perder o meu tempo
19. A crença em mim	me ajuda em períodos difíceis	☐☐☐☐☐☐☐☐	pouco me ajuda em períodos difíceis
20. Os meus objetivos para o futuro são	vagos	☐☐☐☐☐☐☐☐	bem pensados
21. Fazer contato com novas pessoas	é difícil para mim	☐☐☐☐☐☐☐☐	eu tenho facilidade
22. Em momentos difíceis	a minha família mantém uma visão positiva do futuro	☐☐☐☐☐☐☐☐	a minha família tem uma visão negativa do futuro
23. Quando algum membro da minha família entra em crise	eu fico sabendo rapidamente da situação	☐☐☐☐☐☐☐☐	eu sou um dos últimos a ficar sabendo da situação
24. Regras e rotinas fixas	faltam no meu dia-a-dia	☐☐☐☐☐☐☐☐	facilitam o meu dia-a-dia
25. Em adversidades eu tenho tendência a	ver as coisas de um jeito ruim	☐☐☐☐☐☐☐☐	ver de um modo bom para que eu possa crescer
26. Quando estou na presença de outras pessoas	tenho facilidade em rir	☐☐☐☐☐☐☐☐	não consigo rir
27. Em relação a outras pessoas, na nossa família nós	nos apoiamos pouco	☐☐☐☐☐☐☐☐	somos leais
28. Eu tenho apoio	de amigos e familiares	☐☐☐☐☐☐☐☐	não tenho apoio de ninguém
29. Acontecimentos na vida que para mim são difíceis	eu consigo lidar com eles	☐☐☐☐☐☐☐☐	eu estou em constante estado de preocupação
30. Iniciar uma conversa interessante, eu acho	difícil	☐☐☐☐☐☐☐☐	fácil
31. Na minha família nós gostamos	de fazer coisas em conjunto	☐☐☐☐☐☐☐☐	de cada um fazer algo por si próprio
32. Quando preciso	eu não tenho nunca alguém que pode me ajudar	☐☐☐☐☐☐☐☐	tenho sempre alguém que pode me ajudar
33. Os meus amigos/ familiares próximos	valorizam as minhas qualidades	☐☐☐☐☐☐☐☐	veem com maus olhos as minhas qualidades

Figura 12 – Formulário: Resiliência (continuação).
Autoria: Carvalho, Teodoro e Borges (2014)

2.9.3 Escala de Impacto do Evento

Esta é uma escala utilizada no rastreio do transtorno de estresse pós-traumático, traduzida e adaptada para o português do Brasil por Caiuby et al. (2012). Trata-se de uma escala do tipo *likert* na qual o indivíduo, com base nos 7 dias anteriores à aplicação autoavalia-se em 22 questões distribuídas em 3 subescalas (evitação, intrusão e hiper estimulação) que contemplam os critérios de avaliação de transtorno do estres- se pós-traumático.

Cada item vale de 0 a 4 pontos e o cálculo do escore de cada subescala é obtido por meio da média dos itens que compõem as subescalas evitação, intrusão e hiper estimulação, desconsiderando as questões não respondidas. O escore geral é dado pela soma do escore de cada subescala. Caso o valor da soma seja maior ou igual a 5,6 é considerado que o indivíduo está sob efeito do TEPT. As Figuras 13 e 14 mostram as questões que compõem este instrumento.

Listamos abaixo as dificuldades que as pessoas algumas vezes apresentam, após passar por eventos estressantes. Com relação às memórias do evento estressor _____, por favor, leia cada item abaixo e depois marque com um X a coluna que melhor corresponde a seu nível de estresse, nos últimos 7 dias.

	Nem um pouco	Um pouco	Moderadamente	Muito	Extremamente
1. Qualquer lembrança trazia de volta sentimentos sobre a situação	0	1	2	3	4
2. Eu tinha problemas em manter o sono	0	1	2	3	4
3. Outros acontecimentos faziam com que eu ficasse pensando sobre a situação	0	1	2	3	4
4. Eu me sentia irritável e bravo	0	1	2	3	4
5. Eu evitava ficar chateado quando pensava sobre a situação ou era lembrado dela	0	1	2	3	4
6. Eu pensava sobre a situação mesmo quando não tinha intenção de pensar	0	1	2	3	4
7. Eu sentia como se não tivesse passado pela situação ou como se não fosse real	0	1	2	3	4
8. Eu me mantive longe de coisas que pudessem relembrar a situação	0	1	2	3	4
9. Imagens sobre a situação saltavam em minha mente	0	1	2	3	4
10. Eu ficava sobressaltado e facilmente alarmado	0	1	2	3	4

Figura 13 – Formulário: Escala de Impacto de Evento.
Autoria: Caiuby et al. (2012)

11. Eu tentei não pensar sobre a situação	0	1	2	3	4
12. Eu sabia que ainda tinha muitas emoções ligadas à situação, mas as evitei	0	1	2	3	4
13. Meus sentimentos sobre a situação estavam como que entorpecidos	0	1	2	3	4
14. Eu me peguei agindo ou sentindo como se estivesse de volta à situação	0	1	2	3	4
15. Eu tive problemas para dormir	0	1	2	3	4
16. Eu tive ondas de fortes emoções relativas à situação	0	1	2	3	4
17. Eu tentei retirar a situação da minha memória	0	1	2	3	4
18. Eu tive problemas de concentração	0	1	2	3	4
19. Lembranças da situação faziam com que eu tivesse reações físicas, como suores, problemas para respirar, náuseas ou coração disparado	0	1	2	3	4
20. Eu tive sonhos sobre a situação	0	1	2	3	4
21. Eu me sentia atento ou na defensiva	0	1	2	3	4
22. Eu tentei não falar sobre a situação	0	1	2	3	4

Figura 14 – Formulário: Escala de Impacto de Evento. (Continuação)
Autoria: Caiuby et al. (2012)

Neste Capítulo foram apresentados conceitos básicos relacionados à mineração de textos. Foi discutida a importância das redes sociais como fonte de dados e todos os processos envolvidos, desde a coleta à técnicas de classificação de sentimentos. Foram apresentadas técnicas aplicadas ao pré-processamento de documentos, representação de dados, extração de tópicos, algoritmos de classificação e métricas de qualidade. Além disso, foi apresentado o algoritmo utilizado no trabalho para classificação de estresse e relaxamento no texto e formulários validados na literatura para mensurar o nível de estresse percebido e resiliência, além da ocorrência de TEPT em seres humanos. No Capítulo 3 serão apresentados trabalhos relacionados com a presente pesquisa.

Capítulo 3

Trabalhos Relacionados

A seguir serão destacados trabalhos realizados pela comunidade acadêmica relacionados aos efeitos psicológicos das doenças infecciosas sobre a população, identificação de transtornos mentais através de ferramentas adequadas e trabalhos com aplicações de PLN para a identificação automática de sentimentos presentes no texto e sintomas de transtornos.

3.1 Efeitos Psicológicos da COVID-19 sobre a População

De acordo com a Organização Mundial da Saúde (2021), a COVID-19 é uma doença causada pelo novo coronavírus SARS-CoV-2. Os primeiros casos foram notificados em 2019, após a ocorrência de uma pneumonia viral em chineses de Wuhan. Até setembro de 2023, aproximadamente 771 milhões de casos foram notificados em todo o mundo, com mais de 6.9 milhões de óbitos e mais de 13 bilhões de vacinas administradas em todo o mundo. A pandemia teve seu fim decretado pela OMS em maio de 2023.

Embora o isolamento social tenha sido fundamental para minimizar o contágio pelo vírus, os riscos de desenvolver doenças psiquiátricas aumentariam conforme o tempo de isolamento, com consequências a longo prazo incluindo alcoolismo, depressão e estresse pós-traumático (AFONSO, 2020). O autor explica que o estresse associado à pandemia podia ser explicado pelo receio de contrair a doença, bem como por outros fatores associados que aumentariam a vulnerabilidade psicológica das pessoas. Além dos reveses econômicos, outro impacto associado aos abalos psicológicos sofridos, era o luto dos que perderam seus entes. Devido aos protocolos de segurança, cerimônias eram realizadas com poucas pessoas presentes e familiares e amigos são privados do habitual consolo do luto feito em comunidade, acarretando sofrimento aos que perdem pessoas queridas.

O surto de doenças infecciosas é moldado pelas reações psicológicas da população afetada, tanto em termos de propagação da doença, quanto em termos de sofrimento

emocional e desordem social durante e após o surto. As reações psicológicas mais comuns em tempos de pandemia incluem comportamentos inadequados, resposta defensiva e angústia emocional (CULLEN; GULATI; KELLY, 2020). Com o avanço da pandemia, suspeitava-se que, entre outros impactos que careciam de atenção, ocorresse um aumento considerável dos sintomas de ansiedade e depressão entre pessoas sem problemas de saúde mental pré-existentes, com o desencadeamento de transtorno de estresse pós-traumático em uma parcela desses indivíduos (CULLEN; GULATI; KELLY, 2020).

Observações clínicas e pesquisas sugerem que em tempos de pandemia as pessoas tendem a demonstrar respostas relacionadas à estresse e ansiedade, que incluem o medo de se contaminar através de objetos e superfícies, aversão ao contato com estrangeiros, preocupações com consequências socioeconômicas, checagens compulsivas dos sinais vitais, pesadelos e pensamentos intrusivos (BROWN et al., 2020; TAYLOR et al., 2020).

Um estudo realizado na Colômbia avaliou a prevalência de fatores relacionados ao estresse percebido associado à COVID-19 em adultos. Os participantes responderam uma versão da Escala de Estresse Percebido modificada para a COVID-19. Chegou-se à conclusão de que o grupo estudado demonstrou altos níveis de prevalência de estresse associado à pandemia (PEDROZO-PUPO; PEDROZO-CORTÉS; CAMPO-ARIAS, 2020).

Shevlin et al. (2020), investigaram a prevalência de sintomas de trauma, depressão, ansiedade e ansiedade generalizada durante o período inicial da pandemia no Reino Unido, além de estimarem a associação de variáveis que podem influenciar esses sintomas. Os dados avaliados da amostra de 2025 participantes adultos estratificada por idade, gênero e renda forneceu evidências de que ocorreu um aumento na incidência de problemas mentais sobre a população estudada nos estágios iniciais da pandemia.

Em outro estudo, Rehman et al. (2021) avaliaram o estresse sofrido pela população indiana durante o *lockdown*. Ao todo, 403 participantes responderam um questionário que visava mensurar sintomas de estresse e depressão associados com a renda familiar dos respondentes. Os resultados indicaram que as pessoas com menos condições de se sustentarem durante o *lockdown* foram as mais afetadas, de forma que a renda familiar demonstrou uma correlação negativa com a incidência dos problemas mentais avaliados. Entre as pessoas estudadas, estudantes e profissionais de saúde foram os que mais demonstraram sintomas de ansiedade, estresse e depressão.

O distanciamento social, necessário durante o enfrentamento da pademia, é um fator provável de causar sentimentos de alienação e outros sentimentos negativos correlacionados com a incidência dos Transtorno de Estresse Pós-traumático (TEPT). Com base nessa hipótese, Zhu et al. (2021) conduziram um estudo de corte transversal com 7145 participantes através de questionários online. Como resultado, verificou-se uma incidência moderada de transtornos mentais entre os respondentes em decorrência do isolamento social durante a COVID-19, além do suporte para a hipótese de que as emoções negativas e sentimetos de alienação atuam como preditores para sintomas de TEPT, com efeitos

diretos e indiretos desses sintomas moderados por diferentes níveis de ansiedade.

Castelli et al. (2020) avaliaram a prevalência de sintomas de TEPT na população italiana em associação com fatores socioeconômicos, aspectos relacionados com a COVID-19, qualidade de vida e saúde pública, além da ocorrência de sintomas de ansiedade e depressão, comumente observados em pessoas expostas a fatores altamente estressores. Dos 1321 participantes estudados, 12% disseram ter tido contato com pessoas contaminadas pela COVID-19 e 16% disseram ter conhecido alguém que faleceu, vítima da infecção causada pelo vírus. 20% dos participantes apresentaram sintomas de TEPT, experenciando a COVID-19 com traumas psicológicos. Esse estudo evidencia os impactos negativos da pandemia sobre a saúde mental das pessoas estudadas e possíveis efeitos do medo da infecção e das medidas de isolamento adotadas.

Campos et al. (2020) realizaram um estudo com pessoas adultas com o objetivo de avaliar a saúde mental da população brasileira durante a pandemia da COVID-19 em associação com características demográficas e sanitárias. Verificou-se uma alta prevalência de depressão, ansiedade, estresse e impacto psicológico causado pelo distanciamento social. Os indivíduos mais afetados foram os jovens, com sentimentos declarados de insegurança, problemas de saúde anterior à pandemia, percepção de mudança do estado mental devido ao contexto e excessivamente expostos a notícias. Mulheres de baixa renda demonstraram menor probabilidade de desenvolver sintomas depressivos, enquanto níveis inferiores de escolaridade aumentariam as chances da ocorrência desses sintomas.

Um outro efeito adverso de crises como a pandemia da COVID-19, é o desenvolvimento da resiliência, entendida como a capacidade dos indivíduos de se transformarem no enfrentamento e superação de adversidades (PINHEIRO, 2004). Embora os efeitos da COVID-19 ainda sejam pouco conhecidos, evidências históricas sugerem que a resiliência a longo prazo será uma consequência comum entre os indivíduos. Diversos fatores podem afetar a resiliência dos indivíduos durante a pandemia, tais como perda de renda ou a perda de entes queridos, com diferentes consequências, entre elas momentos de tristeza e ocorrência de outros sentimentos negativos, que não devem ser confundidos com a falta de resiliência. A resiliência pode ser aprendida, adquirida ou cultivada em comunidade, além de se constituir um capital social, conferindo um senso de propósito e significado adaptativo associado à sobrevivência de uma crise (PECONGA et al., 2020).

Com o objetivo de identificar variáveis preditoras para a resiliência durante a pandemia, Ferreira, Buttell e Cannon (2020) pesquisaram 374 adultos avaliando possíveis associações entre variáveis demográficas, medidas de resiliência e estresse percebido causado pela COVID-19. Os fatores idade e educação demonstraram associação positiva com a resiliência dos indivíduos, enquanto o inglês como segundo idioma, a necessidade do suporte de familiares e vizinhos, tempo de permanência em *lockdown* e o nível de estresse percebido apresentaram uma associação negativa. Killgore et al. (2020) avaliaram a resiliência psicológica de trabalhadores da saúde e concluiram que a melhora desse aspecto

entre esses indivíduos está associada à qualidade do sono, emoções positivas e níveis de satisfação com a vida.

3.2 Pesquisas Afins: PLN, redes sociais e análise de sentimento

Devido ao potencial das redes sociais na propagação da informação, especialmente em contextos de emergência, Freitas, Borges e Carvalho (2020) consideram que essas mídias podem atuar como importante recurso de gerenciamento de crises, auxiliando as autoridades na tomada de decisão e aumentando a consciência situacional da população. Os autores propuseram um sistema designado para selecionar, classificar e priorizar, através de parâmetros, mensagem com conteúdo relevante para o contexto de emergência. O sistema proposto poderia auxiliar as autoridades a gerar ações de resgate e contribuir com soluções para o contexto de emergência. O sistema proposto foi testado a partir de publicações do *X* (*tweets*) relacionadas a um terremoto na Cidade do México e um incêndio na Califórnia, possibilitando a seleção e priorização de aproximadamente 2% dos *tweets* com conteúdo relevante para a ação das operações de resgaste.

No decorrer da pandemia do coronavírus, as pessoas utilizavam as redes sociais para a troca de diversos tipos de informação, de forma que, identificar o conteúdo postado relacionado ao contexto, ou a informação situacional, pode ajudar as autoridades a responder de forma adequada à crise que estava em curso. Com base nisso, Li et al. (2020) investigaram publicações realizadas na rede social chinesa *Sina Weibo*, identificaram e categorizaram o conteúdo relacionado à pandemia em 7 tipos de informação situacional. Entre os resultados observados verificou-se que: 1) publicações com conteúdo sobre medidas de segurança e notificações, doações e pedidos de ajuda foram realizadas em geral por autoridades e são mais facilmente propagadas de acordo com o uso de *hashtags*; 2) postagens realizadas por usuários não verificados demonstraram um grande alcance, de forma que as autoridades precisam verificar a credibilidade desses usuários; 3) publicações com lançamentos de dúvidas e críticas são mais compartilhadas nos casos em que o autor possui muitos seguidores, reside em cidades mais desenvolvidas ou utiliza palavras menos negativas; 4) para todos os tipos de informação, exceto as que combatiam rumores, quanto maior a quantidade de palavras utilizadas, maior o volume de respostagens; 5) as postagens que combatiam rumores foram mais propagadas se o usuário possuía muitos seguidores e estava localizado em cidades mais desenvolvidas.

Imran et al. (2020) realizaram um estudo com objetivo de analisar a reação da população de diferentes culturas durante a crise do COVID-19, bem como o sentimento sobre as ações realizadas pelas autoridades locais. Com base nas publicações feitas no *X* pelas 3 duplas de países estudados, verificou-se uma correlação entre as polaridades das publicações dos Estados Unidos e Canadá, bem como Paquistão e Índia. No caso da Noruega

e Suécia, essa mesma correlação não ocorreu. Na Noruega houve uma diminuição no volume de *tweets* com sentimentos positivos muito antes que na Suécia. Os autores chamam a atenção para o fato dessa diminuição de postagens positivas em paralelo à medidas de restrição adotas pelos governantes noruegueses. Os autores utilizaram técnicas de *deep learning* para realizar a classificação dos sentimentos e emoções presentes nas publicações. Foram coletados dados do *X* e bases de dados públicas para o desenvolvimento das técnicas de análise de sentimento baseadas em PLN. No *framework* proposto, *emoticons* foram extraídos e utilizados como forma de validar a classificação de sentimento dos *tweets*.

Ebeling et al. (2020) propuseram um *framework* para analisar como a polarização política afeta o comportamento dos grupos com opiniões divergentes no Brasil, de acordo com publicações realizadas no *X*: de um lado pessoas pró isolamento social, denominadas "Quarenteners"e de outro as pessoas que apoiavam as decisões do governo, denominadas "Cloroquiners". O *framework* proposto constitui-se de técnicas para inferir de forma automática a orientação política dos indivíduos, modelagem de tópicos, visando averiguar a homogeneidade do posicionamento de cada grupo, análise de rede e detecção de comunidades com foco na análise de comportamento nos grupos avaliados e análise de características linguísticas para a identificação de aspectos psicológicos dos usuários. Entre outros pontos, concluiu-se que os indivíduos denominados Cloroquiners demonstraram uma inclinação política de direita, além de formarem um grupo mais fechado e conecto. Já os Quarenteners demonstraram uma inclinação política de esquerda com um engajamento político mais diverso. Não houve diferenças significativas em termos de questões cognitivas e emoções negativas.

De acordo com Brum et al. (2020), o *X* se constitui uma das principais fontes de informação e discussão durante a pandemia da COVID-19. Os autores coletaram um conjunto de mais de 56 milhões de *tweets* escritos em língua portuguesa, publicados durante um período de 70 dias. A análise das postagens mostrou um aumento das publicações no mesmo período em que o cenário político brasileiro repercutia a segunda troca do então ministro da saúde. Os tópicos mais frequentes ("Quarentena", "Hidroxicloroquina", "Aglomeração" e "Distanciamento") refletem o debate dos usuários ativos em relação às medidas adotadas pelo governo para o enfrentamento da pandemia.

Analisar o sentimento da população durante a pandemia possibilita informações, *insights* e auxilia a tomada de decisão por parte das autoridades públicas. Wang et al. (2020) analisaram publicações realizadas no *Sina Weibo*. Os autores utilizaram o modelo de Representações de Codificador Bidirecional de Transformadores (BERT) para classificar o sentimento presente nas publicações (positivo, negativo ou neutro), além do modelo de Frequência do Termo–Inverso da Frequência nos Documentos (TF-IDF) para sumarizar os tópicos das publicações. Quatro tópicos principais foram identificados nas postagens: a origem do vírus, sintomas da doença, atividades de produção e controle da saúde pública.

Praveen, Ittamalla e Deepak (2021) utilizaram técnicas de aprendizado de máquina e

PLN para entender o sentimento presente em discussões online da população indiana sobre estresse, ansiedade e traumas decorrentes da COVID-19 e suas causas. Foram coletados e analisados 840.000 *tweets* nesse estudo. Observou-se que os *tweets* apresentaram um sentimento neutro em sua maioria. As mortes e o *lockdown* durante a pandemia foram os tópicos mais associados ao estresse, ansiedade e traumas na população indiana. Os pesquisadores destacam a importância de entidades e profissionais de saúde entende a perspectiva dos cidadãos para a causa de traumas e doenças mentais e formular ações para o tratamento adequado dessas pessoas, o que pode ser facilitado a partir de estudos e aplicações de técnicas de PLN.

Melo e Figueiredo (2020) apresentaram o primeiro conjunto de dados público a partir da coleta de 3.925.366 de *tweets* e 18.413 notícias sobre a pandemia publicadas no portal de notícias UOL. Os *tweets* foram categorizados de acordo com a presença de *hashtags*, mídias e *retweets*. Os pesquisadores ressaltam a importância da base de dados por ser a primeira construída a partir de duas fontes distintas e importantes com discussões online sobre a crise do coronavírus. O conjunto de dados pode contribuir na condução de estudos comparativos sobre a percepção da pandemia além de possibilitar que instituições acadêmicas, agências de saúde, comunidades científicas, pesquisadores e estudantes utilizem os dados para avaliar o sentimento público e os efeitos da pandemia no Brasil.

Souza, Nobre e Becker (2020) propuseram um conjunto de modelos empilhados para a detecção automática de sintomas de ansiedade, depressão e suas comorbidades a partir de um conjunto de dados do *Reddit* com respostas dos usuários para Diagnóstico de Saúde Mental Autodeclarado. O conjunto de modelos se constitui de classificadores fracos para a distinção de usuários controle e diagnosticados e classificadores multiclasse. Os autores identificaram uma forte influência da ansiedade nas decisões tomadas pelos modelos. Porém, devido ao fato de existirem características de transtorno de ansiedade em todos os usuários avaliados, houve limitações na distinção dos usuários diagnosticados, sugerindo a consideração de padrões mais sutis na detecção dos níveis de ansiedade para uma correta identificação dos transtornos ou suas comorbidades.

Santos e Pereira (2023) analisaram um conjunto de dados composto por mensagens divulgadas por autoridades durante a pandemia da COVID-19, como a Agência Nacional de Vigilância Sanitária, o Ministério da Saúde, a Organização Mundial da Saúde e a Sociedade Brasileira de Doenças Infecciosas. Os autores descreveram as características de comunicação dessas entidades nas redes sociais durante a pandemia, considerando elementos sociais, semânticos e temporais.

Sousa e Becker (2022) realizaram uma análise temporal das atitudes a favor e contra a vacinação da COVID-19 nos Estados Unidos da América (EUA), estabelecendo um paralelo com o Brasil com base em dados do X. Os pesquisadores chegaram à conclusão de que o movimento antivacinação nos Estados Unidos tem uma presença mais predominante em comparação com o Brasil. Segundo os autores, apesar de haver semelhanças entre os

dois países, a população dos Estados Unidos é mais direta e enfática ao defender suas posições. No Brasil, notou-se uma urgência da população para se vacinar e um viés político.

Diversas estratégias para identificar notícias falsas divulgadas na internet se baseiam na análise de informações coletadas após sua disseminação, conforme explicam Couto et al. (2022). No entanto, os autores apresentaram uma metodologia que visa detectar essas notícias em estágios iniciais de propagação. Os pesquisadores conduziram uma análise exploratória na qual treinaram milhares de modelos utilizando conjuntos variados de parâmetros e atributos textuais extraídos de notícias suspeitas. Durante esse processo, eles criaram uma base de dados inédita contendo notícias falsas relacionadas à Covid-19 que foram propagadas no Brasil. Os resultados da pesquisa revelaram quais conjuntos de atributos são mais relevantes e a eficácia dos classificadores supervisionados para abordar esse problema específico no contexto brasileiro.

Aguiar et al. (2022) argumentam que lidar com o processamento de linguagem natural em idiomas pouco explorados na literatura apresenta desafios significativos. A escassez de conjuntos de dados extensos afeta o estudo de vários algoritmos nessa área. A fim de contornar essa questão, utilizou-se a técnica de aumento de dados para aumentar a disponibilidade de dados de treinamento para a tarefa de seleção de respostas por *chatbots*. Os pesquisadores fizeram a tradução automática de um conjunto de dados em inglês para o português brasileiro e o utilizaram para treinar uma rede neural profunda. Os resultados indicaram que a estratégia de treinamento combinando o conjunto de dados traduzido seguido por um ajuste fino com o conjunto de dados específico do contexto resultou nos melhores resultados de recuperação para todos os modelos analisados. Além disso, os autores disponibilizaram o conjunto de dados traduzido para acesso.

Criada em abril de 2021, a Comissão Parlamentar de Inquérito (CPI) da COVID-19 teve como propósito investigar possíveis omissões e irregularidades do governo federal durante a pandemia, a qual resultou em um trágico saldo de mais de 660 mil vidas perdidas no Brasil, posicionando o país entre os que mais sofreram com óbitos relacionados à COVID-19 (SANTOS; MARINHO; CAMPELO, 2022). O objetivo do estudo conduzido por Santos, Marinho e Campelo (2022) foi analisar o debate público acerca da CPI da COVID no X, identificando grupos, examinando suas características e interações, e investigando sinais de polarização política nessa plataforma de mídia social. Para alcançar esse objetivo, os autores coletaram 3.397.933 tweets ao longo de um período de 26 semanas, analisados em quatro redes distintas, levando em consideração diferentes tipos de interação dos usuários. Foi aplicado um pré-processamento de linguagem natural para detectar características de grupo e identificar possíveis conteúdos tóxicos. Como resultado, foi possível identificar a existência de três grupos de usuários com base na análise do uso de hashtags e na detecção automatizada de comunidades.

