

INSTITUTO FEDERAL GOIANO - CAMPUS MORRINHOS
CURSO BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

CARLOS EDUARDO ROCHA MIRANDA

**AUTISMO NO REDDIT: UM ESTUDO DE SENTIMENTOS E TÓPICOS
POR MEIO DE MODELOS DE APRENDIZADO DE MÁQUINA**

MORRINHOS - GO
2025

CARLOS EDUARDO ROCHA MIRANDA

**AUTISMO NO REDDIT: UM ESTUDO DE SENTIMENTOS E TÓPICOS
POR MEIO DE MODELOS DE APRENDIZADO DE MÁQUINA**

Monografia apresentada ao Curso Bacharelado em Ciência da Computação do Instituto Federal Goiano - Campus Morrinhos, como requisito parcial para obtenção de título de Bacharel em Ciência da Computação.

Área de Concentração: Metodologia e Técnicas da Computação; Banco de Dados; Modelos Analíticos e de Simulação.

Orientador: Prof. Dr. Jesmmer da Silveira Alves.

**MORRINHOS – GO
2025**

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/IF Goiano Campus Morrinhos

M672a Miranda, Carlos Eduardo Rocha.

Autismo no REDDIT: um estudo de sentimentos e tópicos por meio de modelos de aprendizado de máquina. / Carlos Eduardo Rocha Miranda. – Morrinhos, GO: IF Goiano, 2025.

78 f. : il. color.

Orientador: Dr. Jesmmer da Silveira.

Trabalho de conclusão de curso (graduação) – Instituto Federal Goiano Campus Morrinhos, Bacharelado em Ciências da Computação, 2025.

1. Aprendizado de máquina. 2. Modelagem de Tópicos. 3. Autismo. I. Silveira, Jesmmer da. II. Instituto Federal Goiano. III. Título.

CDU 004.6

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS

NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
☐ Dissertação (mestrado)
☐ Monografia (especialização)
☒ TCC (graduação)

- ☐ Artigo científico
☐ Capítulo de livro
☐ Livro
☐ Trabalho apresentado em evento

☐ Produto técnico e educacional - Tipo:

Nome completo do autor:

Carlos Eduardo Rocha Miranda

Matrícula:

2020104201940144

Título do trabalho:

Autismo no Reddit: Um Estudo de Sentimentos e Tópicos por meio de modelos de Aprendizado de Máquina

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☒ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: 03/03/2025

O documento está sujeito a registro de patente? ☐ Sim ☒ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☒ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

- Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
- Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
- Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Edealina - Goiás

Local

20/02/2025

Data

Assinado por:

Carlos Eduardo Rocha Miranda

4722C3B632B947B...

Assinatura do autor e/ou detentor dos direitos autorais

Ciente e de acordo:

Assinado por:

Jessimer da Silveira Alves

5612544DED224A2...

Assinatura do(a) orientador(a)



SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Ata nº 4/2025 - CCEG-MO/CEG-MO/DE-MO/CMPMHOS/IFGOIANO

ATA DE DEFESA DE TRABALHO DE CURSO (TC)

Aos três dias do mês de fevereiro de 2025, no auditório do Núcleo de Computação do Instituto Federal Goiano, campus Morrinhos (GO), realizou-se a sessão pública de defesa de TCC do Curso Bacharelado em Ciência da Computação, do acadêmico Carlos Eduardo Rocha Miranda, sob orientação do professor Jesmmer da Silveira Alves, intitulada AUTISMO NO REDDIT: UM ESTUDO DE SENTIMENTOS E TÓPICOS. Compuseram a Banca Examinadora os professores:

Orientador

Prof. Dr. Jesmmer da Silveira Alves

Membro 2

Prof.(a) Dra. Leila Roling Scariot da Silva

Membro 3

Prof.(a) Dra. Thelma Maria de Moura Bergamo

Membro 4

Prof. Dr. Fernando Barbosa Matos

Após a exposição oral, o candidato foi arguido pelos membros da banca, os quais reuniram-se reservadamente, e decidiram, **APROVADO**. Para constar, redigi a presente Ata, que aprovada por todos os presentes, vai assinada por mim, Orientador do TC, e pelos demais membros da banca.

Prof. Orientador

Assinado eletronicamente

Prof.(a)/Membro 2

Assinado eletronicamente

Prof.(a)/Membro 3

Assinado eletronicamente

Prof./Membro 4

Assinado eletronicamente

Documento assinado eletronicamente por:

- Jesmmer da Silveira Alves, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 03/02/2025 16:44:03.
- Thelma Maria de Moura Bergamo, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 03/02/2025 17:16:27.
- Leila Roling Scariot da Silva, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 03/02/2025 18:20:25.
- Fernando Barbosa Matos, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 03/02/2025 21:55:26.

Este documento foi emitido pelo SUAP em 03/02/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifgoiano.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código Verificador: 671709

Código de Autenticação: 4943ea6292



INSTITUTO FEDERAL GOIANO

Campus Morrinhos

Rodovia BR-153, Km 633, Zona Rural, SN, Zona Rural, MORRINHOS / GO, CEP 75650-000

(64) 3413-7900

DEDICATÓRIA

Dedico este trabalho à minha mãe, Alcione Lopes da Rocha, ao meu pai, Abílio Eurípedes Polidório Miranda e a minha irmã Maria Olívia Rocha Miranda, por todo amor, apoio e incentivo ao longo dessa jornada.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão ao meu orientador, o Prof. Dr. Jesmmer da Silveira Alves, pela paciência, dedicação e orientação constante durante todo o processo de desenvolvimento deste trabalho. Agradeço também aos residentes da Residência da Amoreira e aos pesquisadores do Laboratório GEDAIA, minha segunda família.

RESUMO

Este estudo teve como objetivo analisar publicações relacionadas ao autismo na rede social Reddit, com foco na comunidade "*autism*", a fim de identificar se essas postagens contêm informações úteis para compreender as necessidades e os problemas da população autista. Dada a importância de compreender experiências dessa população, foi escolhido o Reddit como a plataforma de análise, considerando seu grande volume de dados e a diversidade de opiniões compartilhadas por seus usuários. Inicialmente, foi configurado um banco de dados PostgreSQL para armazenar os dados coletados. A coleta de dados foi realizada utilizando a API oficial do Reddit, ao longo dos meses de junho e julho de 2024. Em seguida, foi realizada uma análise de sentimentos utilizando a biblioteca Transformers e o modelo "*Twitter-roBERTa-base for Sentiment Analysis*", que permitiu classificar as postagens em sentimentos positivos, negativos ou neutros. Posteriormente, o modelo BERTopic foi utilizado para realizar a modelagem de tópicos, visando identificar os temas recorrentes nas postagens e entender os principais desafios enfrentados pelas pessoas autistas. A análise macro investigou os cinco principais tópicos, bem como o impacto do sentimento nas submissões e comentários, revelando padrões e tendências. Foi feita, também, uma análise mais detalhada dos tópicos, com foco nos termos mais relevantes, na distribuição de sentimentos e nos temas dos comentários associados. Os resultados confirmaram o potencial do Reddit como fonte valiosa de informações, especialmente quando combinado com ferramentas avançadas como os modelos BERTopic e RoBERTa, e podem ser utilizados para apoiar a criação de ações mais eficazes para melhorar a qualidade de vida da população autista.

Palavras-chave: Autismo. Análise de Sentimentos. Modelagem de Tópicos. Reddit. Aprendizado de máquina.

ABSTRACT

This study aimed to analyze posts related to autism on the social network Reddit, focusing on the "autism" community, to determine whether these posts contain useful information for understanding the needs and challenges of the autistic population. Given the importance of understanding the experiences of this population, Reddit was chosen as the analysis platform due to its large volume of data and the diversity of opinions shared by its users. Initially, a PostgreSQL database was set up to store the collected data. The data collection was conducted using Reddit's official API throughout June and July 2024. Subsequently, a sentiment analysis was performed using the Transformers library and the "Twitter-roBERTa-base for Sentiment Analysis" model, which classified the posts as positive, negative, or neutral. Following this, the BERTopic model was used for topic modeling to identify recurring themes in the posts and understand the main challenges faced by autistic individuals. The macro analysis examined the five main topics, as well as the impact of sentiment in submissions and comments, revealing patterns and trends. A more detailed analysis of the topics was also conducted, focusing on the most relevant terms, sentiment distribution, and the themes of associated comments. The results confirmed Reddit's potential as a valuable source of information, especially when combined with advanced tools such as the BERTopic and RoBERTa models, and can be used to support the development of more effective initiatives aimed at improving the quality of life of the autistic population.

Keywords: Autism. Sentiment Analysis. Topic Modeling. Reddit. Machine Learning.

LISTA DE ILUSTRAÇÕES

Figura 1 - Página inicial da subreddit "autism" no Reddit.	17
Figura 2 - Esboço da arquitetura Transformers.	19
Figura 3 - Quantidade de submissões diárias coletadas.	29
Figura 4 - Quantidade de comentários diários coletados.	29
Figura 5 - Categorias das submissões.	30
Figura 6 - Redução de dados após limpeza: Submissões, Comentários e Usuários.	32
Figura 7 - Distribuição de sentimentos por título.	33
Figura 8 - Distribuição de sentimentos por corpo de texto.	34
Figura 9 - Distribuição de sentimentos por título e corpo de texto.	35
Figura 10 - Distribuição de sentimentos dos comentários.	35
Figura 11 - Distribuição de sentimentos ao longo de junho e julho.	45
Figura 12 - Semelhança de tópicos com base nos sentimentos.	46
Figura 13 - Proporção dos sentimentos dos comentários por sentimento das submissões.	47
Figura 14 - Nuvem de Palavras: Dificuldades de saúde mental e ansiedade social.	48
Figura 15 - Proporção dos sentimentos: Dificuldades de saúde mental e ansiedade social.	49
Figura 16 - Conexões entre tópicos identificados nos comentários com o tópico Dificuldades de saúde mental e ansiedade social.	49
Figura 17 - Nuvem de Palavras: Acessibilidade no trabalho e saúde mental.	50
Figura 18 - Proporção dos sentimentos: Acessibilidade no trabalho e saúde mental.	51
Figura 19 - Conexões entre tópicos identificados nos comentários com o tópico Acessibilidade no trabalho e saúde mental.	51
Figura 20 - Nuvem de Palavras: A diversidade de hobbies e interesses especiais no autismo.	52
Figura 21 - Proporção dos sentimentos: A diversidade de hobbies e interesses especiais no autismo.	53
Figura 22 - Conexões entre tópicos identificados nos comentários com o tópico A diversidade de hobbies e interesses especiais no autismo.	54

Figura 23 - Nuvem de Palavras: Diagnóstico tardio.	55
Figura 24 - Proporção dos sentimentos: Diagnóstico tardio.	56
Figura 25 - Conexões entre tópicos identificados nos comentários com o tópico Diagnóstico tardio.	56
Figura 26 - Nuvem de Palavras: Desafios sensoriais e maneiras de lidar com eles.	57
Figura 27 - Proporção dos sentimentos: Desafios sensoriais e maneiras de lidar com eles.	58
Figura 28 - Conexões entre tópicos identificados nos comentários com o tópico Desafios sensoriais e maneiras de lidar com eles.	59

LISTA DE TABELAS

Tabela 1 - Submissões.....	25
Tabela 2 - Comentários.....	26
Tabela 3 - Palavras-chave dos usuários	31
Tabela 4 - Tópicos avulsos das submissões.....	38
Tabela 5 - Tópicos das submissões positivas	39
Tabela 6 - Tópicos avulsos das submissões negativas.....	40
Tabela 7 - Três maiores tópicos de comentários.....	41
Tabela 8 - Três maiores tópicos dos comentários positivos.....	41
Tabela 9 - Três maiores tópicos dos comentários negativos	41
Tabela 10 - Tópicos das submissões reorganizados e renomeados.....	42
Tabela 11 - Tópicos das submissões negativas reorganizados e renomeados	43

SUMÁRIO

1 INTRODUÇÃO	11
1.1 METODOLOGIA.....	13
1.2 ORGANIZAÇÃO DO TRABALHO	14
2 REFERENCIAL TEÓRICO.....	15
2.1 TRANSTORNO DO ESPECTRO AUTISTA	15
2.2 REDDIT	16
2.3 PROCESSAMENTO DE LINGUAGEM NATURAL	17
2.3.1 Análise de Sentimentos	17
2.3.2 Modelagem de Tópicos.....	18
2.3.3 Transformers	19
2.4 TRABALHOS RELACIONADOS	22
3 DESENVOLVIMENTO	25
3.1 CONFIGURAÇÃO DO BANCO DE DADOS	25
3.2 COLETA DE POSTAGENS E COMENTÁRIOS DO REDDIT	27
3.3 LIMPEZA E SELEÇÃO DOS DADOS	29
3.4 CLASSIFICAÇÃO DE SENTIMENTOS.....	33
3.5 EXTRAÇÃO DE TÓPICOS.....	36
4 ANÁLISE DOS RESULTADOS	45
4.1 DIFICULDADES DE SAÚDE MENTAL E ANSIEDADE SOCIAL	47
4.2 ACESSIBILIDADE NO TRABALHO E SAÚDE MENTAL	50
4.3 A DIVERSIDADE DE HOBBIES E INTERESSES ESPECIAIS NO AUTISMO ...	52
4.4 DIAGNÓSTICO TARDIO.....	54
4.5 DESAFIOS SENSORIAIS E MANEIRAS DE LIDAR COM ELES	57
5 CONSIDERAÇÕES FINAIS	60
REFERÊNCIAS.....	63
APÊNDICE.....	67

1 INTRODUÇÃO

O autismo é um transtorno de neurodesenvolvimento que afeta aproximadamente 61.8 milhões de pessoas no mundo todo (SANTOMAURO *et al.*, 2024). Ele é caracterizado pela falta de comunicação social e pela presença de padrões de comportamento repetitivos. Dada a expressiva quantidade de pessoas impactadas globalmente, torna-se fundamental que a sociedade desenvolva estratégias eficazes para melhorar a qualidade de vida dessa população. No entanto, identificar e compreender as necessidades específicas dessas pessoas pode ser um desafio significativo.

Este estudo concentrou-se na análise de publicações relacionadas ao autismo na rede social Reddit, buscando identificar se essas postagens contêm informações úteis para compreender as necessidades e os problemas da população autista. Especificamente, a pesquisa analisou a comunidade “*autism*” no Reddit para identificar os sentimentos e temas recorrentes nas publicações, além de avaliar a eficácia dos modelos de *machine learning* aplicados na organização e análise dos dados.

Compreender os desafios enfrentados pelas pessoas autistas é essencial para o sucesso de iniciativas que promovam a melhoria de sua qualidade de vida. Diversos estudos têm utilizado as redes sociais como ferramenta para investigar as vivências da população autista. Beykikhoshk *et al.* (2015), por exemplo, analisaram dados da plataforma X (antigo Twitter) para estudar as preocupações e necessidades de autistas, enquanto Gabarron *et al.* (2023), investigaram os tópicos mais discutidos por essa população na mesma plataforma.

Apesar de iniciativas como a de Thom-Jones, Melgaard e Gordon (2024), que exploraram a experiência de maternidade de mulheres autistas no Reddit, representem avanços importantes, ainda há uma lacuna na literatura no que diz respeito a utilizar o Reddit como fonte de dados para explorar as necessidades da população autista de forma abrangente. Este estudo propõe-se a utilizar a comunidade “*autism*” como fonte de dados, com o objetivo de verificar a viabilidade dessa plataforma para fornecer informações relevantes para a compreensão dos problemas comuns da população autista.

Este trabalho tem como principal objetivo analisar as postagens da comunidade “autism”, visando compreender as necessidades e os desafios enfrentados pelas pessoas autistas. Para cumprir o objetivo proposto, pretende-se:

- Analisar os dados coletados utilizando o modelo de machine learning RoBERTa para classificar os sentimentos das postagens e compreender a dinâmica de sentimentos na comunidade.
- Identificar os principais tópicos mais discutidos na comunidade por meio de modelagem de tópicos com o modelo BERTopic.
- Relacionar os temas e os sentimentos predominantes nas postagens para identificar os principais problemas enfrentados.

Para conduzir este estudo, foi inicialmente configurado um banco de dados PostgreSQL¹ para armazenar todas as informações relevantes. A coleta de dados foi realizada utilizando a API (interface de programação de aplicativos) oficial do Reddit, que permitiu a extração das submissões e seus respectivos comentários da comunidade analisada. Após o armazenamento dos dados, as publicações foram classificadas como neutras, positivas ou negativas por meio da biblioteca *Transformers*². Foi utilizado uma variação do modelo RoBERTa treinado com dados do X, conforme descrito por Loureiro (2022).

Na etapa final, os temas principais foram identificados por meio de uma modelagem de tópicos com o uso do modelo BERTopic, escolhido por sua eficácia na análise de dados textuais do X, conforme demonstrado por Egger (2022). Foram analisados os cinco tópicos mais relevantes, com atenção especial ao impacto do sentimento da submissão no sentimento dos comentários subsequentes. Adicionalmente, realizou-se uma investigação mais detalhada dos tópicos recorrentes, com foco nos termos mais relevantes, a distribuição de sentimentos e os temas dos comentários associados.

¹ PostgreSQL: é um sistema de gerenciamento de banco de dados relacional. Documentação oficial disponível em: <https://www.postgresql.org/docs/current/>. Acesso em: 28 nov. 2024.

² *Transformers*: biblioteca Python que fornece ferramentas para baixar e treinar modelos de inteligência artificial. Documentação oficial disponível em: <https://huggingface.co/docs/transformers/index>. Acesso em: 28 nov. 2024.

1.1 METODOLOGIA

A metodologia deste estudo consistiu em analisar publicações relacionadas ao autismo no Reddit, por meio de uma pesquisa documental de caráter descritivo, abordagem quantitativa e desenho longitudinal com relação ao tempo.

Um computador pessoal do modelo Acer Nitro ANVI 5-51 com 16 GB de memória RAM, processador i7-13620H de 2.40 GHz, e placa de vídeo NVIDIA GeForce RTX 4050 foi utilizado para coletar os dados e aplicar os modelos.

A plataforma Reddit foi escolhida como cenário da pesquisa devido ao anonimato oferecido, o que, segundo Ma, Hancock e Naaman (2016), incentiva o compartilhamento de informações confiáveis. A subreddit “*autism*” foi selecionada por ser a maior comunidade sobre autismo na plataforma, com mais de 400 mil membros.

O público alvo deste estudo compreende usuários autistas da comunidade, incluindo aqueles com diagnóstico confirmado ou que possuem suspeitas sobre sua condição. Todo o trabalho foi desenvolvido utilizando a linguagem de programação Python. O estudo foi conduzido em 6 etapas: configuração do banco de dados; coleta de dados; limpeza e seleção dos dados; análise de sentimentos; modelagem de tópicos; e análise dos resultados.

A primeira etapa consistiu em configurar um banco de dados PostgreSQL para armazenar todas as informações relevantes para a análise de dados. Nessa fase foi necessário mapear os campos das submissões e dos comentários.

Para a coleta de dados foi feito o uso da API oficial do Reddit. A coleta ocorreu entre os meses de junho e julho de 2024, resultando em 11.623 publicações, 135.894 comentários, de 28.249 usuários diferentes. Após essa etapa, os dados foram limpos, removendo publicações e comentários sem conteúdo relevante, bem como aquelas identificadas como mensagens administrativas de moderadores da comunidade. Além disso, foram selecionadas apenas as submissões de usuários que indicaram, de forma explícita em seus perfis, o diagnóstico ou a suspeita de autismo. Ao final do processo de limpeza e seleção, restaram 1.916 publicações, 29.004 comentários, e 10.585 usuários.

A análise de sentimentos foi realizada utilizando a biblioteca *Transformers*, desenvolvida pela *Hugging Face*, uma plataforma de código aberto que tem como objetivo democratizar o acesso a modelos de IA. Para essa análise, foi selecionada

uma variação do modelo RoBERTa chamada “*Twitter-roBERTa-base for Sentiment Analysis*”³ que foi treinada com dados do X (LOUREIRO, 2022). Já a modelagem de tópicos foi conduzida com o uso do modelo BERTopic, escolhido por sua eficácia na análise de dados textuais da mesma plataforma, conforme demonstrado por Egger (2022). O Reddit compartilha características semelhantes ao X, e, por isso, ambos os modelos foram considerados adequados, já tendo se mostrado eficazes com dados desse tipo.

Por fim, a análise dos resultados foi realizada em duas fases: uma análise macro, investigando os cinco principais tópicos e o impacto do sentimento nas submissões e comentários subsequentes; e uma análise mais detalhada dos tópicos recorrentes, com foco nos termos mais relevantes, a distribuição de sentimentos e os temas dos comentários associados.

1.2 ORGANIZAÇÃO DO TRABALHO

O estudo foi dividido em cinco capítulos: o Capítulo 1 – Introdução, que contextualiza e justifica a importância da pesquisa; o Capítulo 2 – Referencial Teórico, que discute os conceitos centrais do estudo; o Capítulo 3 – Desenvolvimento, onde são descritos os processos de coleta, análise de sentimentos e modelagem de tópicos; o Capítulo 4 – Análise dos Resultados, que apresenta as descobertas obtidas a partir das submissões e comentários do Reddit; e o Capítulo 5 – Considerações Finais, que avalia se os objetivos e o problema de pesquisa foram atendidos, discute as limitações enfrentadas e oferece recomendações para estudos futuros.

