Universidade Federal de Pernambuco

Departamento de Sociologia

Departamento de Antropologia e Museologia

Curso de Ciências Sociais - Bacharelado

Análise Sociológica de Dados de Sofrimento Psíquico no Reddit: Uma Abordagem Metodológica Integrativa

Trabalho de Conclusão de Curso de Graduação

por

Nicólly Lira de Albuquerque

Orientador: Prof. Dr. Francisco Jatobá

Recife, Janeiro / 2024

Nicólly Lira de Albuquerque

Análise Sociológica de Dados de Sofrimento Psíquico no Reddit: Uma Abordagem Metodológica Integrativa

Monografia apresentada ao Curso de Ciências Sociais - Bacharelado, como requisito parcial para a obtenção do Título de Bacharel em Ciências Sociais, Centro de Filosofia e Ciências Humanas da Universidade Federal de Pernambuco.

Orientador: Prof. Dr. Francisco Jatobá

Recife

2023

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Albuquerque, Nicolly Lira de .
Análise Sociológica de Dados de Sofrimento Psíquico no Reddit: Uma Abordagem Metodológica Integrativa / Nicolly Lira de Albuquerque. - Recife, 2023.
36 : il., tab.

Orientador(a): Francisco Jatobá de Andrade
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Filosofia e Ciências Humanas, Ciências Sociais - Bacharelado, 2023.
Inclui referências, apêndices, anexos.

1. Proposta metodológica. 2. Interdisciplinaridade. 3. Pré-processamento de texto. 4. Sofrimento psíquico. 5. Sociologia. I. Andrade, Francisco Jatobá de . (Orientação). II. Título.

300 CDD (22.ed.)

Agradecimentos

A minha família, que nunca deixou de apoiar minhas escolhas. Aos meus pais, obrigada por terem batalhado absurdos para que hoje eu pudesse estudar e ser a primeira pessoa da família a concluir uma graduação em uma universidade pública. A minha irmã Nívia, por ser uma das melhores pessoas que existem no mundo.

As pessoas especiais que a vida me deu: meu grande amor e companheiro de vida Daniel, minha grande amiga e também companheira de vida Cecy. Obrigada por terem feito parte não só deste, mas de muitos outros processos da minha vida. Obrigada pelo acolhimento durante esse perído em que passei monotemática, pelo conforto nas horas difíceis, pelos abraços e pelo trabalho (não pago) de revisores (risos). Aos demais amigos que foram acolhimento nessa jornada, eu não teria conseguido sem o apoio emocional de vocês.

Agradeço também ao meu professor e orientador Francisco Jatobá, por ter acreditado em mim e nas minhas propostas para este trabalho e área de estudos que desejo seguir. Por ter sido sempre presente em orientações academicas e eventuais sessões de terapia anti ânsiedade.

Qualquer tecnologia suficientemente avançada é indistinguível de magia.

Arthur C. Clarke

RESUMO

Este trabalho adota uma abordagem metodológica interdisciplinar, situada em um diálogo entre as ciências sociais e a ciência da computação. Dentro deste contexto, busca-se propor caminhos para o acesso e a compreensão das informações disponíveis na internet, reconhecendo que a internet desempenha um papel fundamental na construção de subjetividades e na transformação das maneiras pelas quais as informações de relevância sociológica podem ser acessadas.

Nesse contexto, este estudo se concentra em explorar os relatos espontâneos relacionados ao sofrimento psíquico, que são amplamente disseminados na internet, especialmente em plataformas de redes sociais. Para alcançar esse objetivo, empregamos métodos de coleta e análise de dados em redes sociais digitais, ao mesmo tempo que utilizamos teorias sociológicas relacionadas ao sofrimento psíquico como base para a seleção e interpretação desses dados.

Palavras-chave: Sofrimento psíquico, Sociologia, Ciência de dados, Text rank, Interdisciplinaridade, Metodologia.

ABSTRACT

This work adopts an interdisciplinary methodological approach, situated within a dialogue between social sciences and computer science. Within this context, it seeks to propose pathways for accessing and understanding information available on the internet, recognizing that the internet plays a fundamental role in shaping subjectivities and transforming the ways in which sociologically relevant information can be accessed.

In this context, this study focuses on exploring spontaneous narratives related to psychological distress, which are widely disseminated on the internet, especially on social media platforms. To achieve this goal, we employ methods for collecting and analyzing data on digital social networks, while also drawing upon sociological theories related to psychological distress as a foundation for the selection and interpretation of this data.

Keywords: Psychological suffering, Sociology, Data Science, TextRank, Interdisciplinarity, Methodology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

UFPE	Universidade Federal de Pernambuco.
API	Application Programming Interface.
BERt	Bidirectional Encoder Representations from Transformers.
NLP	Natural Language Processing.
POS	Part Of Speach.
UMAP	Uniform Manifold Approximation and Projection.
TF-IDF	Term Frequency-Inverse Document Frequency.
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise.
LDA	Latent Dirichlet Allocation.

SUMÁRIO

1 INTRODUÇÃO

Em tempos de rede mundial de computadores e redes sociais digitais, em que as informações e socializações são extendidas para um locus virtual de ciberespaço, pensar em formas de acessar e compreender as interações que ocorrem através dessas novas tecnologias se torna indispensável para o estudo das sociedades atuais. Pois, assim como as demais tecnologias, as tecnologias digitais provocam mudanças sociais que alteram as configurações de sociedades e grupos sociais. Só no Brasil, 90% [6] das residências possuíam acesso à internet em 2022. Neste mesmo ano, a maior parte do tempo de utilização de internet pelos brasileiros foi destinado a navegação em sites de redes sociais, tornando o Brasil o terceiro país do mundo com o maior tempo despendido nas redes sociais digitais. [7]

Das diversas experiências compartilhadas em redes sociais, estão aquelas relacionadas ao sofrimento psíquico. Sendo a natureza interativa das redes sociais oportuna para a conexão de indivíduos com outros que enfrentam desafios semelhantes, possibilitando um ambiente online de suporte emocional mútuo. Isso foi corroborado por estudos que destacaram o papel das redes sociais digitais na promoção do apoio mútuo e no compartilhamento de informações em questões relacionadas ao sofrimento psíquico [22], [29], [21], [1], [14]. O subreddit desabafos existe com essa proposta, como diz sua descrição: "Quer desabafar e não tem com quem fazer isso? Quer apenas ser ouvido sem ser julgado? Quer apenas despejar tudo e ficar com o coração mais leve? Ou quer simplesmente pedir conselhos para algo que o aflija? Está no lugar certo."[31]

Isto posto, cabe às ciências sociais desenvolver e adaptar metodologias para a coleta e análise de dados, bem como teorias para compreender as novas formas de pensamento, comportamento e percepção que surgem a partir, ou acontecem através, dessas inovações tecnológicas. Nesse contexto, esse trabalho se propõe a coletar, pré processar e extrair informações de dados de uma rede social digital para analisar esses dados instrumentalizando teorias sociológicas de sofrimento psíquico. Para isto, o segundo capítulo será dedicado a apresentar o conceito de virtualidade, de redes sociais e do sujeito que é atravessado por essa tecnologia e se apresenta nela. O terceiro capítulo detalhará a metodologia empregada, que inclui tópicos que abordam o processamento de linguagem natural, uma descrição densa do site de rede social reddit, a coleta dos dados, o pré pro-

cessamento dos dados, a extração de palavras chaves utilizando text rank, a criação de categorias qualitativas baseadas nas palavras chaves mais proeminentes e uma discussão metodológica sobre as tentativas metodológicas para a realização desse trabalho. Em sequência, o quarto capítulo abordará os resultados obtidos durante a pesquisa. Por fim, o ultimo capítulo trará a conclusão e a perspectiva para o desenvolvimento de trabalhos futuros.