No estudo realizado por Silva, Souza e Oliveira (2022) foi criado o SEMcTrA, um

sistema especialista multicamadas de teleconsulta que visa a classificação, triagem e alocação inteligente de casos relacionados à COVID-19. O sistema é composto por módulos de comunicação assíncrona e síncrona, que validam e facilitam o encaminhamento médico, além de permitir a coleta de dados precisos sobre a progressão da doença. Como resultado, a implementação do SEMcTrA proporcionou suporte à tomada de decisão médica, fornecendo informações de maneira eficiente e rápida para lidar com os desafios da pandemia.

Paula, Oliveira e Moro (2022) examinaram as flutuações nos sentimentos gerais da população brasileira durante a pandemia, utilizando como indicador as músicas consumidas. Foi realizada uma análise do consumo de streaming musical no Brasil entre 2019 e 2021, com ênfase nas datas-chave durante a pandemia. Os resultados revelaram que, durante o período analisado, as pessoas demonstraram preferência por músicas mais animadas e positivas.

Paiva et al. (2020) levantaram m panorama geral de dados de usuários do X, no Brasil, relacionados à COVID-19. Os autores aplicaram técnicas de PLN em um conjunto pré-processado de dados públicos em português. O modelo proposto pelos autores captou comportamentos e tendências relacionados a COVID-19, como similaridades entre palavras, os unigramas e bigramas mais frequentes e hipóteses baseadas em dados estatísticos recolhidos.

O levantamento de trabalhos correlatos na literatura indicou trabalhos com foco: na detecção de publicações com conteúdo relevante em contextos de crises, na categorização de conteúdos publicados durante a pandemia, na análise da reação da população diante de medidas adotadas pelos seus respectivos governos, na análise da polarização das medidas de vacinação, na avaliação do sentimento presente em publicações com conteúdo relacionado a estresse, ansiedade e depressão, na detecção de notícias falsas e na classificação e triagem de casos de COVID-19. Não foram encontrados, no entanto, trabalhos que modelassem o estresse e a resiliência dos indivíduos por meio da aplicação de questionários e da análise de suas publicações nas redes sociais em língua portuguesa.

O Capítulo 3 apresentou trabalhos publicados com alguma relação com este trabalho. Foram elencados trabalhos com o foco em efeitos psicológicos sofridos pela população durante a pandemia da COVID-19, diferentes trabalhos que exploraram técnicas de PLN e extração de sentimento de textos publicados em redes sociais. O Capítulo 4 apresentará a proposta de metodologia deste trabalho.

Capítulo 4

Método para Analisar Estresse e Resiliência a partir de Dados de Redes Sociais

O presente capítulo apresenta em detalhes o método proposto neste trabalho para a coleta de dados, avaliação do estresse, resiliência e Transtorno de Estresse Pós-Traumático (TEPT) de voluntários da UFU em conjunto com características de publicação desses indivíduos nas redes sociais. A partir das bases de dados geradas em cada cenário foi possível caracterizar cada amostra e comprar o comportamento dos indivíduos em diferentes redes sociais.

Para cada usuário da base de dados, cada uma das postagens foi classificada de acordo com o nível de estresse e relaxamento. Este trabalho propõe o uso do TSpt para esta classificação. Como o TSpt foi criado inicialmente para textos em inglês, uma adaptação e um estudo inicial foram necessários a fim de adaptar tal ferramenta. A adaptação do TSpt e os experimentos realizados para confirmarem o seu desempenho em bases em português serão descritos na Seção 4.4.

Validada a aplicação do TSpt, realizou-se a caracterização de usuários voluntários da Universidade Federal de Uberlândia em dois estágios, durante a pandemia da COVID-19 e após a flexibilização do distanciamento social e retorno parcial das atividades na instituição. Em ambos os estágios da pandemia realizou-se a coleta de dados via formulário e foram coletadas publicações realizadas pelos voluntários no *Instagram* e *X*.

Na Figura 15 é ilustrado o método para analisar estresse e resiliência a partir de redes sociais. De maneira geral o método, que será detalhado nas sessões seguintes, pode ser assim descrito:

1. Inicialmente foram divulgadas na comunidade UFU o trabalho que seria realizado para a captação de voluntários;

2. Formulários foram aplicados para a coleta de dados dos participantes;

3. Escores de estresse percebido, resiliência e TEPT foram calculados a partir das respostas informadas;
4. Publicações das redes sociais dos voluntários foram coletadas;
5. As publicações coletadas foram classificadas pelo TSpt em termos de estresse e relaxamento;
6. Bases de dados considerando as respostas dos participantes nos formulários, estresse e relaxamento no texto e estatísticas de publicação foram geradas;
7. Análises estatísticas das informações geradas foram realizadas.

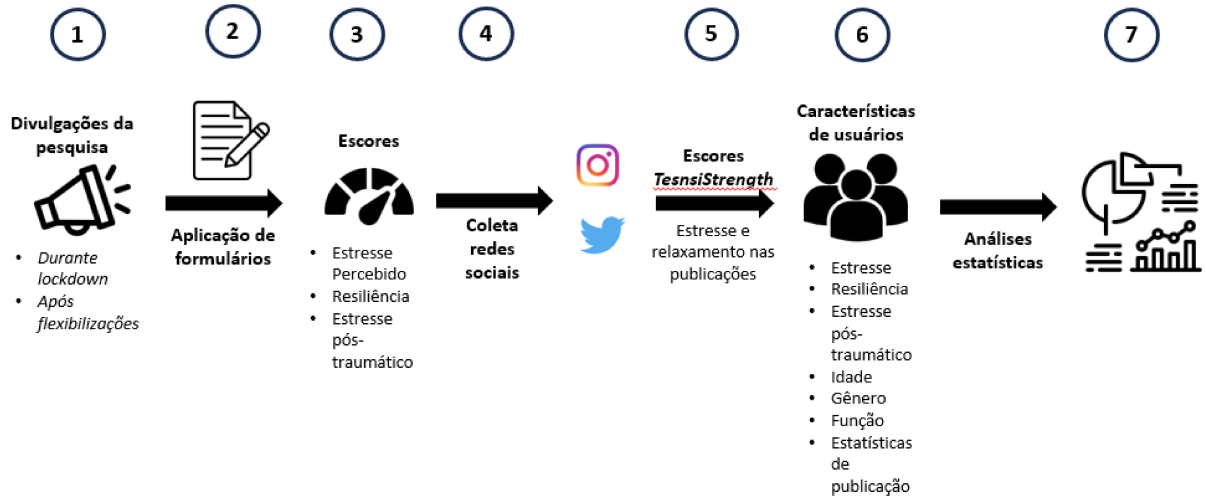

Figura 15 – Método para analisar estresse e resiliência a partir de redes sociais. Autoria própria.

A etapa de análises estatísticas é composta de duas fases. Na primeira fase de análise dos formulários objetivou-se avaliar o nível de estresse percebido e resiliência dos participantes em contraste com o estresse e relaxamento no texto, detectado por meio do TSpt, além de características dos usuários informadas via formulário e estatísticas de publicação nas redes sociais.

Na segunda fase objetivou-se avaliar o nível a ocorrência de TEPT nos participantes também em comparação com o estresse e relaxamento no texto, dados informados via formulário e estatísticas de publicação.

4.1 Coleta de Dados dos Voluntários Participantes

Para encontrar os voluntários participantes foi submetido um projeto de pesquisa ao Comitê de Ética em Pesquisas (CEP) da UFU com todo detalhamento do processo. Após

a apreciação do projeto aprovação do CEP (CAAE 36616620.9.0000.5152) realizou-se a divulgação da pesquisa e coleta de dados dos participantes. A divulgação foi realizada por meio de envio de *e-mails*, anúncio em comunidades da UFU, em redes sociais e durante as aulas com a devida autorização dos professores responsáveis.

A primeira coleta, referente ao período de isolamento social na pandemia, foi realizada entre 17/05 e 26/10/2021. Os voluntários foram convidados a preencherem um formulário criado no *Google Forms* [1]. Ao acessarem o formulário poderiam ler o Termo de Consentimento Livre e Esclarecido (TCLE) com informações da pesquisa e autorização do CEP. Ao dar o aceite poderiam prosseguir com o preenchimento das informações. O formulário constava de questões referente à idade, função (discentes, docentes ou técnicos administrativos), gênero e questões com o objetivo de avaliar nos participantes os níveis de estresse percebido por meio da versão brasileira da escala de estresse percebido (LUFT et al., 2007) e de resiliência por meio da escala de resiliência para adultos (CARVALHO; TEODORO; BORGES, 2014).

A segunda coleta, referente à flexibilização das medidas de distanciamento social, foi realizada entre 19/10 e 01/12/2022. Da mesma forma, os voluntários foram convidados a preencherem um formulário no *Google Forms* [2], tinham acesso ao TCLE, davam o aceite e informavam a idade, função e gênero. Nesse segundo formulário as questões respondidas eram referentes à ocorrência de transtorno de estresse pós-traumático avaliado por meio da escala do impacto do evento revisada (IES-R) publicada por Caiuby et al. (2012).

Após o preenchimento dos dados em cada formulário, os voluntários informaram o usuário do *Instagram* e *X* para a coleta das publicações realizadas nas redes sociais e receberam uma cópia do TCLE via *e-mail*.

Para o *X*, na primeira fase da pesquisa foram coletadas as publicações realizadas em modo público num prazo de 2 anos anteriores à data da coleta e na segunda fase foram coletadas a publicações públicas realizadas num prazo de 30 dias anteriores à data da coleta. Para os usuários do *Instagram* todas as publicações realizadas foram coletadas nas duas fases. Nas bases finais foram considerados os voluntários com dados fornecidos no formulário e coletas de publicações com sucesso, uma vez que uma parte dos voluntários não tinha o perfil aberto ou não possuía perfil ativo em alguma ou nenhuma das redes sociais avaliadas.

4.2 Extração de Variáveis do Usuário

Cada publicação extraída de cada usuário foi classificada em termos de estresse e relaxamento, com base na classificação fornecida pelo TSpt. Em seguida, foi extraída a pontuação média de estresse e relaxamento para cada usuário em cada uma das redes soci-

[1] https://forms.gle/TWrWnyQt2txrHvMK9
[2] https://forms.gle/YXgfop4G6FW2hmwb7

ais avaliadas. Além da pontuação média de estresse e relaxamento nos textos publicados, os voluntários foram caracterizados em termos de função, gênero, idade, características de publicação (tais como total de publicações na rede, tempo médio de publicação em semanas e quantidade de seguidores) juntamente com o escore de estresse percebido e fatores de resiliência ou TEPT a depender do estágio da pesquisa.

Para extrair o conhecimento sobre o comportamento *online* dos usuários do *X* foram definidos 2 grupos de variáveis consideradas na coleta de dados:

- Um primeiro grupo de variáveis caracteriza o comportamento do usuário na rede, tais como: total de publicações, idade do perfil em anos, idade média das publicações em semanas, tamanho médio dos textos em caracteres, quantidade média de *hashtags* por publicação, total de seguidores, total de amigos (ou seguidos), total de listas (permitem personalizar, organizar e priorizar as publicações que o usuário vê na linha do tempo), total de favoritos (lista privada de publicações), moda de horário de publicação, proporção de publicações com imagens, estresse médio TSpt e relaxamento médio TSpt (variáveis de 1 a 14 da Tabela 2);
- Um segundo grupo de variáveis foi considerado com base nas respostas dadas pelos voluntários nos formulários aplicados. Em cada formulário foi questionado a faixa de idade dos participantes (16 a 20 anos, 21 a 25 anos, 26 a 30 anos e assim por diante), o gênero (masculino, feminino ou não informado) e a função exercida na UFU (aluno, professor ou técnico administrativo), conforme as variáveis 15, 16 e 17 da Tabela 2.

Assim como no caso do *X*, os usuários do *Instagram* foram classificados em termos de comportamento na rede social e por meio de características informadas via formulários:

- características dos usuários na rede: total de publicações coletadas, idade média das publicações em semanas, tamanho médio das publicações em caracteres, tamanho médio da descrição do perfil em caracteres, total de seguidos, total de perfis que seguem o voluntário (seguidores), moda do período de publicação, estresse médio TSpt e relaxamento médio TSpt (variáveis de 1 a 9 da Tabela 3). A moda do período de publicação sinaliza se a maioria das publicações dos usuários aconteceram no primeiro período entre 06:00 e 18:00 (p1) ou no segundo período entre 18:00 e 06:00 (p2);
- características informadas no formulário: faixa de idade do voluntário, gênero e função exercida na UFU (variáveis 10, 11 e 12 da Tabela 3).

Além das variáveis mencionadas anteriormente, também foram criadas variáveis relacionadas ao nível de estresse percebido (escore de estresse percebido na Tabela 4), a fatores de resiliência (escore de percepção de si mesmo, futuro planejado, coesão familiar

Tabela 2 – Características extraídas: *X - Fases 1 e 2*

Variável	Descrição
1	Total de publicações coletadas
2	Idade do perfil do usuário em anos com base no fim do ano de coleta
3	Idade média das publicações em semanas com base no fim do ano de coleta
4	Tamanho médio das publicações em caracteres
5	Média de *hashtags* por publicação
6	Tamanho da descrição do perfil em caracteres
7	Total de Seguidores
8	Total de amigos
9	Total de listas
10	Total de favoritos
11	Moda do período de publicação (06:00 - 18:00 ou 18:00 - 06:00)
12	Taxa de *tweets* com imagens
13	Média de estresse por publicação
14	Média de relaxamento por publicação
15	Faixa de idade
16	Gênero
17	Função exercida na UFU

Tabela 3 – Características extraídas: *Instagram*

Variável	Descrição
1	Total de publicações coletadas
2	Idade média das publicações em semanas com base no fim do ano de coleta
3	Tamanho médio das publicações em caracteres
4	Tamanho da descrição do perfil em caracteres
5	Total de seguidores
6	Total de contas seguidas pelo voluntário
7	Moda do período de publicação (06:00 - 18:00 ou 18:00 - 06:00)
8	Média de estresse por publicação
9	Média de relaxamento por publicação
10	Faixa de idade
11	Gênero
12	Função exercida na UFU

e recursos sociais na Tabela 4) ao escore e classificação de TEPT (escore e ocorrência de estresse pós-traumático na Tabela 4) nos usuários, com base nos formulários respondidos.

As variáveis das Tabelas 2 e 3 aparecem em conjunto com as da Tabela 4 de acordo com a fase da pesquisa e disponibilidade de publicações dos voluntários nas redes e períodos considerados. Ao todo 4 bases com características dos voluntários foram geradas: (1) características dos voluntários com publicações no X na primeira fase e escores de estresse e resiliência; (2) características dos voluntários com publicações no X na segunda fase e escores de estresse pós-traumático; (3) características dos voluntários com publicações no *Instagram* na primeira fase e escores de estresse e resiliência e (4) características dos

voluntários com publicações no *Instagram* na segunda fase e escores de estresse pós-traumático.

Tabela 4 – Características extraídas conforme a fase da pesquisa

Fase	Variável	Descrição
1	1	Escore de estresse percebido
1	2	Escore de percepção de si mesmo
1	3	Escore de futuro planejado
1	4	Escore de coesão familiar
1	5	Escore de recursos sociais
2	1	Escore de estresse pós-traumático
2	2	Ocorrência de estresse pós-traumático

As variáveis foram escolhidas por conveniência. Idade e gênero são informações comumente levantadas em pesquisas em geral que envolvem seres humanos. A função dos indivíduos foi selecionada para avaliar possíveis diferenças entre alunos, professores e técnicos administrativos. As estatísticas de publicação nas redes foram escolhidas de acordo com a disponibilidade do método de coleta de cada rede avaliada.

4.3 Análises Estatísticas

Proposto por Karl Pearson, o coeficiente de *correlação de Pearson* permite mensurar a força do grau de relacionamento, ou ainda, a direção e o grau de relação linear entre duas variáveis quantitativas (FILHO; JÚNIOR, 2009). Além do coeficiente de correlação em si é possível testar a *correlação de Pearson* de acordo com o nível de significância desejado. A correlação foi escolhida como medida de avaliação por permitir mensurar a força da interação entre as variáveis extraídas.

Foi realizada o cálculo da *correlação de Pearson* e análise da significância da correlação entre os pares de variáveis quantitativas de cada uma das 4 bases com características de usuários geradas. Em particular, observou-se as correlações entre:

- ❏ estresse percebido e média de estresse usando o TSpt;
- ❏ constructos de resiliência e média de relaxamento no TSpt;
- ❏ constructos de resiliência e estatísticas de publicações;
- ❏ estresse percebido e estatísticas de publicações;
- ❏ estresse médio no TSpt e estatísticas de publicações;
- ❏ relaxamento médio no TSpt e estatísticas de publicações.

O teste *chi-quadrado* é aplicado quando se deseja comparar as frequências experimentais de uma classe com as frequências teóricas, baseadas em alguma hipótese (TALLARIDA et al., 1987). Entre os pares de variáveis categóricas de cada tabela realizou-se o teste de *chi-quadrado* para associação, possibilitando investigar as associações entre ocorrência de transtorno de estresse pós-traumático e:

- ❏ idade;
- ❏ gênero;
- ❏ função exercida na instituição;
- ❏ moda do período de publicação.

Quando se deseja comparar a distribuição de duas amostras, na ausência da suposição de normalidade e assumindo que os dados são provenientes de uma distribuição simétrica, utiliza-se o *teste não paramétrico de Wilcoxon* como alternativa ao teste *t-Student* (BARROS; MAZUCHELI, 2005). Por se tratarem de amostras não probabilísticas, optou-se por realizar o *teste de Wilcoxon*, que possibilitou comparar as distribuições do estresse percebido, constructos de resiliência, estresse pós-traumático, estresse e relaxamento médio no texto de acordo com a faixa de idade, gênero, função e moda de período de publicação em cada base de usuários.

Foi possível também, por meio do mesmo teste, comparar os escores de estresse, resiliência, TEPT, bem como os escores médio de estresse e relaxamento no texto entre as redes para uma mesma fase da pesquisa, além da comparação do estresse e relaxamento médio no texto entre as fases da pesquisa para uma mesma rede.

Em todos os testes aplicados foi considerada uma significância de 10% para a avaliação de resultados extraídos por meio de testes de hipótese. Devido ao tamanho amostral insuficiente das bases de usuários foi necessário adotar uma significância inferior à aplicada na maioria dos trabalhos de 5% ou 1%, embora existam diversos trabalhos na literatura com a significância adotada nesse trabalho.

4.4 Validação e Aplicação do TensiStrength

O TS foi desenvolvido inicialmente para classificar o estresse e relaxamento em textos escritos na língua inglesa. Para utilizar o algoritmo fez necessário adaptá-lo para classificar textos em português. As sessões a seguir tratam do método de validação do algoritmo adaptado, TSpt. Uma aplicação do TSpt foi realizada para entender melhor o nível de estresse e relaxamento durante a pandemia da COVID-19.

Para validar e aplicar o TS na classificação do estresse e em textos em português, realizou-se o seguinte procedimento: 1) coleta de Dados do X em diferentes momentos

da pandemia do coronavírus no Brasil; 2) pré-processamento dos *tweets*; 3) aplicação do algoritmo TSpt para a classificação das publicações em termos de estresse e relaxamento (Seção 4.4.3) e 4) avaliação da proporção de estresse, geração de nuvens de palavras e extração de tópicos dos cenários avaliados. A Figura 16 ilustra o processo realizado na primeira etapa da pesquisa. O procedimento aqui descrito foi apresentado em Peres et al. (2023).

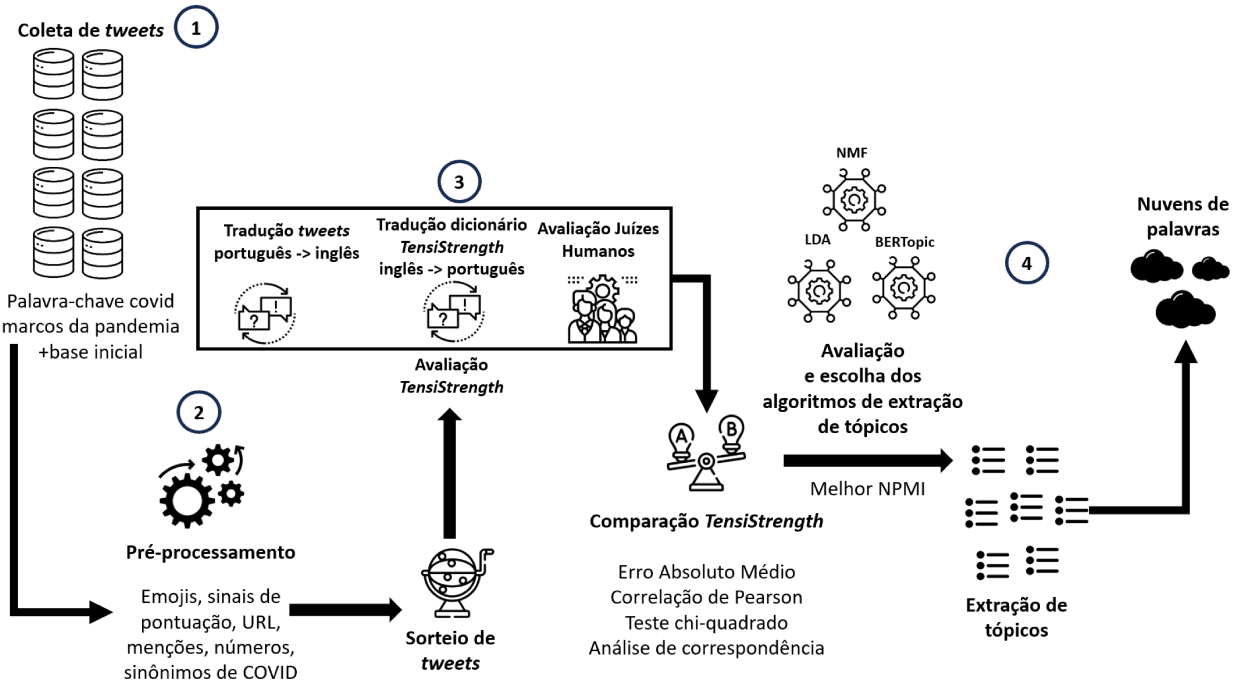

Figura 16 – Validação do TensiStrength em português.
Autoria própria

4.4.1 Coleta de Dados

Para a validação do TSpt optou-se por avaliar os *tweets* publicados ao longo de diferentes momentos da pandemia no Brasil. O *X* foi escolhido como rede social foco da pesquisa devido aos recursos gratuitos disponibilizados pela plataforma para a coleta de dados, realizada por meio da biblioteca *snscrape*[3] no ambiente *Python*. Para a busca das publicações, considerou-se as postagens identificadas por meio da palavra-chave *covid*. Além das bases de avaliação, foram coletadas, num primeiro momento da pesquisa, publicações realizadas entre abril e maio de 2021, selecionadas por meio da mesma palavra-chave. Essa base de dados, composta por 33.703 *tweets* e sujeita às mesmas etapas de pré-processamento que as bases de avaliação, foi mantida no escopo deste trabalho para testes e avaliação de

[3] https://github.com/JustAnotherArchivist/snscrape

parâmetros. Para a definição dos períodos que seriam avaliados, levou-se em consideração marcos históricos da pandemia no Brasil, identificados por meio de acervos online[4,5].

Os períodos considerados em cada coleta, a quantidade de *tweets* e de usuários coletados podem ser verificados na Tabela 5. A fim de obter um volume maior de publicações para serem avaliadas neste trabalho, realizou-se a coleta em um período de 2 meses em torno de cada evento selecionado.

Tabela 5 – Períodos considerados para a coleta de *tweets*.

Período	Descrição	*Tweets*	Usuários.
02/2020 - 03/2020	03/2020: Primeiros registros de transmissão interna do coronavírus no Brasil	26.180	15.182
06/2020 - 07/2020	07/2020: Aproximadamente $\frac{1}{3}$ das mortes registradas desde o início da pandemia ocorreram em julho de 2020	62.624	29.000
11/2020 - 12/2020	12/2020: Início da segunda onda de contágio da COVID-19 no Brasil	45.753	23.104
01/2021 - 02/2021	01/2021: Autorização do uso emergencial das vacinas CoronaVac e Oxford pela Anvisa	42.441	21.366
06/2021 - 07/2021	07/2021: O Brasil atinge a marca de 500.000 mortes causadas pelo coronavírus	33.497	17.280
12/2021 - 01/2022	12/2021: Anúncio da variante Ômicron pela OMS como uma variante de preocupação depois de descoberta na África do Sul	42.702	21.399
04/2022 - 05/2022	04/2022: Desobrigação do uso de máscaras em todos os estados brasileiros, com índice de mortes causadas pela COVID-19 $< 0,3$ mortes por 100.000 habitantes	8.094	5.569

4.4.2 Pré-processamento

Todo o pré-processamento foi realizado através de bibliotecas *Python*. Foram aplicadas as seguintes tarefas em cada uma das bases de dados: conversão de *emojis* em palavras; remoção de sinais de pontuação; conversão do texto para letras minúsculas; remoção de menções a usuários e *URLs*; remoção de caracteres numéricos; remoção de palavras sinônimas e associadas à COVID-19, tais como *morte*, *covid19*, *coronavírus* e *pandemia*,

[4] https://memoriadaeletricidade.com.br/comunicacao-memoria/117830/linha-do-tempo-covid-19
[5] https://www.sanarmed.com/linha-do-tempo-do-coronavirus-no-brasil

remoção de palavras de parada e lematização da base de dados coletada. Foram utilizadas, ainda, nuvens de palavras para avaliar e refinar cada etapa do pré-processamento aplicado.

4.4.3 TensiStrength aplicado a textos escritos em português

Para a detecção dos níveis de estresse e relaxamento nas publicações utilizou-se o algoritmo TS (Seção 2.8) fornecido por Thelwall (2017). Foram executadas, em uma amostra de *tweets*, duas versões do algoritmo: a versão original, em inglês, e uma versão adaptada para o português (TSpt). Para executar o *TensiStrength* foi necessário utilizar uma ferramenta de tradução através da biblioteca *googletrans*[6] para traduzir os *tweets* coletados e o dicionário do algoritmo nas aplicações do TS e TSpt, respectivamente. Na versão original traduziu-se a amostra de *tweets* para o inglês, enquanto na versão TSpt o dicionário do algoritmo foi traduzido para o português. O desempenho das duas versões do algoritmo foi comparado com base nas mesmas métricas escolhidas pelos autores do TS.