³ Modelo *cardiffnlp/twitter-roberta-base-sentiment*: uma versão do RoBERTa-base ajustada para análise de sentimentos em inglês. Disponível em: <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>. Acesso em: 7 dez. 2024.

2 REFERENCIAL TEÓRICO

2.1 TRANSTORNO DO ESPECTRO AUTISTA

Segundo o CID-11⁴, o transtorno do espectro autista é caracterizado por déficits persistentes na habilidade de iniciar e manter interações sociais e comunicação social recíprocas, e por uma gama de padrões de comportamento, interesses ou atividades restritos, repetitivos e inflexíveis, que são claramente atípicos ou excessivos para a idade e o contexto cultural do indivíduo. O início do transtorno ocorre durante o período do desenvolvimento, tipicamente na primeira infância, mas os sintomas podem não se manifestar plenamente até mais tarde, quando as demandas sociais excedem as capacidades limitadas. Os déficits são graves o suficiente para causar prejuízos no funcionamento pessoal, familiar, social, educacional, ocupacional e em outras áreas importantes, e são, geralmente, uma característica generalizada do funcionamento do indivíduo, observável em todos os ambientes, podendo variar de acordo com o contexto social, educacional ou outro. Os indivíduos ao longo do espectro exibem toda a gama de funcionamento intelectual e habilidades de linguagem.

Visto que, conforme explica Lord *et al.* (2018), os sintomas variam entre os indivíduos e não existem biomarcadores específicos, a avaliação comportamental torna-se essencial para o diagnóstico. Com o objetivo de padronizar essa avaliação, o Manual Diagnóstico e Estatístico de Transtornos Mentais (DSM-5) (AMERICAN PSYCHIATRIC ASSOCIATION *et al.*, 2014, p. 50) estabelece critérios diagnósticos padronizados. Entre esses critérios, destacam-se dois principais:

1. Déficit persistentes na comunicação social e na interação social em múltiplos contextos;
2. Padrões restritos e repetitivos de comportamento, interesses ou atividades.

⁴ Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde – CID-11. Genebra: OMS, 2024. Disponível em: <https://icd.who.int/browse/2024-01/mms/en#437815624>. Acesso em: 21 set. 2024.

É importante destacar que este trabalho adota a perspectiva médica em relação às definições, na qual as deficiências são vistas como um conjunto de sintomas e o autismo é tratado como uma condição médica. No entanto, existe também o modelo social, que considera as deficiências como uma variação natural da diversidade humana (DA SILVA; GESSER; NUERNBERG, 2019). Para este estudo, optou-se pela abordagem médica.

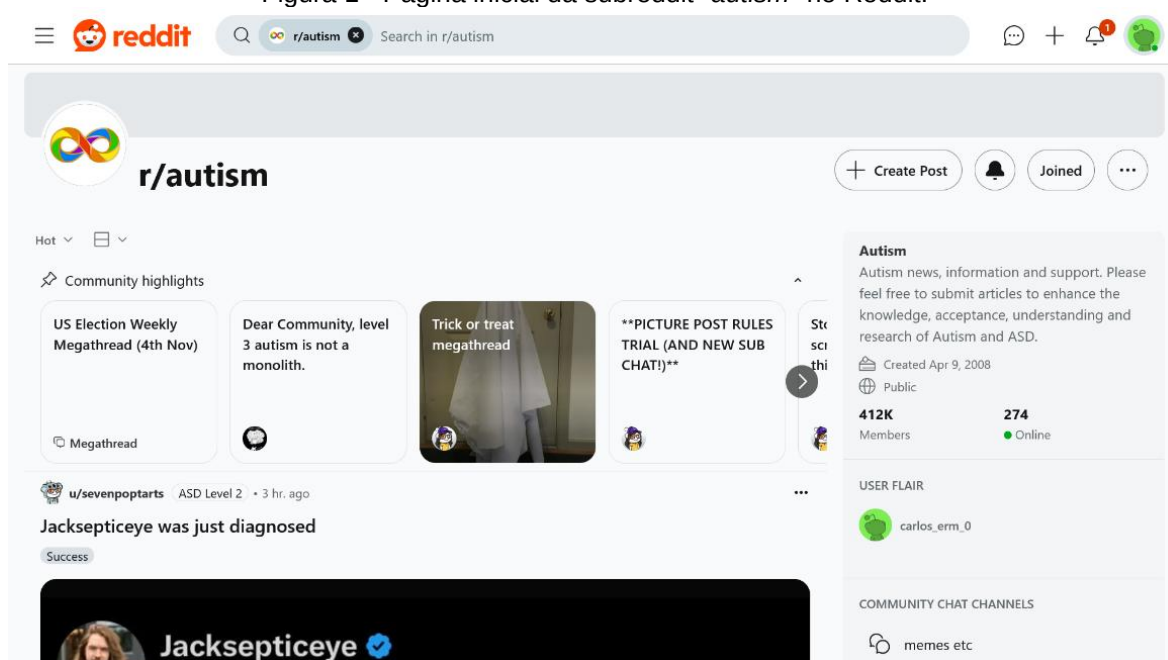
2.2 REDDIT

O Reddit⁵ descreve sua plataforma como: “a casa de milhares de comunidades, conversas intermináveis e conexões humanas autênticas”. É um site que permite a criação de comunidades chamadas *subreddits*. Nelas os membros são capazes de postar, comentar e votar nas postagens mais interessantes. Com mais de 90 milhões de usuários diários em junho de 2024, o Reddit é uma plataforma rica em conteúdo, reunindo mais de 100.000 comunidades sobre temas variados como política, esportes, séries de TV, filmes e livros.

Segundo Anderson (2015), qualquer usuário pode criar uma *subreddit*, e a sua moderação é realizada pelos membros da comunidade do Reddit. Cada *subreddit* possui um tema ou assunto específico, permitindo que o usuário encontre e se inscreva nas comunidades de seu interesse, conforme ilustrado na Figura 1. A experiência na plataforma é customizável, permitindo que os usuários visualizem apenas posts das *subreddits* às quais estão inscritos. É possível ainda filtrar as postagens com as seguintes categorias: em destaque; mais novos; mais votados; e em ascensão. Qualquer usuário, mesmo sem um registro na plataforma, pode visualizar as postagens de uma *subreddit*, desde que seja pública.

⁵ Reddit. Homepage. Disponível em: <https://www.redditinc.com/>. Acesso em: 15 set. 2024.

Figura 1 - Página inicial da subreddit "autism" no Reddit.



2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

A linguagem natural, também chamada de linguagem comum, é aquela falada e escrita por humanos para fins de comunicação cotidiana (CHOPRA; PRASHAR; SAIN, 2013). Segundo esses autores, o processamento de linguagem natural (PLN) é uma subárea da inteligência artificial cujo objetivo é capacitar os computadores a compreenderem e interpretar a linguagem natural. Khurana *et al.* (2023) destacam que o desenvolvimento do PLN surgiu do desejo de facilitar a comunicação com máquinas utilizando a linguagem natural.

Hirschberg e Manning (2015), complementam que o PLN utiliza técnicas computacionais para entender e reproduzir a linguagem humana, possibilitando o desenvolvimento de soluções práticas. Entre as aplicações mais populares do PLN estão os mecanismos de tradução automática, a análise de sentimentos, a modelagem de tópicos, a criação de sistemas de diálogo falado e a extração de informações em redes sociais.

2.3.1 Análise de Sentimentos

A análise de sentimentos (AS) é uma subárea da ciência da computação que estuda as opiniões e emoções das pessoas em relação a diferentes objetos,

indivíduos, eventos ou ideias (MEDHAT; HASSAN; KORASHY, 2014). Segundo Taboada (2016), o objetivo principal da AS é determinar se um texto contém uma opinião subjetiva e, caso contenha, identificar se essa opinião é positiva ou negativa. Para alcançar esses objetivos, são empregados diversos algoritmos de aprendizado de máquina, como *Naive Bayes*, Máquina de Vetores de Suporte (*Support Vector Machines*) e modelos mais avançados baseados em aprendizado profundo (AHMAD *et al.*, 2017).

Wankhade, Rao e Kulkarni (2022) ressaltam que entender o sentimento das pessoas a respeito de produtos, serviços ou assuntos específicos é uma ferramenta poderosa para governos e empresas, pois possibilita a tomada de decisões mais informadas e assertivas. Com o crescimento das redes sociais, onde indivíduos frequentemente expressam opiniões e ideias, a análise de sentimentos tornou-se ainda mais relevante, aumentando a demanda por soluções capazes de processar e interpretar essas emoções de maneira eficaz e automática.

2.3.2 Modelagem de Tópicos

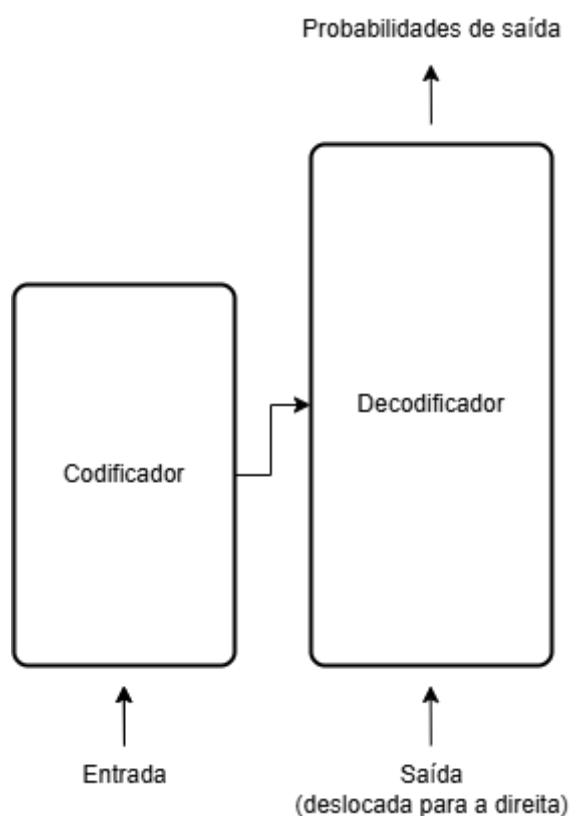
A modelagem de tópicos é uma técnica utilizada para reduzir grandes volumes de dados textuais, identificando suas características principais e temas ocultos (KHERWA; BANSAL, 2019). Segundo Silveira *et al.* (2021), esta técnica é útil em contextos onde o volume de dados textuais é tão grande que se torna inviável realizar uma classificação manual. Ela permite, por exemplo, gerar resumos textuais e agrupar documentos com base em temas em comum.

Nas redes sociais, onde a quantidade de informações geradas pelos usuários é massiva, a modelagem de tópicos é amplamente utilizada para analisar as opiniões predominantes em temas de destaque. Segundo Rohani, Shayaa e Babanrjaddehaki (2016), essa técnica é particularmente eficaz na identificação de tópicos em redes sociais, gerando *insights* valiosos para áreas como marketing, segurança e educação.

2.3.3 Transformers

Os *Transformers* constituem uma arquitetura de redes neurais introduzida pelo seminal “*Attention is All You Need*” (VASWANI *et al.*, 2017), originalmente voltada para a tradução de linguagens. Essa arquitetura tem como base o “*self-attention*”, que permite que cada palavra considere todas as outras palavras, garantindo a representação do contexto ao seu redor de maneira eficiente. A estrutura do Transformers é composta por dois blocos: *Encoder* (Codificador) e *Decoder* (Decodificador), conforme ilustrado na Figura 2.

Figura 2 - Esboço da arquitetura *Transformers*.



De acordo com o curso de Processamento de Linguagem Natural da *Hugging Face*⁶, o *Encoder* recebe uma entrada textual e transforma-a em uma representação numérica que captura o contexto semântico dos dados, também

⁶ *Hugging Face NLP Course*. Disponível em: <https://huggingface.co/learn/nlp-course/chapter1/4?fw=pt>. Acesso em: 19 dez. 2024.

conhecida como *embedding*. O *Decoder* utiliza essas representações, juntamente com outras informações de entrada, para realizar tarefas como geração de texto, prevendo a próxima palavra em uma sequência.

Esses componentes podem ser utilizados de maneira independente. Modelos baseados em *Encoders*, como BERT e RoBERTa, são particularmente eficazes em entender a estrutura da linguagem humana, sendo ideais para tarefas como análise de sentimentos e classificação de textos. Por outro lado, modelos baseados em *Decoders*, como o GPT⁷, destaca-se em tarefas que envolvem a geração de texto, como criação de conteúdo e respostas automatizadas. Quando combinados, os *Encoders* e *Decoders* formam a arquitetura *sequence-to-sequence*, amplamente utilizada em tarefas de tradução automática.

2.3.4 BERT

O BERT (*Bidirectional Encoder Representations from Transformers*) é um modelo de linguagem desenvolvido por pesquisadores do Google em 2018. Ele é conhecido por ser capaz de considerar o contexto através de uma análise bidirecional das palavras, ou seja, ele considera os textos que vem antes e depois de uma palavra (DEVLIN *et al.*, 2019). Essa característica torna o BERT útil em tarefas como análise de sentimentos, perguntas e respostas, e tradução automática.

Um exemplo notável da aplicação do BERT é o trabalho de Müller, Salathé e Kummervold (2023), que desenvolveram o CT-BERT (COVID-Twitter-BERT), um modelo baseado no BERT treinado com tweets relacionados à COVID-19. O CT-BERT mostrou um desempenho excepcional na classificação de conteúdo relacionado à COVID-19, especialmente em dados extraídos de redes sociais. Essas descobertas têm o potencial de contribuir para aplicações práticas, como o monitoramento do sentimento público e o desenvolvimento de *chatbots* capazes de fornecer informações confiáveis sobre a pandemia.

⁷ *Generative Pre-trained Transformer* (Transformador Generativo Pré-treinado). São uma família de grandes modelos de linguagem (LLMs) baseados na arquitetura *transformers* de aprendizado profundo. Disponível em: <https://www.ibm.com/think/topics/gpt>. Acesso em: 09 jan. 2025.

2.3.5 RoBERTa

O RoBERTa (*Robustly Optimized BERT Approach*) é uma versão otimizada do BERT, desenvolvida por pesquisadores do Facebook em 2019. A principal diferença entre os dois modelos é que o RoBERTa foi treinado com um conjunto de dados significativamente maior e empregou técnicas de treinamento mais eficientes. Esse modelo ajusta diversos hiperparâmetros do BERT, incluindo a remoção do objetivo de pré-treinamento de previsão da próxima sentença (*next-sentence prediction objective*), além de usar *mini-batches* (porções dos dados de treinamento) maiores e taxas de aprendizado mais elevadas (LIU *et al.*, 2019). Essas mudanças resultaram em um desempenho melhor para o RoBERTa em várias tarefas de processamento de linguagem natural.

Uma aplicação do RoBERTa foi realizada por Jung (2024), que utilizou o modelo pré-treinado Sentence-RoBERTa (SRoBERTa), especializado em medir a similaridade entre sentenças. O modelo foi ajustado com dados sobre as relações entre produtos para criar o Pro-SRoBERTa, uma versão customizada destinada à área de inteligência de mercado. O objetivo foi aprimorar a delimitação de escopos de mercado e fornecer *insights* mais precisos sobre informações de negócios. O Pro-SRoBERTa foi implementado com sucesso no sistema de informações de mercado utilizando um conjunto de dados real.

2.3.6 BERTopic

O BERTopic é um modelo que extrai tópicos de documentos através de uma variação do TF-IDF⁸ chamada c-TF-IDF, que agrega os termos de todos os documentos dentro de um *cluster* e calcula a importância dos termos para todo o *cluster* (GROOTENDORST, 2022). Inicialmente, os documentos são convertidos em

⁸ *Term Frequency-Inverse Document Frequency* (Frequência do termo–inverso da frequência nos documentos). É uma estatística numérica que quantifica a importância de um termo em relação à sua frequência num documento e sua raridade em outros documentos. Disponível em: <https://medium.com/@abhishekjainindore24/tf-idf-in-nlp-term-frequency-inverse-document-frequency-e05b65932f1d>. Acesso em: 09 jan. 2025.

representações numéricas (vetores) de alta dimensionalidade, chamadas *embeddings*. As direções nesses vetores refletem significados semânticos e contextuais. Em seguida, esses *embeddings* passam por um processo de redução de dimensionalidade, facilitando o agrupamento em *clusters*. Para o agrupamento, o modelo utiliza, como padrão, o algoritmo HDBSCAN que identifica grupos de dados sem a necessidade de especificar o número de *clusters* previamente.

Para identificar os tópicos, o BERTopic calcula a frequência de cada palavra em cada grupo por meio de uma representação *Bag-of-Words* (saco-de-palavras). Em seguida, aplica o c-TF-IDF, ajustado para considerar as características específicas de cada grupo, identificando as palavras que melhor representam o conteúdo de cada *cluster*. Consequentemente, os tópicos identificados refletem as ideias predominantes dentro de cada agrupamento.

2.4 TRABALHOS RELACIONADOS

Diversos estudos analisam publicações em redes sociais como o Reddit e o X, abordando diferentes aspectos, como o trabalho de Almeida (2024) *et al.* que investiga o uso de opioides através do Reddit, ou o trabalho de Hasan, Maliha e Arifuzzaman (2019) que aplica a análise de sentimentos no *Twitter* para analisar a opinião pública sobre um produto. Alguns destes trabalhos tratam especificamente de entender o que é compartilhado por comunidades de difícil acesso, como pessoas autistas, que são relativamente poucas e estão espalhadas pelo mundo. Estes estudos utilizam técnicas de análise de sentimentos e, em alguns casos, modelagem de tópicos para extrair informações relevantes. Nesta seção, serão apresentados alguns estudos que exploram as interações entre pessoas autistas, além de um estudo sobre pessoas com depressão, em diálogo com o propósito deste trabalho.

O primeiro estudo, realizado por Thom-Jones, Melgaard e Gordon (2024), investigou a experiência de maternidade de mulheres autistas na comunidade “*AutisticParents*” do Reddit. Os dados foram coletados via API e analisados qualitativamente com uso do software NVivo⁹. A análise revelou 4 temas: “Ser mãe

⁹ NVivo. O *software* mais citado e poderoso para análise de dados qualitativos. Disponível em: <https://lumivero.com/product/nvivo/>. Acesso em: 30 nov. 2024.

autista é diferente”, “Mães autistas precisam de mães autistas”, “Mães autistas enfrentam estigma”, e “Lições do *lockdown*”.

Outro estudo, de Bellon-Harn, Boyd e Manchaiah (2022), visou entender o conjunto de opiniões e discussões nas redes sociais sobre a ABA (análise do comportamento aplicada), para auxiliar na tomada de decisões de famílias de crianças com autismo. Especificamente, os autores propõem identificar os principais temas das postagens no Reddit sobre intervenções ABA para o TEA (transtorno do espectro autista) utilizando modelagem de tópicos. Além disso, buscam examinar as características linguísticas dessas postagens com a ferramenta LIWC¹⁰, avaliando sua relação com a categoria dos usuários, como, por exemplo, aqueles a favor ou contra a ABA. Como resultado, foram identificados 11 temas principais na modelagem de tópicos, e observou-se que a maioria das postagens se baseia em experiências pessoais.

Outro exemplo relevante é o trabalho de Gabarron *et al.* (2023), que teve como objetivo identificar os tópicos mais discutidos por pessoas autistas no X (antigo Twitter). Entre os dias 10 de fevereiro e 14 de setembro de 2022, os *tweets* com a hashtag “*#ActuallyAutistic*” foram coletados e a modelagem de tópicos foi aplicada utilizando o BERTopic. O modelo gerou automaticamente 35 tópicos, posteriormente reduzidos para 20. Em seguida, os autores classificaram manualmente os tópicos identificados em seis temas principais.

- Experiências de indivíduos autistas;
- Conscientização sobre o autismo, orgulho e financiamento;
- Intervenções e ABA;
- Reações e expressões;
- Estilo de vida autista;
- Símbolos e características.

Os autores destacam que a identificação dos tópicos discutidos por pessoas autistas pode fornecer informações valiosas sobre problemas atuais enfrentados por essa comunidade. Esses resultados contribuem para o desenvolvimento de políticas públicas e direcionamento das pesquisas focadas nas necessidades dessa população.

¹⁰ Linguistic Inquiry and Word Count (LIWC). Um *software* para análise do uso de palavras. Disponível em: <https://www.liwc.app/>. Acesso em: 30 nov. 2024.

Além disso, o estudo de Beykikhoshk *et al.* (2015) tem como objetivo usar a plataforma Twitter como uma fonte de informação para estudar a população afetada pelo TEA, suas preocupações e necessidades. Inicialmente, os pesquisadores coletaram os *tweets* e os separaram em dois grupos: um a respeito de autismo e um grupo de controle composto por dados não relacionados ao autismo. Logo em seguida, foram realizados experimentos para descobrir características comuns e diferentes entre esses dois grupos. Por fim, demonstraram que os *tweets* no conjunto relacionados ao autismo continham informações detalhadas e bem fundamentadas, capazes de auxiliar os gestores de saúde pública na tomada de decisões precisas para atender às necessidades dos indivíduos autistas.

Diferentemente destes últimos trabalhos, o estudo de Liu *et al.* (2023), possui dois objetivos principais. O primeiro é identificar marcadores linguísticos que possam indicar sintomas de depressão, com isto, detectar sintomas normalmente associados à depressão por meio de palavras. Com a ajuda de profissionais da área de saúde, foram selecionados 13 sintomas, como ideação suicida, auto aversão e solidão, entre outros. O segundo objetivo consiste em treinar modelos de inteligência artificial para detectar estes sintomas nas publicações. Estes modelos utilizaram os marcadores e também os *embeddings* do modelo RoBERTa.