.

2 OBJETIVOS

2.1 Objetivo Geral

Instrumentalizar teoria sociologica em um modelo analítico de criação de categorias centrais em textos para utilização metodológica nas ciências sociais.

2.2 Objetivos Específicos

Para atingir o objetivo principal, foram definidos os seguintes objetivos específicos:

- Situar conceitos de virtualidade e redes sociais numa abordagem sociológica.
- Coletar pre processar uma grande quantidade de narrativas textuais sobre sofrimento psíquico.
- Propor um modelo analítico que possibilite a extração de informações importantes desse grande volume de dados.
- Criar categorias gerais que englobem as principais informações extraídas.
- Instrumentalizar teoria sociológica sobre sofrimento psiquico com base nas categoria criadas.

3 O VIRTUAL, AS REDES SOCIAIS E AS PESSOAS DAS REDES

Rede social é uma ideia que muitos autores das ciências sociais já evocavam antes da era digital para fazer referência a estruturas abstratas de comunicação. Essas malhas não materiais, são formadas através do uso da linguagem entre os indivíduos, e tem suas formas de existência condicionadas a um movimento dialético com as materialidades sócio históricas. Como contextualiza Sonia Acioli (2007), desde 1949 há um aceno nas ciências sociais ao que seria compreendido por redes sociais. Lévy Strauss em sua obra "Estruturas elementares do parentesco" já traz a palavra rede para se referir a malha de relações que o antropólogo precisa se atentar para alcançar os sentidos das diferentes estruturas sociais. Posteriormente os textos de Radcliffe Brown foram utilizados pelo Antropólogo J. C. Michel (1969), que propõe um diálogo crítico com o autor para delimitar seu conceito de redes e trazer atualizações ao tema, como é o caso da ideia que essas redes possuem a capacidade de mutabilidade.

As linguagens, por sua vez, são elementos que residem no campo da virtualidade, se definimos virtualidade como "(...) toda entidade "desterritorializada", capaz de gerar diversas manifestações concretas em diferentes momentos e locais determinados, sem contudo estar ela mesma presa a um lugar ou tempo em particular."[15, p.47] Com base nisso, podemos considerar as redes sociais construções virtuais desde a sua gênese por terem, como sua principal ferramenta, o campo abstrato da linguagem. Quando há uma comunicação exclusivamente oral, ainda que a linguagem possa perpetuar informações através de tempos e espaços, ela está restrita a transmissão face a face. A transmissão de informação acontece de forma viva, por meio do vínculo contínuo e permanente das interconexões entre as pessoas. A criação da linguagem escrita, traz uma revolução para as redes sociais, pois, de acordo com Pierre Levy, as mensagens agora passam a se tornar universalizadas. Ou seja, as informações conseguem ser passadas para diversos locais e tempos, sem que seja necessário um indivíduo estar presente num momento e local específico para transmiti-la ou acessá-la. Por ser estática, a linguagem escrita precisa passar por uma adequação das suas informações a novos contextos e sentidos, para que continuamente siga fazendo sentido de forma universalizada. Ou seja, continuem fazendo sentido através de tempo e/ou espaço, sem um componente humano vivo presente para transmitir a mensagem. Como é o caso de traduções, releituras e intérpretes.

Porém, com a criação da internet, uma rede mundial de computadores conectados no mundo inteiro, surge o cyberespaço e uma nova forma de se comunicar, modificando as formas que as redes sociais se estabelecem e são acessadas. Uma nova forma de universalização da linguagem surge, com a característica da linguagem oral de ter uma forma efervescente de renovação de contextos e sentidos pelas contínuas interações e, ao mesmo tempo, com a característica do registro da linguagem escrita, que torna desnecessária a presença de um indivíduo para acessar e transmitir uma mensagem. Aumentando a possibilidade de acesso de informações a nível de velocidade e facilidade de acesso, de forma historicamente nova. Apesar desse cyberespaço já ser formado por uma interação contínua de pessoas por meio da internet, existem sites específicos que foram desenvolvidos com o propósito de ter redes sociais construídas e bem definidas, as redes sociais digitais ou sites de redes sociais. Esses sites podem ser definidos como "um serviço baseado na internet que permite os indivíduos criarem uma conta de perfil público ou semipúblico dentro desse sistema limitado, articular uma lista de outros usuários com quem eles compartilham uma conexão e visualizar e explorar a sua lista de conexões e aquelas feitas por outras pessoas dentro do sistema." [4, p.221] A construção desses perfis é feita com base em como os usuários querem se expressar e serem vistos pelas outras pessoas, sendo então um local rico quando se trata de acessar a construção da auto representação do "eu" explorada por Goffman em suas obras. Apesar de em "A representação do eu na vida cotidiana" de 1959, Goffman deixar claro a necessidade de uma interação face a face para o desenrolar de sua análise, é necessário compreender o momento histórico que sua obra foi escrita e as diversas mudanças sofridas na comunicação citadas anteriormente. Mudanças que afetam diretamente a construção de redes de interação entre pessoas, e consequentemente, as construções do "eu" para serem apresentadas nessas diversas redes. [11] Como apresenta em sua compreensão socio dramatúrgica da construção das formas de "ser", os indivíduos se assemelham a atores que se apresentam de forma a tentar convencer seu público de sua performance. A visualização de papéis sociais como expectativas de um público, traz à luz a adequação dos atores a espaços em que um determinado papel social é evocado, a fim de angariar reações de aprovação ou desaprovação do seu público. Esses papéis sociais são construtos históricos constantemente presentes e transpostos entre si, de forma que, cada espaço ocupado pelos atores, exigirá uma ou mais performances a serem assimiladas. Por ser algo que envolve toda a existência dos indivíduos, os esforços de balancear suas

subjetividades e as normas sociais, não são necessariamente uma caricatura maquiavélica do sujeito, mas o processo de formação das subjetividades em si, o que permite que os sujeitos se auto compreendam, e assim expressem sua própria subjetividade. Isto posto, o que clamam como superficialidade das interações mediadas por redes sociais digitais, é um movimento de auto construção de identidade já previsto na teoria de Goffman nas diversas interações cotidianas. Desde a escolha da imagem de perfil até o que é escolhido diariamente para ser postado é parte dessa construção mais ou menos controlada na forma de se auto projetar, e por consequência, se auto perceber. Para finalizar, é importante ressaltar a maior facilidade na criação desses perfis de redes sociais digitais em comparação com a vida offline, desta forma um mesmo ator pode ter diferentes perfis de um mesmo site de rede social, podendo se apresentar de diferentes maneiras num mesmo palco digital sem gerar um sentimento de estranheza nos espectadores.