A aplicação do TSpt foi ainda comparada com a rotulagem do estresse e do relaxamento presentes nas sentenças a partir da avaliação de juízes humanos. O manual para a rotulação utilizado no trabalho desenvolvido por (THELWALL, 2017) foi traduzido (ver Apêndice A) e disponibilizado para que 2 voluntários, alunos de iniciação científica da instituição e disponíveis para auxiliar a pesquisa, classificassem o nível de estresse e relaxamento na amostra de 386 *tweets*. A pontuação de estresse, varia de -5 a -1, enquanto a pontuação de relaxamento varia de 1 a 5, tanto nas classificações do TS e TSpt quanto na dos juízes. Para cada *tweet* foi gerada a média dos escores de estresse e relaxamento fornecidos pelos juízes. Somadas as duas pontuações, foi possível classificar os textos com maior incidência de estresse, em caso de um resultado negativo para a soma, relaxamento, no caso de uma soma com resultado positivo, ou neutro, caso a soma seja igual a zero. Os rótulos derivados do TSpt e das avaliações dos juízes foram comparados por meio do teste *chi-quadrado de Pearson* (PLACKETT, 1983) e da *análise de correspondência*, com o auxílio do pacote *FactoMineR*[7].

Validada a versão do algoritmo para aplicação em textos na língua portuguesa, seguiu-se com a classificação das publicações em termos de estresse e relaxamento em toda a base de dados coletada. Uma vez realizada a classificação de cada publicação, observou-se a distribuição da proporção de estresse e relaxamento em cada período. Para descrever cada cenário, foram usadas nuvens de palavras e algoritmos de PLN para a extração de tópicos.

[6] https://pypi.org/project/googletrans/
[7] http://factominer.free.fr/

4.4.4 Extração de Tópicos

Para a extração de tópicos foram comparados 3 algoritmos amplamente utilizados na literatura: LDA, NMF e BERTopic (ver Seções 2.6.1, 2.6.2 e 2.6.3). Para a definição do número de tópicos utilizou-se a base de testes inicial de 30 dias extraída entre abril e maio de 2021. O algoritmo inicial, LDA, foi executado sobre essa base de dados, variando o número de tópicos entre 1 e 14, com uma coerência máxima verificada para 5 tópicos. Utilizou-se a mesma quantidade de tópicos entre os 3 algoritmos para efeito de comparação.

Para a quantidade de palavras em cada tópico, levou-se em consideração a literatura correlata, com 10 palavras ou menos por tópico (EBELING et al., 2021), (HABIBABADI; HAGHIGHI, 2019). Optou-se, assim, por utilizar 6 palavras para descrever os tópicos extraídos. Os algoritmos selecionados foram comparados considerando a coerência NPMI (*Normalized Pointwise Mutual Information*) de cada algoritmo em cada um dos 7 períodos. Para a execução do NMF e BERTopic utilizou-se as bibliotecas *Scikit-Learn*[8] e *BERTopic*[9]. A execução do LDA e a verificação da coerência dos algoritmos foi possível graças a funções disponíveis na biblioteca *Gensim*[10].

Nas seções a seguir serão apresentados os resultados da validação do TSpt (Seção 4.4.5); da aplicação do TSpt, análise do estresse e relaxamento por período e extração de tópicos (Seção 4.4.6) e conclusões obtidas nessa etapa da pesquisa (Seção 4.4.7).

4.4.5 Validação do TSpt

Para avaliar a aplicação do TSpt em relação ao algoritmo original, retirou-se uma amostra de 386 *tweets* estratificada de acordo com a quantidade de publicações presentes em cada uma das 7 bases coletadas (Tabela 5). A amostra de publicações foi traduzida para o inglês e submetida ao algoritmo TS da forma como foi fornecido pelos autores. Em paralelo foi feita a tradução do dicionário de palavras presente no código do TS para o português e o algoritmo foi executado considerando as publicações da amostra sem tradução.

As pontuações de estresse e relaxamento das duas versões do TS foram comparadas considerando o *Erro Absoluto Médio* (EAM) (CHAI; DRAXLER, 2014) e a *Correlação de Pearson* (COHEN et al., 2009). De acordo com os autores Thelwall (2017), essas duas métricas são mais adequadas por levarem em consideração o quanto um valor predito está distante de um valor de referência. As pontuações do TS, já validado na literatura, foram usadas como referência para avaliar o desempenho do TSpt. O resultado das métricas elencadas de comparação estão descritos na Tabela 6.

[8] https://scikit-learn.org/stable/
[9] https://maartengr.github.io/BERTopic/index.html
[10] https://radimrehurek.com/gensim/auto$_e$*xamples/index.html*

Tabela 6 – Comparação entre o TS e TSpt

Categoria	EAM	Correlação de *Pearson*
Estresse	0.5881	0.4950
Relaxamento	0.3005	0.6254

Verificou-se que as correlações apresentadas, significativas com 5% de significância, foram superiores às obtidas pelos autores para comparar o desempenho do TS em relação a codificadores humanos. No trabalho publicado por Thelwall (2017), as correlações para estresse e relaxamento foram de 0.465 e 0.422, respectivamente. Os autores defendem, porém, que o EAM é a métrica mais adequada por assumir que a predição e o valor de referência estão na mesma direção. Na amostra selecionada ambos os erros médios absolutos foram menores que os utilizados pelos autores para validar o TS. Os erros verificados indicam que se espera, em média, uma diferença de ± 0.5881 na pontuação do TSpt em relação ao TS em se tratando de estresse. Já as pontuações de relaxamento apresentaram um desvio ainda menor, de ± 0.3005, indicando uma concordância ainda maior entre as duas versões do TS. Considera-se, portanto, que a modificação do TSpt foi próxima ao resultado que seria verificado ao utilizar o algoritmo original.

As pontuações de estresse e relaxamento do TSpt foram também comparadas com as pontuações fornecidas por juízes humanos, treinados a partir do manual fornecido por Thelwall (2017), traduzido para o português. Após a classificação, cada publicação foi rotulada de acordo com a soma das pontuações de estresse e relaxamento. O resultado do teste de chi-quadrado de Pearson pode ser visualizado na Figura 17. Verifica-se que, de acordo com *p-valor* obtido no teste, existe uma associação significativa entre os rótulos derivados do TSpt e os rótulos derivados das médias das avaliações dos juízes, com 5% de significância. Particularmente, o resultado da análise de correspondência na Figura 18 revela que o rótulo de estresse derivado do TSpt está mais associado aos rótulos de neutralidade e estresse derivados da avaliação dos juízes. Já os rótulos de relaxamento e neutro derivados do TSpt estão mais associados ao rótulo de relaxamento derivado da anotação dos juízes. A avaliação do estresse e do relaxamento presentes nas publicações foi realizada, portanto, considerando as pontuações obtidas por meio do TSpt.

4.4.6 Análise da incidência de estresse e extração de tópicos nos *tweets*

Na Figura 19 é possível verificar que, dentre os períodos selecionados, aquele com maior ocorrência de publicações ocorreu entre junho e julho de 2021, período em que ocorreu o pico de mortes no segundo ano da pandemia. Em relação à proporção de estresse nas publicações, de acordo com a Figura 20, observou-se as maiores ocorrências, durante os períodos de fevereiro a março de 2020, de junho a julho de 2020 e de janeiro a fevereiro

LABEL_TS	*LABEL_JZ*			
	ESTRESSE_JZ	NEUTRO_JZ	RELAXAMENTO_JZ	***Total***
ESTRESSE_TS	99	23	10	132
NEUTRO_TS	121	20	32	173
RELAXAMENTO_TS	52	9	20	81
Total	272	52	62	386

χ^2=13.491 · df=4 · Cramer's V=0.132 · p=0.009

Figura 17 – Teste de chi-quadrado: TSpt x Juízes
Autoria própria.

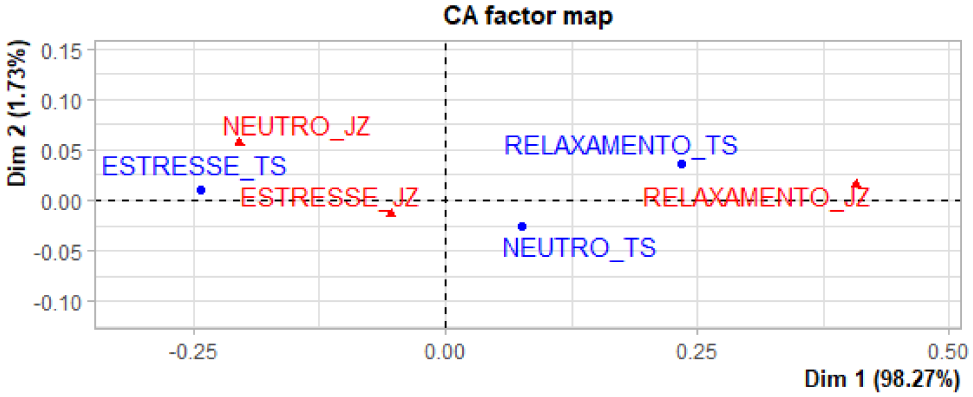

Figura 18 – Análise de correspondência: TSpt x Juízes
Autoria própria

de 2021 com 50%, 49% e 49% das publicações rotuladas como estresse, respectivamente. Nota-se uma maior proporção de publicações rotuladas como estresse no início e no pico de mortes durante o primeiro ano e no início da vacinação contra a COVID-19 no Brasil. O período final selecionado, de abril a maio de 2022, foi o que demonstrou a menor ocorrência de publicações, momento em que ocorreu a flexibilização oficial do uso de máscaras no Brasil, além das menores taxas de óbito desde o início da pandemia.

Para avaliar a extração de tópicos foram comparados os algoritmos LDA, NMF e BERTopic através da coerência NPMI (BOUMA, 2009). Essa métrica de avaliação é uma variação da métrica PMI (*Pointwise Mutual Information*), que avalia a associação entre 2 termos. A NPMI é uma medida normalizada no intervalo $[-1, 1]$, em que os resultados -1, 0 e 1 indicam nenhuma coocorrência, independência e total coocorrência entre os termos, respectivamente (CAMPAGNOLO; DUARTE; BIANCO, 2022). O resultado da coerência NPMI para os algoritmos selecionados em cada uma das bases de dados coletadas por ser verificado na Tabela 7. Devido ao maior valor médio da coerência NPMI relacionada ao NMF, optou-se por caracterizar os tópicos discutidos nos 4 períodos iniciais por meio

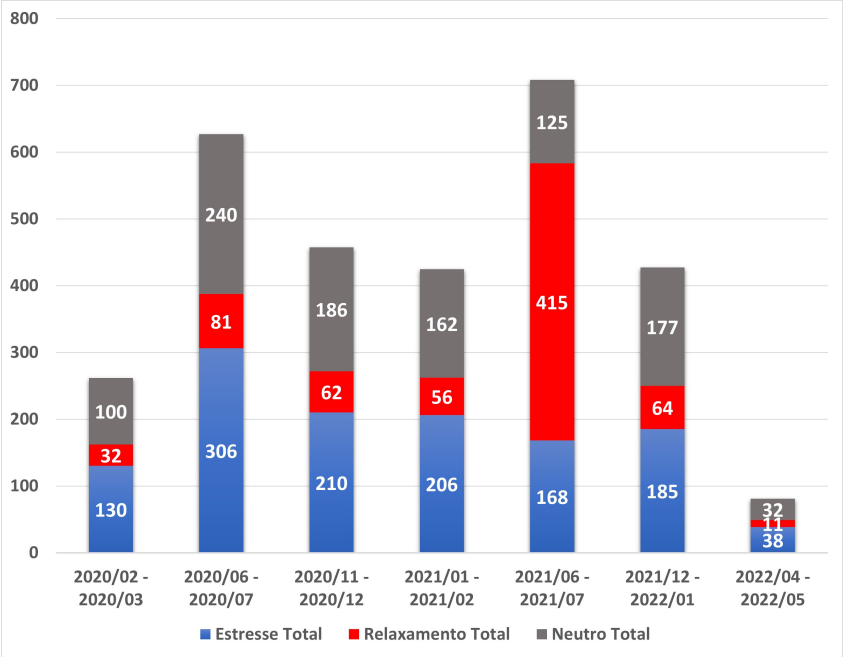

Figura 19 – Total de *tweets* (milhares) por rótulo.
Autoria própria.

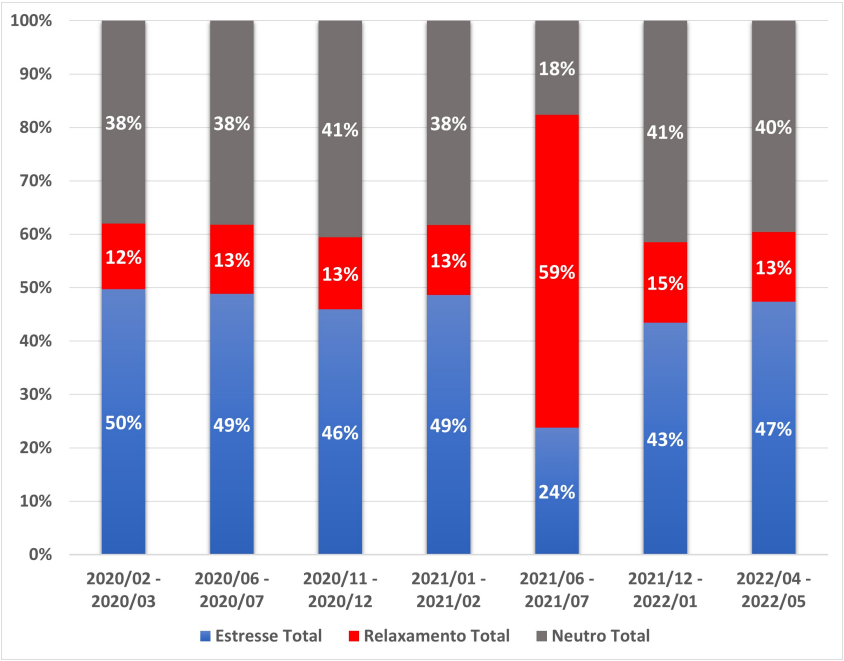

Figura 20 – Proporção de *tweets* por rótulo.
Autoria própria.

deste algoritmo. Para os 3 períodos finais foram considerados os tópicos extraídos através do BERTopic. As Tabelas 8 e 9 apresentam exemplos de tópicos e termos associados a cada conjunto de *tweets* analisado.

Tabela 7 – Coerência NPMI dos algoritmos em cada base

Base	LDA	NMF	BERTopic
02/2020 - 03/2020	-0.0443	**0.0189**	-0.0552
06/2020 - 07/2020	0.0130	**0.0623**	0.0169
11/2020 - 12/2020	0.0022	**0.0400**	0.0148
01/2021 - 02/2021	-0.0382	**0.0809**	0.0301
06/2021 - 07/2021	-0.0079	0.0734	**0.0789**
12/2021 - 01/2022	-0.0189	0.0511	**0.0523**
04/2022 - 05/2022	-0.0585	0.0020	**0.0241**
Média	-0.0218	**0.0469**	0.0231

É possível notar que, no período inicial da pandemia no Brasil, houve discussões relacionadas à deflagração do coronavírus no país e à reação do governo brasileiro frente à pandemia. No período com maior pico de casos no primeiro ano discutiu-se sobre o acesso à vacinação e tratamentos defendidos como alternativos por parte do governo brasileiro. Os 3 períodos seguintes foram marcados por discussões relacionadas ao protocolo de distanciamento social, medidas de prevenção, testagem, vacinação e reverberações no cenário político brasileiro. Os períodos de dezembro de 2021 a janeiro de 2022 e de abril a maio de 2022 foram marcados por discussões ainda relacionadas à vacinação bem como por questões relativas a doenças com sintomas similares aos da COVID-19 e à perda de familiares no decorrer da pandemia por parte da população. No período final avaliado, em particular, nota-se a discussão sobre o protocolo de vacinação em crianças.

Tabela 8 – Tópicos extraídos via NMF

Base	Tópico	Palavras
02/2020-03/2020	1	caso,confirmar,primeiro,número,suspeito,estado
02/2020-03/2020	2	bolsonaro,pegar,presidente,positivo,jair,comitiva
06/2020-07/2020	1	vacina,risos,teste,contra,achar,bom,tomar
06/2020-07/2020	2	bolsonaro,positivo,cloroquina,exame,sintoma,presidente
11/2020-12/2020	1	teste,negativo,positivo,resultado,maõsjuntas,amanhã
11/2020-12/2020	2	vacina,contra,tomar,querer,pfizer,bolsonaro
01/2021-02/2021	1	vacina,contra,tomar,dose,seringa
01/2021-02/2021	2	tratamento,precoce,existir,cloroquina,ivermectina,médico

Para a representação das publicações por meio de nuvens de palavras foram selecionados alguns dos períodos com maior ocorrência de publicações. As Figuras, de autoria própria, 21, 22 e 23 exibem as nuvens de palavras para a base de dados coletada entre junho e julho de 2020 com rótulos de estresse, relaxamento e neutro, respectivamente. A mesma sequência de nuvens de palavras pode ser visualizada nas Figuras 24, 25 e 26, para o período de novembro a dezembro de 2020, e nas Figuras 27, 28 e 29 para o período de dezembro de 2021 a janeiro de 2022.

Tabela 9 – Tópicos extraídos via BERTopic

Base	Tópico	Palavras
06/2021-07/2021	1	teste,nariz,cotonete,horrível,exame,mãosjuntas
06/2021-07/2021	2	máscara,usar,distanciamento,pegar,bolsonaro,queiroga
12/2021-01/2022	1	dose,gripe,vacina,terceiro,influenza,reforço
12/2021-01/2022	2	família,mãe,criança,pai,irmã,perder
04/2022-05/2022	1	gripe,dor,garganta,febre,sinusite,alergia
04/2022-05/2022	2	dose,vacina,criança,gripe,vacina,terceiro

Figura 21 – Estresse: 06/20-07/20

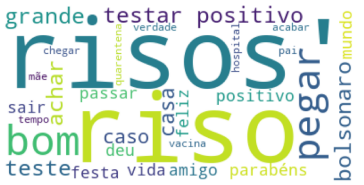

Figura 22 – Relaxamento: 06/20-07/20

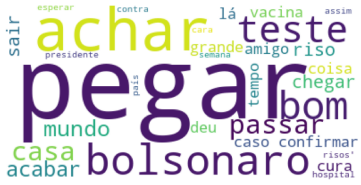

Figura 23 – Neutro: 06/20-07/20

Figura 24 – Estresse: 11/20-12/20

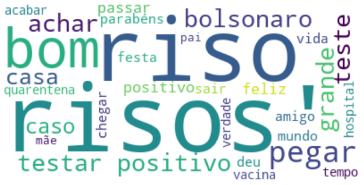

Figura 25 – Relaxamento: 11/20-12/20

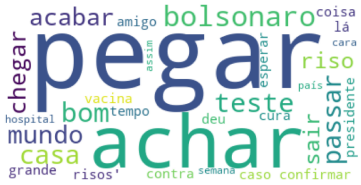

Figura 26 – Neutro: 11/20-12/20

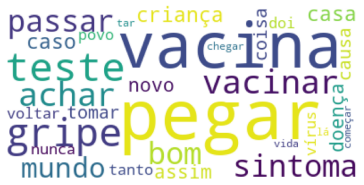

Figura 27 – Estresse: 12/21-01/22

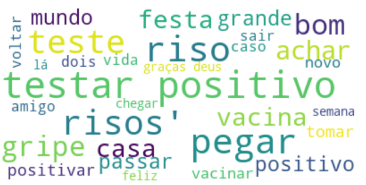

Figura 28 – Relaxamento: 12/21-01/22

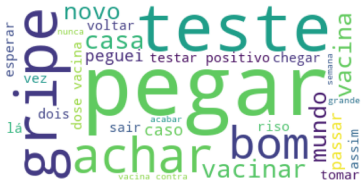

Figura 29 – Neutro: 12/21-01/22

É possível notar que, de acordo com as nuvens de palavras geradas a partir dos *tweets* rotulados como estresse, houve uma reação da população a questões no cenário político brasileiro no período com maior pico de mortes no primeiro ano da pandemia. No período de novembro a dezembro de 2021 o estresse presente nas publicações relaciona-se com o diagnóstico e à morte pelo coronavírus. O estresse nas publicações realizadas entre dezembro de 2021 e janeiro de 2022 está relacionado à vacinação, testagem e ocorrência de doenças com sintomas similares aos da COVID-19. Os *tweets* rotulados como relaxamento apresentaram conteúdo relacionado ao então presidente brasileiro, testes positivos

realizados e com a volta de eventos sociais. As publicações com o rótulo neutro, devido ao empate entre estresse e relaxamento, indicaram conteúdos relacionados a testes e ao contágio da COVID-19 e doenças similares.

4.4.7 Considerações

A partir da busca de publicações realizadas no *X*, identificadas por meio da palavra-chave *covid* e coletadas em diferentes momentos da pandemia do coronavírus, foi possível observar uma variação no volume de postagens capturadas. Dentre os períodos elencados para avaliação observou-se um pico de *tweets* entre junho e julho de 2020, momento em que $\frac{1}{3}$ dos óbitos ocorridos no primeiro ano de pandemia no Brasil foram registrados. Já o período final de coleta, de abril a maio de 2022, demonstrou o menor volume de publicações coletadas, com aproximadamente 13% do volume coletado entre junho e julho de 2020, refletindo o momento de flexibilização da pandemia no Brasil com a desobrigação do uso de máscaras e queda drástica no número de óbitos.

A comparação entre o TS, fornecido pelos autores, e o TSpt a partir do EAM e da *Correlação de Pearson* demonstrou correlações superiores e erros inferiores aos apresentados por (THELWALL, 2017) no trabalho validado na literatura. Ao utilizar a avaliação humana como referência para o TSpt, observou-se uma associação entre as classificações com predominância de estresse e com predominância de relaxamento entre o TSpt e as classificações dos juízes humanos. Os períodos com maior proporção de estresse aconteceram nos 2 primeiros meses da pandemia e durante os períodos com picos de mortes observadas no primeiro e no segundo ano de pandemia no Brasil.

Os tópicos extraídos por meio dos algoritmos NMF e BERTopic demonstraram o conteúdo publicado relacionado a: medidas de prevenção à COVID-19, reverberações no cenário político, testagem, vacinação, óbitos e ocorrência de doenças com sintomas parecidos com os apresentados em decorrência do contágio pelo coronavírus.

Como sugestão para trabalhos futuros seria importante avaliar o desempenho do TSpt na classificação do estresse presente nas sentenças em relação a algoritmos de classificação já consolidados na literatura. Além disso, uma amostra maior de publicações rotuladas por um número superior de juízes treinados para esse fim, a exemplo dos resultados apresentados por (THELWALL, 2017), poderia fornecer mais evidências do desempenho do TSpt. Uma maior volumetria de rótulos gerados por juízes humanos, poderia ainda possibilitar a comparação do TSpt em relação a diversos algoritmos, bem como a avaliação dos escores gerados, ao invés da avaliação somente da tendência de estresse e relaxamento, obtida a partir da soma dos escores.

O Capítulo 4 apresentou o detalhamento das bases de dados com as características extraídas dos voluntários em cada rede social e fase da pesquisa , além do processo utilizado para a validação do TSpt e extração de tópicos do *X* em diferentes momentos da pandemia da COVID-19. No Capítulo 5 serão apresentados os resultados das análises

realizadas a partir de dados coletados de voluntários da UFU durante a pandemia e após a flexibilização das medidas de distanciamento social.

Capítulo 5

Formulários x Redes Sociais

Este capítulo apresenta os resultados encontrados na comparação entre os dados coletados por meio de formulários e dados captados nas redes sociais dos indivíduos. A primeira fase de divulgação e coleta de dados ocorreu durante as medidas de distanciamento social, enquanto a segunda fase ocorreu após a flexibilização dessas medidas. Foram coletados dados pessoais e questões para a avaliação de estresse percebido, resiliência e TEPT através de formulários. Os dados dos formulários foram cruzados com a avaliação das publicações realizadas nas redes sociais. Uma base com características de usuários foi gerada em cada cenário.

A Seção 5.1 apresenta a volumetria de dados resultante de cada etapa de divulgação e rede social. As Seções 5.2, 5.3, 5.4 e 5.5 apresentam os resultados das avaliações dos usuários do *X* na primeira etapa, do *X* na segunda etapa, do *Instagram* na primeira etapa e do *Instagram* na segunda etapa de coleta de dados, respectivamente. Por fim, a Seção 5.7 apresenta as conclusões obtidas a partir das análises realizadas.

5.1 Volumetria de Dados Capturados

Da base inicial de formulários preenchidos em cada fase chegou-se a uma volumetria de usuários válidos, conforme pode ser verificado na Tabela 10.

Tabela 10 – Total de usuários finais em cada base

Fase	Respostas no Formulário	Usuários finais *X*	Usuários finais *Instagram*
1	109	22	9
2	94	18	23

A seguir serão detalhados os resultados encontrados em cada base avaliada.

5.2 *X*: Fase 1

As estatísticas descritivas das variáveis quantitativas referentes aos usuários com dados coletados no *X* na primeira fase da pesquisa estão apresentadas na Tabela 11. Dos 22 indivíduos dessa base verifica-se uma ocorrência de 14 usuários (64%) com predominância de publicações no período entre 18:00 e 06:00, 12 usuários (55%) com idade até 25 anos no momento da coleta, 10 usuários (45%) do gênero feminino e 16 discentes (73%).