A descoberta mais relevante do estudo foi que sintomas como ideação suicida, auto aversão, distúrbios alimentares e solidão apresentam fortes associações com os marcadores linguísticos analisados. O modelo RoBERTa demonstrou ser mais eficaz na detecção da maioria destes sintomas.

Esses estudos destacam como as redes sociais podem ser ferramentas valiosas para compreender as experiências de diferentes populações. Inspirado por esses trabalhos, o presente estudo tem como objetivo investigar a comunidade “autism” no Reddit, utilizando técnicas de análise de sentimentos e modelagem de tópicos para identificar os temas mais recorrentes compartilhados pelos usuários autistas sobre seu cotidiano. O estudo busca mapear os desafios enfrentados por essa população e contribuir para a criação de ações que buscam construir um ambiente mais inclusivo e direcionado às suas necessidades.

3 DESENVOLVIMENTO

3.1 CONFIGURAÇÃO DO BANCO DE DADOS

Um banco de dados PostgreSQL 17 foi utilizado para armazenar os dados coletados do Reddit. A escolha desse banco deve-se à sua robustez e por ser uma solução de código aberto. Através dele, foi possível organizar os dados em duas tabelas: submissões e comentários. A primeira, conforme apresentado na Tabela 1, armazena o conteúdo da submissão, incluindo o título (*title*) e corpo do texto (*selftext*), bem como a quantidade de curtidas (*score*) e sua categoria (*link_flair_text*). Além disso, a tabela contém informações sobre o autor (*author*) e a descrição de seu perfil (*author_flair_text*), entre outros atributos importantes.

Tabela 1 - Submissões

(continua)

Coluna	Tipo de dado	Descrição
<i>id</i>	TEXT	Identificador da submissão (chave primária).
<i>name</i>	TEXT	Nome da submissão.
<i>subreddit</i>	TEXT	Nome da <i>subreddit</i> (comunidade).
<i>author</i>	TEXT	Nome do usuário.
<i>author_flair_text</i>	TEXT	Descrição do autor.
<i>title</i>	TEXT	Título da submissão.
<i>selftext</i>	TEXT	O conteúdo da submissão.
<i>created_utc</i>	BIGINT	Momento de criação da submissão, em Unix Time.
<i>score</i>	INTEGER	O número de votos positivos da submissão.
<i>upvote_ratio</i>	REAL	A porcentagem de votos positivos da submissão.
<i>is_original_content</i>	BOOLEAN	Indica se a submissão é um conteúdo original.
<i>distinguished</i>	TEXT	Indica se a submissão é distinta das outras.
<i>is_self</i>	BOOLEAN	Indica se submissão é uma postagem própria (somente texto)
<i>link_flair_text</i>	TEXT	A categoria da submissão.
<i>num_comments</i>	INTEGER	O número de comentários da submissão.

Tabela 1 - Submissões

(conclusão)

Coluna	Tipo de dado	Descrição
<i>over_18</i>	BOOLEAN	Indica se a submissão é proibida para menores de 18 anos.
<i>permalink</i>	TEXT	Um link permanente da submissão.
<i>stickied</i>	BOOLEAN	Indica se a submissão está fixada.
<i>url</i>	TEXT	A URL ao qual a submissão está vinculada ou o link permanente, caso seja uma postagem própria.

Na outra tabela (Tabela 2) foram armazenados os conteúdos dos comentários, juntamente com a identificação da submissão (*submission_id*), que é utilizada como chave estrangeira. Essa chave permite estabelecer a relação entre cada comentário e sua respectiva submissão, facilitando a realização de pesquisas que associam uma submissão a todos os seus comentários.

Tabela 2 - Comentários

Coluna	Tipo de dado	Descrição
<i>id</i>	TEXT	Identificador do comentário (chave primária).
<i>body</i>	TEXT	Conteúdo do comentário.
<i>created_utc</i>	BIGINT	Momento de criação do comentário, em Unix Time.
<i>distinguished</i>	TEXT	Indica se o comentário é distinto dos outros.
<i>link_id</i>	TEXT	Identificador da submissão deste comentário.
<i>parent_id</i>	TEXT	O identificador da submissão ou do comentário pai.
<i>permalink</i>	TEXT	Um link permanente do comentário.
<i>score</i>	INTEGER	O número de votos positivos do comentário.
<i>stickied</i>	BOOLEAN	Indica se o comentário está fixado.
<i>subreddit</i>	TEXT	Nome da subreddit (comunidade).
<i>author</i>	TEXT	Nome do usuário.
<i>subreddit_id</i>	TEXT	Identificador da submissão. (chave estrangeira)

3.2 COLETA DE POSTAGENS E COMENTÁRIOS DO REDDIT

A biblioteca PRAW¹¹ foi utilizada para coletar os dados dessa comunidade. Essa ferramenta simplifica o processo de interação com a API oficial do Reddit, abstraindo a complexidade das requisições. Além disso, ela fornece instâncias que representam as *subreddits* (comunidades), as submissões e os comentários, permitindo uma manipulação eficiente dos dados.

Durante a coleta, foi utilizada uma versão gratuita da *Reddit Data API*¹², que é limitada a até 1000 requisições a cada 10 minutos. Para respeitar essas limitações, foi implementado um temporizador no código, configurado para pausar as requisições por 10 minutos ao atingir o limite, como pode ser visto no Código 1.

O algoritmo foi estruturado da seguinte forma: inicialmente, o horário da execução é atribuído à variável *today_time*. Para calcular o intervalo de um dia, o valor equivalente a 24 horas em segundos (86.400) é armazenado na variável *one_day_seconds*. Subtraindo-se esse valor de *today_time*, obtém-se o *timestamp* correspondente ao mesmo horário 24 horas atrás.

Ao utilizar o método *subreddit.new*, com o limite definido como *None*, a API retorna todas as submissões mais recentes. Contudo, apenas aquelas que se enquadram no intervalo das últimas 24 horas são filtradas e coletadas. Assim, todas as postagens e comentários publicados dentro desse período são devidamente salvos no banco de dados.

Caso o número de requisições atinja o limite de 1.000, o programa entra em pausa por 600 segundos (10 minutos) utilizando o método *time.sleep*. Esse ciclo se repete até que todas as submissões e comentários publicados no intervalo de um dia tenham sido coletados. Quando as submissões excedem o limite temporal definido, o processo de coleta é encerrado.

As colunas do banco de dados apresentadas na Tabela 1 e Tabela 2 contêm todas as informações extraídas das submissões e dos comentários. A

¹¹ Python Reddit API Wrapper (PRAW) fornece uma interface para interagir com a API oficial do Reddit. Disponível em: <https://praw.readthedocs.io/en/stable/index.html>. Acesso em: 12 dez. 2024.

¹² Reddit Data API é a API oficial para dados do Reddit. Documentação oficial disponível em: <https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki> Acesso em: 28 nov. 2024.

estratégia adotada foi coletar mais dados do que o estritamente necessário, garantindo que, caso haja a necessidade de utilizar informações adicionais, elas já estejam disponíveis. Embora o Reddit permita que os usuários utilizem apelidos na plataforma, o que garante o anonimato, este trabalho tomou o cuidado de encriptar o nome de cada um dos autores para dificultar a sua identificação.

Código 1 - Função de coleta de dados

```

1. def collect_data(subreddit):
2.     today_time = math.floor(time.time())
3.     one_day_seconds = 86400 # Um dia atrás
4.     one_day_ago = today_time - one_day_seconds
5.     cont = 0
6.
7.     for submission in tqdm(subreddit.new(limit = None), desc="Coletando Submissões"):
8.         if submission.created_utc >= one_day_ago and submission.created_utc <=
today_time:
9.             get_submission(submission)
10.            cont = get_comments(submission, cont)
11.
12.            if cont != 999:
13.                cont += 1
14.            else:
15.                print("Você atingiu 1000 requisições. Espere 10 minutos!")
16.                cont = 0
17.                time.sleep(600)
18.        else:
19.            break

```

Este código foi configurado para ser executado uma vez ao dia, às 8:00 da manhã, por meio do Agendador de Tarefas do Windows 11. Essa coleta foi realizada entre os meses de junho e julho de 2024, como resultado, foi possível coletar 11.623 publicações e 135.894 comentários de 28.249 usuários. Foram coletadas, em média, 190,54 publicações por dia, com uma mediana de 215, conforme apresentado na Figura 3. Já os comentários apresentaram uma média de 2.227,77 por dia, com uma mediana de 2.474, como pode ser visto na Figura 4.

Figura 3 - Quantidade de submissões diárias coletadas.

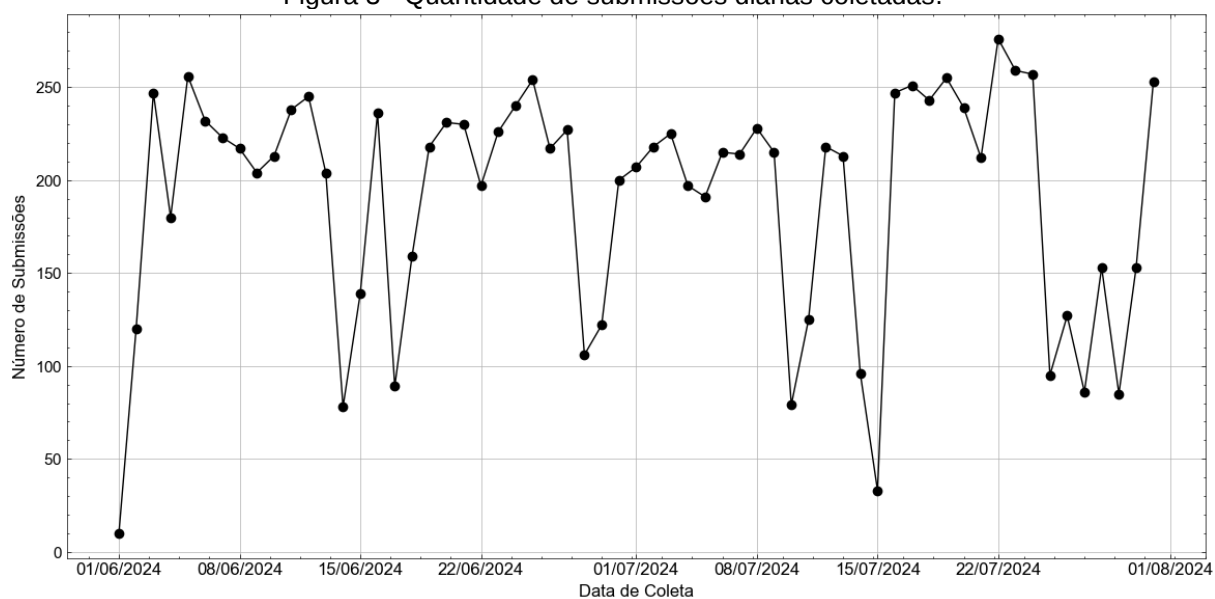
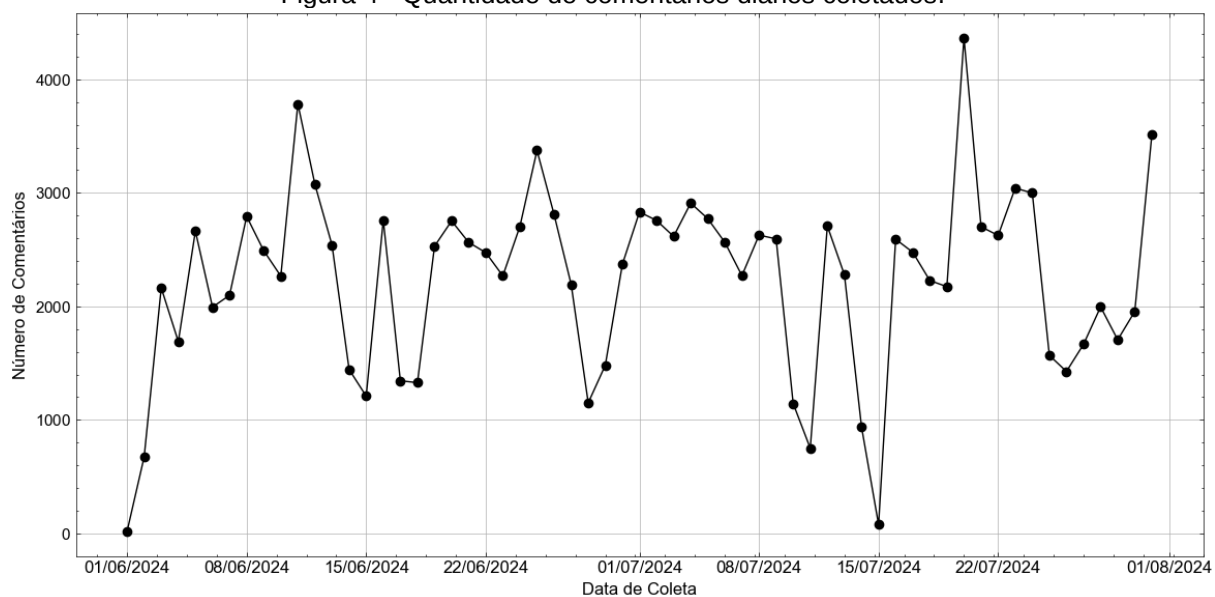


Figura 4 - Quantidade de comentários diários coletados.

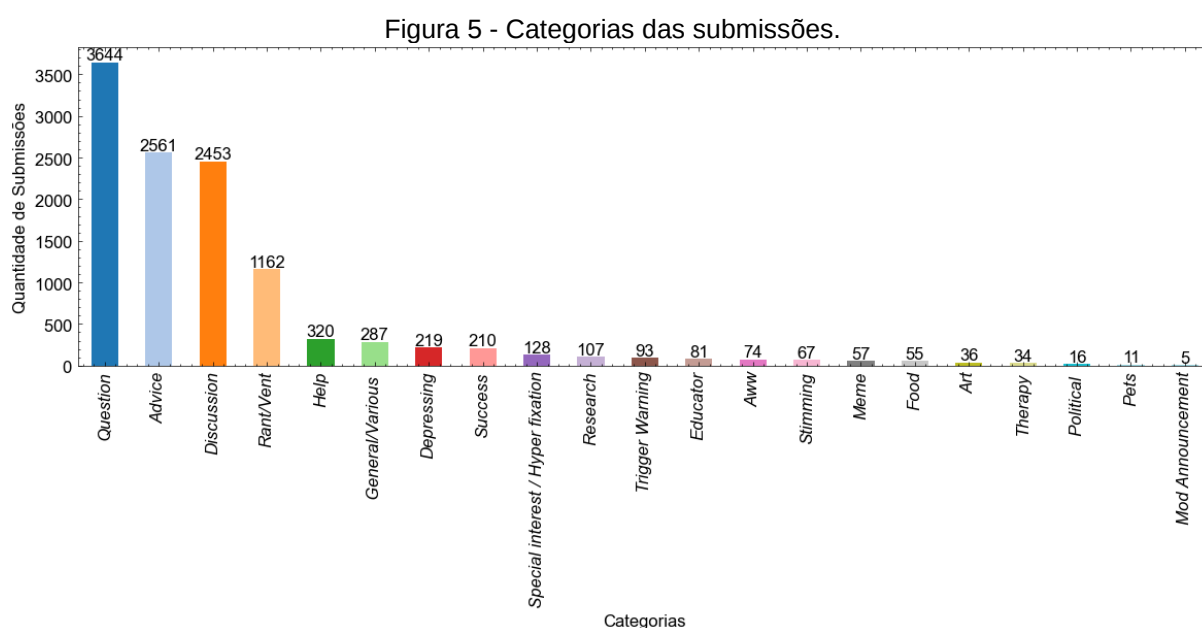


3.3 LIMPEZA E SELEÇÃO DOS DADOS

3.3.1 Das submissões

Para garantir uma análise precisa, foi necessário realizar uma etapa de limpeza nos dados, excluindo informações irrelevantes e selecionando apenas as submissões de usuários que se identificam como autistas em seus perfis. Essa escolha se deve ao objetivo do estudo, que é compreender os desafios comuns

enfrentados por pessoas autistas a partir de suas próprias perspectivas. Por esse motivo, não foram consideradas as percepções de pais e familiares sobre o autismo. Primeiramente, as publicações sem um corpo de texto (*selftext*) foram deletadas, priorizando as publicações com maior volume de informações, pois isso contribui nas previsões dos modelos. Posteriormente, as categorias das publicações (*link_flair_text*) foram listadas, conforme ilustrado na Figura 5. A categoria *Mod Announcement*, referente a anúncios dos moderadores da comunidade “autism”, foi excluída, por não agregar informações relevantes para esse estudo.



Para identificar as publicações de usuários autistas, foi utilizada a propriedade *author_flair_text*, que contém uma pequena biografia fornecida pelos próprios usuários. As biografias foram analisadas manualmente e palavras-chave que indicam a presença ou suspeita de autismo foram selecionadas. A Tabela 3 apresenta todas as palavras-chave identificadas.

Após a filtragem, alguns autores foram excluídos, pois embora apresentassem palavras-chave relevantes, eram pais de crianças autistas, e não indivíduos autistas. Por fim, o campo *created_utc*, que indica o instante de postagem da publicação em Unix Time (número de segundos desde 1º de janeiro de 1970), foi convertido em uma data e separado em cinco novos campos: *created_datetime* (Data e hora), *created_year* (Ano), *created_month* (Mês), *created_day* (Dia), e *created_date* (Data).

Tabela 3 - Palavras-chave dos usuários

Palavra-chave	Descrição
<i>audhd</i>	Transtorno do Déficit de Atenção e Hiperatividade (TDAH) e autismo.
<i>asd</i>	Sigla para Transtorno do Espectro Autista (<i>Autism Spectrum Disorder</i>).
<i>asperger</i>	Síndrome de Asperger.
<i>autism</i>	Transtorno do espectro autista
<i>autistic</i>	Adjetivo para descrever uma pessoa com autismo.
<i>self-diagnosed</i>	Indica que o usuário acredita ser autista, mas não possui um diagnóstico formal.
<i>self-suspecting</i>	Indica que o autor suspeita ser autista, mas não tem certeza.
<i>diagnosed</i>	Possui um diagnóstico formal de autismo.
<i>aspie</i>	Gíria usada para se referir a pessoas com Síndrome de Asperger.
<i>masking</i>	Ato de "mascarar" comportamentos autistas para se encaixar em normas sociais.
<i>austisms</i>	Transtorno do Espectro Autista.
<i>support</i>	Indivíduo autista que precisa de suporte.
<i>auuuutism</i>	Variação estilística de " <i>autism</i> ".
<i>diagnosis</i>	Diagnóstico formal de autismo.
<i>level</i>	Indica o nível de gravidade do autismo.
<i>neurodivergent</i>	Indivíduo neurodivergente, pode ser utilizado para descrever uma pessoa autista.
<i>labelled</i>	Rotulado como autista, seja pelo diagnóstico formal ou informal.
<i>tism</i>	Abreviação informal de " <i>autism</i> ".
<i>nuerodivergent</i>	Erro ortográfico de " <i>neurodivergent</i> ", que significa neurodivergente.
<i>undiagnosed</i>	Indivíduos que acreditam ser autistas, mas que não têm um diagnóstico formal.
<i>aut</i>	Abreviação de " <i>autism</i> ".
<i>functioning</i>	Refere-se ao nível de funcionamento de uma pessoa autista.
<i>auocdh</i>	Transtorno obsessivo-compulsivo (TOC) e autismo.