4 METODOLOGIA

Esse trabalho se lança ao desafio de produzir uma análise sociológica, de abordagem mista, utilizando técnicas de NLP (Natural language Preprocess), referenciando-se em conceitos de teoria social sobre sofrimento psíquico. Isto é realizado a partir de uma coleta de dados online do subreddit "desabafos"do fórum reddit Brasil. Após a coleta, é incorporada a criação de um filtro fundamentado a partir de um léxico que visa a identificação de indícios de depressão em postagens de redes sociais. Onde, é mantido para análise posterior, apenas as postagens que possuem cinco ou mais palavras contidas neste léxico. Posteriormente, é utilizado o método de NLP denominado text rank, que se encarrega de extrair palavras com uma maior relevância para a compreensão das temáticas centrais das postagens. Com essas palavras-chave, devidamente selecionadas, as classes são construídas de forma qualitativa. Por fim, as teorias sociológicas de sofrimento psíquico são instrumentalizadas para contextualizar e compreender as razões subjacentes ao sofrimento coletivamente experimentado, imbuídas no contexto social. [3]

4.1 O processamento de linguagem natural: Uma breve contextualização

Linguagem natural, em oposição as linguagens de máquina, é como a ciência da computação chama as diversas linguagens humanas. Processar essas linguagens consiste em fazer com que um computador seja capaz de interpretar as palavras que as compõe, associando essas palavras à números, para que seja possível manipular essas palavras na forma de dados, com técnicas matemáticas e computacionais. Existem diversos algoritimos que podem ser utilizados nessa numerificação de palavras. Eles podem ser mais simples como é o exemplo do bag of words, que utiliza um vetor em que cada componente é correspondente ao número de ocorrências de uma palavra em um documento [25], até os mais complexos utilizando redes neurais pronfundas, como é o caso do BERt [5].

A preocupação com o processamento de linguagem natural surge voltada para a tradução de idiomas, tendo a sua primeira aparição em 1949 [12]. Nesse período, a abordagem girava em torno da sintaxe e a tradução era feita verbete por verbete. Posteriormente, nos anos 60, os cientistas da área se debruçam em uma empreitada para construir um

[3] Todos os códigos utilizados na construção das etapas de coleta, tratamento e análise de dados desse trabalho foram escritos em scrip python e estão disponíveis para livre acesso e utilização: https://www.github.com/Nico-lly/depressao_reddit/

algorítmo capaz de um diálogo interativo, onde, consequêntemente a semântica é trazida a tona como o marco a ser conquistado. A inteligência artificial começa a ser cogitada com entusiasmo como principal aposta para potencializar a interpretabilidade das palavras. Mas, apesar dos esforços e avanços, os resultados de um diálogo interativo não foram tão satisfatórios. Finalmente, nos anos 70 e 80 foram desenvolvidos modelos para as análises de texto, com teoria da gramática computacional impulsionada pelo surgimento das linguagens lógicas de programação e de teorias de estruturas gramaticais que surgiram na época.

NLP é uma área que segue em contínuo aprimoramento e expansão até os dias atuais, muitos desafios ainda precisam ser ultrapassados, mas já é possível utilizar uma série de modelos para apreender representações textuais. Hoje os conhecimentos da área são utilizados para análises de discursos, análises de sentimentos, análises de tópicos, entre outros. Específicamete nas ciências sociais, essa área da ciência da computação já foi absorvida em métodos para produzir estudos como a determinação da influência dos principais filósofos da ciência na sociologia [10], estudos de polarização política [8, 17], a vizualisação do processo civilizador de Norbert Elias enquanto fenômeno em textos jurídicos [13], identificação de tendência de termos chave em estudos sociológicos [2].

4.2 **O reddit**

O *reddit* é uma rede social online de texto, imagem e vídeo criada em 2005 nos Estados Unidos. Foi a rede social escolhida para a coleta de dados neste trabalho por ser uma rede social que possui um grande número de postagens construídas exclusiva ou majoritariamente de texto, uma API gratuita que permite acesso aos dados de um grande número de posts e por ter muitos usuários ativos.

Ela tem o formato de fórum com divisões de grupos em torno de temas, chamados *subreddits*. Seu formato é de rolagem infinita, ou seja, não existe paginação, o usuário desce continuamente a barra de rolagem e mais postagens são carregadas. Assim, as postagens mais recentes ou relevantes aparecem primeiro, ou em cima, na hierarquia da barra de rolagem. Não é necessário ter conta *reddit* para visualizar conteúdos de *subreddits* e perfís públicos, no entanto, para ser possível realizar interações nessa rede social, é necessário fazer um cadastro para a criação de um perfil identificável por meio de um nome de usuário e uma imagem. As interações possíveis são: adicionar outro perfil

como amigo, a criação de novos *subreddits*, postagens dentro dos subreddits ou dentro de seu próprio perfil, comentários dentro de postagens e *ups* das postagens. As postagens são divididas em título, corpo e *tag*, onde, o título contém apenas texto com no máximo trezentos caractéres, o corpo pode ser composto de textos com limite de caractere não informado, fotos, vídeos, *links* e enquetes e as *tags* são palvras ou siglas que sinalizem um assunto específico. Os comentários, por sua vez, podem existir em forma de texto com limite de caractere não informado, fotos ou *gifs*. E, por fim, os *ups* são o que tornam as postagens mais ou menos relevantes, consiste em um botão que ao ser clicado sinaliza que a postagem importa para o usuário que clicou e ele quer que ela siga aparecendo na parte mais acima do fórum independente de sua data de criação.

Até 2021, segundo números oficiais da própria plataforma, o reddit possuia cerca de 57 milhões de usuários ativos, 100 mil comunidades ativas e 13 bilhões de postagens e comentários[28]. Não existem números oficiais da quantidade de usuários brasileiros disponível, mas no *subreddit* Brasil, até a data de escrita deste trabalho, existem 1.6 milhões de usuários membros [26]. Apesar de não ser a rede social mais popular do Brasi, o *reddit* é uma das redes sociais com mais usuários ativos do mundo.