Tabela 11 – *X* Fase 1: Estatísticas Descritivas dos Usuários (Variáveis Quantitativas)

Variável	Mín.	Q1	Mediana	Média	Q3	Máx.
Quantidade de Publicações	2.0	18,5	114,0	351,1	462,2	1.582,0
Idade do Perfil	1.0	3.3	6.8	6.7	9.5	12.8
Idade Média das Publicações	18.7	55.4	99.8	165.1	176.5	623.2
Tamanho Médio das Publicações	46.5	65.1	71.2	76.8	87.5	146.4
Média de *Hashtags*	0.0	0.0	0.1	0.5	0.6	2.3
Tamanho da Descrição	0.0	23.2	45.5	58.7	94.5	148.0
Total de Seguidores	3.0	23.2	75.5	98.7	122.5	417.0
Total de Amigos	18.0	54.2	174.5	234.4	334.2	854.0
Total de Listas	0.0	0.0	0.0	0.6	0.8	6.0
Total de Favoritos	4.0	88.8	1216.0	9313.3	10925.5	69673.0
Média de *Tweets* com Imagens	0.0	0.0	0.1	0.1	0.2	0.7
Estresse Médio por *Tweet*	-1.8	-1.4	-1.3	-1.3	-1.2	-1.0
Relaxamento Médio por *Tweet*	1.1	1.2	1.3	1.4	1.4	2.0
Estresse Percebido	14.0	29.5	34.5	33.7	39.8	50.0
Percepção de Si Mesmo	10.0	21.0	28.5	28.5	34.7	44.0
Futuro Planejado	5.0	16.0	20.0	20.2	26.5	31.0
Coesão Familiar	14.0	22.2	34.0	31.7	40.2	47.0
Recursos Sociais	19.0	37.0	43.0	41.2	47.2	56.0

Na Figura 30 é possível checar todas as correlações significativas entre os pares de variáveis quantitativas da primeira base de usuários. A partir das correlações aferidas é possível notar que:

- O Estresse Percebido (EP) apresentou correlações negativas com os constructos percepção de si mesmo (percepcao), futuro planejado (futuro), tempo médio das publicações (Idade_posts) e o escore médio de estresse gerado pelo TSpt. Indicando que quanto maior o estresse percebido, menores são a percepção de si mesmo e o futuro planejado nos indivíduos. É esperado que indivíduos com maior estresse percebido apresentem pontuações mais baixa para os constructos de resiliência, uma vez que o estresse em excesso pode afetar a resiliência aos estímulos estressores dos indivíduos, conforme explicam Galvão-Coelho, Silva e Sousa (2015). Perfis mais antigos demonstraram menor estresse percebido. Além disso, o estresse percebido por

meio das questões no formulário foi inversamente proporcional ao estresse medido nas publicações através do TSpt;

- O estresse médio das publicações gerados pelo TSpt (Estresse_TS) apresentou correlação positiva com a taxa de imagens publicadas junto com os textos coletados (TX_imagens). Pode-se dizer que, na amostra, indivíduos com maior publicação de imagens demonstraram maior estresse no texto;

- O relaxamento médio medido pelo TSpt (Relaxamento_TS) demonstrou uma correlação negativa com o constructo de recursos sociais (rec_soc) e uma correlação positiva com a idade média das publicações em semanas (Idade_posts), permitindo concluir que quanto maiores são os recursos sociais desses indivíduos, menor é o relaxamento no texto e que publicações mais antigas apresentaram maior relaxamento;

- O total de amigos na rede (Qtd_amigos) apresentou correlações negativas com os constructos futuro planejado (futuro) e coesão familiar. Assim, quanto mais amigos os indivíduos dessa amostra têm na rede, menores são o seu futuro planejado e a sua coesão familiar;

- A idade do perfil na rede em anos (Idade_perfil) demonstrou uma correlação negativa com o constructo percepção de si mesmo (percepcao). Assim, quanto mais antigo é o perfil da rede nessa amostra menor é a percepção de si mesmo;

- A idade das publicações em semanas ou o tempo médio das publicações em semanas (Idade_posts) apresentou uma correlação negativa com a escore de recursos sociais (rec_soc), indicando que usuários com publicações realizadas num período mais distante no tempo demonstraram menor escore de recursos sociais.

A realização do teste de Wilcoxon para a comparação da distribuição do estresse percebido de acordo com a função dos indivíduos foi significativa (p-valor = 0.09679) indicando que existem diferenças na distribuição do estresse percebido de acordo com o grupo. Conforme a Figura 31 mostra, é possível notar que os alunos foram os que demonstraram maior estresse percebido em comparação com os servidores do grupo. Não ocorreram diferenças significativas nas distribuições do estresse médio no texto considerando as funções dos indivíduos.

Verificou-se também uma diferença no relaxamento médio obtido por meio do TSpt em relação à moda de publicação dos indivíduos (p-valor = 0.005692). Conforme pode ser observado na Figura 32, os indivíduos que realizaram as publicações predominantemente entre 18:00 e 06:00 (P2) tiveram menores escores de relaxamento nos textos.

O escore de percepção de si mesmo apresentou distribuições divergentes segundo o teste de Wilcoxon (p-valor = 0.09811) conforme o gênero dos indivíduos. A Figura 33

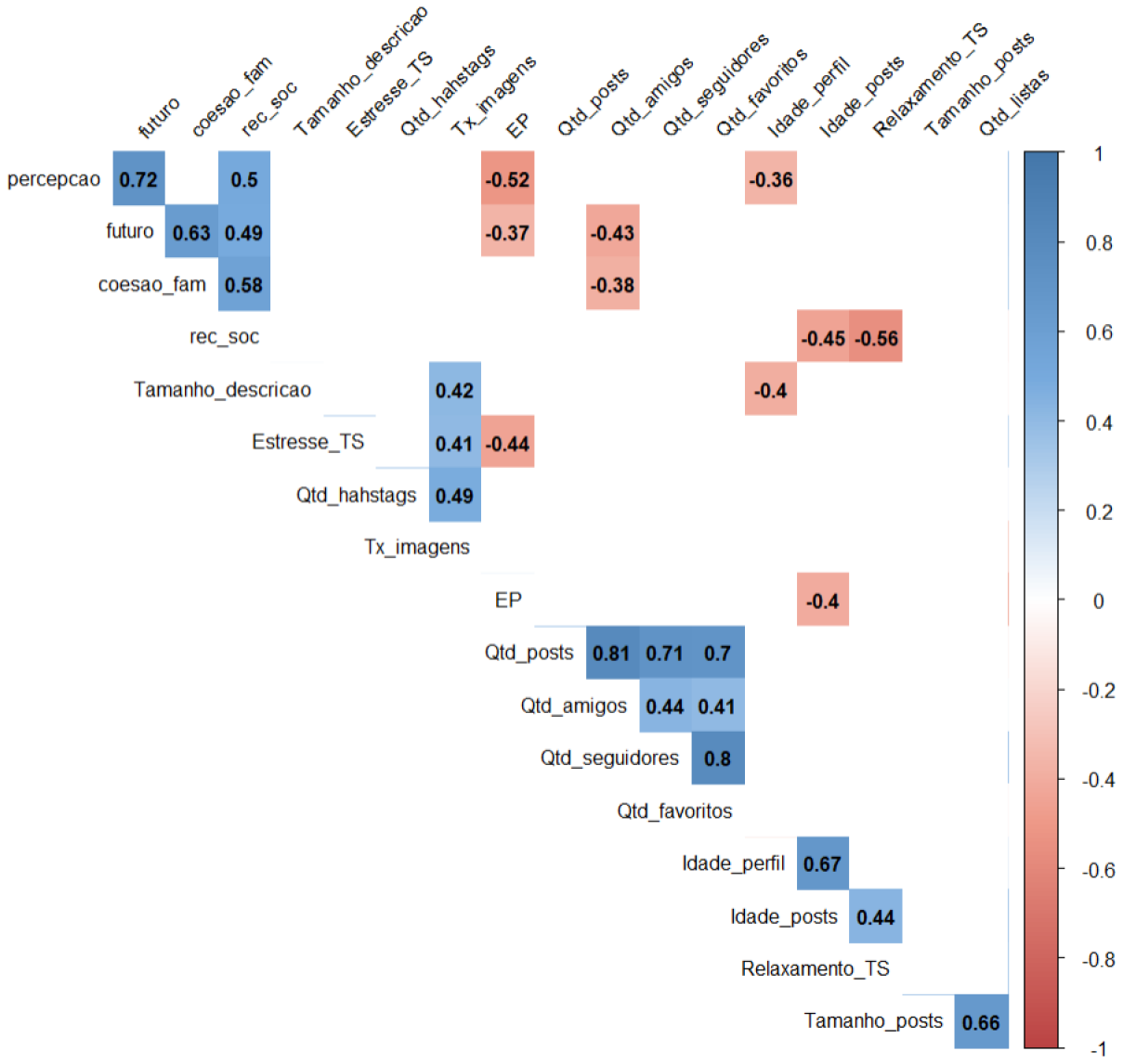

Figura 30 – *X* Fase 1: Correlações de Pearson Significativas
Autoria própria

mostra que os indivíduos do gênero masculino apresentaram, em geral, maior percepção de si mesmo.

5.3 *X* Fase 2

As estatísticas descritivas das variáveis quantitativas dos usuários do *X* na segunda fase da pesquisa podem ser verificadas na Tabela 12. Dos 18 voluntários dessa base de dados: 14 (78%) realizaram mais publicações entre 18:00 e 06:00, 11 (61%) tinham até 25 anos no momento da coleta, 8 (44%) são do gênero feminino, 16 (88%) são discentes e 9 (50%) apresentaram escore mínimo para a detecção do transtorno de estresse pós-

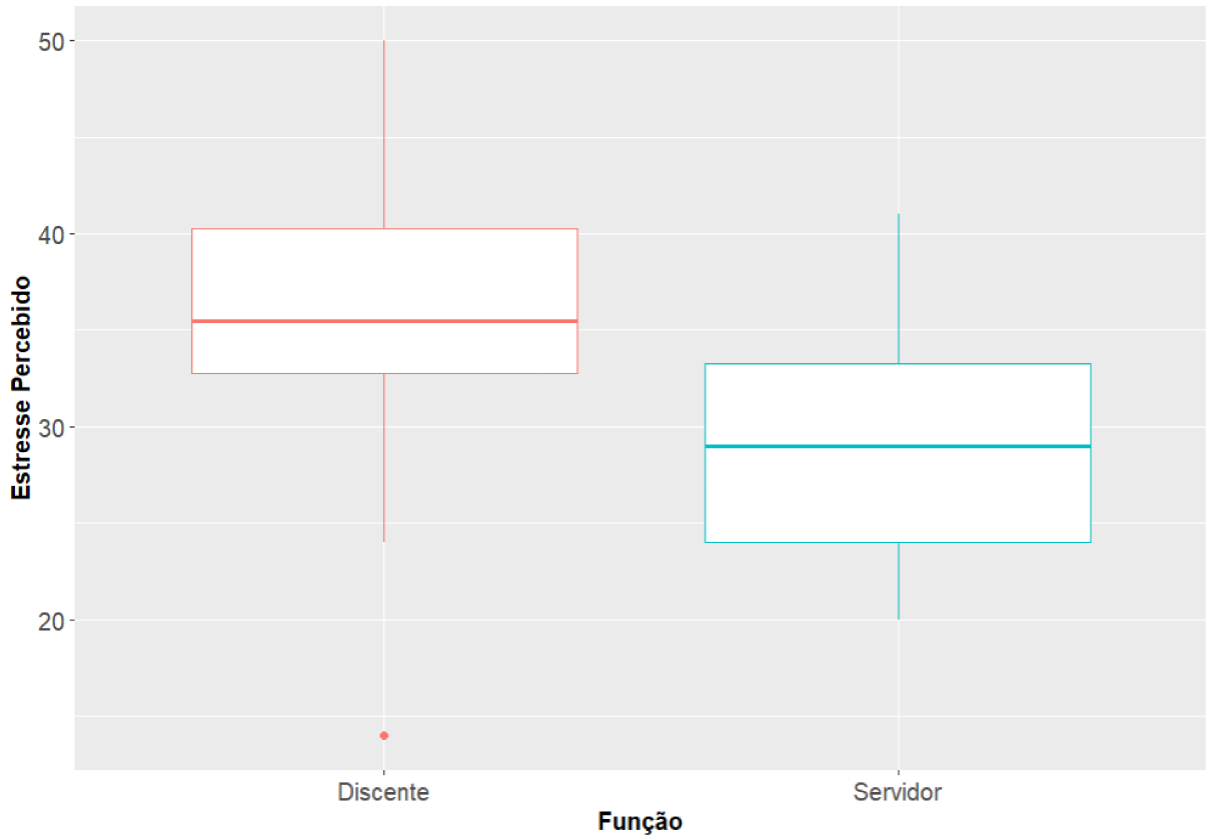

Figura 31 – *X* Fase 1: Distribuição do Estresse Percebido por Função
Autoria própria

traumático.

Tabela 12 – X Fase 2: Estatísticas Descritivas dos Usuários (Variáveis Quantitativas)

Variável	Mín.	Q1	Mediana	Média	Q3	Máx.
Quantidade de Publicações	1.0	4.0	14.5	57.0	71.0	344.0
Idade do Perfil	0.6	3.1	6.6	7.1	12.1	12.8
Idade Média das Publicações	0.0	5.4	7.0	6.5	7.8	9.3
Tamanho Médio das Publicações	41.0	56.5	69.0	89.7	101.5	296.0
Média de *Hashtags*	0.0	0.0	0.0	0.2	0.1	2.5
Tamanho da Descrição	0.0	0.7	31.0	47.1	100.2	131.0
Total de Seguidores	2.0	27.5	170.5	458.4	584.2	2638.0
Total de Amigos	16.0	64.0	212.5	559.3	752.2	4290.0
Total de Listas	0.0	0.0	0.0	0.9	0.8	8.0
Total de Favoritos	23.0	1439.0	5660.0	22591	21063.0	195517.0
Média de *Tweets* com Imagens	0.0	0.0	0.0	0.2	0.4	1.0
Estresse Médio por *Tweet*	-2.0	-1.6	-1.3	-1.3	-1.0	-1.0
Relaxamento Médio por *Tweet*	1.0	1.0	1.3	1.3	1.5	2.0
IES-R	0.7	3.8	5.2	5.5	7.6	10.7

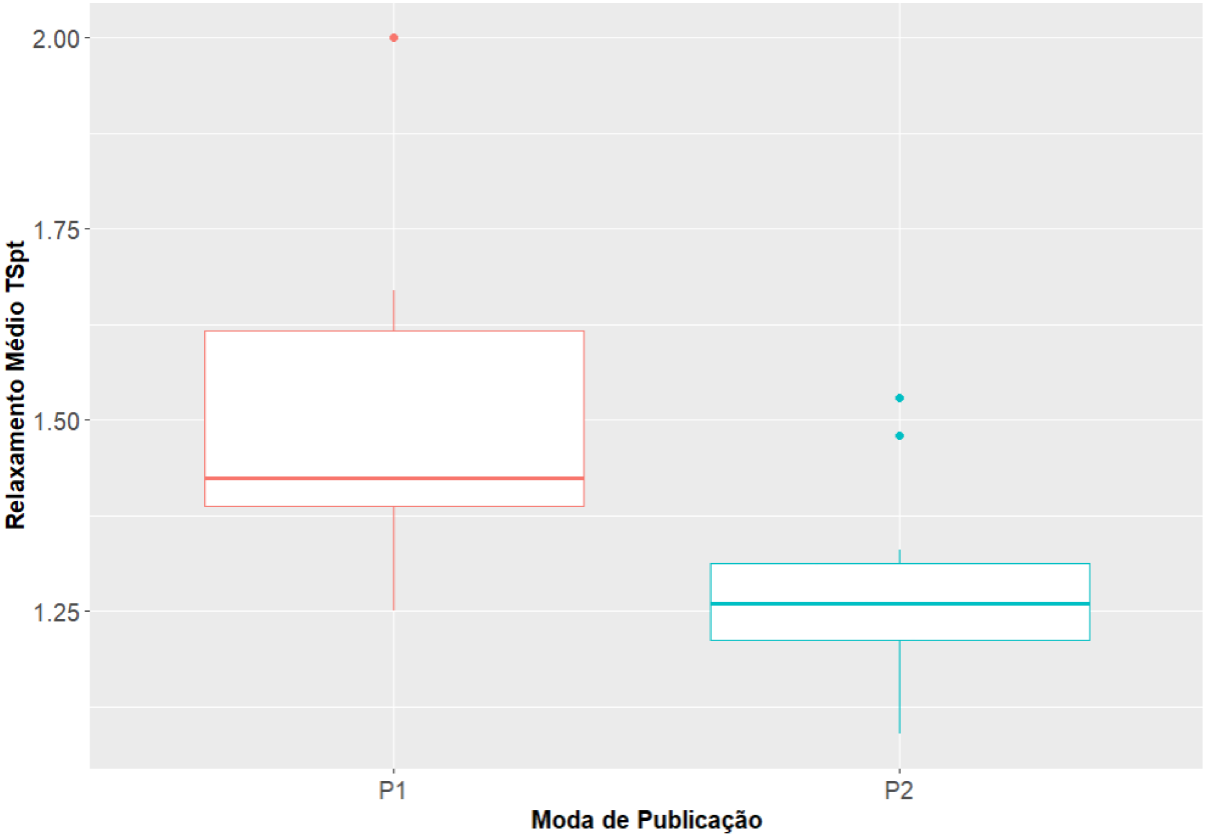

Figura 32 – X Fase 1: Distribuição do Relaxamento Médio por Período
Autoria própria.

As correlações significativas apresentadas na Figura 34 permitiram as seguintes conclusões:

- O escore para a ocorrência de transtorno de estresse pós-traumático (IES-R) na amostra indicou correlações positivas com a quantidade de seguidores nas redes (Qtd_seguidores), quantidade de amigos (Qtd_amigos) e o tamanho da descrição no perfil em caracteres (Tamanho_descricao). Assim, quanto maior a quantidade de seguidores, amigos e a descrição do perfil, maior é o indicador IES-R, em geral.

- O estresse médio das publicações (Estresse_TS) apresentou correlações positivas com a quantidade de seguidores (Qtd_seguidores), quantidade de amigos (Qtd_amigos) e idade do perfil (Idade_perfil). Pode-se dizer que a tendência é que o estresse médio no texto tenha menor incidência em perfis com mais seguidores, amigos e mais antigos. Nota-se um contraponto em relação às quantidades de amigos e seguidores em termos de IES-R e estresse médio no texto.

As diferenças entre as distribuições do IES-R foram significativas, segundo o teste de Wilcoxon de acordo com o gênero dos indivíduos (p-valor = 0.01638). A Figura 35 mostra

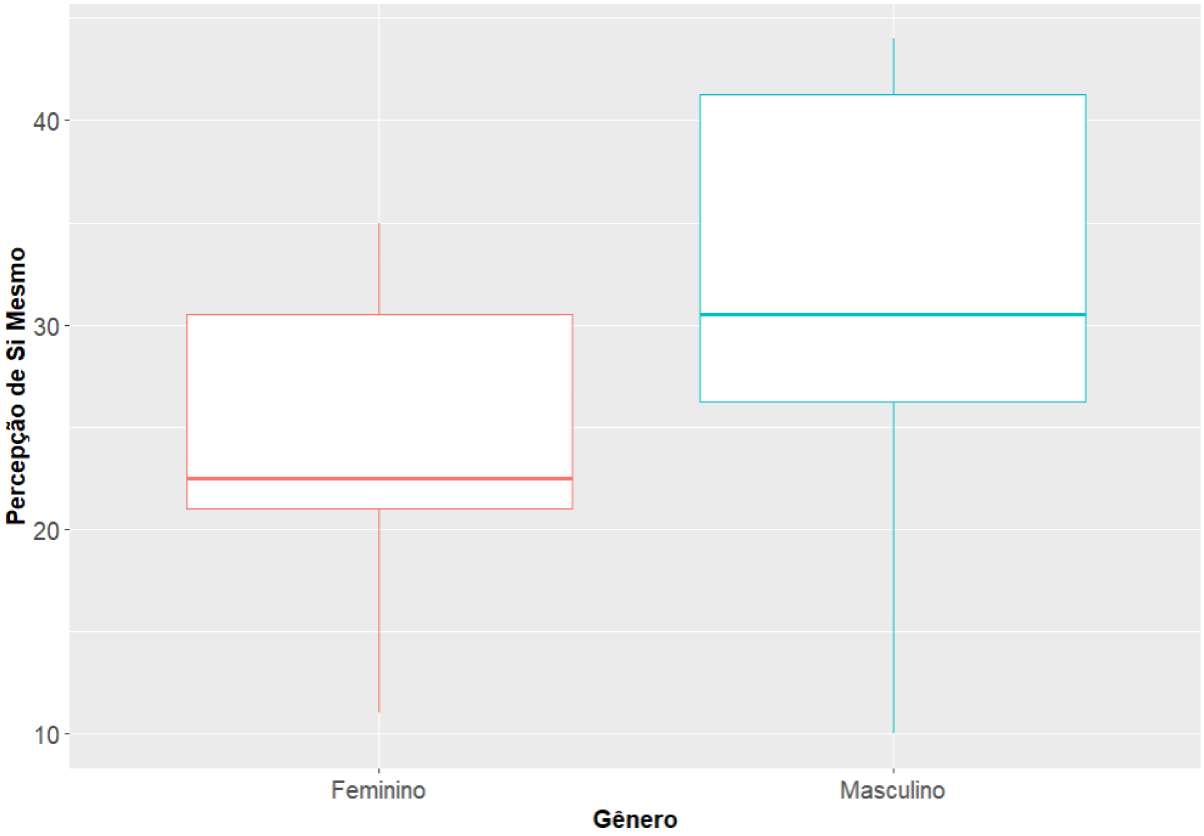

Figura 33 – X Fase 1: Distribuição do Escore de Percepção por Gênero
Autoria própria

que, em geral, os indivíduos do gênero feminino apresentaram maior escore de estresse pós-traumático. Este resultado é corroborado por pesquisas afins realizadas durante a pandemia, que demonstraram a maior incidência de TEPT em pessoas do gênero feminino (SILVA, 2023; RODRIGUES, 2023; KAMEO et al., 2023). As distribuições do estresse médio no texto não foram significativas, de acordo com o gênero.

Outro resultado importante verificado nessa base é o de que todos os indivíduos com moda de publicação entre 18:00 e 06:00 (P2) apresentaram escore suficiente para o diagnóstico de transtorno de estresse pós-traumático. A tabela cruzada entre a moda de publicação (a) e o diagnóstico de transtorno de estresse pós-traumático (b) pode ser verificada na Figura 36. A Figura 37 mostra os resíduos padronizados ajustados para cada associação que, por serem maiores em módulo que o *z-score* para uma significância de 10% (1.645), são todos significativos.

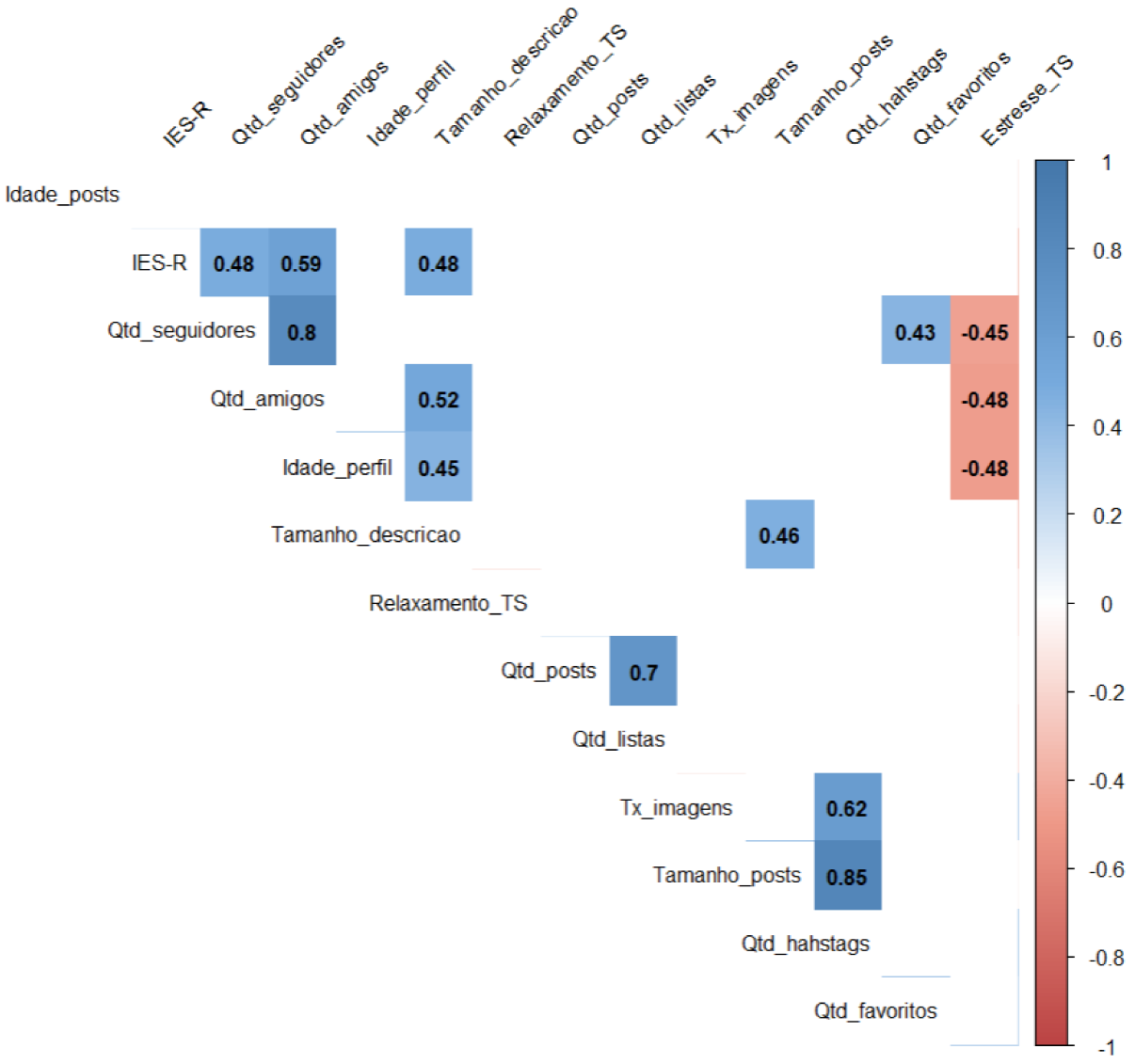

Figura 34 – X Fase 2: Correlações de Pearson Significativas
Autoria própria

5.4 *Instagram* Fase 1

Dos 9 voluntários com dados coletados no *Instagram* na primeira fase da pesquisa: 7 (78%) realizaram mais publicações entre 06:00 e 18:00, todos tinham mais de 25 anos, 6 (67%) são do gênero feminino e 6 (67%) são docentes. As estatísticas descritivas relacionada a esse grupo de voluntários consta na Tabela 13.