3.3.2 Dos comentários

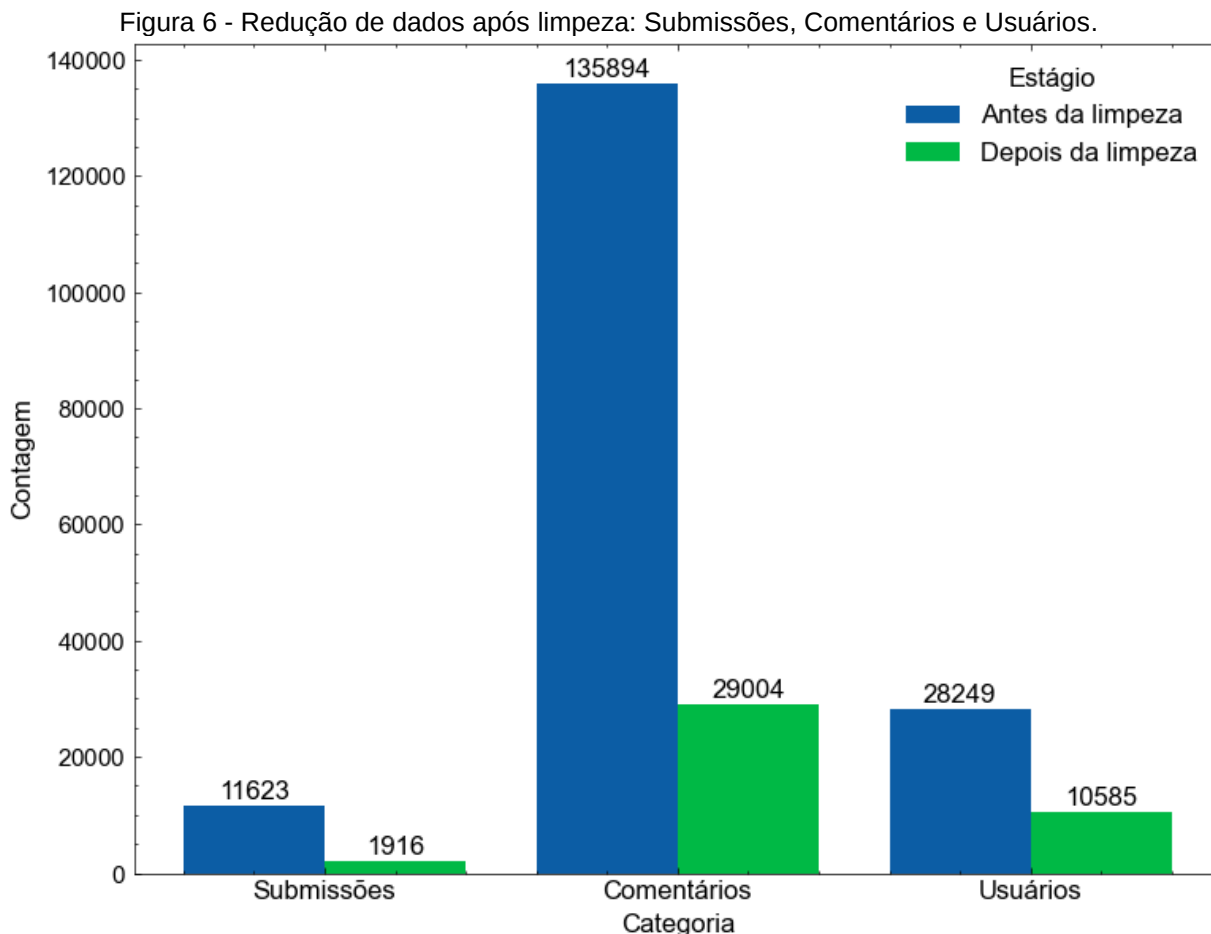
Após a limpeza das submissões, os comentários relacionados às publicações descartadas foram excluídos. Diferentemente das submissões, os comentários não foram filtrados para incluir apenas os de pessoas autistas. Essa abordagem teve como objetivo ampliar o conjunto de informações disponíveis e

enriquecer a análise considerando como a comunidade de forma geral responde às submissões e quais sentimentos são demonstrados em relação a elas.

Em seguida, os comentários sem corpo de texto e também aqueles enviados por moderadores (*distinguished*) foram excluídos, pois não apresentavam informações relevantes para a análise. Por fim, o campo *created_utc* recebeu o mesmo tratamento das submissões, sendo convertido em cinco novos campos: *created_datetime* (Data e hora), *created_year* (Ano), *created_month* (Mês), *created_day* (Dia), e *created_date* (Data).

4.1.1 Submissões e comentários antes e depois da filtragem

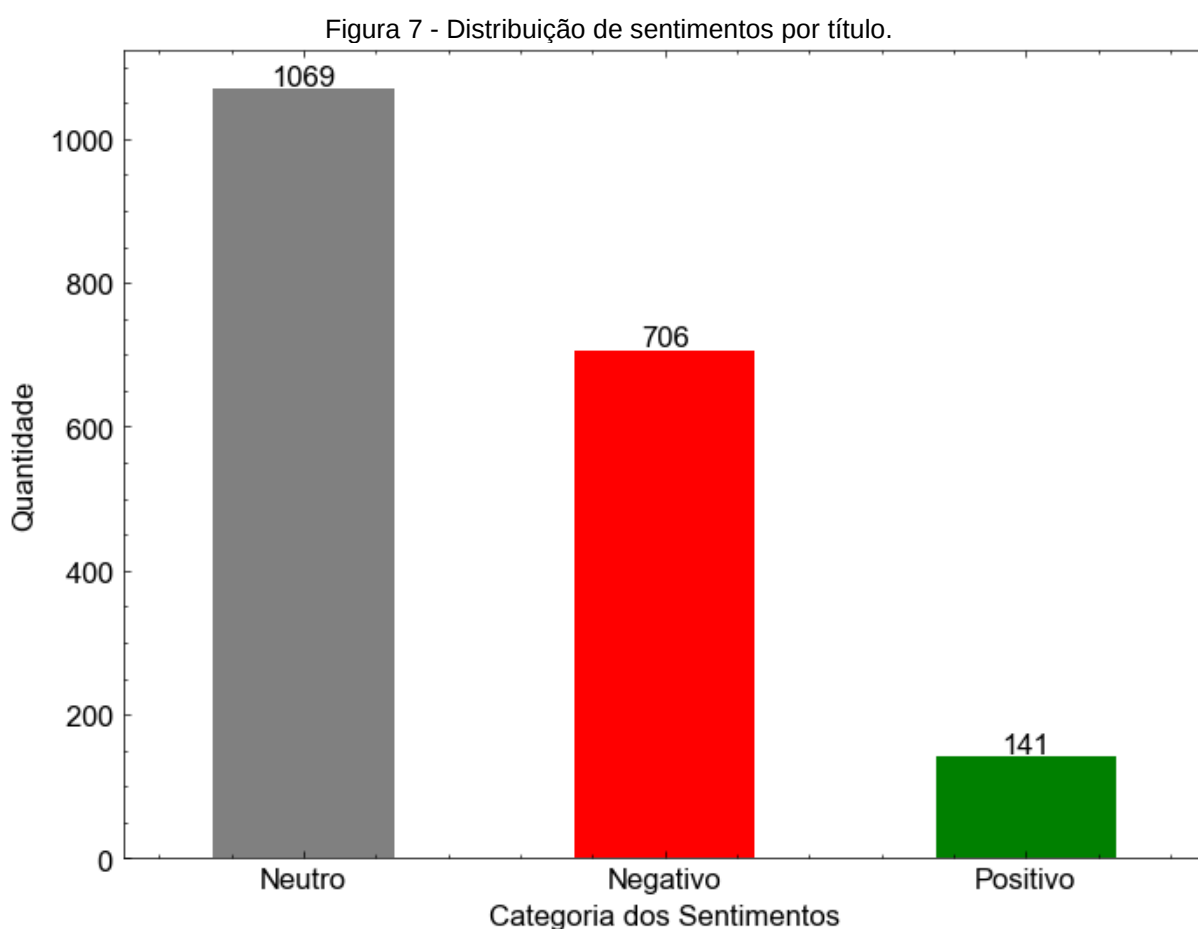
Após a etapa limpeza, as 11.623 publicações foram reduzidas para 1.916, enquanto o número de comentários diminuiu de 135.894 para 29.004. Além disso, o total de usuários foi reduzido de 28.249 para 10.585, como pode ser observado na Figura 6.



3.4 CLASSIFICAÇÃO DE SENTIMENTOS

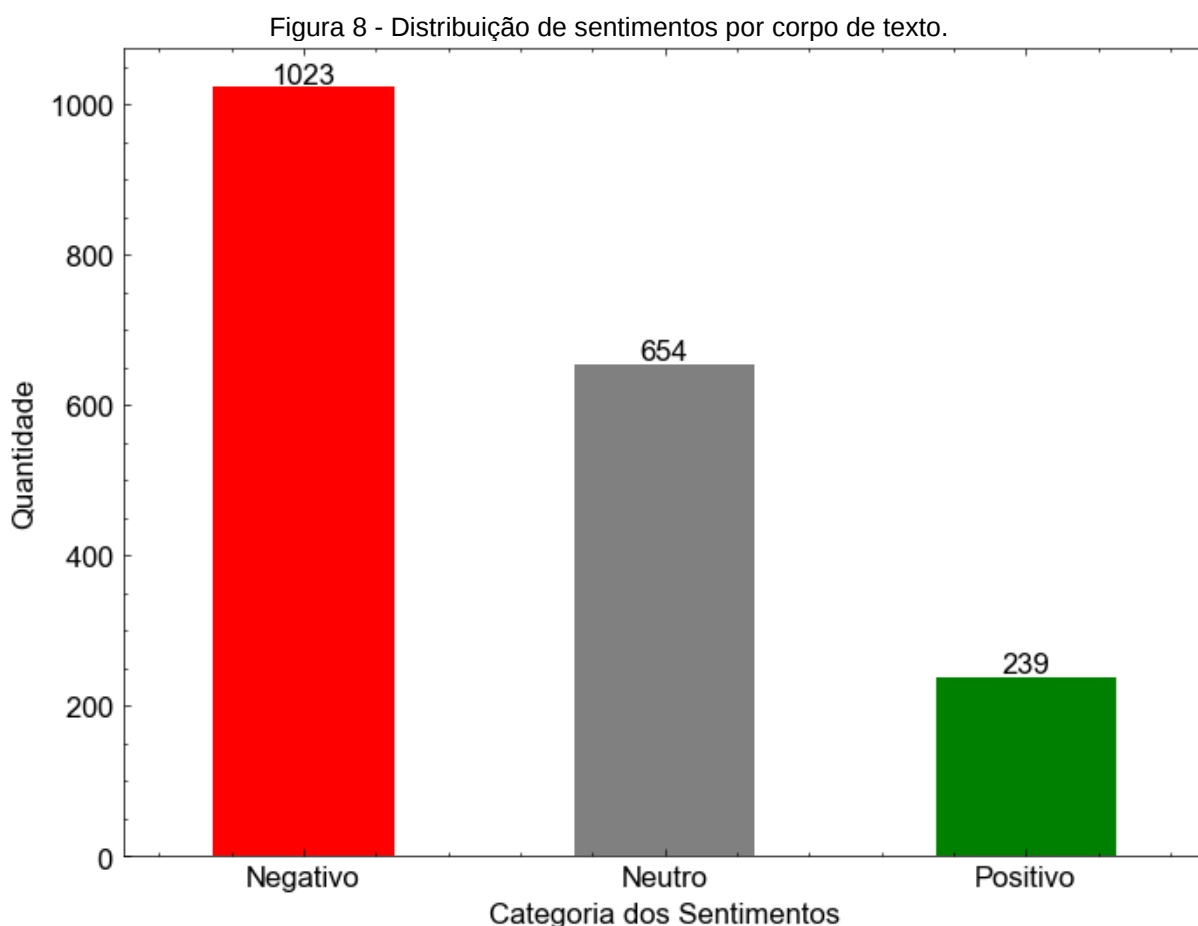
Conforme apresentado no Apêndice A, o modelo “*Twitter-roBERTa-base for Sentiment Analysis*” foi utilizado para realizar a classificação dos sentimentos em três categorias: negativo, neutro e positivo. As submissões selecionadas para análise contêm dois componentes textuais relevantes: o título (*title*); e o corpo de texto (*selftext*).

Para verificar a possibilidade de inferir os sentimentos de uma submissão apenas pelo seu título, foi realizada uma análise inicial utilizando apenas o campo *title*. Os resultados indicaram que 1.069 publicações foram classificadas como neutras, 706 como negativas e 141 como positivas (veja na Figura 7). A predominância de publicações neutras sugere que os títulos, isoladamente, não possuem as informações contextuais necessárias para a análise.



Diante das limitações identificadas no *title*, decidiu-se utilizar o campo *selftext* das submissões. Essa propriedade, por ser mais desenvolvida, oferece

informações mais completas e potencialmente mais precisas sobre os sentimentos predominantes. Como resultado, foram observadas 1.023 publicações identificadas como negativas, 654 como neutras e 239 como positivas (veja na Figura 8).



Ao final, embora o corpo do texto apresentasse um contexto adequado para inferência de sentimentos, optou-se por combinar o título e o corpo. Essa escolha foi realizada com o objetivo de assegurar que informações relevantes não fossem omitidas. Os resultados finais indicaram 1.059 submissões identificadas como negativas, 639 como neutras e 218 como positivas (veja na Figura 9).

Após a análise das submissões, os comentários foram classificados. Diferentemente das submissões, que possuíam tanto título quanto corpo de texto, os comentários contêm apenas o campo *body*, o que simplifica o processo de predição. Eles foram classificados em 11.364 neutros, 11.216 negativos e 6.424 positivos (veja na Figura 10).

Figura 9 - Distribuição de sentimentos por título e corpo de texto.

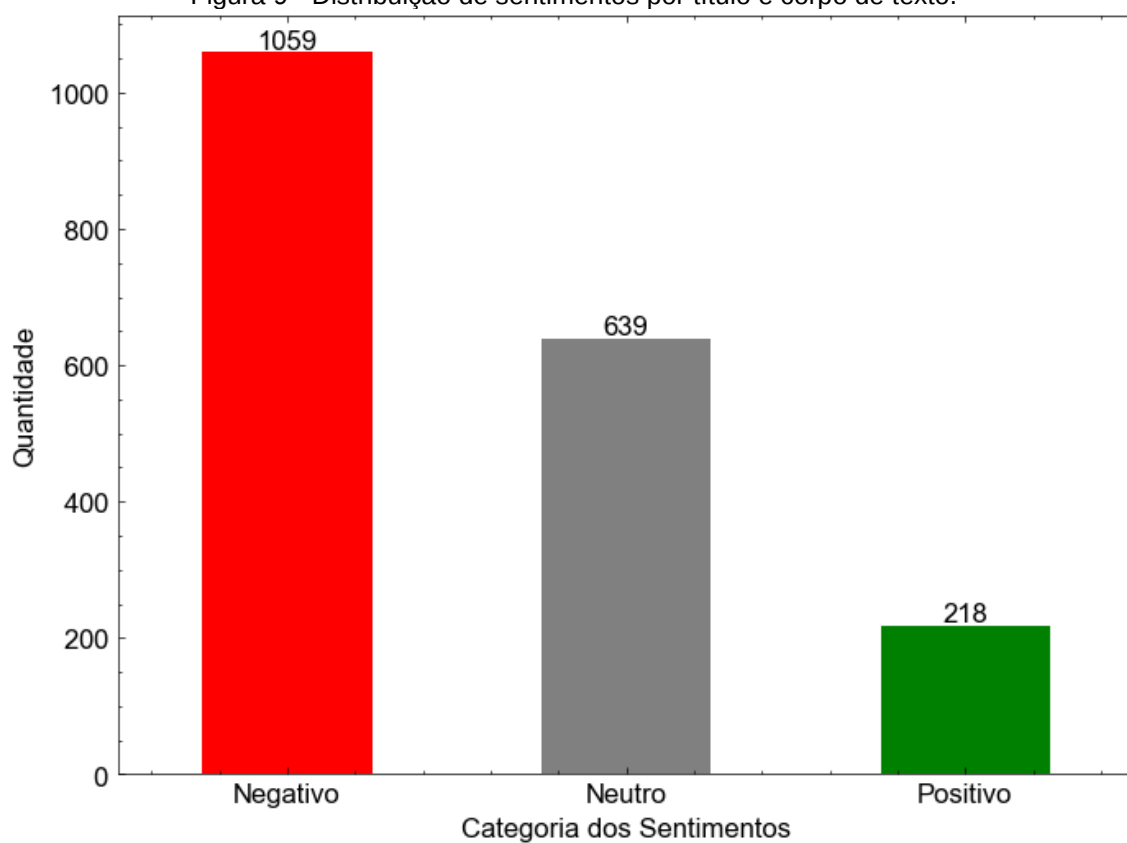
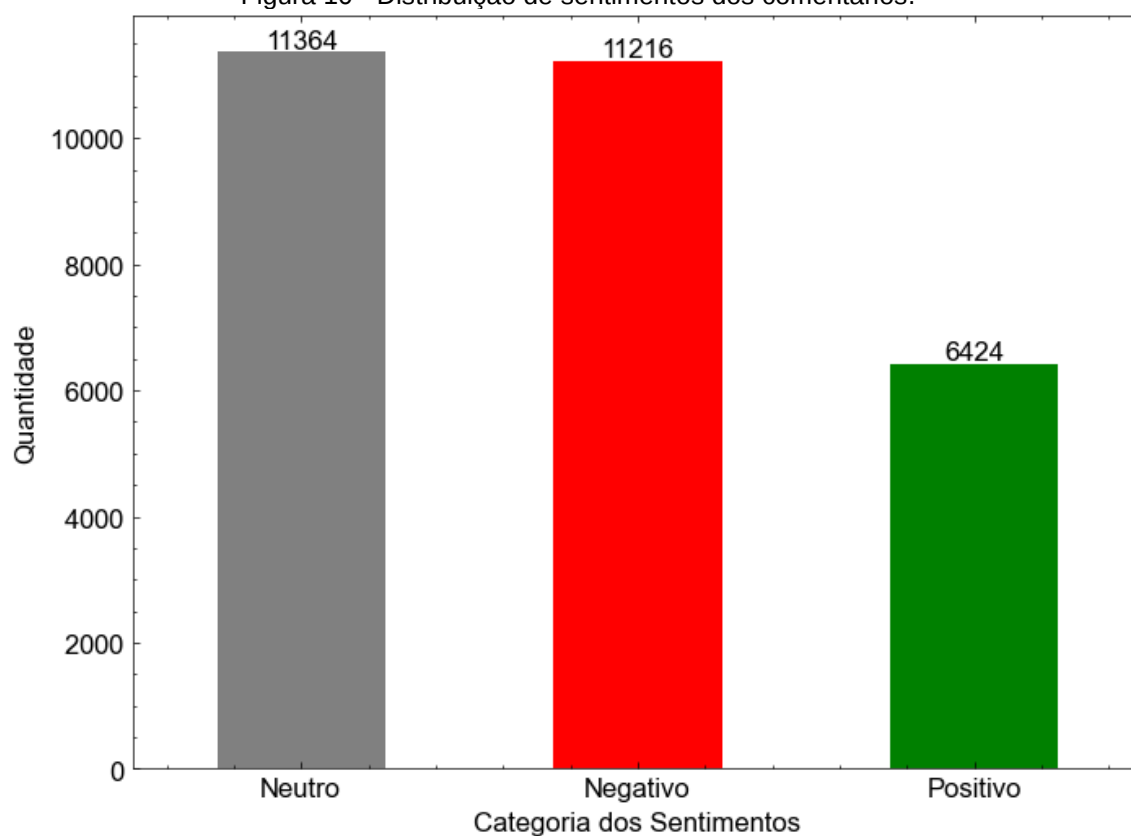


Figura 10 - Distribuição de sentimentos dos comentários.



3.5 EXTRAÇÃO DE TÓPICOS

3.5.1 Preparação inicial

A modelagem de tópicos, como pode ser visto no Apêndice B, foi realizada utilizando o BERTopic, uma ferramenta altamente versátil que permite o uso de diversos modelos para a extração de tópicos. O processo começou pelas submissões, sendo necessário, antes disso, combinar o título e o corpo de texto da publicação, conforme realizado anteriormente na fase de análise de sentimentos, para maximizar as informações contextuais disponíveis.

Logo em seguida, durante os testes iniciais, verificou-se que certas palavras nos tópicos extraídos não contribuíam para a análise. Palavras como “autism”, “autistic” e “autistics”, eram frequentemente destacadas pelo algoritmo e apareciam na maioria dos tópicos encontrados. Considerando que os dados foram coletados de uma comunidade específica sobre autismo, o tópico principal sempre girava em torno do transtorno, dificultando a identificação de subtópicos mais relevantes. Para contornar esse problema, as palavras encontradas foram removidas através do Código 2, permitindo uma análise mais profunda dos temas discutidos por pessoas autistas.

Após a remoção, foi encontrado um tema com a palavra “http”, o que exigiu a remoção de todas as URLs do conjunto de texto. Concluído o pré-processamento, iniciou-se a fase de configuração e escolha dos modelos que serão utilizados pelo BERTopic.

Código 2 – Função que remove palavras.

```
1. def remove_words(data):
2.     stop_words = ["autism", "autistic", "autistics", "aspergers"]
3.     pattern = r'\b(?:' + '|'.join(map(re.escape, stop_words)) + r')\b'
4.     # Substitui as palavras por uma string vazia.
5.     return re.sub(pattern, '', data, flags=re.IGNORECASE).strip()
```

A primeira etapa foi a escolha de um bom modelo de *Embedding*, que fosse capaz de representar os dados textuais de forma eficiente. Com o objetivo de encontrar um modelo que apresentasse um bom desempenho com dados do Reddit,

consultou-se o ranking do *Massive Text Embedding Benchmark* (MTEB)¹³, que contém a classificação dos melhores modelos que passaram pelo processo de avaliação proposto por Muennighoff *et al.* (2023). Nesse ranking, o modelo que apresentou melhor desempenho com dados do Reddit e compatibilidade com o hardware utilizado neste estudo foi o *hkunlp/instructor-large*, um modelo capaz de gerar *embeddings* para qualquer tarefa, incluindo agrupamento (SU *et al.*, 2023), especialidade do BERTopic.

Para reduzir a dimensionalidade das *embeddings* geradas pelo modelo *instructor-large* e realizar o agrupamento dos dados, foram utilizados o UMAP e o HDBSCAN, respectivamente, que são os modelos padrões do BERTopic. A tokenização foi realizada utilizando o modelo *CountVectorizer*¹⁴, com o parâmetro *stop_words* configurado como "*english*" para remover automaticamente termos de alta frequência e baixo valor semântico (palavras de parada).

Por fim, após o processamento final do BERTopic utilizando o c-TF-IDF para criar uma representação precisa dos tópicos, o "*ChatGPT 3.5 Turbo 0125*" foi empregado para o refinamento e aprimoramento dos tópicos identificados, visando uma análise mais detalhada e precisa. Para otimizar a performance do modelo, foi usado o Código 3, com o prompt recomendado pelo BERTopic.