4.3 **A Coleta de dados**

As postagens do *reddit* foram coletadas em uma única passagem, via *API*. [27], por meio da realização de cadastro e autenticação de identidade para acesso ao banco de dados. Devido a uma limitação da *API*, que permite coletar dados de apenas 100 postagens por *endpoints*, foram utilizados vários *endpoints*, dos quais: *hot*, *news*, *rising* e *top*. Além disso, também utilizou-se uma busca em *looping* de cem em cem postagens para cada endpoint, a partir da data mais antiga dos primeiros cem dados coletados até seus respectivos limites de retorno, para com isso, aumentar a quantidade de dados por endpoit e coletar a maior quantidade possível de dados em um curto período de tempo. No total, foram obtidos 2028 dados da API, onde, todos os dados são de postagens feitas entre 04/2020 e 04/2023 e pertencem ao *subreddit* desabafos. Os dados coletados foram os títulos dos textos das postagens, os textos das postagens, os números de identificação única de cada postagem, datas de publicação, números de UPs e números de comentários. Nesse trabalho, apenas os dados de texto serão utilizados, sendo eles os títulos das postagens e os escritos do corpo das postagens.

Tabela 1: Exemplos de dados brutos coletados da API reddit

coments	**created**	**id**	**selftext**	**title**	**up**
118	1.61066e+09	t3_kxfnzb	Olá, meu nome é Gabriel, tenho 18 anos e moro em Manaus. Se você tá acompanhando as notícias, você sabe como as coisas estão por aqui [...]	Hoje é o dia mais sombrio da minha cidade.	1850
808	1.658598e+09	t3$_w$69jxb	Em 2019 reencontrei minha primeira namorada, v[...]	Meu casamento me destruiu financeiramente.	1558

O léxico utilizado foi coletado diretamente do github dos autores. Ele foi escrito originalmente em 2017 para um estudo que tem como propósito identificar e monitorar sintomas clínicos de depressão nas redes sociais. O léxico está disponível apenas em língua inglesa e possui um total de 1299 palavras divididas entre 10 sintomas clínicos.[32] Pela dificuldade de encontrar um léxico semelhante em português, foi realizada uma tradução livre, totalmente manual, visando preservar ao máximo o significado das palavras na passagem de uma língua para outra. Dessa maneira, mantivemos as palavras em inglês que possuem flexão de gênero neutra e adicionamos as flexões de gênero necessárias para abranger todas as traduções possíveis para o português. Palavras como gírias locais foram retiradas, para minimizar a quantidade de viés. Por fim, esta tradução reduziu o léxico utilizado neste trabalho para um total de 1265 palavras.

4.4 **Pré-processamento**

Via de regra, é necessário fazer um tratamento do texto a priori, antes de realizar análises posteriores, visando melhorar a qualidade do resultado. Apenas alguns algoritmos que utilizam redes neurais possuem na literatura registros de indiferença com a não realização de um pré-processamento [9]. Desta forma, por utilizar um algoritmo estatístico, as etapas a seguir foram necessárias para a conformidade dos dados das postagens:

- Junção do campo de título e texto: Para que o fossem analisados em conjunto, como parte de um único campo de texto;

- Tokenização: Transformação de texto corrido em lista de palavras separadas por vírgula;
- Transformação de todas as letras em minúsculas, retirada de acentuação, pontuação e caraceteres especiais: Para evitar a duplicação de palavras iguais com ortografias diferentes;
- Tradução de palavras escritas com gírias de internet para o português formal: Para, novamente, evitar a duplicação de palavras;
- Remoção de todas as palavras contidas lista de stop words: Conectivos e palavras cuja sua repetição pode ser prejudicial na análise;
- Contagem de quantas palavras do léxico estavam presentes no texto: Para a seleção de textos com cinco ou mais palavras do léxico.

Além das tradicionais palavras definidas como stop words (conectivos), foi incluída como palavras a serem retiradas desta análise todas as palavras que não são substantivos e adjetivos. Essa decisão foi tomada, pois, essas palavras quando isoladas, não trazem nenhum significado intrínseco que pode ser interpretado posteriormente em uma análise sociológica. A construção de stop words para este trabalho foi realizada da seguinte forma:

- Utilização da biblioteca de processamento de texto spacy, para coletar todas as POS tags (part of speech) das palavras dos textos, separando em uma lista todas as palavras que não são adjetivos e substantivos;
- Donwload da lista pré definida de stop words da biblioteca de processamento de texto nltk;
- Unificação da lista de palavras selecionadas com as POS do spacy a lista de stop words padrão da nltk;

O léxico foi traduzido e armazenado manualmente em formato de lista, por ser composto apenas de termos e palavras significativas, precisou passar por um pré-processamento simples, com os seguintes passos:

- Retirada de acentuação
- Transformação de todas as letras em minúsculas.

4.5 **Extração de palavras chaves: Text rank**

O text rank é um algoritmo estatístico baseado no pagerank da google, que, por sua vez, foi criado para recomendar páginas da internet nas buscas dos usuários da plataforma. De forma que, um dado conjunto de páginas forma um grafo, onde, cada nó é uma página. Se uma página contém um link direto para outra, esta página possui uma relação com a outra, e em um grafo, essa relação é representada com uma aresta unidirecional ligando um nó ao outro. No caso do text rank, o grafo é formado por um conjunto de textos, em que os nós são as palavras, e quando há co-ocorrências entre as palavras, surgem as arestas. Após essa formação inicial do grafo, é calculada o peso dos nós, ou seja, a quantidade de relações que ele possui com outras palavras [20]. Sendo assim, esse algoritmo leva em consideração a interação das palavras umas com as outras para calcular a relevância da palavra para a compreensão dos textos, tendo uma proximidade maior com seu contexto e, logo, com seu sentido semântico, ao invés de apenas considerar sua forma gramatical.

O text rank utilizado nesse trabalho foi construído manualmente com base no código disponibilizado no github [16], da seguinte forma: Para cada texto, foi criado um conjunto de palavras únicas, e um conjunto de pares de palavras para registrar as coocorrências entre elas. Para os pares de co-ocorrência foi definida uma janela de cinco palavras, ou seja, a cada grupo de cinco palavras foram gerados vinte e cinco pares de coocorência sendo cada uma das cinco palavras coocorrentes entre sí. Posteriormente, é contruída uma matriz para cada texto, preenchida integralmente com zeros, onde, a quantidade de linhas e colunas é igual a quantidade de palavras únicas. Então, para cada par de coocorrência, é identificado a posição de ambas as palavras no conjunto de palavras únicas, e a posição da primeira palavra é assinalado como o das linhas da matriz e o da segunda palavra como as colunas da matriz. Desta forma, é preenchido com 1 (um) apenas as posições da matriz em quem as palavras co-ocorrem na linha em relação a coluna. Feito isto, é necessário agora tornar esta matriz simétrica, ou seja, fazer com que a relação entre duas palavras seja a mesma, independentemente da ordem em que essas palavras são consideradas. Para isso, é feita a soma da matriz original com a sua transposta e a subtração da matriz resultante da soma pela matriz diagonal principal. Para finalizar a construção da matriz, ela é normalizada por coluna. Isto é, cada elemento da coluna da matriz é dividido pela soma de valores da coluna. Com a matriz de similaridade pronta, é o momento de calcular os pesos de cada palavra (vertices).

$$S(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \tag{4.1}$$

A equação 2.1, é uma função que calcula o peso dos vértices, onde:

- S(Vi) = Peso do vértice que está sendo calculado;
- In(Vi) = O número de vértices que apontam para o vértice Vi;
- Out(Vj) = O número de vértices que o vértice Vj aponta;
- S(Vj) = Peso de um dado vertice que Vi aponta;
- d = Fator de amortecimento, número entre 0 e 1 que pode ser escolhido;

Essa equação vai sendo continuamente interada até que os valores resultantes estabilizem, ou seja, não possuam uma mudança maior que 1e-5. O resultado disto é o peso de cada palavra para cada texto.