A matriz com as correlações significativas entre os pares variáveis quantitativas pode ser verificada na Figura 38, que permite tirar as seguintes conclusões:

- O escore de estresse percebido (EP) apresenta correlações significativas somente com os constructos percepção de si mesmo (percepcao) e futuro planejado (futuro).

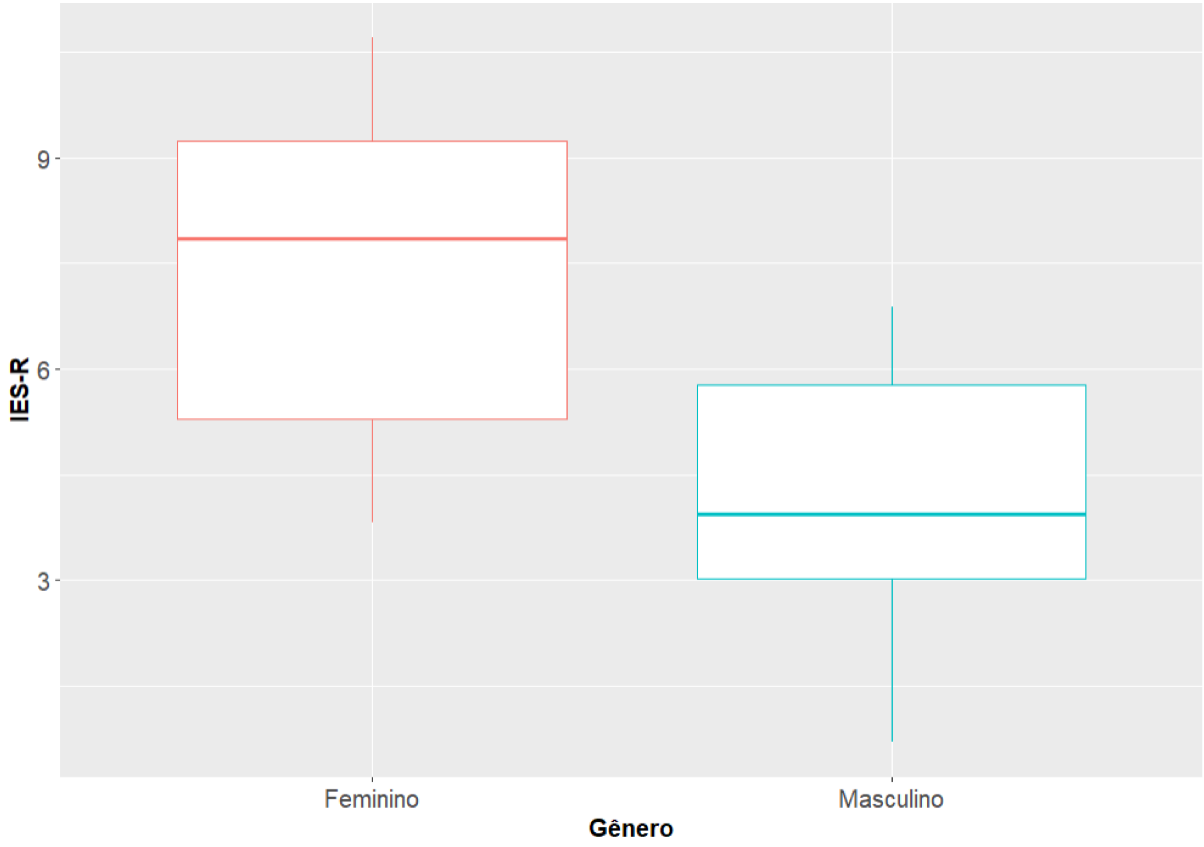

Figura 35 – X Fase 2: Distribuição IES-R por Gênero
Autoria própria

a	*b* NÃO	SIM	***Total***
P1	4	0	4
P2	5	9	14
Total	9	9	18

χ^2=2.893 · df=1 · φ=0.535 · Fisher's p=0.082

Figura 36 – X Fase 2: Teste de Associação entre Gênero e Estresse Pós-traumático
Autoria própria

Assim, na amostra quanto maior o escore EP, menores são a percepção de si mesmo e o futuro planejado nos indivíduos;

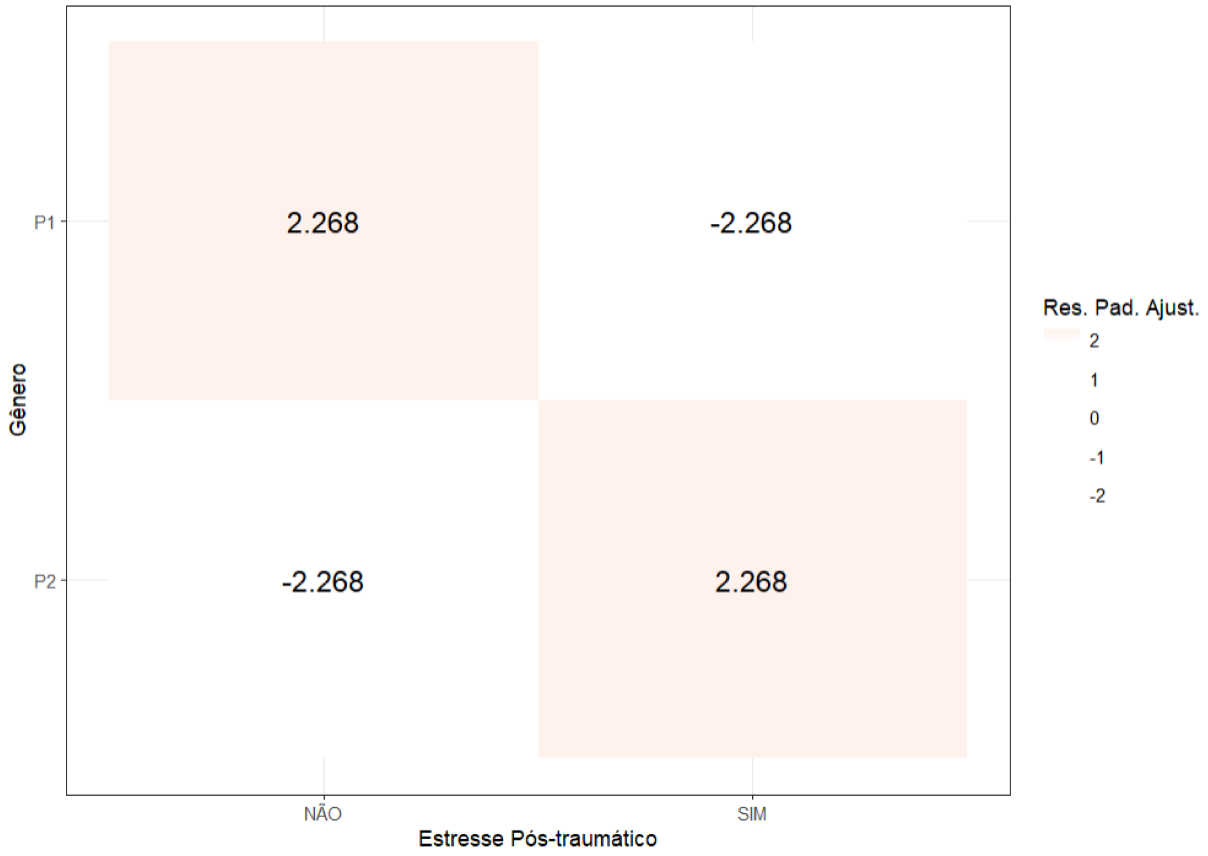

Figura 37 – X Fase 2: Gênero x Estresse Pós-Traumático - Resíduos Padronizados Ajustados. Autoria própria

Tabela 13 – Instagram Fase 1: Estatísticas Descritivas dos Usuários (Variáveis Quantitativas)

Variável	Mín.	Q1	Mediana	Média	Q3	Máx.
Quantidade de Publicações	5.0	12.0	22.0	24.2	28.0	68.0
Média de *Likes*	11.9	24.4	40.9	39.3	49.9	70.4
Idade Média das Publicações	20.3	27.0	75.2	119.2	122.4	397.2
Tamanho Médio das Publicações	24.5	49.3	66.4	122.1	165.7	290.2
Tamanho da Descrição	0.0	0.0	0.0	35.7	63.0	132.0
Total de Seguidores	167.0	303.0	451.0	1123.0	650	6119.9
Total Seguindo	79.0	288.0	530.0	615.9	1127.0	1252.0
Estresse Médio por Publicação	-1.3	-1.2	-1.2	-1.1	-1.0	-1.0
Relaxamento Médio por Publicação	1.2	1.3	1.5	1.6	1.6	2.4
Estresse Percebido	12.0	23.0	28.0	26.0	32.0	37.0
Percepção de Si Mesmo	19.0	25.0	33.0	31.9	37.0	41.0
Futuro Planejado	10.0	21.0	23.0	22.3	25.0	28.0
Coesão Familiar	14.0	28.0	37.0	33.4	40.0	47.0
Recursos Sociais	20.0	39.0	46.0	41.9	50.0	52.0

- A média de estresse no texto (Estresse_TS) não apresentou correlações significativas com nenhuma das demais variáveis;
- O relaxamento médio no texto (Relaxamento_TS) não apresentou correlações significativas com nenhuma das demais variáveis;
- O constructo futuro planejado (futuro) apresentou uma correlação negativa com o a idade das publicações em semanas (Idade_posts). Assim, a tendência é de que os indivíduos com publicações mais antigas demonstrem uma menor visão de futuro planejado;
- Os constructos de recursos sociais (rec_soc) e coesão familiar (coesao_fam) apresentaram correlações negativas com a quantidade de perfis seguidos pelos voluntários dessa amostra (Qtd_seguindo). Pode-se dizer que quanto maior o número de perfis seguidos por essas pessoas, menores são as suas percepções de recursos sociais e coesão familiar.

A aplicação do teste de Wilcoxon para comparação das distribuições permitiu verificar que existe uma diferença significativa entre a percepção de si mesmo entre alunos e servidores na amostra (p-valor = 0.09192). A Figura 39 mostra que os alunos da amostra demonstraram menor escore de percepção de si mesmo em comparação aos professores.

Outra diferença observada na amostra relacionada à função das pessoas avaliadas foi o nível de recursos sociais (p-valor = 0.05182). A Figura 40 mostra que os alunos da amostra também demonstraram um escore de recursos sociais em geral menor que o escore dos professores.

5.5 *Instagram* Fase 2

Dos 23 indivíduos com dados coletados do *Instagram* na segunda fase da pesquisa: 19 (83%) realizou a maioria das publicações entre 18:00 e 06:00, 17 (74%) tinham até 25 anos no momento da coleta, 11 (48%) são do gênero feminino, 19 (83%) são discentes e 10 (43%) apresentaram escore mínimo para a incidência de transtorno de estresse pós-traumático. Na Tabela 14 constam as estatísticas descritivas desses indivíduos.

A avaliação da matriz de correlações entre as variáveis quantitativas dessa base de dados está apresentada na Figura 41. Conforme pode ser observado:

- O escore IES-R não apresentou correlações significativas com nenhuma das demais variáveis;
- O estresse médio verificado no texto (Estresse_TS) apresentou correlações negativas com o relaxamento médio do texto (Relaxamento_TS) e com o tamanho das publicações em caracteres (Tamanho_posts). De forma que a tendência é de que

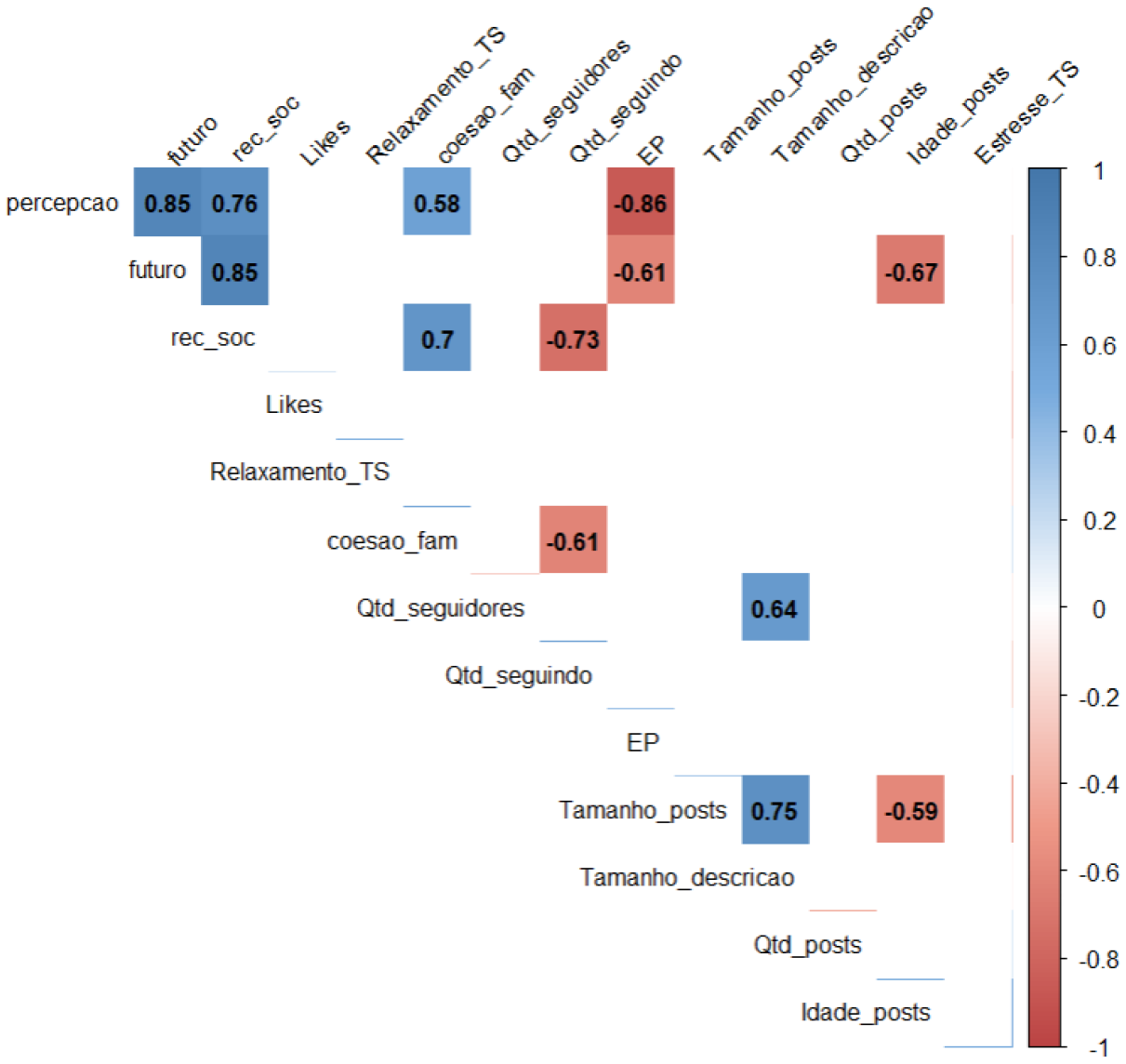

Figura 38 – Instagram Fase 1: Correlações de Pearson Significativas
Autoria própria

quanto maio o estresse médio no texto menor seja o relaxamento médio e de que usuários com textos mais longos na rede demonstrem menor estresse no texto;

- O relaxamento médio no texto (Relaxamento_TS) apresentou uma correlação positiva com o tamanho médio do texto (Tamanho_posts). Em geral, usuários com textos mais longos expressaram maior relaxamento nessa base.

Verificou-se também diferenças significativas nas distribuições do estresse médio no texto de acordo com a idade dos participantes, por meio do teste de Wilcoxon (p-valor = 0.003757). A Figura 42 mostra que na amostra avaliada os indivíduos com mais de 25 anos expressaram maior estresse médio no texto.

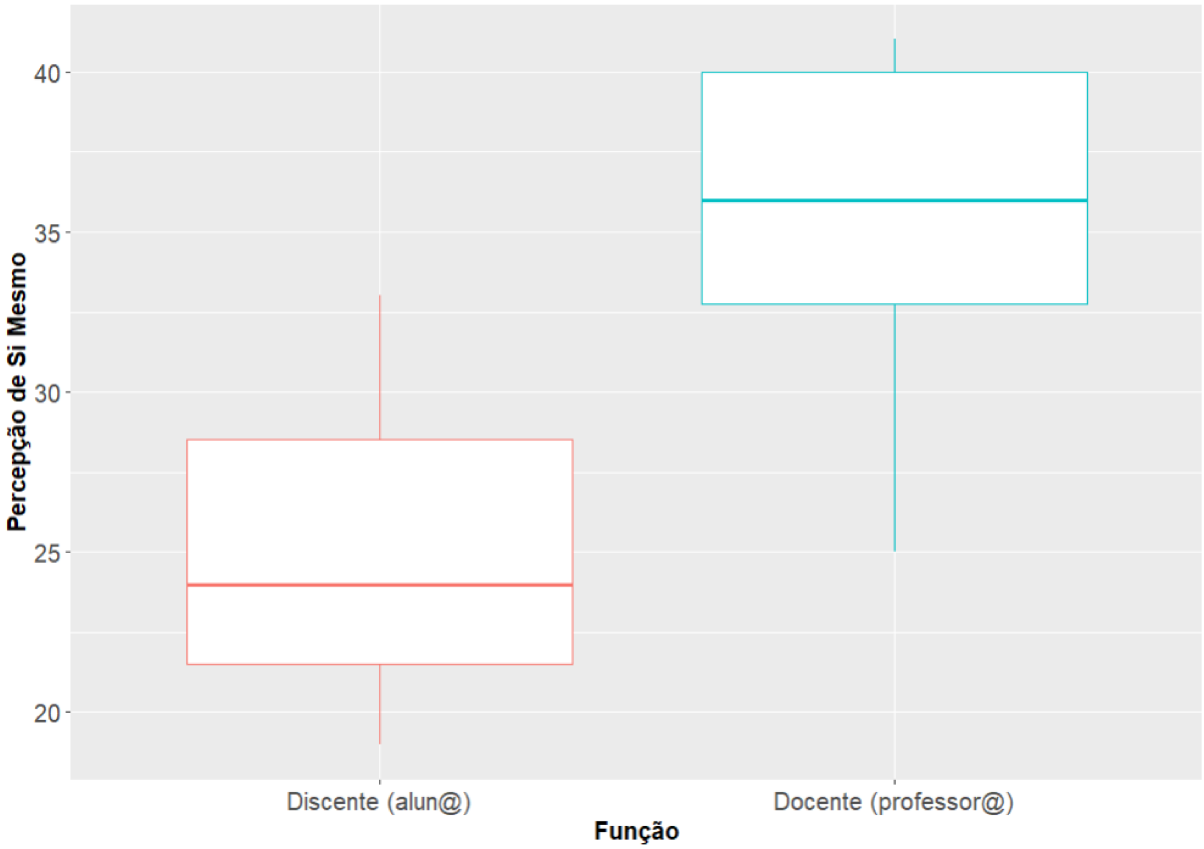

Figura 39 – Instagram Fase 1: Percepção de Si Mesmo por Função
Autoria própria

Tabela 14 – Instagram Fase 2: Estatísticas Descritivas dos Usuários (Variáveis Quantitativas)

Variável	Mín.	Q1	Mediana	Média	Q3	Máx.
Quantidade de Publicações	2.0	7.0	42.0	97.6	100.0	642.0
Idade Média das Publicações	0.0	3.0	129.1	100.0	177.6	264.6
Tamanho Médio das Publicações	1	24.3	45.9	76.0	84.0	524.7
Tamanho da Descrição	0.0	26.0	47.0	45.3	60.5	117.0
Total de Seguidores	122.0	406.0	809.0	1279.0	1758.0	4756.0
Total Seguindo	155.0	491.5	797.0	976.6	1261.5	3531.0
Estresse Médio por Publicação	-1.5	-1.3	-1.2	-1.2	-1.0	-1.0
Relaxamento Médio por Publicação	1.0	1.2	1.3	1.4	1.5	2.6
IES-R	0.0	2.7	5.0	5.0	7.4	9.4

O estresse médio no texto apresentou diferenças significativas considerando a função dos participantes (p-valor = 0.02186). A Figura 43 mostra que os servidores da amostra demonstraram maior estresse médio no texto.

O relaxamento médio no texto também apresentou diferenças de distribuições de acordo com a faixa de idade dos participantes (p-valor = 0.002264). Em particular,

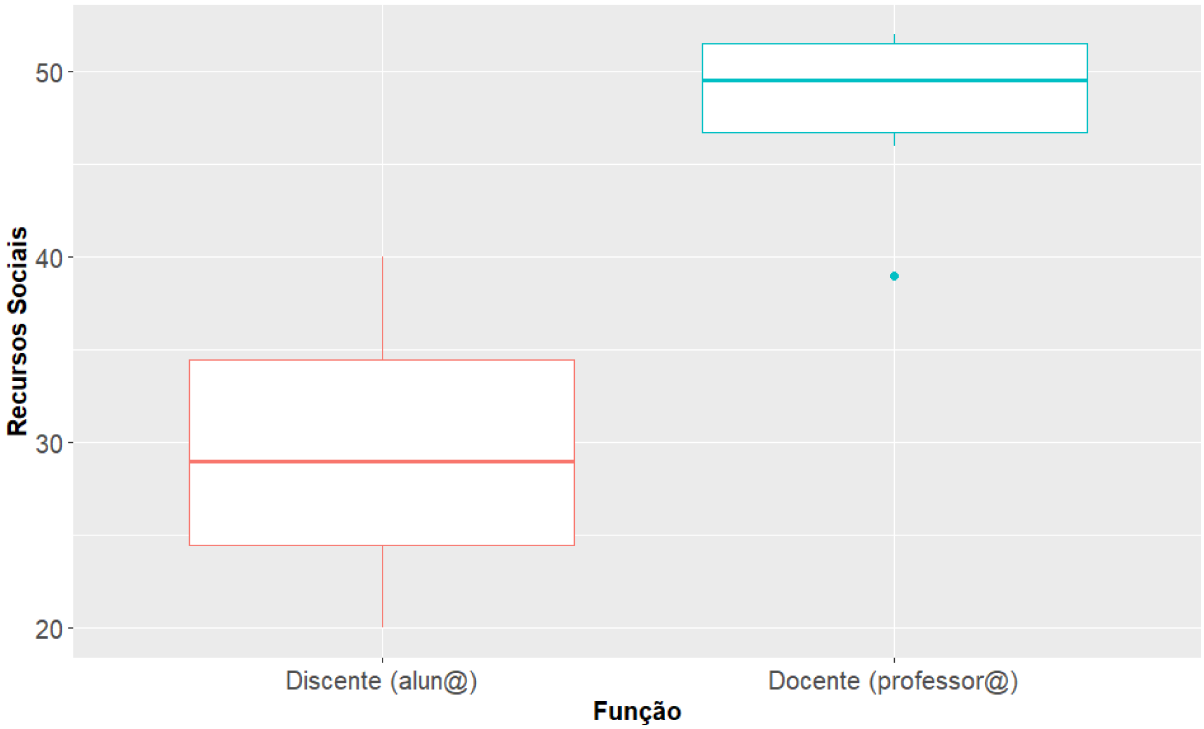

Figura 40 – Instagram Fase 1: Recursos Socias por Função
Autoria própria

os indivíduos com mais de 25 anos demonstraram maior relaxamento no texto, conforme a Figura 44.

Houve diferenças significativas no relaxamento médio no texto conforme a função dos indivíduos (p-valor = 0.01314). A comparação das distribuições na Figura 45 mostra que os servidores da amostra expressaram maior relaxamento nas publicações.