3.5.2 Tópicos das Submissões

Ao aplicar o BERTopic em todas as submissões, foram identificados 27 tópicos distintos, abrangendo temas como: saúde mental, ansiedade, diagnósticos tardios, problemas sensoriais, entre outros. Para ilustrar os resultados, três tópicos selecionados estão representados na Tabela 4.

¹³ *Massive Text Embedding Benchmark* (MTEB): uma avaliação comparativa do desempenho de modelos de embeddings de texto em várias tarefas de processamento de linguagem natural. Disponível em: <https://huggingface.co/spaces/mteb/leaderboard>. Acesso em: 7 dez. 2024.

¹⁴ *CountVectorizer*: É um modelo da biblioteca *scikit-learn* que converte dados textuais em uma matriz de tokens. Disponível em: https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html. Acesso em: 7 dez. 2024.

Código 3 - Prompt do ChatGPT.

```

1. client = openai.OpenAI(api_key=chat_gpt_key)
2. prompt = ""
3.     I have a topic that contains the following documents:
4.     [DOCUMENTS]
5.     The topic is described by the following keywords: [KEYWORDS]
6.     Based on the information above, extract a short but highly descriptive
topic label of at most 5 words. Make sure it is in the following format:
7.     topic: <topic label>
8. ""
9. representation_model = OpenAI(client, model="gpt-3.5-turbo-0125",
exponential_backoff=True, chat=True, prompt=prompt)

```

Quando aplicado exclusivamente às submissões classificadas como positivas, foram identificados apenas 2 tópicos principais: um relacionado ao ato de desmascarar¹⁵ e outro sobre músicas, conforme apresentado na Tabela 5.

Tabela 4 - Tópicos avulsos das submissões

(continua)

Nome	Quantidade	Documento representativo
<i>Mental Health Struggles and Social Anxiety</i>	697	<i>"I feel unlovable Every time I think I've finally made a friend, they cancel plans on me. It just happened again and I don't know what I'm doing wrong. I want to have friends. I love people. I really really want friends. I don't think I've ever had a real friend before and it's starting to get to me."</i>

Nota: A tradução da Tabela 4 pode ser encontrada no Apêndice C (Tabela C.1)

¹⁵ Desmascaramento: é o processo de revelar características autistas que foram escondidas por meio da camuflagem.

Tabela 4 - Tópicos avulsos das submissões

(conclusão)

Nome	Quantidade	Documento representativo
<i>Late Diagnoses and Support Needs</i>	158	“Autism eval Tuesday at 215 CST The only reason I can have this done is because it is online and they take my Medicaid. I have researched so much at least since 2016 when I was 27. I really feel like I match the traits especially what they say about female adults. After learning about masking, I realized that was what I had been doing like automatically. \n\nAnyhow, I am nervous and anxious. I'm afraid that I cannot answer something just right and that I will not get the diagnosis I need. I could get so nervous that I can't stim and what if I mask and they cannot tell that I am autistic. Granted it *is* getting harder to mask as I get older but still. \n\nSomeone in my discord gave me good advice but I would like to hear more. Oh and btw I have a binder with 10 sheets written front and back of why I feel like I am autistic. I also have print outs that I relate to and also my RAADS-R score too. I am trying to be as prepared as possible. I have anxiety disorders too. My husband will try to attend with me if he gets out of work in time. I hope they are knowledgeable about autism in female adults.”
<i>Social Skills and Communication Struggles</i>	55	“Do you prefer talking to adults rather than children or teens? Me, (15M) has a hard time talking to people of my age. But with adults, i can talk for hours nonstop without any issue. \n\nAre there any other young people who have the same as me? \n\nAnd do you think it is Autism-related?”

Nota: A tradução da Tabela 4 pode ser encontrada no Apêndice C (Tabela C.1)

Ao aplicar o BERTopic exclusivamente às submissões negativas, surgiram 9 tópicos que abordam temas como comunicação não verbal, problemas sensoriais e bullying na escola. A Tabela 6 apresenta três desses tópicos como exemplo. O Apêndice C contém a versão traduzida das tabelas dos comentários.

Tabela 5 - Tópicos das submissões positivas

Nome	Quantidade	Documento representativo
<i>Unmasking exhaustion in travel</i>	190	“Getting to use to unmasking I'm taking baby steps and today was a good day. When I started to get overwhelmed I communicated it to my family and later explained that my speaking tone will probably be flat/monotone and my facial expressions will be less expressive than usually for the rest of the day because I just don't have the energy to actively focus on them. They luckily understood it and thanked me for letting them know so we wouldn't misunderstand each other.”
<i>Music Favorites and Poetic Lines</i>	75	“Anyone else obsessed with listening to music? Me personally there is not a second where I'm not listening to music. I personally love aurora, aviva, and neoni. Any suggestions for music artists and does anyone else experience this?”

Nota: A tradução da Tabela 5 pode ser encontrada no Apêndice C (Tabela C.2)

Tabela 6 - Tópicos avulsos das submissões negativas

Nome	Quantidade	Documento representativo
<i>Emotional Detachment and Social Apathy</i>	873	<i>"DAE hate absolutely ALL forms of showing/receiving love? from what i understand, disliking or being uncomfortable with physical touch is a common trait of autism. however, not only do i hate physical touch, but i hate people telling me im loved, i hate the thought of people caring about me, i hate people complimenting me, and the list goes on.\n\nnot only that, but i dont like giving those things out either. if someone needs comfort, i will likely just go blank, because the thought of dealing with it makes me squirm. i feel like a bad person sometimes, but most of the time my behaviour doesnt really bother me. is this normal for anyone else?"</i>
<i>Nonverbal Communication and Language Challenges</i>	46	<i>"semi-nonverbal since i got diagnosed 8 years ago, i've been experimenting with going nonverbal. some days it's more helpful to not speak, but i tend to get stuck being nonverbal for the rest of the day. it literally feels like i'll puke if i do speak, so im just wondering if anyone else has this experience. i'm also open to hearing why anyone else goes nonverbal, especially from folks that have decided to stop speaking permanently"</i>
<i>Special Interests and Social Anxiety</i>	31	<i>"People not talking to you. Whenever I bring up one of my interests, the conversation ends. I don't get interested in anything other than my special interests. This leads to a hard social life because I am not interested in many subjects. I'm mostly interested in Rain World and Bugs. One is a niche indie game and the other is something often considered gross.\n\nSorry if my language is bad \n\nsincerely, a level two autistic \n"</i>

Nota: A tradução da Tabela 6 pode ser encontrada no Apêndice C (Tabela C.3)

3.5.3 Tópicos dos Comentários

Os comentários foram classificados utilizando o mesmo tratamento e configuração aplicados na modelagem das submissões. Como resultado, ao considerar todos os comentários, foram identificados 247 tópicos, abrangendo temas como agradecimento e aprovação, músicas e problemas na infância, conforme apresentado na Tabela 7. Quando analisados apenas os comentários positivos, emergiram 48 tópicos, que abordam, entre outros assuntos, personagens de ficção, música e dinâmicas sociais (Tabela 8). Já a análise dos comentários negativos revelou 68 tópicos, incluindo desafios do autismo, problemas sensoriais e, de forma inesperada, até mesmo casos de má conduta policial (Tabela 9).

Tabela 7 - Três maiores tópicos de comentários

Nome	Quantidade	Documento representativo
<i>Expressing Gratitude and Approval</i>	4116	<i>"Couldn't agree more."</i>
<i>Diverse Music Selection</i>	1540	<i>"Linkin Park and a few other rock/metal artists"</i>
<i>Early Childhood Social Challenges</i>	693	<i>"I thought I was ostracised as a teenager because I was gay. Turns out I was ostracised from my very first day at school years earlier due to ASD.\n\nJust wonderful"</i>

Nota: A tradução da Tabela 7 pode ser encontrada no Apêndice C (Tabela C.4)

Tabela 8 - Três maiores tópicos dos comentários positivos

Nome	Quantidade	Documento representativo
<i>Favorite Characters and Relationship Support</i>	3582	<i>"Cherry blossom + Reki from Sk8 infinity, Sanji and Ussop from one piece :3"</i>
<i>Eclectic Music Taste</i>	675	<i>"for me it's Runaway and Through The Eyes by AURORA. (and maybe Warrior Of Love from the same artist too ! by the way if you haven't listened to her please do, her voice is so angelic !)\n\nand there's also I Don't Need Your Love from Six The Musical !\xa0"</i>
<i>Social dynamics in neurodiverse individuals</i>	340	<i>"I'm the center of the party, outgoing and funny. But only when it's a small group of really close friends.\n\nI'm the "normal" sit in the corner and be miserable in any other social situation."</i>

Nota: A tradução da Tabela 8 pode ser encontrada no Apêndice C (Tabela C.5)

Tabela 9 - Três maiores tópicos dos comentários negativos

Nome	Quantidade	Documento representativo
<i>Challenges and Perspectives on Autism</i>	8416	<i>"I've been told several times I don't look autistic. I think they expect someone who is obviously impaired intelectually. It's just ignorance."</i>
<i>Sensory Sensitivities to Textures</i>	288	<i>"When you touch dry rocks (like on a beach) and they leave the dusty/chalky feeling that dries out your hands absolutely kills me"</i>
<i>Corrupt and Abusive Police Culture</i>	211	<i>"Like I've met amazing cops, but I'm also terrified to meet a terrible one\n\nANYONE in a position of power and is a shitty person is dangerous \n\nThat's why it's important to have checks and balances \n\nCurrently, there aren't good checks or balances for law enforcement."</i>

Nota: A tradução da Tabela 9 pode ser encontrada no Apêndice C (Tabela C.6)

3.5.4 Análise e Refinamento dos Tópicos Extraídos

Os 27 tópicos extraídos a partir da análise de todas as submissões foram revisados manualmente. Durante esse processo, os tópicos *"Emotional Response to*

Puzzle Symbol” e *“LGBTQIA+ Community Polls”* foram removidos. O primeiro foi descartado por apresentar amostras inconsistentes com o tema, já o segundo por se tratar de postagens de enquetes destinadas a descobrir a sexualidade de outros usuários, um tipo de informação que não contribui para a proposta deste estudo.

Por outro lado, os tópicos *“Challenging Job Accessibility Issues”* e *“Coping with Work Burnout”* foram agrupados em um único tópico, denominado “Acessibilidade no Trabalho e Saúde Mental”, pois ambos abordam os desafios enfrentados por pessoas autistas no ambiente de trabalho. Em outros casos, optou-se por simplesmente traduzir o nome do tópico, como no caso de *“Social Skills and Communication Struggles”*, traduzido para “Dificuldades nas habilidades sociais e na comunicação”.

Ao final do processo, todas as submissões foram reorganizadas e agrupadas em 15 tópicos principais, conforme apresentado na Tabela 10. Para essa união, foram considerados os nomes originais e uma análise das amostras representativas de cada grupo. O agrupamento dos tópicos foi realizado com o propósito de simplificar a compreensão e facilitar a análise dos dados.

Tabela 10 - Tópicos das submissões reorganizados e renomeados

(continua)

	Tópico Original	Novo tópico
1	<i>Challenging Job Accessibility Issues</i> <i>Coping with Work Burnout</i>	Acessibilidade no Trabalho e Saúde Mental
2	<i>Clothes and Sensory Preferences</i> <i>Unwanted Touch and Familial Conflict</i> <i>Noise-Canceling Headphone Recommendations</i> <i>Sensory Food Sensitivity Challenges</i> <i>Sunscreen Sensory Nightmare</i> <i>Sensory Issues with Hygiene</i>	Desafios sensoriais e maneiras de lidar com eles
3	<i>Diverse Passionate Hobbies</i> <i>Cosplay and Tokusatsu fandom</i> <i>Diverse Special Interests and Books</i> <i>Complexities of Shifting Special Interests</i>	A diversidade de hobbies e interesses especiais no autismo
4	<i>Late Diagnoses and Support Needs</i> <i>Celebrating ADHD Diagnosis Journey</i>	Diagnóstico Tardio
5	<i>Coping with Upsetting Family Event</i> <i>Family Manipulation and Relationship Struggles</i>	Lidando com eventos familiares desafiadores e dificuldades nos relacionamentos

Tabela 10 - Tópicos das submissões reorganizados e renomeados

(conclusão)

	Tópico Original	Novo tópico
6	<i>Mental Health Struggles and Social Anxiety</i>	Dificuldades de Saúde Mental e Ansiedade Social
7	<i>Social Skills and Communication Struggles</i>	Dificuldades nas habilidades sociais e na comunicação
8	<i>Rediscovering passion through music</i>	Redescobindo a Paixão Através da Música
9	<i>Drinking Habits and Water Preferences</i>	Hábitos de Consumo e Preferências por Água
10	<i>Online Community Interaction Frustrations</i>	Frustrações nas Interações em Comunidades Online
11	<i>Living with Multi-Disability Conditions</i>	Vivendo com Múltiplas Deficiências
12	<i>Overcoming fears and achieving goals</i>	Superando medos e alcançando objetivos
13	<i>Teenage School Bullying Experiences</i>	Experiências de bullying escolar na adolescência
14	<i>Medication Effects Study Analysis</i>	Análise do estudo dos efeitos da medicação
15	<i>Memory and Forgetfulness Experiences</i>	Experiências com Memória e Esquecimento

No caso das submissões positivas, o tópico “*Unmasking Exhaustion in Travel*” foi excluído, por apresentar inconsistências durante a análise manual. Assim, sobrou apenas o tópico “*Music Favorites and Poetic Lines*”, que por sua vez foi traduzido para “Músicas Favoritas”, simplificando sua identificação e alinhando-o ao contexto da análise.

Ao levar em consideração apenas as submissões negativas, foi necessário realizar apenas a tradução e adaptação dos nomes após a avaliação manual. O tópico “*Challenges in Finding Remote Work*”, por exemplo, foi traduzido e adaptado para “Desafios Acerca do Trabalho” (Tabela 11), pois nas amostras analisadas diversos usuários discutiam problemas relacionados ao ambiente de trabalho no geral, e não especificamente questões relacionadas a encontrar empregos ou ao trabalho remoto.

Tabela 11 - Tópicos das submissões negativas reorganizados e renomeados

(continua)

	Tópico Original	Novo tópico
1	<i>Emotional Detachment and Social Apathy</i>	Desconexão Emocional e Apatia Social
2	<i>Nonverbal Communication and Language Challenges</i>	Comunicação não verbal e desafios linguísticos
3	<i>Special Interests and Social Anxiety</i>	Interesses Especiais e Ansiedade Social

Tabela 11 - Tópicos das submissões negativas reorganizados e renomeados
(conclusão)

	Tópico Original	Novo tópico
4	<i>Challenges in Finding Remote Work</i>	Desafios acerca do Trabalho
5	<i>Sensory Sensitivity and Noise Challenges</i>	Sensibilidade Sensorial e Desafios com Ruídos
6	<i>Humanity's Struggle with Evil</i>	A luta da humanidade contra o mal
7	<i>School Bullying Traumas</i>	Traumas de Bullying Escolar
8	<i>Noise-Cancelling Ear Protection Options</i>	Opções de proteção auricular com cancelamento de ruído
9	<i>Memory and hydration challenges</i>	Desafios de Memória e Hidratação

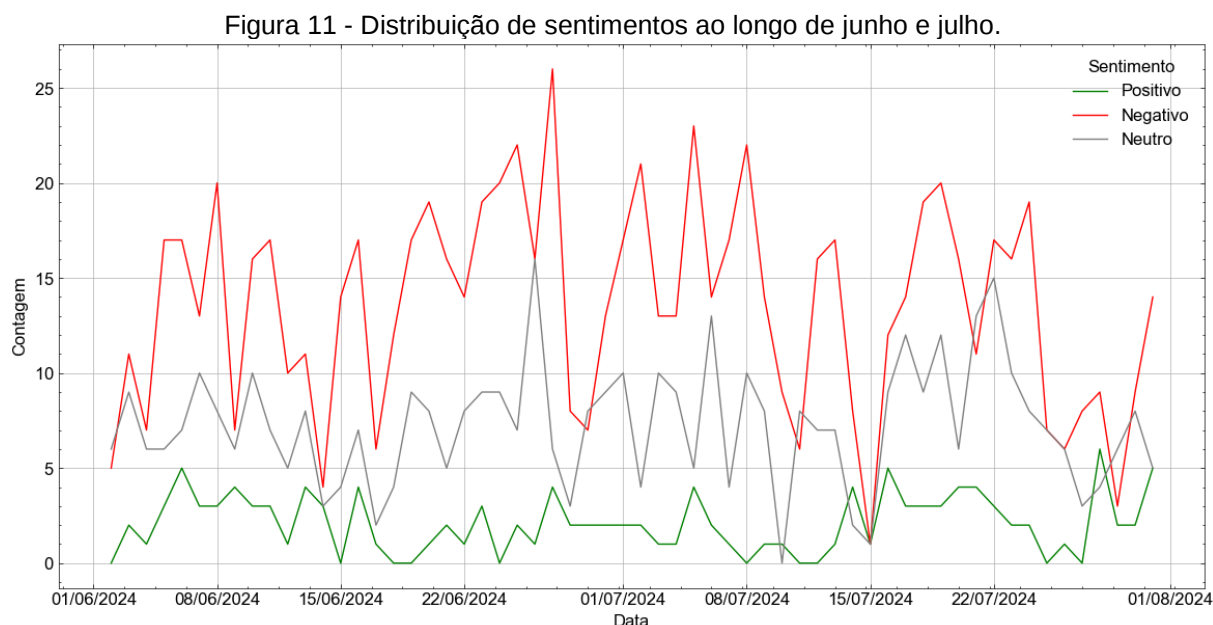
Devido à grande quantidade de tópicos encontrados nos comentários (247 no total, considerando todos os comentários, 48 com apenas os positivos e 68 nos negativos), com o intuito de assegurar a concisão deste trabalho, optou-se por considerar apenas os 10 tópicos mais frequentes relacionados a cada tema de submissão. Na análise dos resultados (Capítulo 4), serão explorados os 5 tópicos mais relevantes das submissões e os 10 tópicos mais relevantes dos comentários associados a cada tema. Ressalta-se que os temas dos comentários apresentados foram traduzidos para facilitar o entendimento.

4 ANÁLISE DOS RESULTADOS

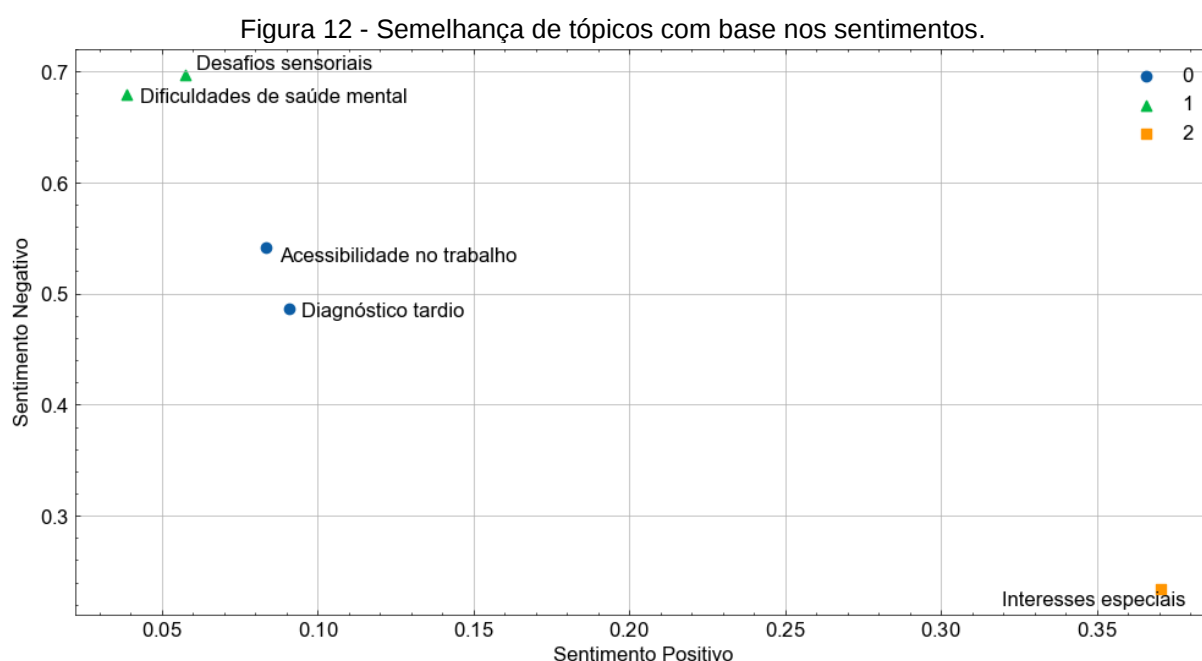
A partir da análise de sentimentos e da modelagem de tópicos, foram identificadas informações relevantes sobre a comunidade “autism”. Este capítulo apresenta as principais descobertas e características dos temas discutidos, demonstrando o potencial do Reddit como fonte valiosa de dados para estudos voltados à compreensão de experiências de pessoas autistas.

Conforme mencionado anteriormente, devido à vasta quantidade de informações coletadas, decidiu-se concentrar a análise nos cinco tópicos mais representativos das submissões: Dificuldades de saúde mental e ansiedade social; Acessibilidade no trabalho e saúde mental; A diversidade de hobbies e interesses especiais no autismo; Diagnóstico Tardio e Desafios sensoriais e maneiras de lidar com eles.