4.6 **Criando categorias**

Com os pesos dos vértices uma vez definidos por texto, somamos os pesos das palavras identicas, resultando em uma visão geral de quais palavras são mais importantes para o conjunto geral de textos. Selecionamos então as cem palavras com maior peso, estas por sua vez são substituídas pelas suas respectivas categorias, e os valores de pesos das palavras de uma mesma categoria somados, resultando no peso da categoria para cada texto. Neste momento alguns ajustes foram feitos com uma amostra aleatória de 20 textos para avaliar qualitativamente a relação dos textos com suas categorias mais pontuadas. Esse ajuste foi feito retirando palavras que faziam categorias aparecerem de forma aleatória, sem retratar os temas principais dos texto, como palavras que podem assumir multiplos significados a depender de seu contexto ou que não formam conjuntos categoricos com nenhuma outra palavra. Para exemplificar, a categoria tempo foi inicialmente criada, mas precisou ser desfeita. Os textos revisados apontaram as palavras encontradas nesta categoria enquanto marcações de uma narrativa linear que está sendo contada, e não percepções ou discussões com relação ao conceito do fenômeno tempo como temática central. Isto exposto, as categorias criadas para utilização neste trabalho são:

- familia = pai (181.51), casa (157.98), irmao (40.77), filho (39.47), lar (38.71), crianca (31.76)
- amigos = amigo (128.99), amizade (29.81)
- labor/sustento = trabalho (89.71), escola (59.07), faculdade (45.92), emprego (37.62), dinheiro (59.63)
- corpo fisico = dor (51.13), corpo (45.34), feio (33.6), pau (29.07), fisico (26.0), medico (26.0), velho (26.13)
- sociabilidade = conversa (38.38), social (35.14), contato (29.41), mensagem (47.47)
- amor = relacionamento (68.85), garota (34.12), amor (31.7), homem (52.57), mulher (75.18)

4.7 **Discussões metodológicas**

Este trabalho teve como ideia inicial a segmentação de tópicos por meio da criação de embeddings e algoritmos de clusterização. Para isto, uma série de conjuntos de algoritmos foi explorada [5] [19] [18] [30] [23] [3], resultando nas seguintes descobertas:

- Bert + UMAP + HDBSCAN: Inicialmente, o HDBSCAN foi cogitado, mas a visualização dos dois componentes principais da redução de dimensionalidade revelou que os dados não apresentavam densidade suficiente para que o algoritmo funcionasse de maneira eficaz.
- Bert + UMAP + Kmeans + TFIDF: Sendo assim, o Kmeans como solução, e foi escolhido devido a sua facil compreensão e aplicação. Ao adotar a combinação de Bert, K-means e TF-IDF, observou-se uma melhoria visual na separação dos tópicos. No entanto, a seleção de palavras-chave se apresentou de forma aleatória, sem um padrão bem definido por cluster/categorias.
- Bert + Kmeans + UMAP + Text rank: O algorítmo de seleção de palavras-chave então foi trocado, no entanto, a seleção permaneceu desafiadora, levando a questionamentos sobre a qualidade da redução de dimensionalidade realizada pelo UMAP. (ver se mudou parametros do kmeans)

- Bert + Kmeans + UMAP + Text rank (apenas títulos): A clusterização parecia interessante demais para não dar uma outra chance. Dessa vez, a abordagem alternativa consistiu em analisar apenas os títulos dos textos, que frequentemente fornecem uma introdução aos tópicos subsequentes, e são mais curtos. Embora a clusterização ainda apresentasse características interessantes, a seleção de palavras-chave permaneceu confusa.

- Bert + LDA + Text rank: A ultima tentativa consistiu na mudança de um algoritmo de clusterização para um modelo criado especificamente para a separação de tópicos. O LDA (Latent Dirichlet Allocation), tem a capacidade de escolher automaticamente o número de tópicos com base nos dados, eliminando a necessidade de especificar previamente o número de clusters. Como resultado, este modelo apontou para um único cluster contendo todos os textos. Isso explicaria a dificuldade de selecionar palavras chaves para os clusters separados anteriormente. Com um teste de log likehood aplicado ao LDA, foi possível observar a qualidade da segmentação entre os clusters diminuindo à medida que o número de tópicos aumentou.

Com base nisso, a abordagem de dividir os textos em vários tópicos foi rejeitada. Como alternativa, adotou-se a ideia de que todos os textos pertenciam a um único tópico. Consequentemente, a extração de palavras-chave foi realizada em todos os textos de forma unificada, considerando que todos fossem de um único grupo. Somente após a extração das palavras-chave desse amplo grupo unificado de textos, é que as classes foram construídas de forma qualitativa, conforme descrito no passo a passo metodológico.

5 CLASSES DE SOFRIMENTO PSÍQUICO

Analisando os dados coletados deste subreddit, sob uma ótica do sofrimento psíquico, foi possível extrair as principais temáticas das narrativas sobre o tema.