5.6 Comparações entre Redes

A comparação entre redes na primeira fase da pesquisa permitiu concluir que:

- Houve diferenças significativas sobre o estresse percebido nos participantes entre redes (p-valor = 0.03102). A Figura 46 mostra que os participantes com perfil no *X* demonstraram maior estresse percebido;

- O estresse médio nos textos publicados pelos indivíduos nas redes apresentou diferenças significativas (p-valor = 0.03049). O estresse médio no texto foi maior entre usuários do *X* (Figura 47). Estes dois resultados iniciais da comparação entre redes é corroborado por pesquisas como a proposta por Waterloo et al. (2018), que demonstrou que usuários do *X* expressam mais emoções negativas;

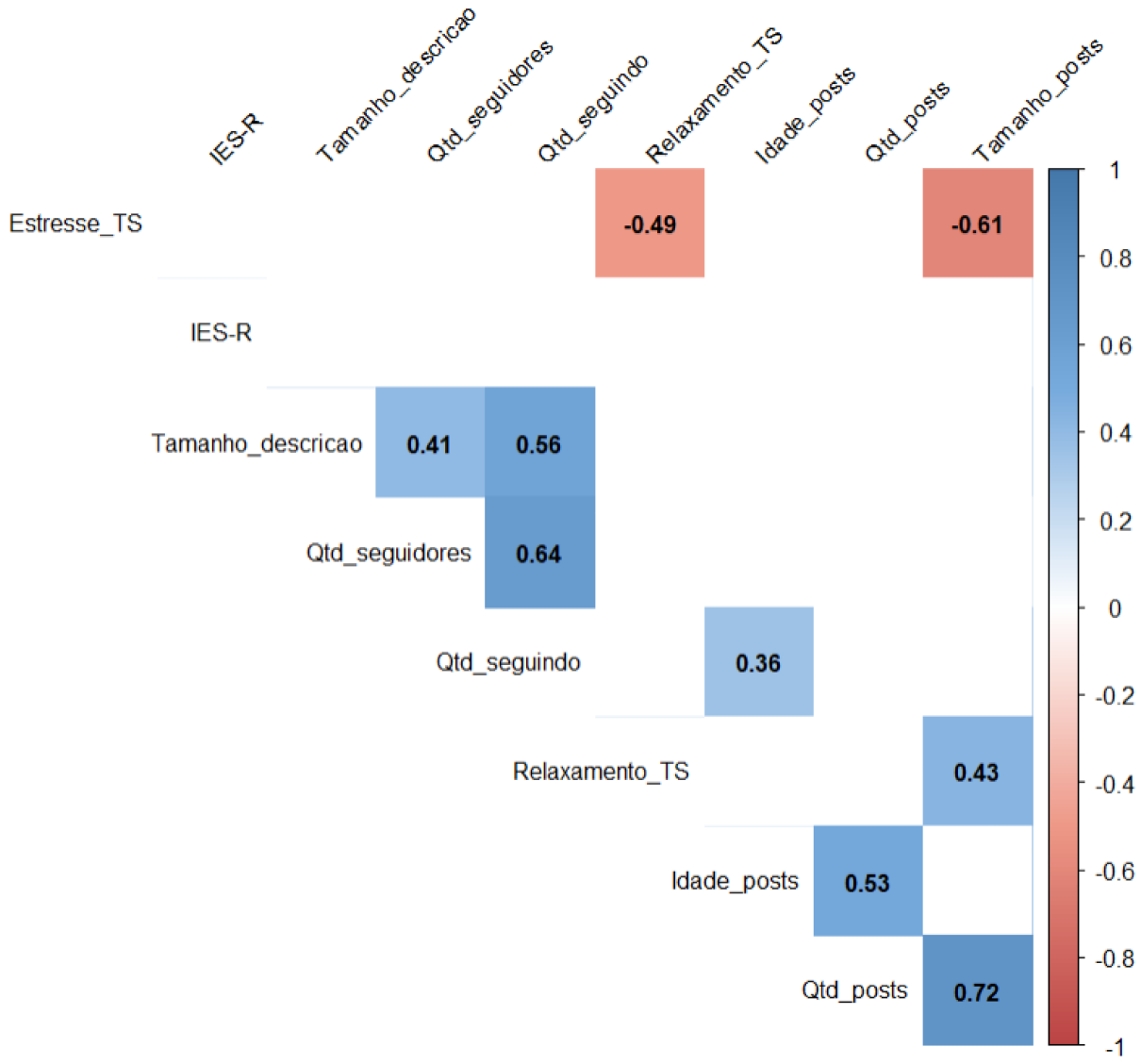

Figura 41 – Instagram Fase 2: Correlações de Pearson Significativas
Autoria própria

- O relaxamento médio nos textos publicados foi significativamente diferente entre as redes (p-valor = 0.07783). Conforme a Figura 48 mostra, os usuários do *Instagram* apresentaram maior relaxamento médio. Novamente este resultado está de acordo com Waterloo et al. (2018), que demonstra que esta é uma rede com maior incidência de emoções positivas.

A comparação entre redes na segunda fase da pesquisa evidenciou que somente o estresse médio no texto apresentou diferenças significativas (p-valor = 0.08697). A Figura 49 mostra que houve usuários do X com valores inferiores de estresse no texto, embora a dispersão tenha sido maior para esse indicador nesse grupo de usuários.

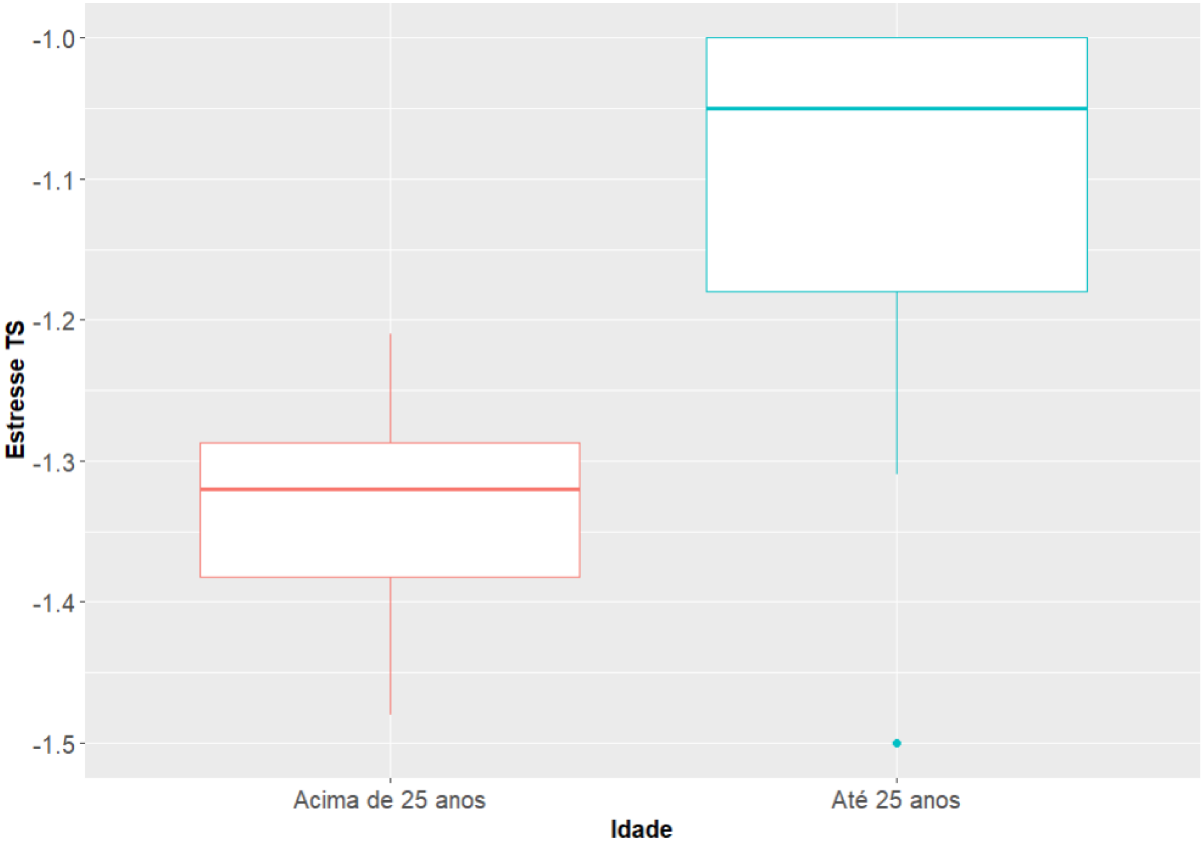

Figura 42 – Instagram Fase 2: Estresse TS por Idade
Autoria própria

5.7 Considerações

Verificou-se uma baixa volumetria de usuários com perfis válidos, conforme pode ser observado na Tabela 10, o que prejudicou a execução de técnicas mais avançadas para a comparação das características entre usuários e redes. O baixo volume de usuários com dados extraídos pode ser explicado por configurações de privacidade habilitadas pelos usuários, por perfis informados inválidos e por uma resistência dos indivíduos abordados em fornecer dados de redes sociais.

Na primeira fase da pesquisa o estresse percebido apresentou correlações significativas com os fatores de resiliência *percepção de si mesmo* e *futuro planejado*, fornecendo evidências em ambas as redes de que indivíduos com maior percepção de si mesmo e com maior nível de futuro planejado demonstraram menor estresse percebido.

As correlações apresentadas entre o estresse percebido e o estresse médio no texto evidenciam que os usuários com o estresse mais elevado não apresentaram um estresse igualmente proporcional no texto. Na amostra do X verificou-se uma tendência de baixo estresse no texto e na amostra do *Instagram* não foi verificada uma correlação significativa. Da mesma maneira, o escore IES-R não demonstrou qualquer correlação com o estresse

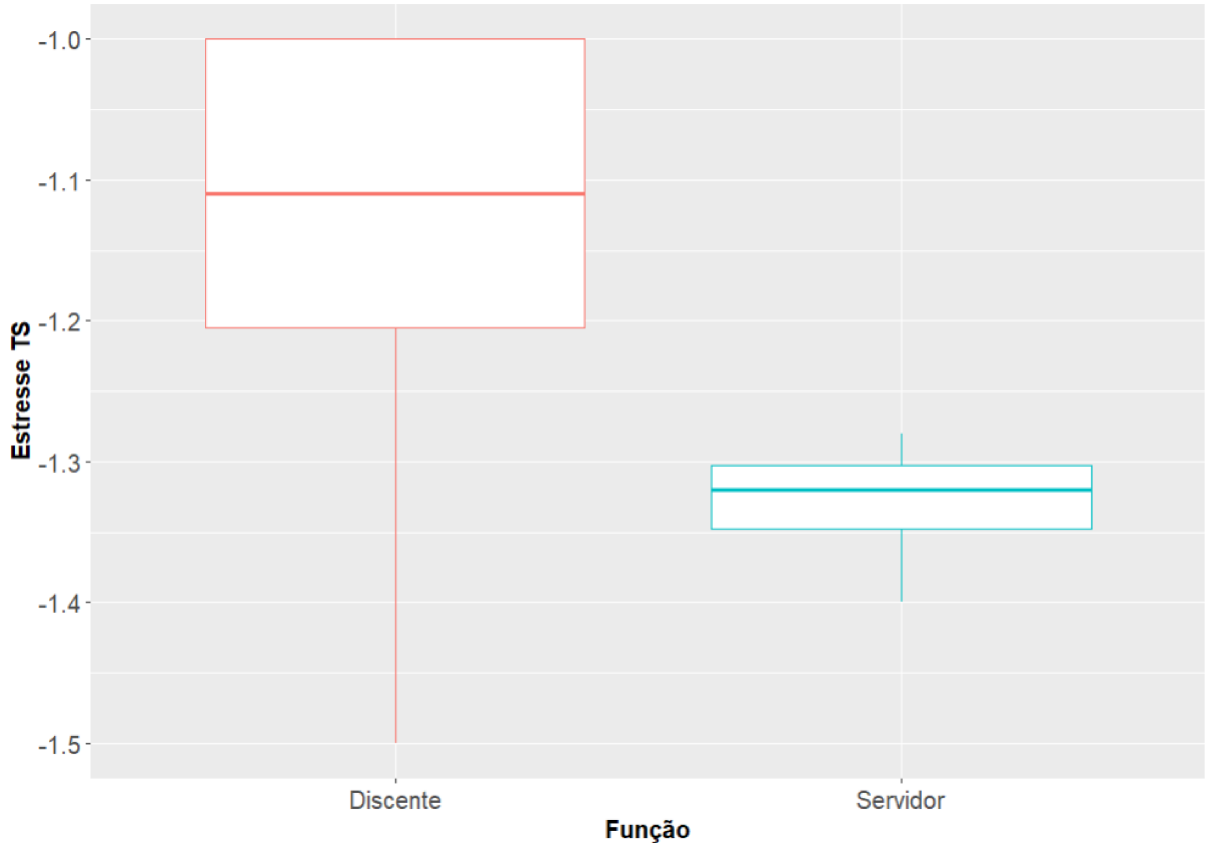

Figura 43 – Instagram Fase 2: Estresse TS por Função
Autoria própria

médio no texto em ambas as redes avaliadas. Os resultados amostrais não indicam, portanto, que o estresse verificado no texto reflete o estresse do usuário.

O estresse percebido foi menor em indivíduos com contas mais antigas no *X* na primeira fase da pesquisa. Na segunda fase, o TEPT foi maior em perfis com mais amigos e seguidores e em perfis com descrições mais longas no *X*. As estatísticas de publicação dos usuários não apresentaram correlações similares com o estresse percebido e o TEPT em usuários do *Instagram*.

O estresse médio de acordo com o TSpt foi maior em indivíduos com maior uso de imagens nas publicações na primeira fase e com mais seguidores, amigos e com perfis mais antigos no *X*. No *Instagram* a amostra da segunda fase indicou menor estresse médio em indivíduos com publicações mais longas e com maior relaxamento médio nas publicações. Nota-se uma distinção em relação ao estresse manifestado em ambas as redes.

No *X* o relaxamento médio no texto foi maior em indivíduos com menor escore de recursos sociais e com publicações mais antigas na primeira fase, ou seja, nesse cenário os usuários com maior relaxamento no texto foram aqueles com maior incidência de postagem em um período mais distante da pandemia. Outro ponto interessante de observar nesse

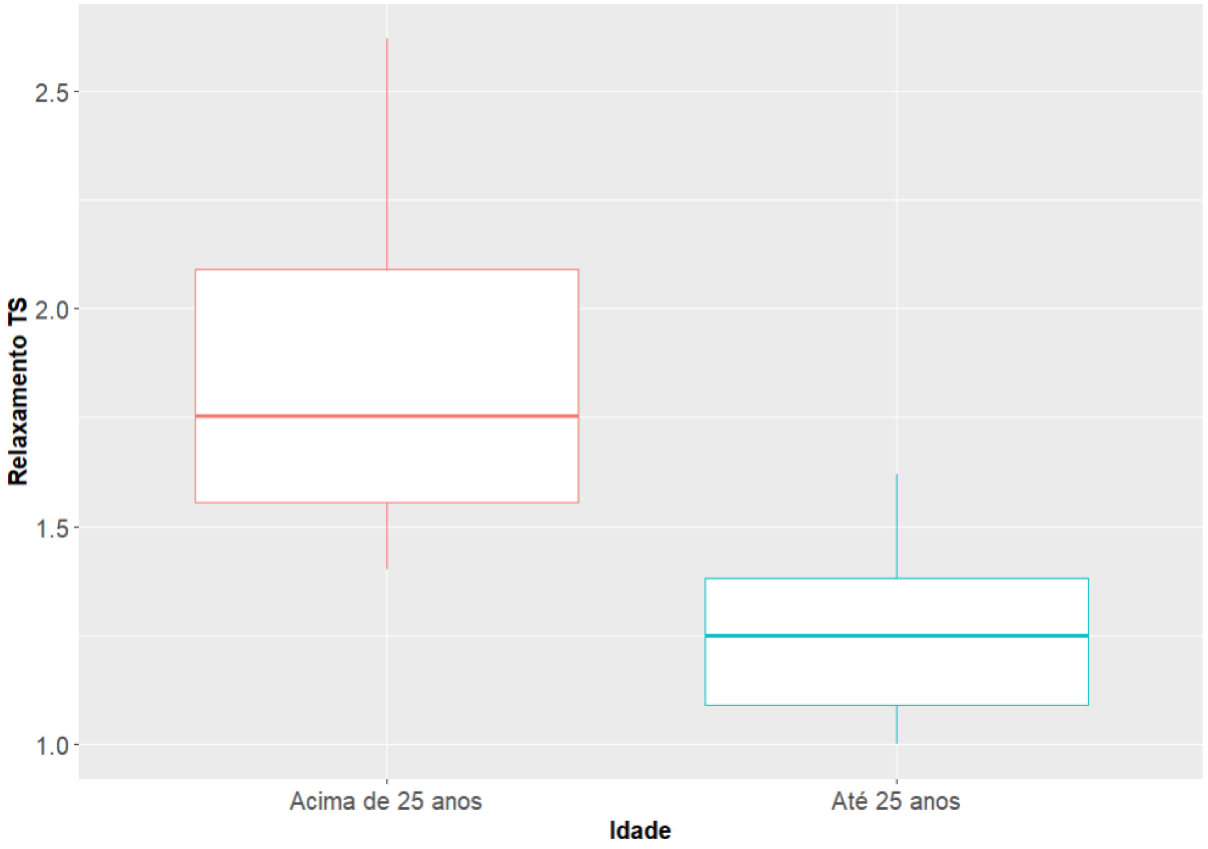

Figura 44 – Instagram Fase 2: Relaxamento TS por Idade
Autoria própria

mesmo cenário é que indivíduos com maior nível de recursos sociais apresentaram menor relaxamento textual. No *Instagram* o relaxamento médio do texto foi maior em usuários com textos mais longos na segunda fase da pesquisa.

A amostra do X na primeira fase indicou que os usuários com mais amigos na rede demonstraram menor nível de futuro planejado e coesão familiar. No *Instagram* os níveis de recursos sociais e coesão familiar foram menores em indivíduos que seguem um número maior de perfis. Percebe-se uma diferença nos indicadores de resiliência em ambas as redes. No primeiro caso existe uma aparente compensação da falta de resiliência na quantidade de contatos seguidos pelos usuários, enquanto no último caso essa compensação se dá pela quantidade de perfis seguidos pelos voluntários.

Quanto ao tempo de publicação em semanas, no X os indivíduos com publicações mais antigas apresentaram menor nível de recursos sociais. Já no *Instagram* os usuários com publicações mais antigas demonstraram menor nível de futuro planejado. Percebe-se que a resiliência dos indivíduos menos ativos nas redes foi afetada negativamente por diferentes fatores durante a pandemia.

O gênero dos usuários foi discriminante na percepção de si mesmo (gênero masculino

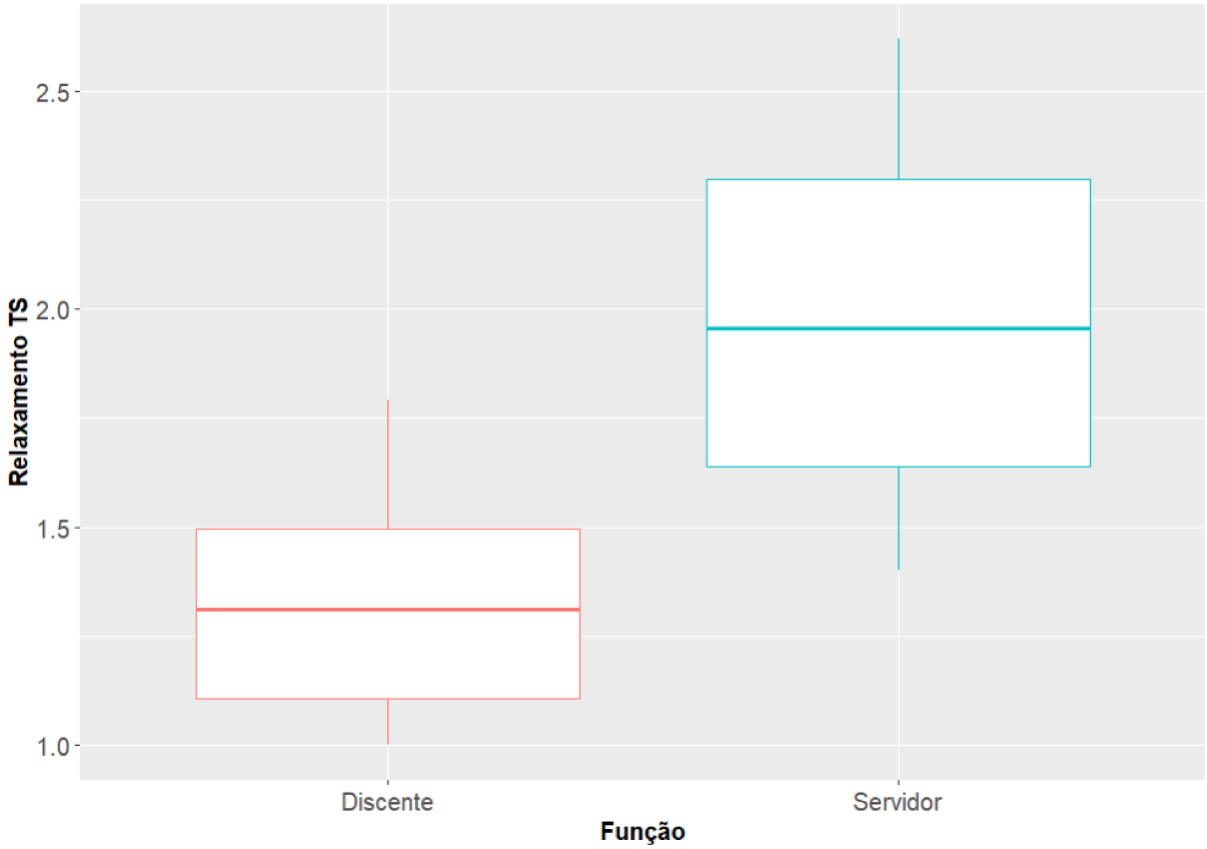

Figura 45 – Instagram Fase 2: Relaxamento TS por Função
Autoria própria

com maior escore) na primeira fase e no nível de TEPT na segunda fase (gênero feminino com maior escore) em usuários do *X*.

A faixa de idade dos participantes foi significativa para distinguir o relaxamento médio no texto em usuários do *Instagram* na segunda fase da pesquisa, indicando maior relaxamento médio no grupo de indivíduos com mais de 25 anos.

A função exercida pelos indivíduos permitiu diferenciar o estresse percebido dos usuários do *X* na primeira fase (alunos com maior escore). Para os usuários do *Instagram* a função dos indivíduos indicou maior percepção de si mesmo e maiores recursos sociais entre servidores na primeira fase e maior estresse e relaxamento no texto entre servidores na segunda fase da pesquisa.

Quanto à moda de publicação, os usuários do *X* com moda de publicação no segundo período demonstraram menor relaxamento médio no texto na primeira fase e 100% dos indivíduos com predominância de publicação nesse período apresentaram escore mínimo para a ocorrência de TEPT no momento de flexibilização da pandemia.

Quanto às redes sociais como fator de comparação, notou-se um maior estresse percebido e estresse médio no texto entre usuários do *X*, enquanto os usuários do *Instagram*

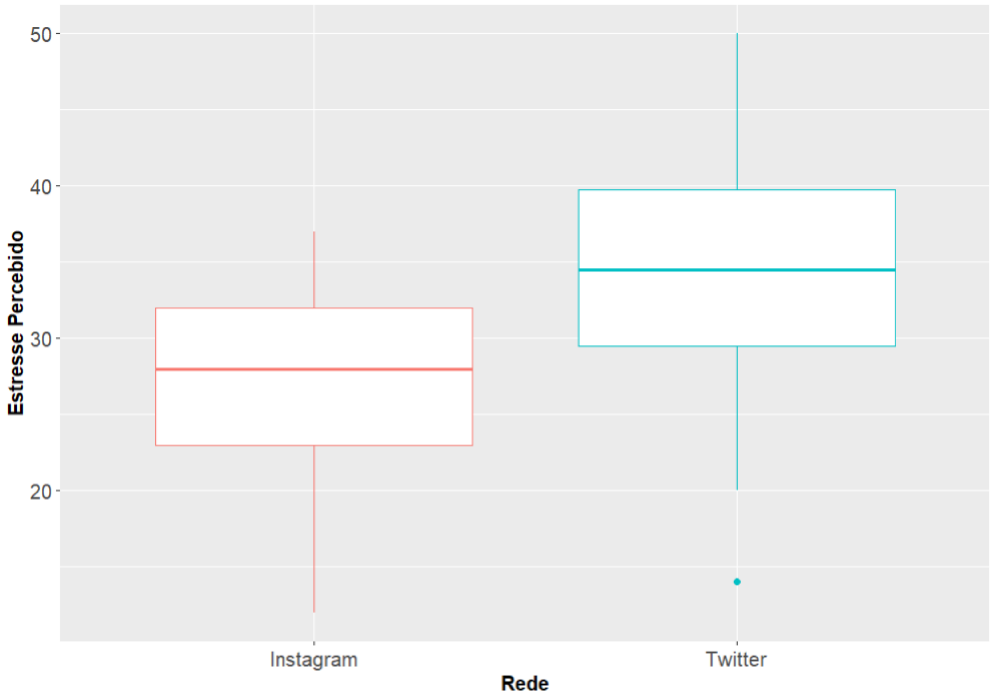

Figura 46 – Comparação entre Redes Fase 1: Estresse Percebido
Autoria própria

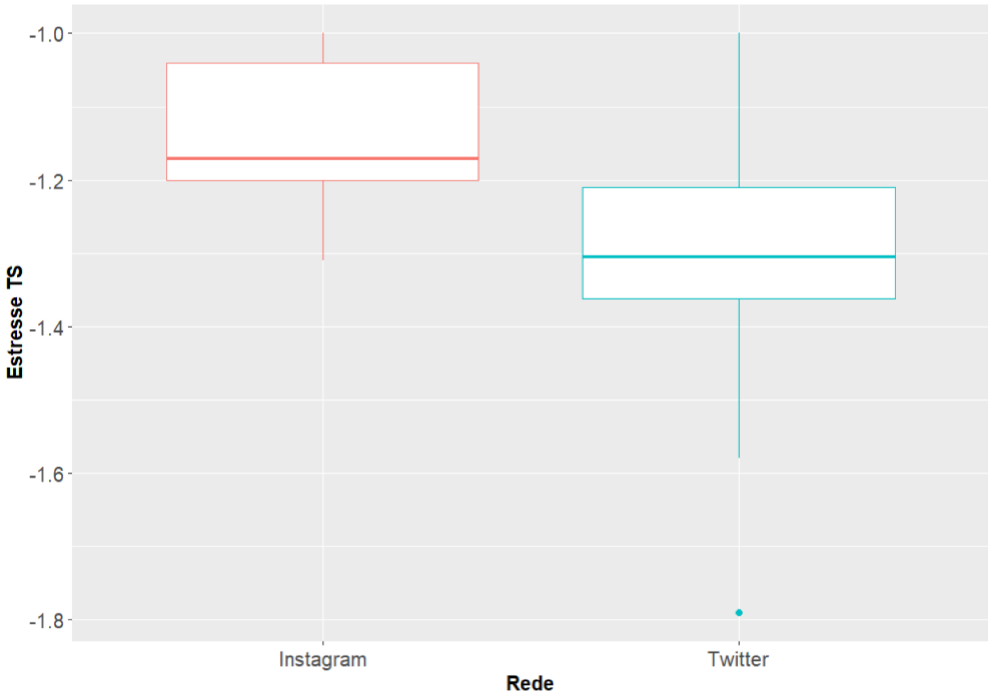

Figura 47 – Comparação entre Redes Fase 1: Estresse TSpt
Autoria própria

apresentaram maior relaxamento médio no texto.