Uma das primeiras características observadas foi a distribuição das submissões ao longo do tempo. Ao analisá-las em um gráfico, observou-se uma predominância de sentimentos negativos durante os dois meses de coleta dos dados, em junho e julho. Essa tendência está representada na Figura 11.



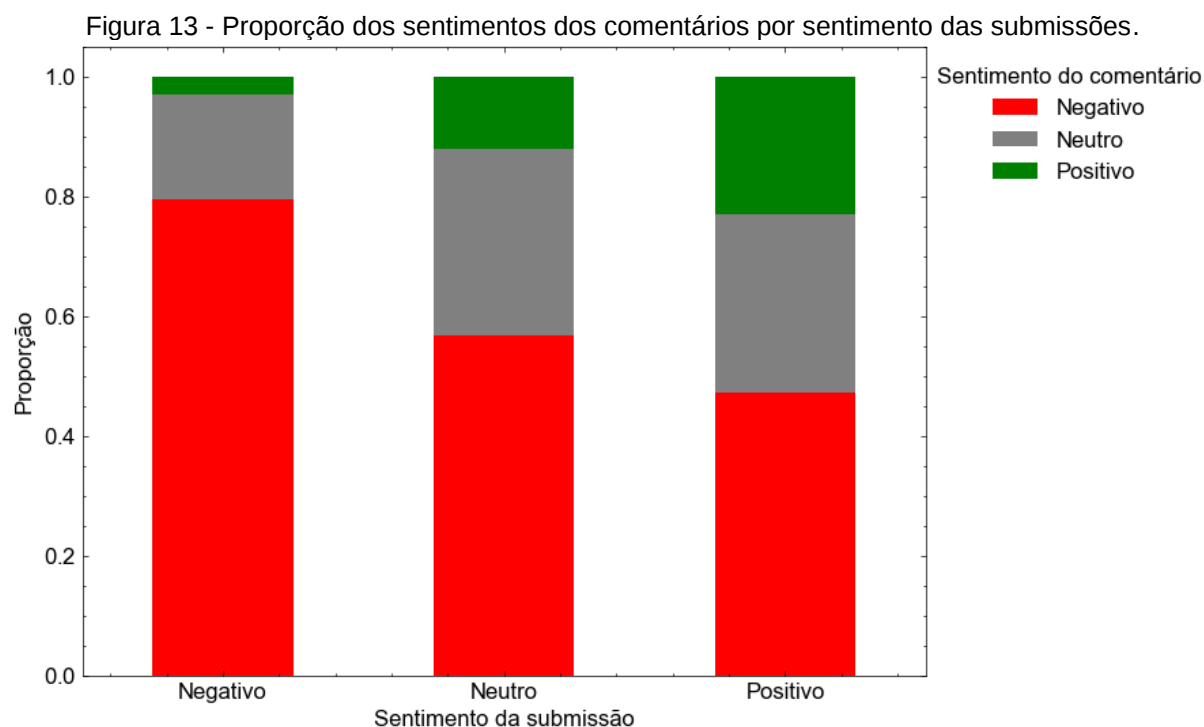
A aplicação do algoritmo de agrupamento *K-means*¹⁶ possibilitou a identificação de tópicos com padrão de sentimentos semelhantes. Os tópicos relacionados a dificuldades de saúde mental e aos desafios sensoriais apresentaram uma forte semelhança, ambos com padrão de predominância de sentimentos negativos. O mesmo padrão foi observado nos tópicos sobre acessibilidade no trabalho e diagnóstico tardio. Por outro lado, o único tópico que apresentou uma proporção maior de sentimentos positivos foi o relacionado aos interesses especiais, conforme ilustrado na Figura 12.



Observou-se que os sentimentos expressos nas submissões têm influência direta sobre os sentimentos expressos nos comentários. De maneira geral, uma submissão negativa tende a gerar comentários negativos. No caso das submissões neutras, embora haja uma quantidade significativa de comentários neutros, a maioria dos comentários ainda apresenta um tom negativo. Por sua vez, as submissões positivas aumentam a proporção de comentários positivos, mas, os comentários negativos continuam prevalecendo. Essa dinâmica está ilustrada na Figura 13.

¹⁶ *K-means*: é uma técnica utilizada para agrupamento de dados, disponível na biblioteca *scikit-learn*. Documentação oficial disponível em: <https://scikit-learn.org/stable/modules/clustering.html#k-means>. Acesso em: 16 dez. 2024.

Para aprofundar a análise de cada um dos cinco tópicos selecionados, eles serão examinados em detalhes nas próximas seções deste capítulo. Cada tópico terá uma investigação dos termos mais recorrentes dentro de cada grupo, a proporção dos sentimentos e os principais temas dos comentários associados.



4.1 DIFICULDADES DE SAÚDE MENTAL E ANSIEDADE SOCIAL

O tópico “Dificuldades de saúde mental e ansiedade social” é marcado por submissões que abordam questões relacionadas a problemas mentais e dificuldades em situações sociais. Na Figura 14, podem ser observados os termos mais recorrentes nas submissões relacionadas a esse tema. Palavras como *"feel"* e *"felt"* indicam um foco no compartilhamento de sentimentos pessoais. Termos como *"people"*, *"friend"* e *"school"* destacam tanto as pessoas quanto os ambientes que podem, em algumas situações, provocar desconforto para indivíduos com autismo.

Os termos e relatos examinados revelaram sentimentos predominantes de tristeza, estresse e desconforto social em pessoas autistas, o que frequentemente resulta em emoções negativas. Como ilustrado na Figura 15, a análise dos sentimentos associados a esse tópico indicou que 473 submissões foram classificadas como negativas, 197 como neutras e 27 como positivas.

Figura 15 - Proporção dos sentimentos: Dificuldades de saúde mental e ansiedade social.

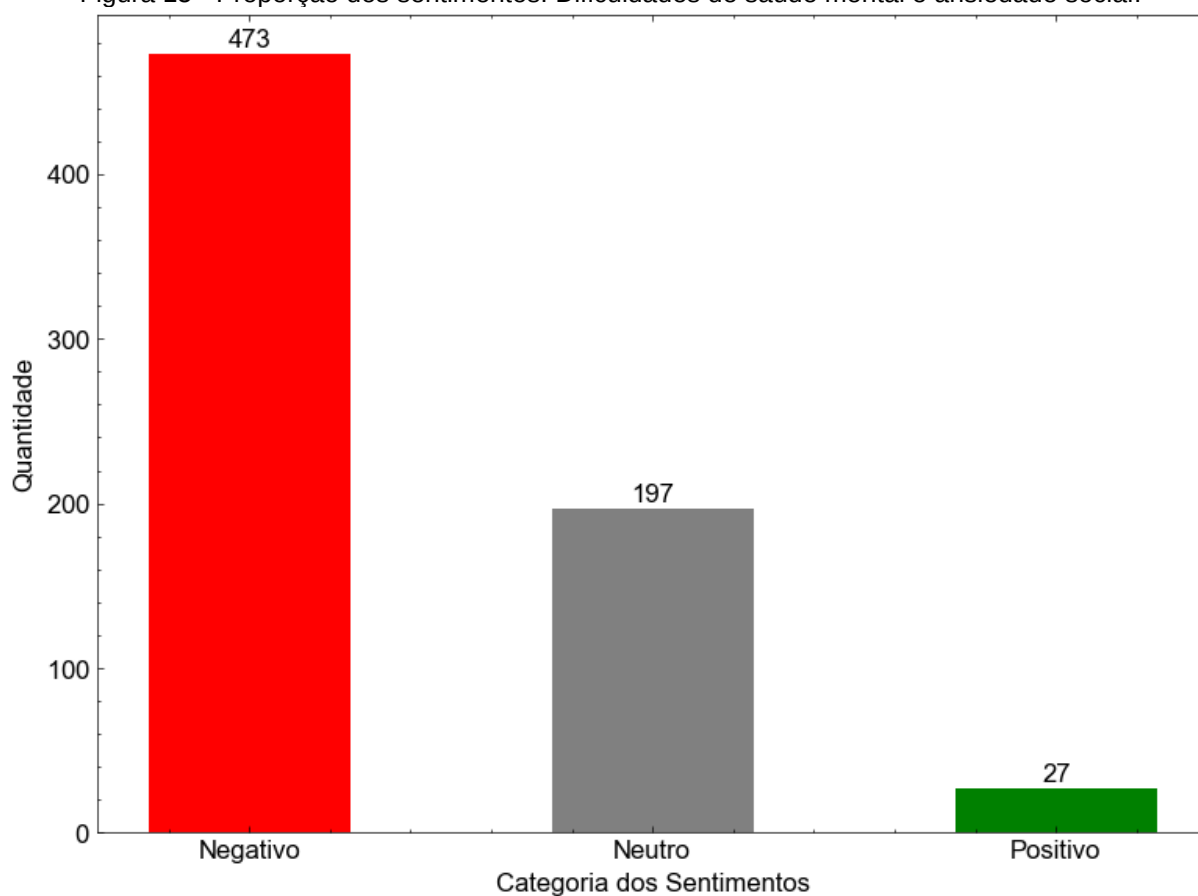
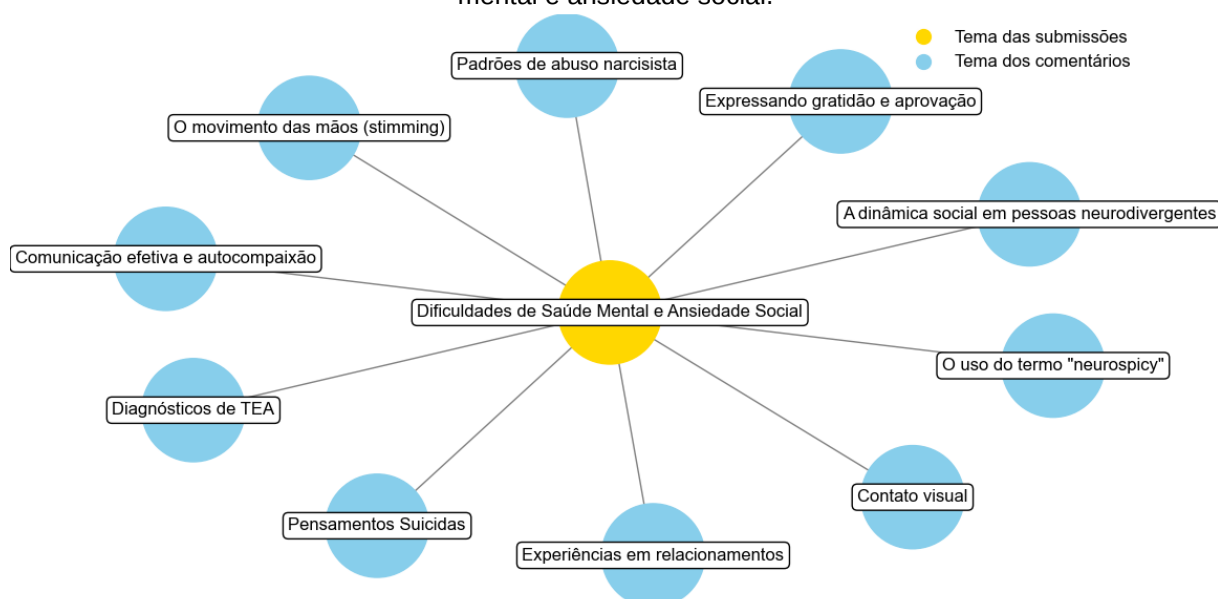


Figura 16 - Conexões entre tópicos identificados nos comentários com o tópico Dificuldades de saúde mental e ansiedade social.



Outro aspecto relevante discutido foi o ato de revelar deficiências durante entrevistas de emprego, como ilustrado na Figura 19.

Figura 18 - Proporção dos sentimentos: Acessibilidade no trabalho e saúde mental.

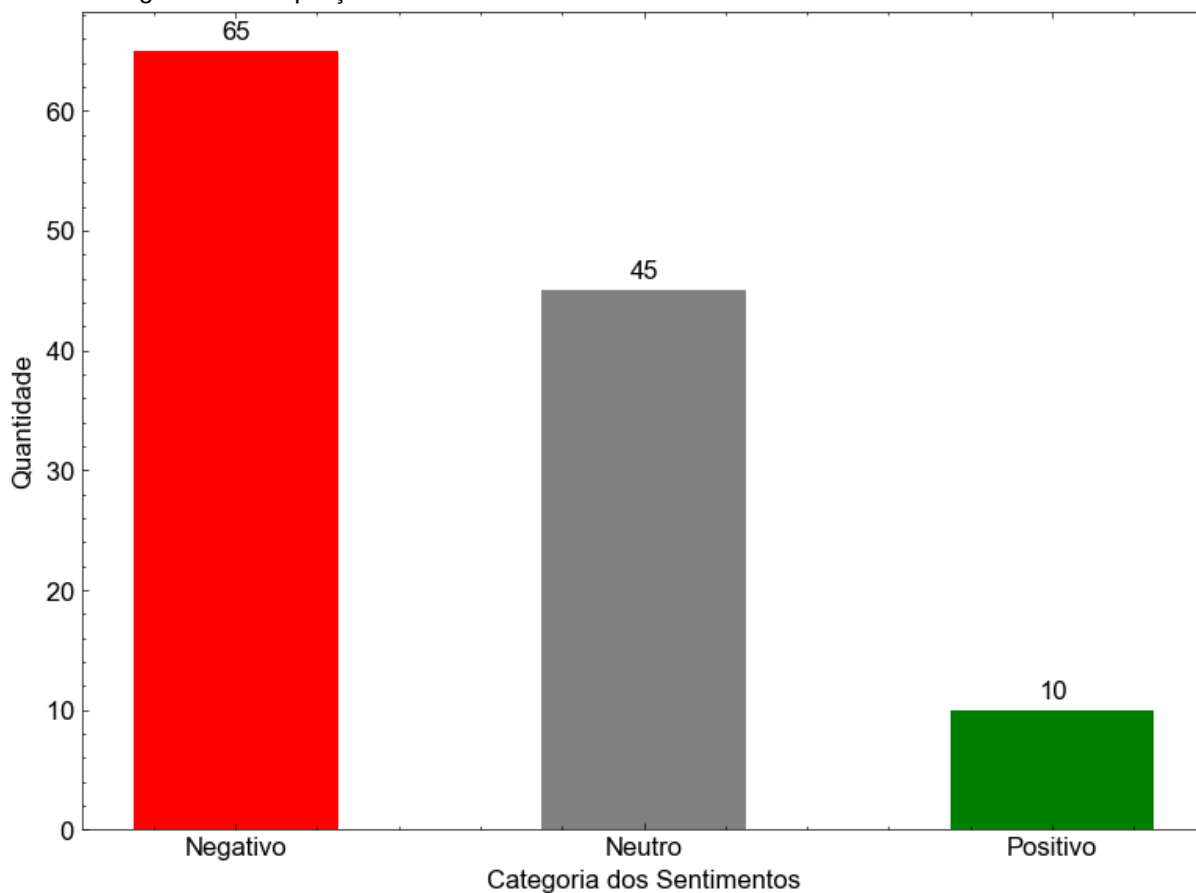
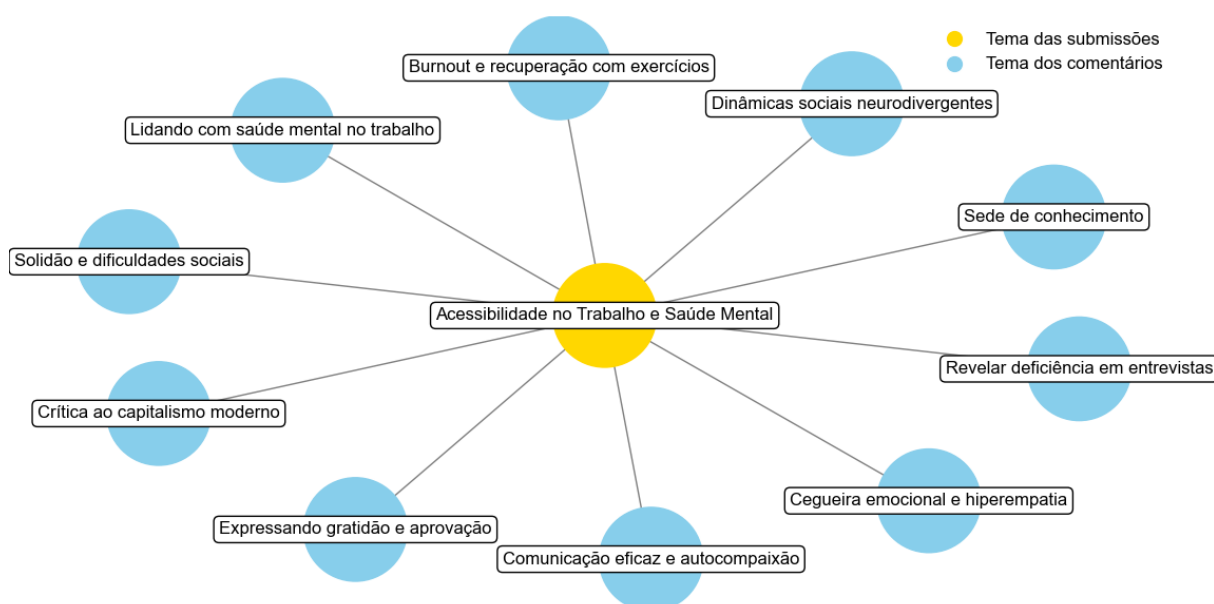
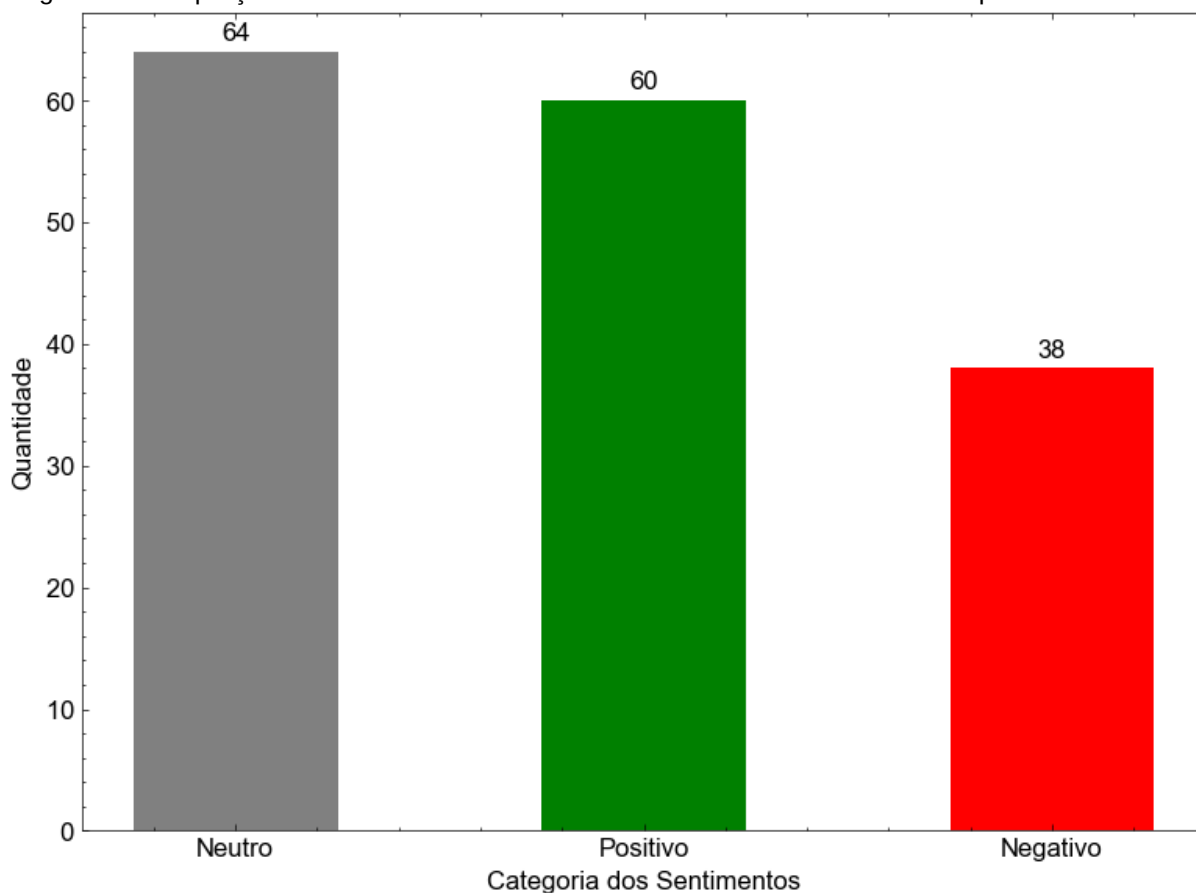


Figura 19 - Conexões entre tópicos identificados nos comentários com o tópico Acessibilidade no trabalho e saúde mental.



Diferentemente dos outros tópicos analisados, este tem como tema os hobbies e hiperfocos dos usuários, que geralmente não estão associados a sentimentos negativos. Essa tendência é confirmada pela distribuição de sentimentos apresentada na Figura 21, que apresenta 64 comentários neutros, 60 positivos e 38 negativos, com predominância de emoções mais neutras e positivas em comparação a outros tópicos.