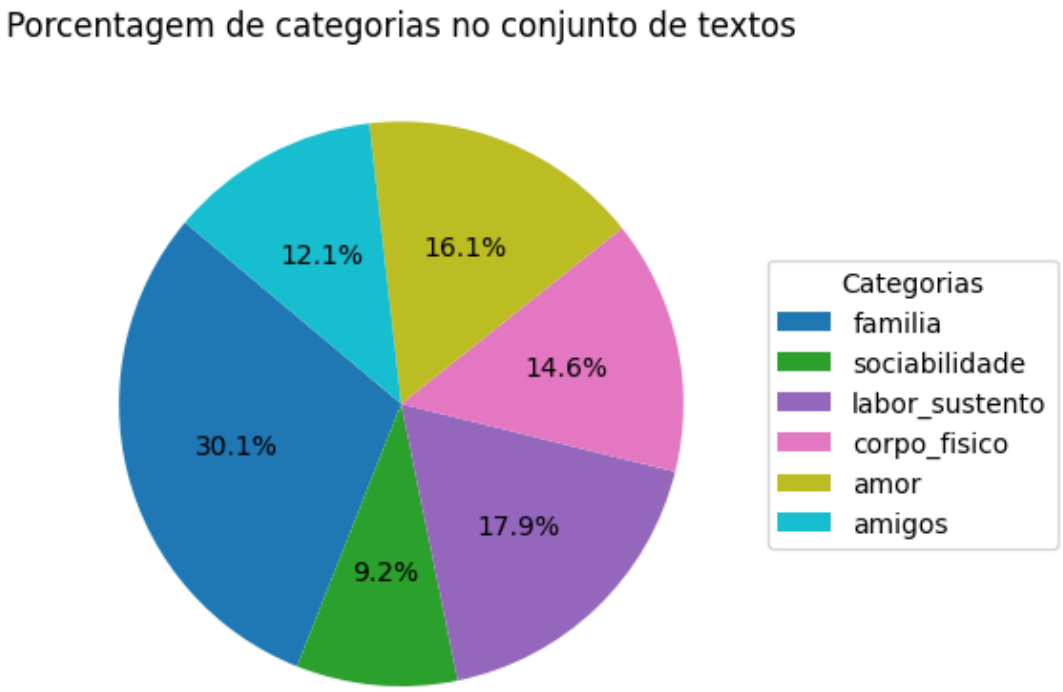

Figura 1: Porcentagem de categorias no conjunto de textos

Como é apresentado na imagem 1, o conjunto de textos tem como tema mais importantes e presentes a categoria família, seguida por corpo físico e labor/sustento. As temáticas porém, se espalham pelos textos de forma não padronizada. Explico, a categoria dominante, em geral, é compatível com o que é descrito nos textos. Porém as categorias subsequêntes dão o contexto para o grande tema da categoria dominante, situando e guiando a narrativa para os desdobramentos variados que o tema central pode confluir.

Classes	Texto
‘corpo_fisico’: 9.552, ‘familia’: 3.213, ‘amor’: 3.098, ‘labor_sustento’: 0.757, ‘sociabilidade’: 0.655	‘Sexo é vida, né? Acho que todos que leram alguma revista de grande circulação no Brasil já devem ter visto essa propaganda. A Boston Medical Group é famosa no ramo de saúde sexual masculina, mundialmente.[...] Em agosto de 2019 comecei um relacionamento. A relação era boa, o sexo também, mas eu não conseguia manter uma ereção prolongada.[...]’
‘amor’: 10.902, ‘familia’: 3.120, ‘sociabilidade’: 2.160, ‘labor_sustento’: 1.622, ‘amigos’: 0.669	‘Eu já vi alguns desabafos aqui sobre a clássica história: menino bom, tímido e estudioso gosta da menina. Menina não gosta dele, mas sim do ”bad boy”[...] Agora vamos lá: o que cria um nice guy? No meu caso, pressão da masculinidade tóxica. Eu tenho uma família onde os homens são metidos a ”HETERO FODA[...]’
‘corpo_fisico’: 2.976, ‘amigos’: 0.771, ‘labor_sustento’: 0.702	Quando eu tinha 15 anos, umas meninas da minha sala me disseram que eu era feio, e ficavam jogando na minha cara que eu era feio. [...] Cresci e fiquei a maior parte da minha vida traumatizado com isso. Não gostava de tirar fotos, pois de fato eu acabava me achando feio. Pois eu cheguei aos 40 anos totalmente marcado por isso, e resolvi que não vou mais dar vazão a esses pensamentos [...]

Tabela 2: Exemplo de texto com suas respectivas classes e pesos para cada classe.

Como podemos verificar nos exemplos de classificação apresentados na tabela 2, a mistura não padronizada de categorias em cada texto, faz com que os relatos tenham uma característica pouco generalizável de explicações e sentidos para a categoria principal. Isso pode ser atribuído, não apenas às limitações inerentes a um trabalho de monografia, mas também à natureza espontânea dos discursos produzidos e ao entrecruzamento das experiências de diferentes campos da vida de um indivíduo.

No contexto metodológico de uma coleta de discursos espontâneos, os indivíduos estão menos dispostos a se limitarem a um eixo temático e mais inclinados a narrar situações que envolvem diversas esferas de suas vidas, buscando justificar suas falas ou criar sentido nas suas narrativas. Em contraste com dados coletados em pesquisas mais tradicionais na área (questionários, etnografia, observação participante, descrição densa, grupo focal e etc), há a ausência de interferência do pesquisador na condução dos relatos. Essa técnica se mostra interessante porém, justamente por não existem perguntas ou roteiros, materiais de pesquisa que cause algum tipo de desconfiança, ou até mesmo uma pessoa estranha tentando se inserir em um ambiente no qual não é nativo. Mas se por um lado, a ausência de um pesquisador pode evitar influenciar ou tensionar a expressão dos indivíduos, por outro se mostra desafiador organizar essas falas.

Além dos desafios das novas metodológias em contrução e refinamento, os resultados nos leva também a uma reflexão sobre como os diversos campos da vida de um indivíduo se entrelaçam de maneira intrincada. Mesmo quando tentamos isolar uma temática, neste caso, sofrimento psíquico. As diversas interações das esferas da vida de um indivíduo se tangigenciam ou até mesmo se atravessam na construção, não apenas de seu discurso livre, mas no constante processo formativo do próprio indivíduo. É possível vislumbrar essa proposta de um indivíduo que transborda campos da vida na abordagem da micro sociologia proposta por Bernard Lahire [24]. Esta perspectiva destaca como diferentes vivências das variadas esferas da vida impactam diretamente o nosso repertório de esquemas de ação. Sendo, esquemas de ação, por definição de Lahire, conjuntos de comportamentos e reações adquiridos socialmente ao longo da vida. Porém, a chave de compreensão de sua teoria se concentra em como as disposições aprendidas em um contexto podem ser habilmente aplicadas em outros, dependendo das exigências das situações em que os indivíduos se encontram. Lançando luz sobre a complexa teia de experiências que moldam a vida de um indivíduo, como essas experiências são internalizadas e como

elas não se restringem a esfera de vida em que foi primeiramente apresentada. Dessa forma, a abordagem sociológica de Lahire não negligencia a importância do contexto em sua análise, e destaca como a interseção de experiências e esquemas de ação molda a complexa tapeçaria da vida de uma pessoa. Isso posto, a sociologia de Lahire nos ajuda a compreender como essas narrativas multifacetadas refletem a complexidade das formas que os seres humanos pensam, sente e agem no mundo. Ao explorar a interação dinâmica entre experiências e esquemas de ação, podemos vislumbrar como uma situação pode ser influenciada por uma variedade de fatores contextuais e como os indivíduos respondem de maneira criativa a essas complexidades. Por fim, para a busca pela compreensão sociológica de possíveis fatores que contribuem para um evento tão subjetivo quanto o sofrimento, mesmo esses eventos se apresentando em grande escala, se faz necessário lembrar com mais esmero dessa proposta teórica Lahireana e da criatividade latente dos indivíduos para continuamente criar e lidar com cenários potencialmente novos.

6 CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho foram explorados desafios de uma abordagem interdisciplinar que busca munir a sociologia de métodos e teorias que englobem o cyberespaço.