O Capítulo 5 apresentou os resultados das avaliações das características extraídas dos

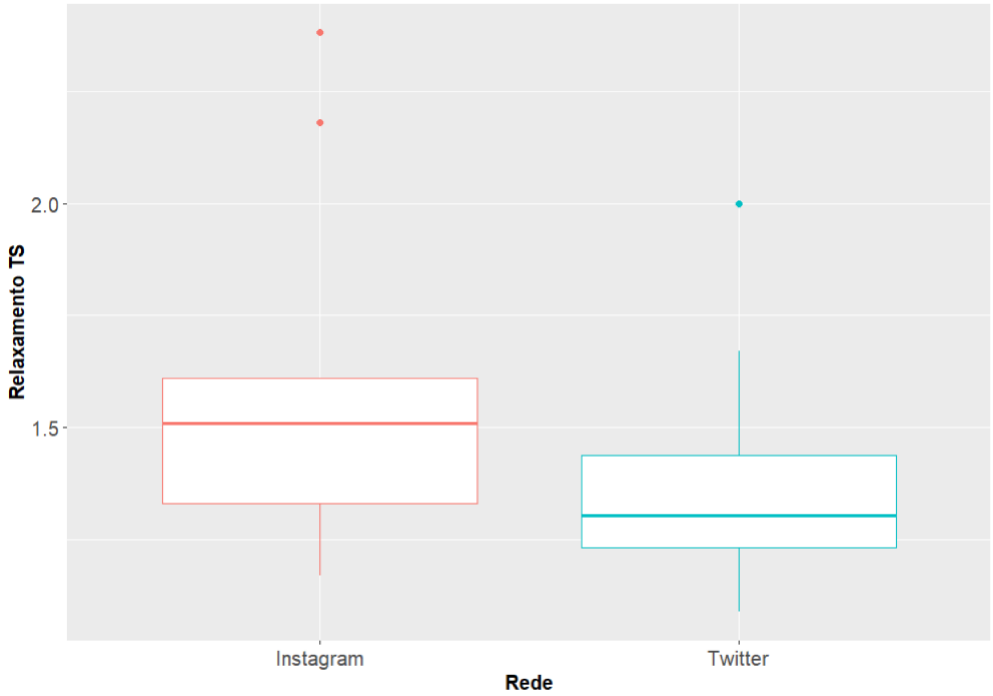

Figura 48 – Comparação entre Redes Fase 1: Relaxamento TSpt
Autoria própria

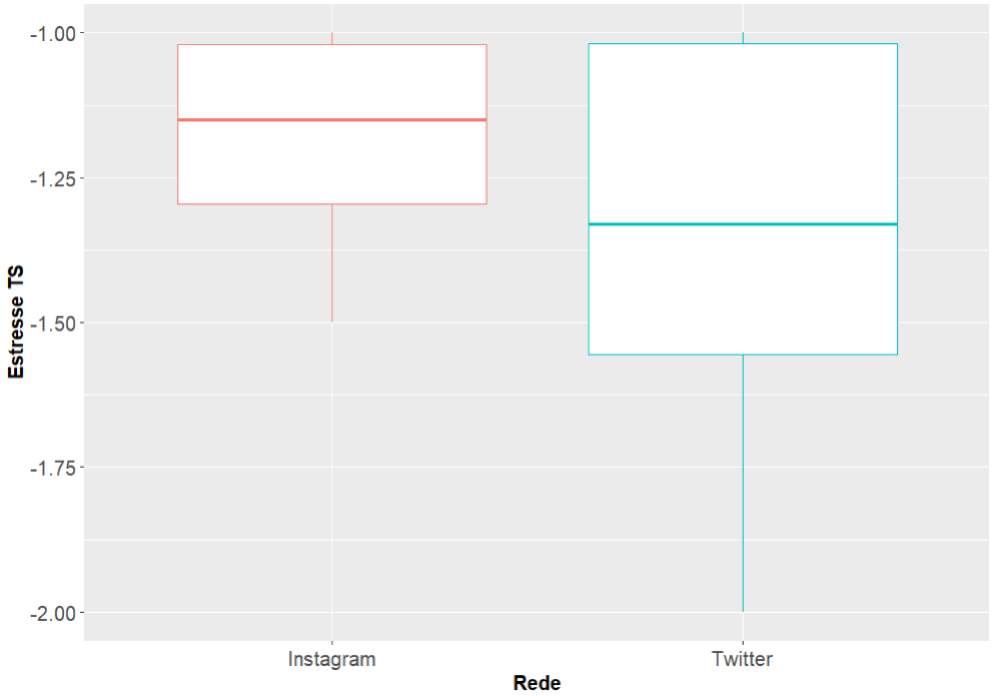

Figura 49 – Comparação entre Redes Fase 2: Estresse TS
Autoria própria

voluntários em cada fase da pesquisa e rede social. As características avaliadas foram importantes para diferenciar os usuários de cada rede e fase, embora os achados não

tenham se refletido em todos os cenários avaliados. No Capítulo 6 serão apresentadas as principais conclusões, limitações da pesquisa e sugestões de trabalhos futuros.

Capítulo 6

Conclusão

Este trabalho desenvolveu um método computacional que permite avaliar o impacto da COVID-19 sobre estresse o percebido, resiliência e o Transtorno de Estresse Pós-Traumático (TEPT) nos participantes usando dados de redes sociais. Para isso foram aplicados formulários autoavalitivos para a coleta de dados da comunidade acadêmica da UFU referentes à idade, função exercida, gênero e escores de estresse percebido e resiliência durante as medidas de distanciamento social em decorrência da pandemia da COVID-19. As mesmas características pessoais dos voluntários foram coletadas em conjunto com escores e classificação para a ocorrência de Transtorno de Estresse Pós-Traumático (TEPT) após a flexibilização das medidas de distanciamento social.

Para contrastar os dados capturados via formulários com as dados das redes sociais dos voluntários foram coletadas publicações realizadas no *X* e *Instagram*. As publicações coletadas foram classificadas através do TSpt de acordo com o nível de estresse e relaxamento na sentença.

A média de estresse e relaxamento no texto foi gerada para cada voluntário, juntamente com características de publicações em cada rede social. Por fim, 4 bases foram geradas resultantes do cruzamento entre os dados capturados através dos formulários, estresse e relaxamento médio nas publicações e estatísticas de publicação: 1) usuários do *X* na primeira fase da pesquisa; 2) usuários do *X* na segunda fase da pesquisa; 3) usuários do *Instagram* na primeira fase de pesquisa e 4) usuários do *Instagram* na segunda fase de pesquisa.

O processo final com a junção de dados dos formulários, estresse e relaxamento médio nos textos e estatísticas de publicação em cada rede e momento de coleta possibilitou criar uma base de dados anonimizada para a avaliação de dados brutos de redes sociais e de questionários de estresse e resiliência.

Para executar o TSpt foi necessário traduzir para o português o dicionário do algoritmo original, TensiStrength (TS), fornecido por Thelwall (2017). A validação do TSpt, um dos objetivos desde trabalho, foi apresentada na Seção 4.4. Após a validação foi realizada uma aplicação do algoritmo adaptado que permitiu distinguir os períodos com maior estresse

em textos publicados no *X* com a palavra *covid* em 7 períodos distintos da pandemia. As publicações coletadas também foram caracterizadas por meio de técnicas de extração de tópicos e nuvens de palavras.

A avaliação das bases de características dos voluntários permitiu enxergar diferentes relações entre os dados coletados de acordo com a fase de coleta e rede social considerada.

O estresse percebido, avaliado na primeira fase de coleta no *X*, apresentou uma tendência de que quanto maior o estresse no texto, menor o estresse percebido nos indivíduos. Nas demais bases não foram verificadas correlações entre o estresse percebido e o estresse médio no texto. O escore e classificação de TEPT não apresentaram correlações significativas com o estresse médio no texto.

No *X* o estresse percebido foi menor em indivíduos que realizaram publicações em um período mais distante da pandemia. O escore de TEPT foi mais elevado em perfis com mais amigos, seguidores e com descrições de perfis mais longas. O estresse médio no texto foi maior para indivíduos com maior uso de imagens nas publicações, com mais amigos, seguidores e perfis mais antigos. O relaxamento médio no texto foi maior em indivíduos com menor escoragem de recursos sociais e com publicações mais antigas. Os usuários com mais amigos demonstraram um menor fator observado de futuro planejado e coesão familiar. Os voluntários do gênero masculino apresentaram maior percepção de si mesmo e os do gênero feminino maior escore de TEPT. Os alunos demonstraram maior estresse percebido e 100% dos usuários com moda de publicação entre 18:00 e 06:00 avaliados apresentaram escore mínimo para a detecção de TEPT.

A avaliação dos usuários do *Instagram* permitiu observar que, após as flexibilizações das medidas de distanciamento social, o estresse médio no texto foi menor em usuários com publicações mais longas e em indivíduos com maior relaxamento médio no texto. Observou-se também um maior relaxamento médio no texto em usuários com publicações mais longas. Usuários com um maior número de seguidos apresentaram escores mais baixos para os fatores recursos sociais e coesão familiar. Aqueles com publicações mais antigas demonstram menores escores de futuro planejado. Na segunda coleta o relaxamento médio no texto foi maior em indivíduos com mais de 25 anos. Os fatores de recursos sociais e percepção de si mesmo, bem como o estresse e relaxamento médio no texto foram maiores entre os servidores em comparação aos alunos.

A comparação entre redes indicou maior estresse percebido e estresse médio no texto entre os usuários do *X* e maior relaxamento médio no texto entre os usuários do *Instagram*.

As avaliações do estresse percebido, de fatores de resiliência e do escore e classificação do TEPT em associação com o estresse e relaxamento no texto e estatísticas de publicação dos usuários nas diferentes redes sociais possibilitaram alcançar os objetivos previstos neste trabalho de criar modelos para descrever sintomas psicológicos de estresse e resiliência por meio das postagens em redes sociais.

Quanto a limitações enfrentadas na execução da pesquisa têm-se que:

- ❏ Para adaptar de forma eficiente o TS ao classificar o estresse em textos publicados em português uma base suficientemente representativa deveria ser rotulada manualmente por seres humanos para possibilitar a comparação da avaliação do estresse medido pelo algoritmo original, pelo algoritmo adaptado e pelos juízes humanos. Nesta etapa foi possível contar com a participação de somente 2 voluntários. Um número maior de juízes humanos permitiria utilizar uma amostra de tamanho maior, possibilitando análises mais avançadas e maior nível de significância dos resultados;

- ❏ A formação de uma amostra suficientemente grande para a coleta de dados das redes sociais de voluntários da UFU foi prejudicada. Para a formação das bases consideradas os voluntários deveriam aceitar os termos estabelecidos no Termo de Consentimento Livre e Esclarecido (TCLE) e ter algum perfil nas redes sociais elencadas com publicações realizadas em modo público. Mesmo após uma extensa divulgação da pesquisa em momentos distintos da pandemia não foi possível obter uma amostra suficientemente grande de voluntários habilitados e que atendessem aos critérios estabelecidos. Uma amostragem efetiva de usuários com publicações nas redes sociais consideradas permitiria ainda avaliar os mesmos usuários em momentos distintos da pandemia.

6.1 Principais Contribuições

O método apresentado neste trabalho pode contribuir com futuras pesquisas que visem aprofundar as discussões sobre a modelagem do estresse e resiliência dos usuários a partir de dados de redes sociais. Além disso pode fomentar futuras pesquisas que visem avaliar a relação entre o estresse do usuário e o estresse presente em publicações desses usuários nas redes sociais. Além disso, pesquisas futuras que busquem avaliar a relação entre outros fenômenos psicológicos pode se beneficiar da metodologia apresentada.

O cruzamento entre dados informados pelos voluntários através de formulários para a avaliação do estresse percebido e resiliência durante o período de isolamento social, do TEPT após a flexibilização dessas medidas e de dados das redes sociais dos voluntários avaliados permitiu observar que a pandemia da COVID-19 impactou de maneira diferente os usuários do *X* e *Instagram* em termos de estresse percebido, fatores de resiliência e TEPT.

A base de características avaliadas por fase da pesquisa e rede social indicou que existe uma relação entre o estresse médio no texto e o estresse percebido somente para o *X* na primeira fase de coleta. O TEPT foi significativo quando correlacionado com diferentes características dos usuários do *X*. O estresse e o relaxamento médio no texto apresentaram correlações significativas com diferentes características dos usuários de acordo com a rede considerada.

6.2 Trabalhos Futuros

Trabalhos futuros com uma amostra maior de publicações com classificações de estresse e relaxamento realizadas por juízes humanos forneceriam mais evidências da validade do TSpt ao classificar o estresse e relaxamento em textos escritos na língua portuguesa bem como comparar o algoritmo adaptado com algoritmos de classificação diversos.

Uma amostra representativa de usuários das redes sociais avaliadas também possibilitaria aprofundar ainda mais a discussão sobre a relação entre estresse e resiliência do usuário e o estresse presente no texto. Modelos de probabilidade de ocorrência de transtornos psicológicos poderiam ser estudados a partir da extração dos dados das redes sociais. Diferentes dados de publicação extraídos, além dos textuais, como imagens, áudios e vídeos publicados poderiam fornecer novas visões para entender os fenômenos psicológicos estudados em relação ao comportamento do usuário nas redes sociais.

6.3 Contribuições em Produção Bibliográfica

O processo de validação do TSpt e aplicação do algoritmo adaptado para classificar o estresse e relaxamento em publicações realizadas no X em diferentes momentos da pandemia em conjunto com técnicas de extração de tópicos para descrever as discussões relacionadas COVID-19 foi apresentado em Peres et al. (2023).

Referências

AFONSO, P. The impact of the covid-19 pandemic on mental health. **Acta medica portuguesa**, v. 33, n. 5, p. 356–357, 2020. Disponível em: <https://doi.org/10.20344/amp.13877>.

AGGARWAL, C. C.; ZHAI, C. **Mining text data**. Springer Science & Business Media, 2012. Disponível em: <https://doi.org/10.1007/978-1-4614-3223-4>.

AGUIAR, L. A. et al. Large-scale translation to enable response selection in low resource languages: A covid-19 chatbot experiment. In: SBC. **Anais do XXXVII Simpósio Brasileiro de Bancos de Dados**. 2022. p. 203–215. Disponível em: <https://doi.org/10.5753/sbbd.2022.224329>.

AHMAD, M.; AFTAB, S.; ALI, I. Sentiment analysis of tweets using svm. **Int. J. Comput. Appl**, v. 177, n. 5, p. 25–29, 2017. Disponível em: <https://doi.org/10.5120/ijca2017915758>.

ALLAHYARI, M. et al. A brief survey of text mining: Classification, clustering and extraction techniques. **arXiv preprint arXiv:1707.02919**, 2017.

ANGUITA, D. et al. The 'k'in k-fold cross validation. In: I6DOC. COM PUBL. **20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)**. [S.l.], 2012. p. 441–446.

BARROS, E. A. C.; MAZUCHELI, J. Um estudo sobre o tamanho e poder dos testes t-student e wilcoxon. **Acta Scientiarum. Technology**, Universidade Estadual de Maringá, v. 27, n. 1, p. 23–32, 2005. Disponível em: <https://doi.org/10.4025/actascitechnol.v27i1.1495>.

BATRINCA, B.; TRELEAVEN, P. C. Social media analytics: a survey of techniques, tools and platforms. **Ai & Society**, Springer, v. 30, n. 1, p. 89–116, 2015. Disponível em: <https://doi.org/10.1007/s00146-014-0549-4>.

BAYHAQY, A. et al. Sentiment analysis about e-commerce from tweets using decision tree, k-nearest neighbor, and naïve bayes. In: IEEE. **2018 international conference on orange technologies (ICOT)**. 2018. p. 1–6. Disponível em: <https://doi.org/10.1109/ICOT.2018.8705796>.

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. **Porto Alegre: Sociedade Brasileira de Computação**, p. 22, 2011.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **the Journal of machine Learning research**, JMLR. org, v. 3, p. 993–1022, 2003.

BOUMA, G. Normalized (pointwise) mutual information in collocation extraction. **Proceedings of GSCL**, Potsdam, v. 30, p. 31–40, 2009.

BROWN, S. M. et al. Stress and parenting during the global covid-19 pandemic. **Child abuse & neglect**, Elsevier, v. 110, p. 104699, 2020. Disponível em: <https://doi.org/10.1016/j.chiabu.2020.104699>.

BRUM, P. V. et al. A characterization of portuguese tweets regarding the covid-19 pandemic. In: SBC. **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. 2020. p. 177–184. Disponível em: <https://doi.org/10.5753/kdmile.2020.11974>.

CAIUBY, A. V. S. et al. Adaptação transcultural da versão brasileira da escala do impacto do evento-revisada (ies-r). **Cadernos de Saúde Pública**, SciELO Brasil, v. 28, p. 597–603, 2012. Disponível em: <https://doi.org/10.1590/S0102-311X2012000300019>.

CAMPAGNOLO, J. M.; DUARTE, D.; BIANCO, G. D. Topic coherence metrics: How sensitive are they? **Journal of Information and Data Management**, v. 13, n. 4, 2022. Disponível em: <https://doi.org/10.5753/jidm.2022.2181>.

CAMPOS, J. A. D. B. et al. Early psychological impact of the covid-19 pandemic in brazil: a national survey. **Journal of Clinical Medicine**, Multidisciplinary Digital Publishing Institute, v. 9, n. 9, p. 2976, 2020. Disponível em: <https://doi.org/10.3390/jcm9092976>.

CARVALHO, V. D.; TEODORO, M. L. M.; BORGES, L. de O. Escala de resiliência para adultos: aplicação entre servidores públicos. **Avaliação Psicológica**, Instituto Brasileiro de Avaliação Psicológica, v. 13, n. 2, p. 287–295, 2014.

CASTELLI, L. et al. The spread of covid-19 in the italian population: anxiety, depression, and post-traumatic stress symptoms. **Canadian journal of psychiatry. Revue canadienne de psychiatrie**, SAGE Publications, v. 65, n. 10, p. 731, 2020. Disponível em: <https://doi.org/10.1177/0706743720938598>.

CHAI, T.; DRAXLER, R. R. Root mean square error (rmse) or mean absolute error (mae). **Geoscientific model development discussions**, v. 7, n. 1, p. 1525–1534, 2014. Disponível em: <https://doi.org/10.5194/gmdd-7-1525-2014>.

CHAKRABORTY, K.; BHATTACHARYYA, S.; BAG, R. A survey of sentiment analysis from social media data. **IEEE Transactions on Computational Social Systems**, IEEE, v. 7, n. 2, p. 450–464, 2020. Disponível em: <https://doi.org/10.1109/TCSS.2019.2956957>.

CHONG, W. Y.; SELVARETNAM, B.; SOON, L.-K. Natural language processing for sentiment analysis: an exploratory analysis on tweets. In: IEEE. **2014 4th International Conference on Artificial Intelligence with Applications**

in Engineering and Technology. 2014. p. 212–217. Disponível em: <https://doi.org/10.1109/ICAIET.2014.43>.

CIRIBELI, J. P.; PAIVA, V. H. P. Redes e mídias sociais na internet: realidades e perspectivas de um mundo conectado. **Revista Mediação**, 2011.

COHEN, I. et al. Pearson correlation coefficient. **Noise reduction in speech processing**, Springer, p. 1–4, 2009. Disponível em: <https://doi.org/10.1007/978-3-642-00296-0_5>.

COHEN, S.; KAMARCK, T.; MERMELSTEIN, R. A global measure of perceived stress. **Journal of health and social behavior**, JSTOR, p. 385–396, 1983. Disponível em: <https://doi.org/10.2307/2136404>.

COUTO, J. et al. Automatic detection of covid-19 misinformation in brazil. In: **Proceedings of the 37th Brazilian Symposium on Databases**. Porto Alegre, RS, Brasil: SBC, 2022. p. 164–176. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/21804>.

CULLEN, W.; GULATI, G.; KELLY, B. Mental health in the covid-19 pandemic. **QJM: An International Journal of Medicine**, Oxford University Press, v. 113, n. 5, p. 311–312, 2020. Disponível em: <https://doi.org/10.1093/qjmed/hcaa110>.

EBELING, R. et al. The effect of political polarization on social distance stances in the brazilian covid-19 scenario. **Journal of Information and Data Management**, v. 12, n. 1, 2021. Disponível em: <https://doi.org/10.5753/jidm.2021.1889>.

_____. Quarenteners vs. cloroquiners: a framework to analyze the effect of political polarization on social distance stances. In: SBC. **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. [S.l.], 2020. p. 89–96.

EGGER, R.; YU, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. **Frontiers in sociology**, Frontiers Media SA, v. 7, 2022. Disponível em: <https://doi.org/10.3389/fsoc.2022.886498>.

Facebook. **Acreditamos no potencial das pessoas quando elas se unem.** 2021. <https://about.fb.com/br/>, Acesso em 21.08.2021.

Facebook for Developers. **Graph API do Instagram.** 2021. <https://developers.facebook.com/docs/instagram-api>, Acesso em 21.08.2021.

_____. **Learn about the programs educating and connecting innovators.** 2021. <https://developers.facebook.com/>, Acesso em 21.08.2021.

FALEIROS, T. d. P.; LOPES, A. d. A. et al. Modelos probabilísticos de tópicos: desvendando o latent dirichlet allocation. São Carlos, SP, Brasil., 2016.

FEINERER, I. **A text mining framework in R and its applications**. Tese (Doutorado) — WU Vienna University of Economics and Business, 2008.

FEITOSA, M. C. et al. Uso de escalas/testes como instrumentos de coleta de dados em pesquisas quantitativas em enfermagem. **SANARE-Revista de Políticas Públicas**, v. 13, n. 2, 2014.

FERREIRA, R. J.; BUTTELL, F.; CANNON, C. Covid-19: Immediate predictors of individual resilience. **Sustainability**, Multidisciplinary Digital Publishing Institute, v. 12, n. 16, p. 6495, 2020. Disponível em: <https://doi.org/10.3390/su12166495>.

FILHO, D. B. F.; JÚNIOR, J. A. S. Desvendando os mistérios do coeficiente de correlação de pearson (r). **Revista Política Hoje**, v. 18, n. 1, p. 115–146, 2009.

FITRI, V. A.; ANDRESWARI, R.; HASIBUAN, M. A. Sentiment analysis of social media twitter with case of anti-lgbt campaign in indonesia using naïve bayes, decision tree, and random forest algorithm. **Procedia Computer Science**, Elsevier, v. 161, p. 765–772, 2019. Disponível em: <https://doi.org/10.1016/j.procs.2019.11.181>.

FREITAS, D. P.; BORGES, M. R.; CARVALHO, P. V. R. d. A conceptual framework for developing solutions that organise social media information for emergency response teams. **Behaviour & Information Technology**, Taylor & Francis, v. 39, n. 3, p. 360–378, 2020. Disponível em: <https://doi.org/10.1080/0144929X.2019.1621933>.

GALVÃO-COELHO, N. L.; SILVA, H. P. A.; SOUSA, M. B. C. d. Resposta ao estresse: Ii. resiliência e vulnerabilidade. **Estudos de Psicologia (Natal)**, SciELO Brasil, v. 20, p. 72–81, 2015. Disponível em: <https://doi.org/10.5935/1678-4669.20150009>.

GROOTENDORST, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. **arXiv preprint arXiv:2203.05794**, 2022.

HABIBABADI, S. K.; HAGHIGHI, P. D. Topic modelling for identification of vaccine reactions in twitter. In: **Proceedings of the Australasian Computer Science Week Multiconference**. [s.n.], 2019. p. 1–10. Disponível em: <https://doi.org/10.1145/3290688.3290735>.

HAROON, M. Comparative analysis of stemming algorithms for web text mining. **no. September**, p. 20–25, 2018. Disponível em: <https://doi.org/10.5815/ijmecs.2018.09.03>.

HEARST, M. What is text mining. **SIMS, UC Berkeley**, v. 5, 2003.

HOTHO, A.; NÜRNBERGER, A.; PAASS, G. A brief survey of text mining. **Journal for Language Technology and Computational Linguistics**, v. 20, n. 1, p. 19–62, 2005. Disponível em: <https://doi.org/10.21248/jlcl.20.2005.68>.

HUQ, M. R.; ALI, A.; RAHMAN, A. Sentiment analysis on twitter data using knn and svm. **International Journal of Advanced Computer Science and Applications**, v. 8, n. 6, p. 19–25, 2017. Disponível em: <https://doi.org/10.14569/IJACSA.2017.080603>.

IKONOMAKIS, M.; KOTSIANTIS, S.; TAMPAKAS, V. Text classification using machine learning techniques. **WSEAS transactions on computers**, Citeseer, v. 4, n. 8, p. 966–974, 2005.

IMRAN, A. S. et al. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. **IEEE Access**, IEEE, v. 8, p. 181074–181090, 2020. Disponível em: <https://doi.org/10.1109/ACCESS.2020.3027350>.

INDRA, S.; WIKARSA, L.; TURANG, R. Using logistic regression method to classify tweets into the selected topics. In: IEEE. **2016 international conference on advanced computer science and information systems (icacsis)**. 2016. p. 385–390. Disponível em: <https://doi.org/10.1109/ICACSIS.2016.7872727>.

Instagram. **Aproximando você das pessoas e das coisas que ama.** 2021. <https://about.instagram.com/pt-br>, Acesso em 21.08.2021.

JIVANI, A. G. et al. A comparative study of stemming algorithms. **Int. J. Comp. Tech. Appl**, v. 2, n. 6, p. 1930–1938, 2011.

JU, Z.; WANG, J.; ZHU, F. Named entity recognition from biomedical text using svm. In: IEEE. **2011 5th international conference on bioinformatics and biomedical engineering**. 2011. p. 1–4. Disponível em: <https://doi.org/10.1109/icbbe.2011.5779984>.

Julia Silge and David Robinson. **Topic modeling**. 2022. <https://www.tidytextmining.com/topicmodeling.html>, Acesso em 10.01.2022.

KAMEO, S. Y. et al. Estresse pós-traumático em casos confirmados de covid-19: estudo observacional. **CONTRIBUCIONES A LAS CIENCIAS SOCIALES**, v. 16, n. 8, p. 13704–13718, 2023. Disponível em: <https://doi.org/10.55905/revconv.16n.8-283>.