Figura 21 - Proporção dos sentimentos: A diversidade de hobbies e interesses especiais no autismo

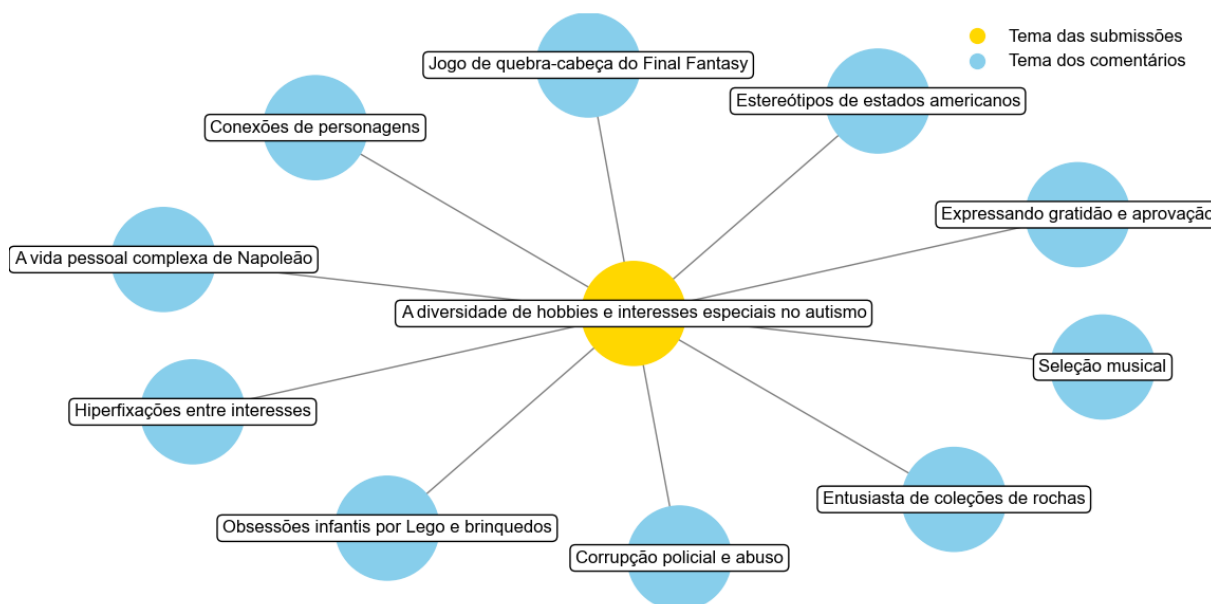


Os tópicos identificados em comentários associados a essas submissões abordam temas variados relacionados aos hiperfocos dos usuários, como Lego, brinquedos e até mesmo coleções de rochas. Além disso, há menções frequentes sobre seleções de músicas e personagens favoritos, como ilustrado na Figura 22. De forma inesperada, foi identificado um tópico relacionado a discussões sobre corrupção policial, destoando dos demais temas mais leves e voltados a interesses pessoais.

O entendimento dos interesses de pessoas autistas pode, por exemplo, auxiliar na criação de ações voltadas para a saúde mental que utilizem interesses pessoais como ferramentas para reduzir o estresse. Além disso, pode contribuir para

o desenvolvimento de terapias ocupacionais personalizadas, que levem em consideração atividades preferidas pelos indivíduos, promovendo o bem-estar emocional e o aumento da eficácia das intervenções.

Figura 22 - Conexões entre tópicos identificados nos comentários com o tópico A diversidade de hobbies e interesses especiais no autismo.



4.4 DIAGNÓSTICO TARDIO

Essa seção é caracterizada por relatos de usuários que receberam o diagnóstico de autismo em fases mais avançadas da vida. O diagnóstico tardio é uma realidade para muitas pessoas autistas, já que a condição não possui características físicas visíveis, mas se manifesta por meio de um conjunto de sintomas diversos. Em muitos casos, esses sintomas se tornam evidentes na idade adulta, quando as expectativas de independência e produtividade superam as capacidades do indivíduo, levando-o a crises (LUPINDO; MAW; SHABALALA, 2023). A Figura 23 destaca os termos mais recorrentes, como “*diagnosis*”, “*ADHD*” e “*autism*”, reforçando a importância do tema diagnóstico e a relação com o Transtorno de Déficit de Atenção e Hiperatividade (TDAH). Conforme apontado por Reiersen e Todd (2008), a coocorrência do TEA e do TDAH é comum, e estudos sugerem que certos genes podem influenciar ambos os transtornos.

Figura 24 - Proporção dos sentimentos: Diagnóstico tardio.

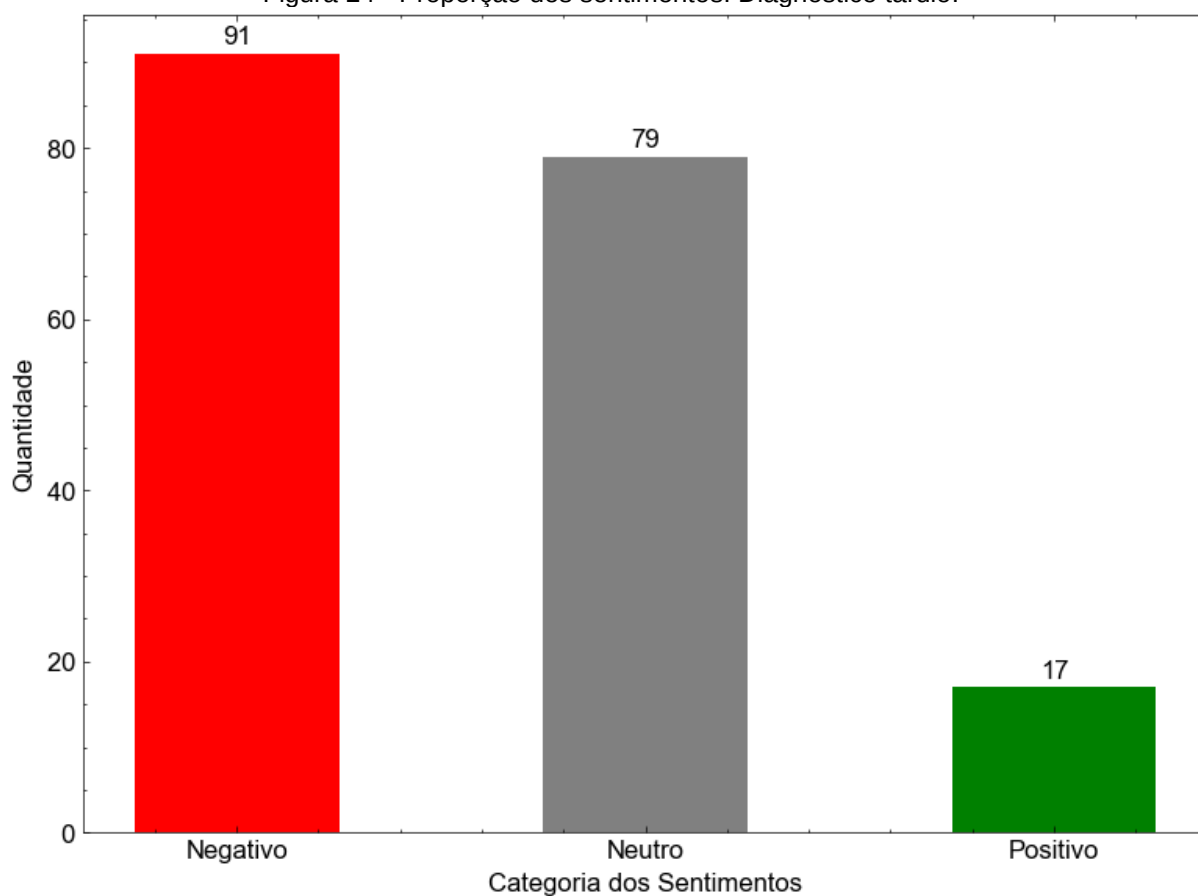
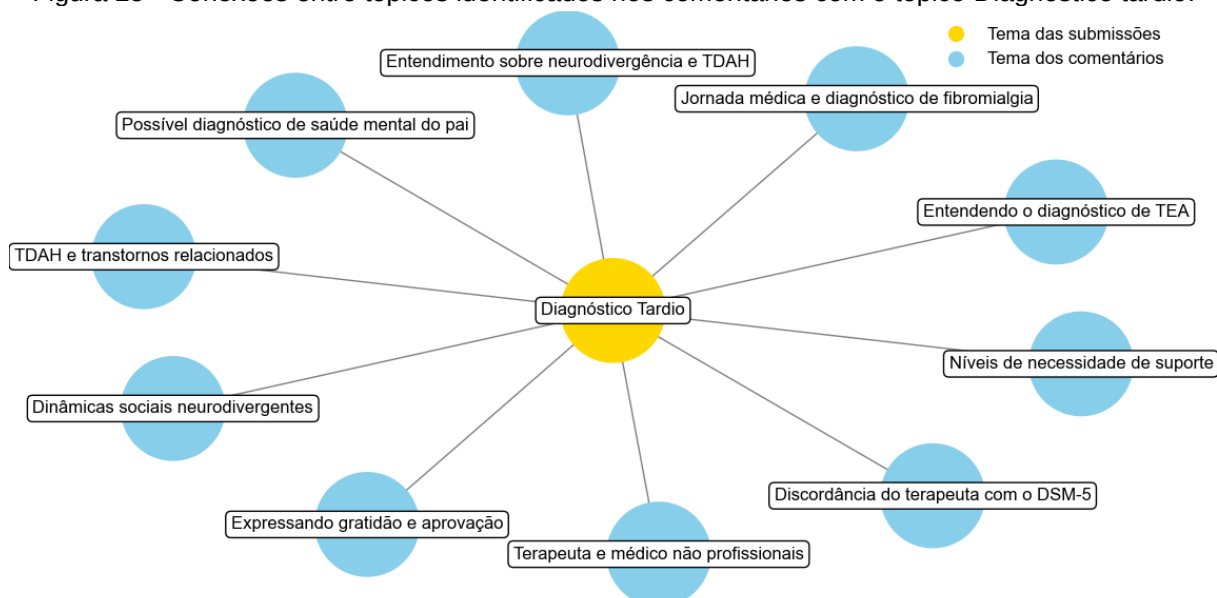
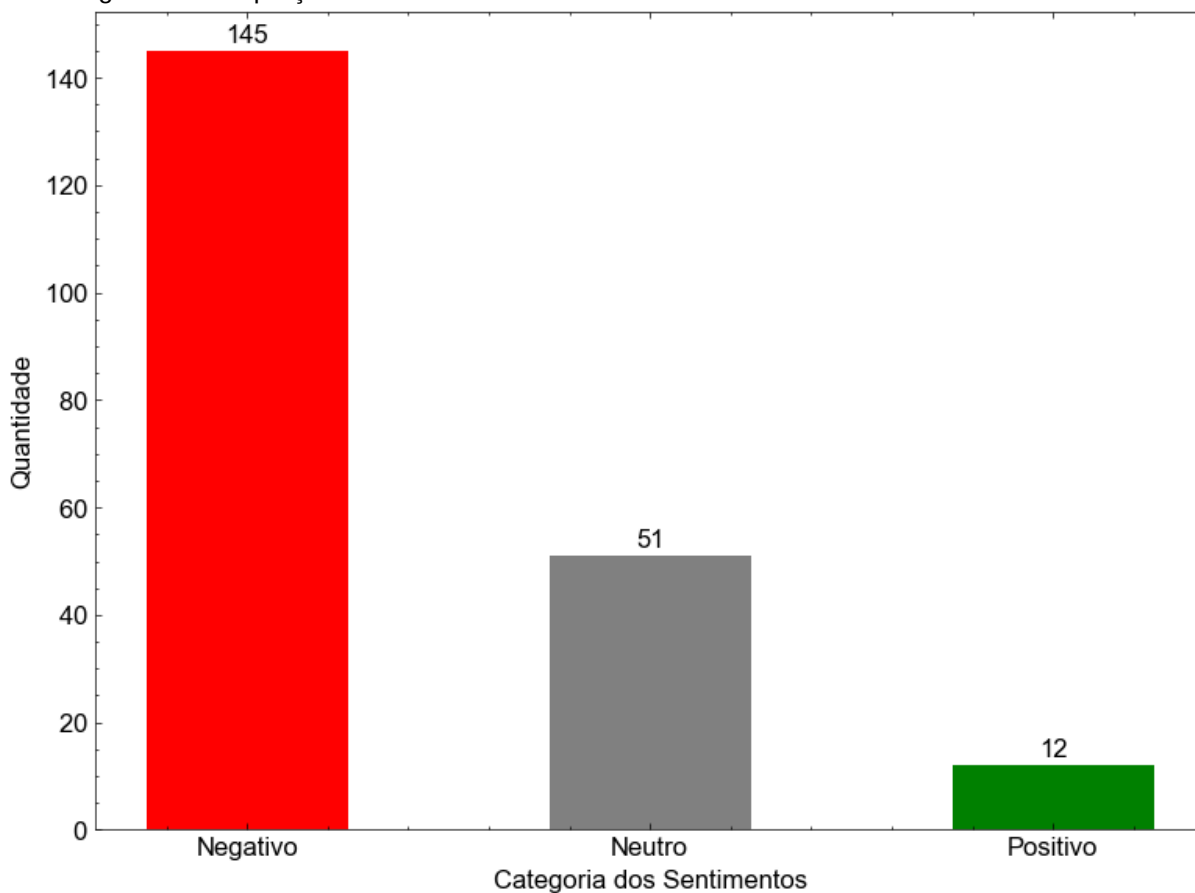


Figura 25 - Conexões entre tópicos identificados nos comentários com o tópico Diagnóstico tardio.



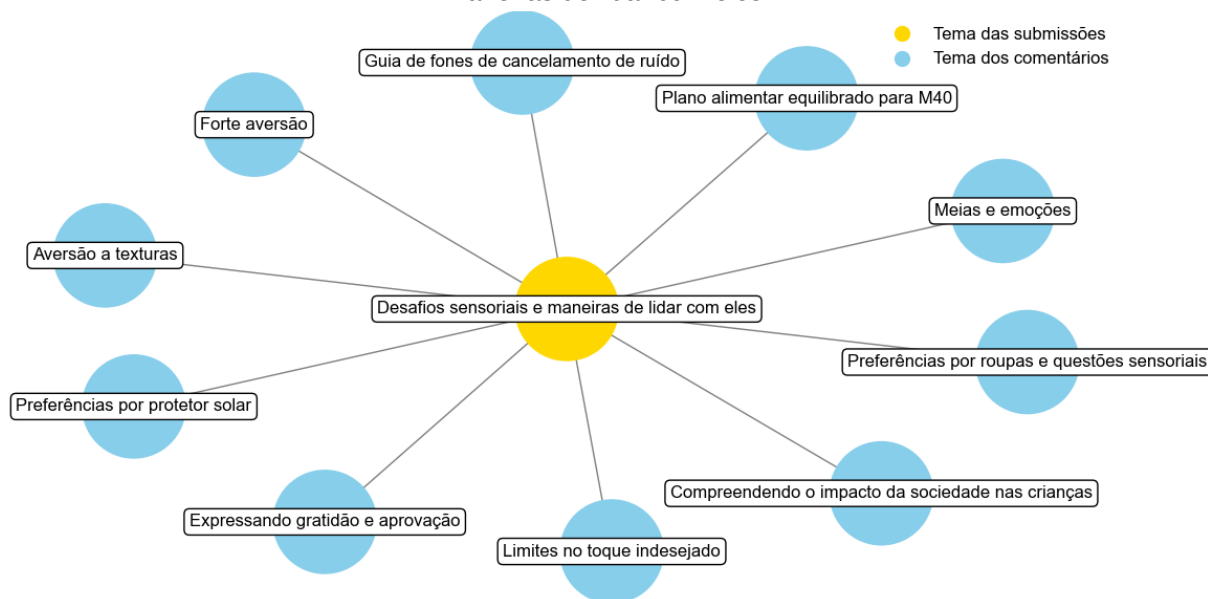
forte aversão a estímulos sensoriais variados, refletindo a complexidade das experiências sensoriais vivenciadas por pessoas com TEA.

Figura 27 - Proporção dos sentimentos: Desafios sensoriais e maneiras de lidar com eles.



Essa análise pode contribuir para a criação de ambientes públicos mais acessíveis para pessoas com hiposensitividade, incluindo espaços com áreas silenciosas e estímulos reduzidos. Além disso, pode subsidiar iniciativas para tornar os transportes públicos mais inclusivos e incentivar empresas a promoverem adaptações em ambientes de trabalho, garantindo condições sensoriais adequadas para pessoas autistas.

Figura 28 - Conexões entre tópicos identificados nos comentários com o tópico Desafios sensoriais e maneiras de lidar com eles.



5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo investigar a possibilidade de compreender, a partir de postagens no Reddit, os desafios comuns enfrentados por pessoas autistas, além de verificar a confiabilidade da plataforma como fonte de informações, considerando os métodos de análise aplicados. A comunidade “autism” foi selecionada como objeto de análise para investigar a viabilidade de extrair informações genuínas sobre autistas.

Os resultados confirmaram o potencial do Reddit como fonte de informações, especialmente quando associado a ferramentas avançadas como os modelos BERTopic e RoBERTa. Esses modelos desempenharam um papel essencial na organização e análise do grande volume de dados coletados. Como previsto na hipótese, os dados revelaram temas relevantes, incluindo desafios sensoriais, saúde mental, relações sociais e interesses especiais, além de uma notável prevalência de sentimentos negativos.

A metodologia seguiu um fluxo estruturado. Os dados foram coletados do Reddit durante os meses de junho e julho de 2024, armazenados em um banco de dados e processados para remover dados irrelevantes e selecionar informações mais significativas. Posteriormente, foram aplicados modelos de *machine learning* para inferir sentimentos e agrupar os dados em tópicos específicos. Esse processo permitiu extrair *insights* importantes sobre as experiências relatadas por membros da comunidade.

No entanto, algumas limitações foram identificadas ao longo do estudo. Em primeiro lugar, a quantidade de dados coletados e o período de análise relativamente curto (dois meses) pode ter restringido o alcance e a profundidade dos resultados. A realização de uma coleta de dados mais extensa, realizada ao longo de um período mais prolongado, pode oferecer uma visão mais abrangente e sólida. Além disso, a validação explícita de que os usuários eram, de fato, autistas poderia contribuir significativamente na precisão dos achados, já que as biografias utilizadas como critério de seleção podem ser imprecisas ou incompletas. Cabe destacar que este estudo considerou também usuários que, apesar de suspeitarem ser autistas, ainda não tinham um diagnóstico formal.

A realização de uma análise manual criteriosa das publicações e comentários pode constituir um aspecto relevante para validar a precisão dos modelos e aprimorar a interpretação dos tópicos. A inclusão de métricas claras para avaliar o desempenho do modelo RoBERTa poderia fortalecer ainda mais a confiabilidade dos resultados apresentados.

Com base nas análises temáticas realizadas no Capítulo 4, vários caminhos podem ser tomados para entender as necessidades das pessoas autistas e propor soluções práticas para auxiliar no seu estilo de vida. Por exemplo, no tema das dificuldades de saúde mental e ansiedade social, futuros estudos podem explorar a criação de plataformas digitais ou aplicativos para gerenciar a ansiedade social.

No que diz respeito à acessibilidade no ambiente de trabalho, os achados podem contribuir para o desenvolvimento de estudos que avaliem taxas de retenção e produtividade em iniciativas de inclusão, fornecendo subsídios para políticas mais eficazes. Em relação à diversidade de hobbies e interesses especiais, os resultados podem fundamentar investigações que busquem identificar hobbies comuns entre autistas de diferentes faixas etárias e culturas, além de embasar intervenções terapêuticas que utilizem esses interesses para reduzir o estresse e a ansiedade. Por fim, os achados relacionados ao diagnóstico tardio podem inspirar estudos futuros que explorem os desafios sociais e emocionais enfrentados por pessoas diagnosticadas em etapas avançadas da vida, propondo estratégias para um diagnóstico mais acessível, preciso e confiável, bem como programas de apoio focados nas necessidades específicas desse grupo.

No que tange aos desafios sensoriais, futuras pesquisas poderiam explorar o desenvolvimento de ferramentas de apoio personalizadas. Por exemplo, aplicativos ou plataformas baseadas em inteligência artificial poderiam ser criados para ajudar pessoas autistas a identificar rotas mais adequadas no transporte público, levando em conta horários com menor lotação ou áreas de menor ruído.

Por fim, este estudo destaca o potencial dos dados provenientes de redes sociais como uma fonte valiosa para uma compreensão mais profunda das experiências vividas por pessoas autistas. Com o apoio de iniciativas públicas e privadas, bem como o desenvolvimento de ferramentas específicas para esse público, essas análises podem contribuir significativamente para a identificação de suas necessidades e desafios. Além disso, os insights obtidos podem servir de subsídio

para diversas iniciativas que, baseadas em dados reais, buscam aprimorar a qualidade de vida da comunidade autista.

REFERÊNCIAS

AHMAD, M. *et al.* Machine learning techniques for sentiment analysis: A review. **International Journal of Multidisciplinary Sciences and Engineering**, v. 8, n. 3, p. 27-32, 2017.

ALMEIDA, A. *et al.* The Use of Natural Language Processing Methods in Reddit to Investigate Opioid Use: Scoping Review. **JMIR infodemiology**, v. 4, p. e51156, 2024.

ANDERSON, K. E. Ask me anything: what is Reddit? **Library Hi Tech News**, v. 32, p. 8-11, 2015.