Inicialmente, estabeleceu-se uma conexão entre a teoria de Pierre Levy sobre cyberespaço com a sócio-dramaturgia de Goffman para produzir uma imagem de como o indivíduo existe nesse ambiente criado por novas ferramentas comunicacionais. Como o ciberespaço existe dentro, e não em oposição, da vida fora da internet. E, desta forma, como as informações produzidas pelos indivíduos nessa ferramenta de comunicação são válidas, por serem reflexo de uma, entre tantas outras, formas de se apresentar e se compreender nos diferentes palcos da vida.

Na etapa de análise dos dados, ao organizar as informações coletadas da internet para a criação de classes de sofrimento psíquico, foi entendida a impossibilidade de categorizar com apenas uma classe cada relato. Visto que, o conjunto de classes se mostrou mais representativo para uma descrição da temática tratada em cada texto. Apesar da metodologia utilizada ter possibilitado a classificação desses textos, com categorias relevantes para a compreensão de seus principais temas, não foram encontrados padrões de coocorrência entre as classes criadas nos relatos coletados dos usuários do fórum desabafos no reddit. Este resultado trouxe consigo uma reflexão sobre a pertinencia de se pensar o objetivismo e o subjetivismo na sociologia. E com isto, a análise seguiu através das lentes teóricas de Bernard Lahire, na tentativa de contribuir com a construção de uma sociologia do sofrimento psíquico. A teoria de Lahire, foi então instrumentalizada, para trazer a atenção para a práxis de teorias que se enveredam por objetos de estudos centralmente subjetivos, em especial para as teorias sociológicas de sofrimento psíquico, mas o fazem em uma escala em que a subjetividade é perdida de vista. Por mais desafiador que se apresente o trabalho de construir teorias sociológicas a nível de compreensão de aspectos centrados na subjetividade, e mesmo que o resultado de uma pesquisa aponte para uma generalização possível, a sua construção precisa dispor da devida importância para o indivíduo como um agente dotado de disposições sociais diversas.

Se tratando de trabalhos futuros, uma perspectiva promissora envolve a continuação do desenvolvimento e validação do método proposto. Este método visa oferecer uma solução alternativa para lidar com textos que apresentam desafios significativos na

separação por tópicos. No entanto, é fundamental a aplicação de métricas de avaliação adequadas. Recomenda-se a criação de um conjunto de dados de controle, onde as categorias são atribuídas manualmente, permitindo a comparação dos resultados gerados pelo modelo com essas categorias predefinidas para os mesmos textos. Um segundo ponto que pode ser desenvolvido é uma análise cronológica, objetivando identificar variações ou estabilidades na aparição das classes de sofrimento ao longo de alguns anos. Uma terceira sugestão é a utilização da metodologia proposta para outras temáticas.

Por fim, essa abordagem interdisciplinar permitiu traçar um quadro que abrange e permite aprofundamento das dinâmicas sociais que se inserem nesses ambientes digitais. A metodologia proposta visa, ao derivar da ciência de dados, possibilitar a coleta e o manuseio de informações em volumes antes improváveis sem uma grande estrutura de pesquisa. Ao mesmo tempo, a interdisciplinaridade deste método entre computação e a sociologia, apresenta o teor factível do levantamento de relatos para a extração de informações de caráter subjetivo. Como resultado, este trabalho reforça a pertinência da apropriação da internet enquanto ferramenta sociológica. Incentivando a interdisciplinaridade com a computação como via para articulação de métodos e teorias sociológicas de compreenção não só o ambiente digital, como também questionamentos sociológicos que existem há mais tempo que este espaço de comunicação característico do mundo contemporâneo.