KANNAN, S. et al. Preprocessing techniques for text mining. **International Journal of Computer Science & Communication Networks**, v. 5, n. 1, p. 7–16, 2014.

KAUR, J.; BUTTAR, P. K. Stopwords removal and its algorithms based on different methods. **International Journal of Advanced Research in Computer Science**, v. 10, n. 5, 2018. Disponível em: <https://doi.org/10.26483/ijarcs.v9i5.6301>.

KILLGORE, W. D. et al. Psychological resilience during the covid-19 lockdown. **Psychiatry research**, Elsevier, v. 291, p. 113216, 2020. Disponível em: <https://doi.org/10.1016/j.psychres.2020.113216>.

KOWSARI, K. et al. Text classification algorithms: A survey. **Information**, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 150, 2019. Disponível em: <https://doi.org/10.3390/info10040150>.

KUMAR, L.; BHATIA, P. K. Text mining: concepts, process and applications. **Journal of Global Research in Computer Science**, v. 4, n. 3, p. 36–39, 2013.

KUMARI, U.; SHARMA, A. K.; SONI, D. Sentiment analysis of smart phone product review using svm classification technique. In: IEEE. **2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)**. 2017. p. 1469–1474. Disponível em: <https://doi.org/10.1109/ICECDS.2017.8389689>.

LAKSONO, R. A. et al. Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes. In: IEEE. **2019 12th International Conference on Information & Communication Technology and System (ICTS)**. 2019. p. 49–54. Disponível em: <https://doi.org/10.1109/ICTS.2019.8850982>.

LI, L. et al. Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. **IEEE Transactions on Computational Social Systems**, IEEE, v. 7, n. 2, p. 556–562, 2020. Disponível em: <https://doi.org/10.1109/TCSS.2020.2980007>.

LIU, Y. Y. et al. **The advances of stemming algorithms in text analysis from 2013 to 2018**. Tese (Doutorado) — University of Pretoria, 2019.

LUFT, C. D. B. et al. Versão brasileira da escala de estresse percebido: tradução e validação para idosos. **Revista de Saúde Pública**, SciELO Public Health, v. 41, n. 4, p. 606–615, 2007. Disponível em: <https://doi.org/10.1590/S0034-89102007000400015>.

MAJUMDER, S.; AICH, A.; DAS, S. Sentiment analysis of people during lockdown period of covid-19 using svm and logistic regression analysis. **Available at SSRN 3801039**, 2021. Disponível em: <https://doi.org/10.2139/ssrn.3801039>.

MATSUBARA, E. T.; MARTINS, C. A.; MONARD, M. C. Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. **Techinical Report**, v. 209, n. 4, p. 10–11, 2003.

MCCALLUM, A.; NIGAM, K. et al. A comparison of event models for naive bayes text classification. In: CITESEER. **AAAI-98 workshop on learning for text categorization**. [S.l.], 1998. v. 752, n. 1, p. 41–48.

MCINNES, L.; HEALY, J.; ASTELS, S. hdbscan: Hierarchical density based clustering. **J. Open Source Softw.**, v. 2, n. 11, p. 205, 2017. Disponível em: <https://doi.org/10.21105/joss.00205>.

MCINNES, L.; HEALY, J.; MELVILLE, J. Umap: Uniform manifold approximation and projection for dimension reduction. **arXiv preprint arXiv:1802.03426**, 2018. Disponível em: <https://doi.org/10.21105/joss.00861>.

MELO, T. de; FIGUEIREDO, C. M. A first public dataset from brazilian twitter and news on covid-19 in portuguese. **Data in brief**, Elsevier, v. 32, p. 106179, 2020. Disponível em: <https://doi.org/10.1016/j.dib.2020.106179>.

Miguel Garcia. **How to Make a Twitter Bot in Python With Tweepy.** 2021. <https://realpython.com/twitter-bot-python-tweepy/#what-is-tweepy>, Acesso em 21.08.2021.

MOHBEY, K. K.; TIWARI, S. Preprocessing and morphological analysis in text mining. **International Journal of electronics communication and computer engineering, ISSN**, 2011.

NOVENDRI, R. et al. Sentiment analysis of youtube movie trailer comments using naïve bayes. **Bulletin of Computer Science and Electrical Engineering**, v. 1, n. 1, p. 26–32, 2020. Disponível em: <https://doi.org/10.25008/bcsee.v1i1.5>.

OPAS. **Histórico da pandemia de COVID-19**. 2023. Disponível em: <https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19>. Acesso em: 17 out 2023.

Organização Mundial da Saúde. **Coronavirus disease (COVID-19)**. 2021. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19>, Acesso em 28.11.2021.

ÖZTÜRK, N.; AYVAZ, S. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. **Telematics and Informatics**, Elsevier, v. 35, n. 1, p. 136–147, 2018. Disponível em: <https://doi.org/10.1016/j.tele.2017.10.006>.

PAIVA, G. P. M. et al. Covid 19: O que sentem os brasileiros de acordo com o twitter? **Journal of Health Informatics**, v. 12, 2020.

PAULA, B. C.; OLIVEIRA, G. P.; MORO, M. M. Mood analysis during the covid-19 pandemic in brazil through music. In: SBC. **Anais Estendidos do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web**. 2022. p. 53–56. Disponível em: <https://doi.org/10.5753/webmedia_estendido.2022.227063>.

PECONGA, E. K. et al. Resilience is spreading: Mental health within the covid-19 pandemic. **Psychological Trauma: Theory, Research, Practice, and Policy**, Educational Publishing Foundation, v. 12, n. S1, p. S47, 2020. Disponível em: <https://doi.org/10.1037/tra0000874>.

PEDROZO-PUPO, J. C.; PEDROZO-CORTÉS, M. J.; CAMPO-ARIAS, A. Perceived stress associated with covid-19 epidemic in colombia: an online survey. **Cadernos de saude publica**, SciELO Brasil, v. 36, 2020. Disponível em: <https://doi.org/10.1590/0102-311x00090520>.

PERES, D. et al. Análise do estresse e tópicos discutidos no twitter durante a pandemia da covid-19 no brasil. In: **Anais do XII Brazilian Workshop on Social Network Analysis and Mining**. Porto Alegre, RS, Brasil: SBC, 2023. p. 43–54. ISSN 2595-6094. Disponível em: <https://doi.org/10.5753/brasnam.2023.229752>.

PINHEIRO, D. P. N. A resiliência em discussão. **Psicologia em estudo**, SciELO Brasil, v. 9, p. 67–75, 2004. Disponível em: <https://doi.org/10.1590/S1413-73722004000100009>.

PLACKETT, R. L. Karl pearson and the chi-squared test. **International statistical review/revue internationale de statistique**, JSTOR, p. 59–72, 1983. Disponível em: <https://doi.org/10.2307/1402731>.

PLISSON, J. et al. A rule based approach to word lemmatization. In: **Proceedings of IS**. [S.l.: s.n.], 2004. v. 3, p. 83–86.

PRASTYO, P. H. et al. Tweets responding to the indonesian government's handling of covid-19: Sentiment analysis using svm with normalized poly kernel. **Journal of Information Systems Engineering and Business Intelligence**, v. 6, n. 2, p. 112–122, 2020. Disponível em: <https://doi.org/10.20473/jisebi.6.2.112-122>.

PRAVEEN, S.; ITTAMALLA, R.; DEEPAK, G. Analyzing indian general public's perspective on anxiety, stress and trauma during covid-19-a machine learning study of 840,000 tweets. **Diabetes & Metabolic Syndrome: Clinical Research & Reviews**, Elsevier, v. 15, n. 3, p. 667–671, 2021. Disponível em: <https://doi.org/10.1016/j.dsx.2021.03.016>.

RAMADHAN, W.; NOVIANTY, S. A.; SETIANINGSIH, S. C. Sentiment analysis using multinomial logistic regression. In: IEEE. **2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)**. 2017. p. 46–49. Disponível em: <https://doi.org/10.1109/ICCEREC.2017.8226700>.

RAMOS, J. et al. Using tf-idf to determine word relevance in document queries. In: CITESEER. **Proceedings of the first instructional conference on machine learning**. [S.l.], 2003. v. 242, n. 1, p. 29–48.

RATHI, M. et al. Sentiment analysis of tweets using machine learning approach. In: IEEE. **2018 Eleventh international conference on contemporary computing (IC3)**. 2018. p. 1–3. Disponível em: <https://doi.org/10.1109/IC3.2018.8530517>.

RAULJI, J. K.; SAINI, J. R. Stop-word removal algorithm and its implementation for sanskrit language. **International Journal of Computer Applications**, Foundation of Computer Science, v. 150, n. 2, p. 15–17, 2016. Disponível em: <https://doi.org/10.5120/ijca2016911462>.

REHMAN, U. et al. Depression, anxiety and stress among indians in times of covid-19 lockdown. **Community mental health journal**, Springer, v. 57, n. 1, p. 42–48, 2021. Disponível em: <https://doi.org/10.1007/s10597-020-00664-x>.

REUTER, C.; STIEGLITZ, S.; IMRAN, M. Social media in conflicts and crises. **Behaviour & Information Technology**, Taylor & Francis, v. 39, n. 3, p. 241–251, 2020. Disponível em: <https://doi.org/10.1080/0144929X.2019.1629025>.

RODRIGUES, D. C. C. B. A prevalência de transtorno depressivo, transtorno de ansiedade generalizada e de transtorno de estresse pós-traumático nos bombeiros militares do cbmdf após a pandemia de covid-19. 2023.

SAHAYAK, V.; SHETE, V.; PATHAN, A. Sentiment analysis on twitter data. **International Journal of Innovative Research in Advanced Engineering (IJIRAE)**, v. 2, n. 1, p. 178–183, 2015.

SANTOS, L. R. J.; MARINHO, L. B.; CAMPELO, C. E. C. Uniting politics and pandemic: a social network analysis on the covid parliamentary commission of inquiry in brazil. In: SBC. **Anais do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web**. [S.l.], 2022. p. 105–113.

SANTOS, M. H. dos; PEREIRA, F. S. Describing covid-19 pandemic by means of tweets from official entities in brazil. In: SBC. **Anais do XII Brazilian Workshop on Social Network Analysis and Mining**. 2023. p. 175–186. Disponível em: <https://doi.org/10.5753/brasnam.2023.230780>.

SARICA, S.; LUO, J. Stopwords in technical language processing. **Plos one**, Public Library of Science San Francisco, CA USA, v. 16, n. 8, p. e0254937, 2021. Disponível em: <https://doi.org/10.1371/journal.pone.0254937>.

SCHONLAU, M.; GUENTHER, N. Text mining using n-grams. **Schonlau, M., Guenther, N. Sucholutsky, I. Text mining using n-gram variables. The Stata Journal**, v. 17, n. 4, p. 866–881, 2017.

SHAMRAT, M. et al. Sentiment analysis on twitter tweets about covid-19 vaccines using nlp and supervised knn classification algorithm. **Indonesian Journal of Electrical Engineering and Computer Science**, v. 23, n. 1, p. 463–470, 2021. Disponível em: <https://doi.org/10.11591/ijeecs.v23.i1.pp463-470>.

SHARMA, D.; CSE, M. Stemming algorithms: a comparative study and their analysis. **International Journal of Applied Information Systems**, v. 4, n. 3, p. 7–12, 2012. Disponível em: <https://doi.org/10.5120/ijais12-450655>.

SHEVLIN, M. et al. Anxiety, depression, traumatic stress and covid-19-related anxiety in the uk general population during the covid-19 pandemic. **BJPsych Open**, Cambridge University Press, v. 6, n. 6, 2020. Disponível em: <https://doi.org/10.1192/bjo.2020.109>.

SHI, L. et al. Rough set based decision tree ensemble algorithm for text classification. **Journal of Computational Information Systems**, Citeseer, v. 6, n. 1, p. 89–95, 2010.

SILVA, C.; RIBEIRO, B. On text-based mining with active learning and background knowledge using svm. **Soft Computing**, Springer, v. 11, n. 6, p. 519–530, 2007. Disponível em: <https://doi.org/10.1007/s00500-006-0080-8>.

SILVA, H. A.; SOUZA, D. F.; OLIVEIRA, L. C. Semctra: a multilayer specialized system for covid-19 remote triage, resource allocation and teleconsultation. In: **Proceedings of the Brazilian Symposium on Multimedia and the Web**. [s.n.], 2022. p. 267–271. Disponível em: <https://doi.org/10.1145/3539637.3556996>.

SILVA, L. O. T. d. Transtorno de estresse pós-traumático entre universitários brasileiros no retorno das aulas presenciais durante a pandemia de covid-19. Universidade Federal de São Paulo, 2023.

SONG, F.; LIU, S.; YANG, J. A comparative study on text representation schemes in text categorization. **Pattern analysis and applications**, Springer, v. 8, n. 1, p. 199–209, 2005. Disponível em: <https://doi.org/10.1007/s10044-005-0256-3>.

SOUSA, A.; BECKER, K. Comparing positions for/against covid vaccination in the united states of america and brazil. In: **Proceedings of the 37th Brazilian Symposium on Databases**. Porto Alegre, RS, Brasil: SBC, 2022. p. 65–77. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/21796>.

SOUZA, Q.; QUANDT, C. Metodologia de análise de redes sociais. **O tempo das redes. São Paulo: Perspectiva**, p. 31–63, 2008.

SOUZA, V. B.; NOBRE, J.; BECKER, K. Characterization of anxiety, depression, and their comorbidity from texts of social networks. 2020. Disponível em: <https://doi.org/10.5753/sbbd.2020.13630>.

SUMATHY, K.; CHIDAMBARAM, M. Text mining: concepts, applications, tools and issues-an overview. **International Journal of Computer Applications**, Citeseer, v. 80, n. 4, 2013. Disponível em: <https://doi.org/10.5120/13851-1685>.

TALLARIDA, R. J. et al. Chi-square test. **Manual of pharmacologic calculations: With computer programs**, Springer, p. 140–142, 1987. Disponível em: <https://doi.org/10.1007/978-1-4612-4974-0_43>.

TAN, S. An effective refinement strategy for knn text classifier. **Expert Systems with Applications**, Elsevier, v. 30, n. 2, p. 290–298, 2006. Disponível em: <https://doi.org/10.1016/j.eswa.2005.07.019>.

TAYLOR, S. et al. Development and initial validation of the covid stress scales. **Journal of Anxiety Disorders**, Elsevier, v. 72, p. 102232, 2020. Disponível em: <https://doi.org/10.1016/j.janxdis.2020.102232>.

THELWALL, M. Tensistrength: Stress and relaxation magnitude detection for social media texts. **Information Processing & Management**, Elsevier, v. 53, n. 1, p. 106–121, 2017. Disponível em: <https://doi.org/10.1016/j.ipm.2016.06.009>.

THELWALL, M. et al. Sentiment strength detection in short informal text. **Journal of the American society for information science and technology**, Wiley Online Library, v. 61, n. 12, p. 2544–2558, 2010. Disponível em: <https://doi.org/10.1002/asi.21416>.

TONG, Z.; ZHANG, H. A text mining research based on lda topic modelling. In: **International Conference on Computer Science, Engineering and Information Technology**. [s.n.], 2016. p. 201–210. Disponível em: <https://doi.org/10.5121/csit.2016.60616>.

Tweepy. **An easy-to-use Python library for accessing the Twitter API.** 2021. <https://www.tweepy.org/>, Acesso em 21.08.2021.

Twitter. **Twitter is what's happening and what people are talking about right now.** 2021. <https://about.twitter.com/en>, Acesso em 21.08.2021.

Twitter API. **Programmatically analyze, learn from, and engage with the conversation on Twitter.** 2021. <https://developer.twitter.com/en/docs/twitter-api>, Acesso em 21.08.2021.

TYAGI, A.; SHARMA, N. Sentiment analysis using logistic regression and effective word score heuristic. **International Journal of Engineering and Technology (UAE)**, v. 7, p. 20–23, 2018. Disponível em: <https://doi.org/10.14419/ijet.v7i2.24.11991>.

VERMA, T.; RENU, R.; GAUR, D. Tokenization and filtering process in rapidminer. **International Journal of Applied Information Systems**, v. 7, n. 2, p. 16–18, 2014. Disponível em: <https://doi.org/10.5120/ijais14-451139>.

VIJAYARANI, S.; JANANI, R. et al. Text mining: open source tokenization tools-an analysis. **Advanced Computational Intelligence: An International Journal (ACII)**, v. 3, n. 1, p. 37–47, 2016. Disponível em: <https://doi.org/10.5121/acii.2016.3104>.

WANG, T. et al. Covid-19 sensing: negative sentiment analysis on social media in china via bert model. **Ieee Access**, IEEE, v. 8, p. 138162–138169, 2020. Disponível em: <https://doi.org/10.1109/ACCESS.2020.3012595>.

WATERLOO, S. F. et al. Norms of online expressions of emotion: Comparing facebook, twitter, instagram, and whatsapp. **New media & society**, Sage Publications Sage UK: London, England, v. 20, n. 5, p. 1813–1831, 2018. Disponível em: <https://doi.org/10.1177/1461444817707349>.

WEBSTER, J. J.; KIT, C. Tokenization as the initial phase in nlp. In: **COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics**. [s.n.], 1992. Disponível em: <https://doi.org/10.3115/992424.992434>.

WITTEN, I. H. **Text Mining.** 2004.

WONGKAR, M.; ANGDRESEY, A. Sentiment analysis using naive bayes algorithm of the data crawler: Twitter. In: IEEE. **2019 Fourth International Conference on Informatics and Computing (ICIC)**. 2019. p. 1–5. Disponível em: <https://doi.org/10.1109/ICIC47613.2019.8985884>.

ZHU, Y. et al. The impact of social distancing during covid-19: A conditional process model of negative emotions, alienation, affective disorders, and post-traumatic stress disorder. **Journal of affective disorders**, Elsevier, v. 281, p. 131–137, 2021. Disponível em: <https://doi.org/10.1016/j.jad.2020.12.004>.

Apêndices

APÊNDICE **A**

Tradução do Manual para Rotulação

A.1 Estresse

Aqui, estamos interessados em identificar as publicações que, direta ou indiretamente, indicam que o usuário ou outra pessoa estão experenciando um sentimento de pressão ou tensão.

Estresse: codifique cada publicação de acordo com o grau em que ela diretamente descreve estresse, preocupação, medo ou raiva. A escala de estresse é:

[nenhuma descrição de estresse] -1 · -2 · -3 · -4 · -5 [descrição de altos níveis de estresse]

- ❏ Aloque -1 se a publicação não contém nenhuma indicação de estresse, medo, preocupação ou raiva.
- ❏ Aloque -5 se o comentário contém altos níveis de estresse, medo, preocupação ou raiva.
- ❏ Aloque um número entre -2 e -4 se o comentário contém níveis intermediários de estresse, medo, preocupação ou raiva, mas não em altos níveis. Use o seu julgamento a respeito do exato nível presente.

Exemplos:

- ❏ Declarações com teor de estresse: Eu estou estressado(a) no momento; Ele está sob muita pressão.
- ❏ Declarações com conteúdo de raiva, medo ou preocupação: Eu estava com raiva por receber tanto spam; nós estamos preocupados com o resultado das eleições.

Estressores: codifique cada publicação de acordo com o grau em que ela diretamente descreve coisas ou situações que podem ser gatilhos para o estresse. A escala de estressores é:

[nenhuma descrição de estressores] -1 · -2 · -3 · -4 · -5 [descrição de fortes estressores]

- ❏ Aloque -1 se a publicação não contém estressores.
- ❏ Aloque -5 se o comentário apresenta estressores bem definidos que provavelmente causarão altos níveis de estresse.
- ❏ Aloque um número entre -2 e -4 se o comentário contém estressores em um grau intermediário. Use o seu julgamento a respeito do exato nível de estresse presente.

Exemplos:

- ❏ Situações estressantes (estressores): Nós estamos atrasados; Eu estou sobrecarregado(a); eles(as) estão presos(as) em um engarrafamento; ela está com pressa; eles(as) estão muito doentes; Eu tenho que terminar de cozinhar antes das 6; Eu corri para o ônibus; Eu estou muito ocupado(a); ela está com pressa.
- ❏ Coisas ruins que podem ser causas de estresse: Este carro é terrível.

Sentimentos negativos: codifique cada publicação de acordo com o grau em que ele descreve ou sugere sentimentos negativos que não sejam estresse, medo, preocupação ou raiva. A escala de sentimentos negativos é:

- ❏ Aloque -1 se a publicação não contém indicação de sentimentos negativos.
- ❏ Aloque -2 se o comentário contém indicação moderada de sentimentos negativos.
- ❏ Aloque -3 se a publicação contém uma forte indicação de sentimento negativos.

Exemplos:

- ❏ Sentimentos ou emoções negativas: Eu odeio Mark; Tony está triste; Eu estou deprimido(a).
- ❏ Opiniões negativas: Eu não gosto da câmera;
- ❏ Insultos (normalmente pontue estes com -2): Sarah é uma idiota; Nigel é estúpido.
- ❏ Contextos sugerindo emoções negativas: Ele está chorando; Nós resmungamos.

O número mais importante é a pontuação geral de estresse – as 3 (três) categorias separadas existem, principalmente, para guiar os seus julgamentos. Não se preocupe com as pontuações individuais se você não tiver certeza de qual das (categorias) acima se aplica – contando que a pontuação geral seja a mesma, não importa.

A.2 Relaxamento

Relaxamento se refere à um estado corporal de prontidão reduzida para atividades, ou um sentimento de paz ou calma. Dormir e praticar yoga são exemplos de extremo relaxamento. Aqui nós estamos interessados em identificar relaxamento, ausência de estresse (i.e., informações que não só não contém indicações de estresse, mas também contém sugestões de que o estresse está ausente, como "Hoje é um bom dia."), assim como confiança na habilidade de lidar com o estresse (e.g., "Eu vou vencer a corrida!").

Relaxamento: codifique cada publicação de acordo com o grau em que ele descreve um estado relaxado. A escala de relaxação é:

[nenhuma descrição de um estado relaxado] $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5$ [descrição de um estado altamente relaxado]

- ❏ Aloque 1 se a publicação não contém descrição alguma de um estado relaxado.
- ❏ Aloque 5 se a publicação descreve um estado altamente relaxado.
- ❏ Aloque um número entre 2 e 4 se o comentário descreve um estado de relaxamento em níveis intermediários. Use o seu julgamento a respeito da exata potência da relaxação.

Exemplos: Eu estou muito relaxado(a); eles(as) estão dormindo; Eu me sinto em paz; eles(as) estão calmos(as).

Relaxadores: codifique cada publicação de acordo com o grau em que ela descreve atividades ou contextos relaxantes. A escala de relaxadores é:

[nenhuma descrição de atividades relaxantes] $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5$ [descrição de atividades altamente relaxantes]

- ❏ Aloque 1 se a publicação não contém descrição alguma de atividades relaxantes.
- ❏ Aloque 5 se a publicação descreve atividades altamente relaxantes.
- ❏ Aloque um número entre 2 e 4 se a publicação descreve atividades relaxantes em níveis intermediários. Use o seu julgamento a respeito da exata potência.

Exemplos:

- ❏ Atividades relaxantes: Eu saí para dar uma volta; nós vagueamos pela cidade; ela está lendo um bom livro; nós estávamos meditando; ela cochilou.
- ❏ Contextos relaxantes: Eu estou em uma sauna; nós temos bastante tempo antes do trem partir; é um dia tranquilo; alguma música lenta está tocando silenciosamente no fundo.

Sentimentos positivos: codifique cada publicação de acordo com o grau em que ele descreve ou sugere sentimentos positivos. A escala de sentimentos positivos é:

[nenhuma indicação de sentimentos positivos] $1 \cdot 2 \cdot 3$ [forte indicação de sentimentos positivos]

- ❏ Aloque 1 se a publicação não contém indicação de sentimentos positivos.
- ❏ Aloque 2 se a publicação contém indicação moderada de sentimentos positivos.
- ❏ Aloque 3 se a publicação contém indicação de fortes sentimentos positivos.

Exemplos:

- ❏ Descrições de emoções ou sentimentos positivos: Eu estou feliz; Eu me sinto maravilhoso(a).
- ❏ Opiniões positivas: O carro é excelente; isso é encantador.
- ❏ Situações ou contextos associados a sentimentos positivos: Ela está sorrindo; Eles(as) se beijaram; É aniversário dela.

O número mais importante é a pontuação geral de relaxamento – as 3 (três) categorias separadas existem, principalmente, para guiar os seus julgamentos. Não se preocupe com as pontuações individuais se você não tiver certeza de qual das (categorias) acima se aplica – contando que a pontuação geral seja a mesma, não importa.

A.3 Exemplo de rotulação para estresse e relaxamento a partir de publicações reais presentes na base de dados coletada.

Cada publicação contém uma pontuação para estresse e relaxamento a partir dos critérios apresentados anteriormente.

Exemplos:

- ❏ “Quero pagodearrrrrr Covid 19 sua estranha Já pode ir embora” (-2,4): nível intermediário de estresse (-2) em “sua estranha” e “já pode ir embora” e nível intermediário de relaxamento (4) em “quero pagodear”;
- ❏ • “Esqueci até do Covid-19 kkkk” (-1,2): ausência de estresse (-1) e nível intermediário de relaxamento (2) em “esqueci da covid-19”;

- ❏ • "Segunda-feira: nada de reclamar e sempre agradecer por mais um dia de vida e por estar empregado em meio dessa crise de covid-19. vamos trabalhar, bom dia a todos e ótima semana!" (-1,5): ausência de estresse (-1) e indício de estado altamente relaxado (5) em "sempre agradecer por mais um dia de vida e por estar empregado" e "bom dia a todos e ótima semana;

- ❏ • "Eu tô com medo real do covid 19. É pra ter medo!!!!! Pessoas jovens estão morrendo! Não vou vacilar mais!" (-5,1): nível alto de estresse (-5) em "eu tô com medo real do covid-19", "pessoas jovens estão morrendo" e "não vou vacilar mais" e ausência de relaxamento (1).