AMERICAN PSYCHIATRIC ASSOCIATION *et al.* **DSM-5 - Manual Diagnóstico e Estatístico de Transtornos Mentais**. 5^a. ed. Porto Alegre: Editora Artmed, 2014.

BEYKIKHOSHK, A. *et al.* Using Twitter to learn about the autism community. **Social Network Analysis and Mining**, v. 5, p. 1-17, 2015.

CHOPRA, A.; PRASHAR, A.; SAIN, C. Natural Language Processing. **International journal of technology enhancements and emerging engineering research**, v. 1, n. 4, p. 131-134, 2013.

DA SILVA, S. C.; GESSER, M.; NUERNBERG, A. H. A contribuição do modelo social da deficiência para a compreensão do Transtorno do Espectro Autista. **Revista educação, artes e inclusão**, v. 15, n. 2, p. 187-207, 2019.

DEVLIN, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, p. 4171–4186, 2018.

EGGER, R.; YU, J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. **Frontiers in sociology**, v. 7, 2022.

GABARRON, E. *et al.* What do autistic people discuss on Twitter? An approach using BERTopic modelling. **Caring is Sharing–Exploiting the Value in Data for Health and Innovation**, p. 403-407, 2023.

GROOTENDORST, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. **arXiv preprint arXiv:2203.05794**, 2022.

HASAN, M. R.; MALIHA, M.; ARIFUZZAMAN, M. Sentiment analysis with NLP on Twitter data. **2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2)**, p. 1-4, 2019.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. **Science**, v. 349, n. 6245, p. 261-266, 2015.

JUNG, Y. L. Market intelligence applications leveraging a product-specific Sentence-RoBERTa model. **Applied Soft Computing**, v. 165, p. 112077, 2024.

KHERWA, P.; BANSAL, P. Topic modeling: a comprehensive review. **EAI Endorsed transactions on scalable information systems**, v. 7, n. 24, 2019.

KHURANA, D. *et al.* Natural language processing: state of the art. **Multimedia Tools and Applications**, v. 82, n. 3, p. 3713-3744, 2023.

LIU, T. *et al.* Detecting symptoms of depression on reddit. **Proceedings of the 15th ACM Web Science Conference 2023**, p. 174-183, 2023.

LIU, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. **CoRR**, abs/1907.11692, 2019.

LORD, C. *et al.* Autism spectrum disorder. **The Lancet**, v. 392, n. 10146, p. 508–520, 2018.

LOUREIRO, D. *et al.* TimeLMs: Diachronic Language Models from Twitter. **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**, p. 251–260, 2022.

LUPINDO, B. M.; MAW, A.; SHABALALA, N. Late diagnosis of autism: Exploring experiences of males diagnosed with autism in adulthood. **Current Psychology**, v. 42, n. 28, p. 24181-24197, 2023.

MA, X.; HANCOCK, J.; NAAMAN, M. Anonymity, intimacy and self-disclosure in social media. **Proceedings of the 2016 CHI conference on human factors in computing systems**, p. 3857-3869, 2016.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams engineering journal**, v. 5, n. 4, p. 1093-1113, 2014.

MUENNIGHOFF, N. *et al.* MTEB: Massive Text Embedding Benchmark. **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, p. 2014–2037, 2023.

MÜLLER, M.; SALATHÉ, M.; KUMMERVOLD, P. E. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. **Frontiers in artificial intelligence**, v. 6, p. 1023281, 2023.

REIERSEN, A. M.; TODD, R. D. Co-occurrence of ADHD and autism spectrum disorders: phenomenology and treatment. **Expert review of neurotherapeutics**, v. 8, p. 657-669, n. 4, 2008.

ROHANI, V. A.; SHAYAA, S.; BABANEJADDEHAKI, G. Topic modeling for social media content: A practical approach. **2016 3rd international conference on computer and information sciences (ICCOINS)**, p. 397-402, 2016.

SANTOMAURO, D. F. *et al.* The global epidemiology and health burden of the autism spectrum: findings from the Global Burden of Disease Study 2021. **The Lancet Psychiatry**, 2024.

SILVEIRA, R. *et al.* Topic modelling of legal documents via legal-bert. **Proceedings of the First International Workshop RELATED - Relations in the Legal Domain 2021**, 2021.

SU, H. *et al.* One Embedder, Any Task: Instruction-Finetuned Text Embeddings. **Findings of the Association for Computational Linguistics: ACL 2023**, p. 1102–1121, 2023.

TABOADA, M. Sentiment analysis: An overview from linguistics. **Annual Review of Linguistics**, v. 2, n. 1, p. 325-347, 2016.

THOM-JONES, S.; MELGAARD, I.; GORDON, C. S. Autistic Women's Experience of Motherhood: A Qualitative Analysis of Reddit. **Journal of Autism and Developmental Disorders**, p. 1-11, 2024.

VASWANI, A. *et al.* Attention Is All You Need. **Advances in Neural Information Processing Systems**, 2017.

WANKHADE, M.; RAO, A. C. S.; KULKARNI, C. A survey on sentiment analysis methods, applications, and challenges. **Artificial Intelligence Review**, v. 55, n. 7, p. 5731-5780, 2022.

Apêndice A - Código da Análise de Sentimento

O Código 4 foi utilizado para classificar as submissões e comentários em positivos, neutros e negativos. O modelo “*cardiffnlp/twitter-roberta-base-sentiment-latest*” foi utilizado para analisar os sentimentos das submissões com base no título, corpo de texto e, por fim, com uma combinação de título e corpo de texto.

Código 4 - Análise de sentimento utilizando RoBERTa

(continua)

```

1. import torch
2. import pandas as pd
3. from tqdm.auto import tqdm
4. from datasets import Dataset
5. from transformers import pipeline
6. from transformers.pipelines.pt_utils import KeyDataset
7.
8. if torch.cuda.is_available():
9.     print("CUDA is available!")
10. else:
11.     print("CUDA is not available.")
12.
13. ## 1. Submissions
14.
15. submissions_df = pd.read_csv("output/autism_submissions.csv")
16. submissions_dataset = Dataset.from_pandas(submissions_df, split = "test")
17.
18. ## 2. Configure Twitter Roberta Base and Dataset
19.
20. MODEL_PATH = "cardiffnlp/twitter-roberta-base-sentiment-latest"
21.
22. pipe = pipeline("sentiment-analysis", model = MODEL_PATH, tokenizer = MODEL_PATH,
max_length=512, truncation=True, device = 0)
23.
24. ## 3. Classification of sentiment based only on the title
25.
26. sentiment_dict = {
27.     "sentiment_title": [],
28.     "sentiment_score_title": [],
29. }
30.
31. for classification in tqdm(pipe(KeyDataset(submissions_dataset, "title"))):
32.     sentiment_dict["sentiment_title"].append(classification["label"])
33.     sentiment_dict["sentiment_score_title"].append(classification["score"])
34.
35. submissions_df = submissions_df.assign(**sentiment_dict)

```

Código 4 - Análise de sentimento utilizando RoBERTa

(continua)

```

36.
37. submissions_df["sentiment_title"] = submissions_df["sentiment_title"].astype('category')
38.
39. ## 4. Classification of sentiment based only on the selftext
40.
41. sentiment_dict = {
42.     "sentiment_selftext": [],
43.     "sentiment_score_selftext": [],
44. }
45.
46. for classification in tqdm(pipe(KeyDataset(submissions_dataset, "selftext"))):
47.     sentiment_dict["sentiment_selftext"].append(classification["label"])
48.     sentiment_dict["sentiment_score_selftext"].append(classification["score"])
49.
50. submissions_df = submissions_df.assign(**sentiment_dict)
51.
52.                 submissions_df["sentiment_selftext"] =
submissions_df["sentiment_selftext"].astype('category')
53.
54. ## 5. Classification of sentiment based only on the title and selftext
55.
56. submissions_df["input"] = submissions_df["title"] + " " + submissions_df["selftext"]
57. submissions_dataset = Dataset.from_pandas(submissions_df, split = "test")
58.
59. sentiment_dict = {
60.     "sentiment_title_selftext": [],
61.     "sentiment_title_score_selftext": [],
62. }
63.
64. for classification in tqdm(pipe(KeyDataset(submissions_dataset, "input"))):
65.     sentiment_dict["sentiment_title_selftext"].append(classification["label"])
66.     sentiment_dict["sentiment_title_score_selftext"].append(classification["score"])
67.
68. submissions_df = submissions_df.assign(**sentiment_dict)
69.
70. submissions_df["sentiment_title_selftext"] =
submissions_df["sentiment_title_selftext"].astype('category')
71.
72. ## 6. Save findings
73.
74. submissions_df = submissions_df.drop(['sentiment_title', 'sentiment_score_title',
'sentiment_selftext', 'sentiment_score_selftext'], axis=1)
75. submissions_df = submissions_df.rename(columns={"sentiment_title_selftext": "sentiment",
"sentiment_title_score_selftext": "sentiment_score"})

```

Código 4 - Análise de sentimento utilizando RoBERTa

(conclusão)

```
76.
77. submissions_df.to_csv('output/submissions_autism_sentiment_analysis.csv', index=False)
78.
79. ## 7. Comments
80.
81. comments_df = pd.read_csv("output/autism_comments.csv")
82. comments_dataset = Dataset.from_pandas(comments_df, split = "test")
83.
84. sentiment_dict = {
85.     "sentiment": [],
86.     "sentiment_score": [],
87. }
88.
89. for classification in tqdm(pipe(KeyDataset(comments_dataset, "body"))):
90.     sentiment_dict["sentiment"].append(classification["label"])
91.     sentiment_dict["sentiment_score"].append(classification["score"])
92.
93. comments_df = comments_df.assign(**sentiment_dict)
94.
95. comments_df["sentiment"] = comments_df["sentiment"].astype('category')
96.
97. comments_df.to_csv('output/autism_comments_sentiment_analysis.csv', index=False)
```

Apêndice B - Pequeno trecho da modelagem de tópicos

O Código 5 foi utilizado para realizar o agrupamento de todas as submissões selecionadas utilizando o BERTopic. Em um primeiro momento todos os links e palavras acerca de autismo foram removidos, o *CountVectorizer* foi utilizado para remover as palavras de parada, OpenAI para representar os tópicos, “*hkunlp/instructor-large*” para criar as embeddings e o UMAP para reduzir a dimensionalidade.

Código 5 – Modelagem de tópicos com todas as submissões

(continua)

```

1. import os
2. import re
3. import openai
4. import pandas as pd
5. from umap import UMAP
6. from bertopic import BERTopic
7. from bertopic.representation import OpenAI
8. from sklearn.feature_extraction.text import CountVectorizer
9. from sentence_transformers import SentenceTransformer
10. import torch
11.
12. ## 1. All Submissions
13.
14. def removeAllLinks(data):
15.     return re.sub(r'http\S+', '', data)
16.
17. def remove_words(data):
18.     stop_words = ["autism", "autistic", "autistics", "aspergers"]
19.     pattern = r'\b(?:' + '|'.join(map(re.escape, stop_words)) + r')\b'
20.     return re.sub(pattern, '', data, flags=re.IGNORECASE).strip()
21.
22. submission_df = pd.read_csv("output/submissions_autism_sentiment_analysis.csv")
23. docs = submission_df["input"].apply(remove_words).apply(removeAllLinks)
24.
25. vectorizer_model = CountVectorizer(stop_words="70english", min_df=2, ngram_range=(1, 2))
26.
27. ### Representation Model
28.
29. chat_gpt_key = os.getenv("CHATGPT_KEY")
30.
31. client = openai.OpenAI(api_key=chat_gpt_key)
32. prompt = ""
33.     I have a topic that contains the following documents:
34.     [DOCUMENTS]
```

Código 5 – Modelagem de tópicos com todas as submissões

(continua)

```

35.     The topic is described by the following keywords: [KEYWORDS]
36.
37.     Based on the information above, extract a short but highly descriptive topic label of at
most 5 words. Make sure it is in the following format:
38.     topic: <topic label>
39.     """
40. representation_model = OpenAI(client, model="gpt-3.5-turbo-0125", exponential_backoff=True,
chat=True, prompt=prompt)
41.
42. ### Embedding Model
43.
44. # Pre-calculate embeddings
45. embedding_model = SentenceTransformer("hkunlp/instructor-large").cuda()
46.
47. sentences = [{"Represent the Reddit Post:", sentence} for sentence in docs]
48.
49. embeddings = embedding_model.encode(sentences, show_progress_bar=True)
50.
51. ### UMAP Model
52.
53. umap_model = UMAP(n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine',
random_state=42)
54.
55. topic_model = BERTopic(
56.     # Pipeline models
57.     embedding_model=embedding_model,
58.     umap_model=umap_model,
59.     vectorizer_model=vectorizer_model,
60.     representation_model=representation_model,
61.     # Hyperparameters
62.     top_n_words=5,
63.     verbose=True,
64.     nr_topics="auto"
65. )
66.
67. # Train model
68. topics, probs = topic_model.fit_transform(docs, embeddings)
69.
70. # Reduce outliers with pre-calculate embeddings instead
71. new_topics = topic_model.reduce_outliers(docs, topics, strategy="embeddings",
embeddings=embeddings)
72.
73. topic_model.update_topics(
74.     docs,

```

Código 5 – Modelagem de tópicos com todas as submissões

(conclusão)

```
75.     topics=new_topics,
76.     vectorizer_model=vectorizer_model,
77.     representation_model=representation_model
78. )
79.
80. submission_df["topic"] = new_topics
81.
82. # Get topic information
83. topic_info = topic_model.get_topic_info()
84.
85. # Map topic themes (keywords) to topic numbers
86. topic_labels = {
87.     row["Topic"]: row["Name"]
88.     for _, row in topic_info.iterrows()
89. }
90.
91. # Annotate themes in the DataFrame
92. submission_df["theme"] = submission_df["topic"].map(topic_labels)
93.
94. submission_df.to_csv('output/submissions_autism_topic_modelling.csv', index=False)
```

Apêndice C – Tabelas traduzidas.

Tabela C.1 - Tópicos avulsos das submissões traduzidas

Nome	Quantidade	Documento representativo
Dificuldades de Saúde Mental e Ansiedade Social	697	“Toda vez que acho que finalmente fiz um amigo, ele cancela os planos. Isso acabou de acontecer novamente e não sei o que estou fazendo de errado. Quero ter amigos. Eu amo as pessoas. Eu realmente quero ter amigos. Acho que nunca tive um amigo de verdade antes e isso está começando a me afetar.”
Diagnóstico Tardio e Necessidades de Suporte	158	“A avaliação para autismo será na terça-feira às 14h15 CST. A única razão pela qual consigo fazer isso é porque é online e eles aceitam meu <i>Medicaid</i> . Pesquiso sobre isso há muito tempo, pelo menos desde 2016, quando eu tinha 27 anos. Realmente sinto que me encaixo nos traços, especialmente no que dizem sobre mulheres adultas. Depois de aprender sobre mascaramento, percebi que era exatamente o que eu vinha fazendo automaticamente. De qualquer forma, estou nervosa e ansiosa. Tenho medo de não conseguir responder algo corretamente e não receber o diagnóstico de que preciso. Posso ficar tão nervosa que não consiga fazer <i>stimming</i> e, se eu mascarar, e eles não conseguirem perceber que sou autista? Claro, mascarar tem se tornado mais difícil com o passar dos anos, mas ainda assim... Alguém no meu Discord me deu um bom conselho, mas gostaria de ouvir mais opiniões. Ah, e a propósito, eu tenho um fichário com 10 folhas escritas frente e verso explicando por que sinto que sou autista. Também tenho impressões de materiais com os quais me identifico, além da minha pontuação no RAADS-R. Estou tentando me preparar o máximo possível. Também tenho transtornos de ansiedade. Meu marido tentará participar comigo, se conseguir sair do trabalho a tempo. Espero que eles tenham conhecimento sobre autismo em mulheres adultas.”
Dificuldades nas habilidades sociais e na comunicação	55	“Você prefere conversar com adultos em vez de crianças ou adolescentes? Eu (15H) tenho dificuldade em conversar com pessoas da minha idade. Mas com adultos, consigo falar por horas sem nenhum problema. Há outras pessoas jovens que passam pelo mesmo que eu? E vocês acham que isso está relacionado ao autismo?”

Tabela C.2 - Tópicos das submissões positivas traduzidas

(continua)

Nome	Quantidade	Documento representativo
Exaustão ao Deixar de Mascarar Durante Viagens	190	“Estou me acostumando a deixar de mascarar, estou dando pequenos passos e hoje foi um bom dia. Quando comecei a me sentir sobrecarregado(a), comuniquei isso à minha família e, mais tarde, expliquei que meu tom de voz provavelmente ficaria mais plano/monótono e minhas expressões faciais seriam menos expressivas do que o normal pelo resto do dia, porque simplesmente não tenho energia para me concentrar ativamente nisso. Felizmente, eles entenderam e me agradeceram por avisar, para que não houvesse mal-entendidos entre nós.”

Tabela C.2 - Tópicos das submissões positivas traduzidas

(conclusão)

Nome	Quantidade	Documento representativo
Músicas Favoritas	75	"Mais alguém é obcecado por ouvir música? Eu, pessoalmente, não passo um segundo sem ouvir música. Eu adoro Aurora, Aviva e Neoni. Alguma sugestão de artistas musicais? E mais alguém passa por isso?"

Tabela C.3 - Tópicos avulsos das submissões negativas traduzidas

Nome	Quantidade	Documento representativo
Desconexão Emocional e Apatia Social	873	"Alguém mais odeia absolutamente TODAS as formas de demonstrar/receber amor? Pelo que entendo, não gostar ou se sentir desconfortável com o toque físico é um traço comum do autismo. No entanto, eu não apenas odeio o toque físico, mas também odeio que me digam que sou amado(a), odeio a ideia de que as pessoas se importam comigo, odeio receber elogios, e a lista continua. Além disso, também não gosto de oferecer essas coisas. Se alguém precisa de conforto, provavelmente fico em branco, porque a ideia de lidar com isso me faz sentir desconfortável. Às vezes me sinto uma pessoa ruim, mas na maior parte do tempo esse comportamento não me incomoda. Isso é normal para mais alguém?"
Comunicação não verbal e desafios linguísticos	46	"Semi-não verbal desde que fui diagnosticado(a) há 8 anos, tenho experimentado ficar não verbal. Em alguns dias, é mais útil não falar, mas acabo ficando preso(a) no estado não verbal pelo resto do dia. Literalmente sinto que vou vomitar se tentar falar, então estou me perguntando se mais alguém passa por essa experiência. Também estou aberto(a) a ouvir por que outras pessoas ficam não verbais, especialmente aqueles que decidiram parar de falar permanentemente."
Interesses Especiais e Ansiedade Social	31	"As pessoas não falam comigo. Sempre que menciono um dos meus interesses, a conversa acaba. Não me interessa por nada além dos meus interesses especiais. Isso torna minha vida social difícil, porque não tenho interesse na maioria dos assuntos. Me interessa principalmente por Rain World e insetos. Um é um jogo indie de nicho, e o outro é algo frequentemente considerado nojento. Desculpe se meu idioma não estiver bom. Atenciosamente, um autista nível dois."

Tabela C.4 -Três maiores tópicos de comentários traduzidos

(continua)

Nome	Quantidade	Documento representativo
Expressando Gratidão e Aprovação	4116	"Não poderia concordar mais."
Seleção Musical Diversificada	1540	"Linkin Park e alguns outros artistas de rock/metal."
Desafios Sociais na Primeira Infância	693	"Eu achava que fui ostracizado na adolescência por ser gay. Acontece que fui ostracizado desde o meu primeiro dia de escola, anos antes, devido ao Transtorno do Espectro Autista (TEA). Maravilhoso."

Tabela C.5 - Três maiores tópicos dos comentários positivos traduzidos

Nome	Quantidade	Documento representativo
Personagens Favoritos e Apoio em Relacionamentos	3582	"Flor de cerejeira + Reki de Sk8 Infinity, Sanji e Usopp de One Piece :3"
Gosto Musical Eclético	675	"Para mim, são <i>Runaway</i> e <i>Through The Eyes</i> da AURORA. (e talvez <i>Warrior Of Love</i> da mesma artista também! A propósito, se você ainda não ouviu ela, por favor, ouça, a voz dela é tão angelical!) E também tem <i>I Don't Need Your Love</i> do <i>Six The Musical</i> !"
Dinâmicas Sociais em Indivíduos Neurodiversos	340	"Sou o centro da festa, extrovertido e engraçado. Mas só quando estou em um pequeno grupo de amigos muito próximos. Em qualquer outra situação social, sou a pessoa 'normal' que senta no canto e fica miserável."

Tabela C.6 - Três maiores tópicos dos comentários negativos traduzidos

Nome	Quantidade	Documento representativo
Desafios e Perspectivas sobre o Autismo	8416	"Já me disseram várias vezes que não pareço autista. Acho que esperam alguém com uma deficiência intelectual óbvia. Isso é apenas ignorância."
Hipersinesia a texturas	288	"Quando você toca em rochas secas (como na praia) e elas deixam aquela sensação empoeirada/giz que resseca suas mãos, isso me incomoda demais."
Cultura Policial Corrupta e Abusiva	211	"Já conheci policiais incríveis, mas também tenho medo de encontrar um que seja terrível. QUALQUER pessoa em uma posição de poder e que seja uma pessoa ruim é perigosa. É por isso que é importante ter freios e contrapesos. Atualmente, não existem bons mecanismos de controle para a aplicação da lei."