7 APENDICE

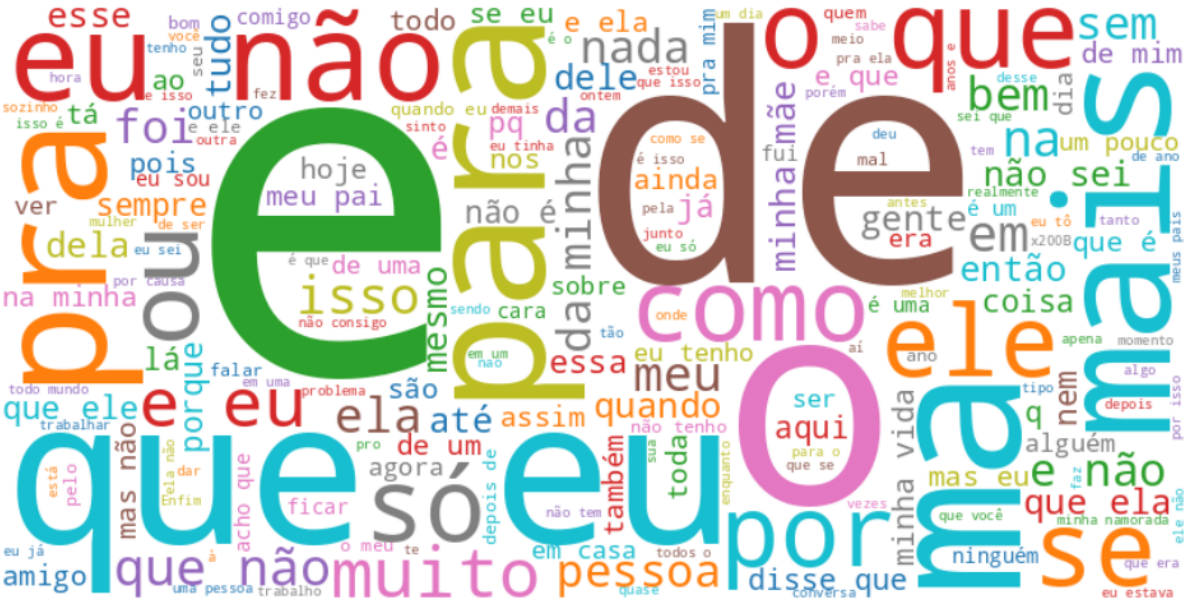

Figura 2: Nuvem de Palavras sem tratamento

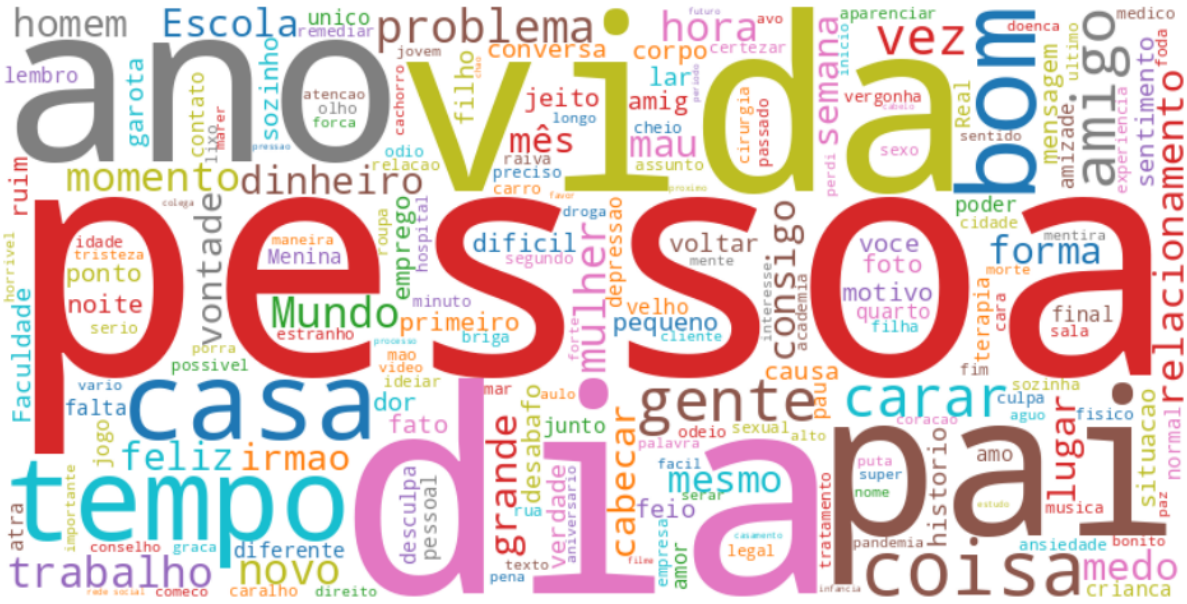

Figura 3: Nuvem de Palavras com tratamento

8 ANEXO

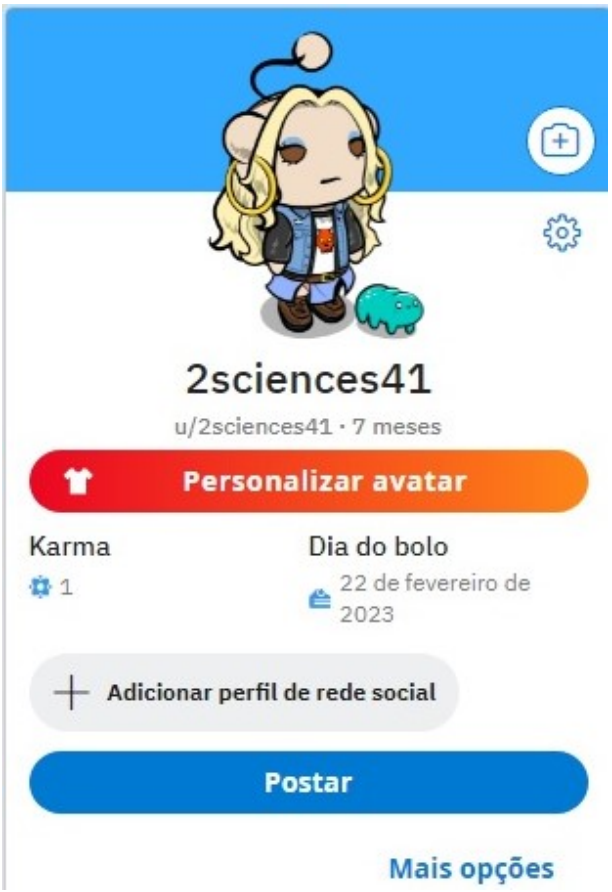

Figura 4: Perfil do Reddit

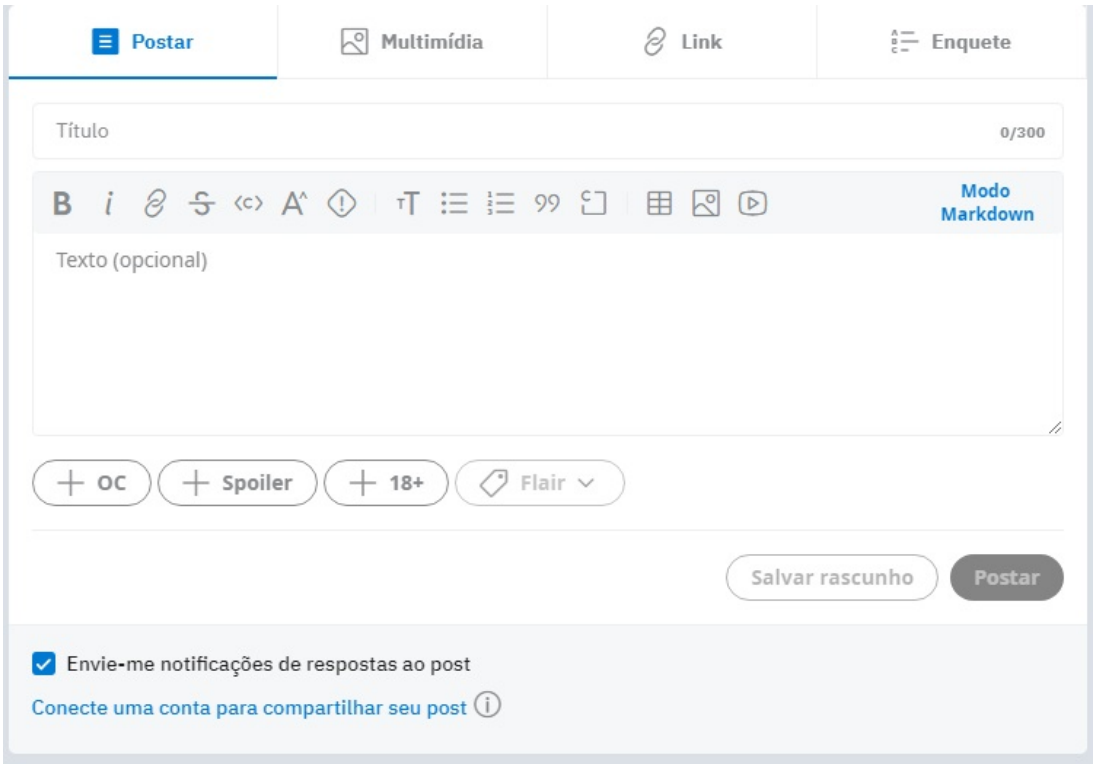

Figura 5: Bloco de criação de postagens do Reddit

9 REFERÊNCIAS

[1] Dane Acena e Guo Freeman. ""in my safe space": Social support for lgbtq users in social virtual reality". Em: *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–6.

[2] John A. Bernau. "Text Analysis with JSTOR Archives". Em: *Socius: Sociological Research for a Dynamic World* 4 (2018). DOI: 10.1177/2378023118809264.

[3] David M Blei, Andrew Y Ng e Michael I Jordan. "Latent dirichlet allocation". Em: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[4] Danah Boyd e Nicole Ellison. "Social network sites: Definition, history, and scholarship." Em: *Journal of Computer-Mediated Communication, 13(1), article 11.* (2007), pp. 210–230.

[5] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Em: 2018. URL: https://arxiv.org/abs/1810.04805.

[6] *Educa IBGE*. IBGE. URL: https://educa.ibge.gov.br/jovens/materias-especiais/21581-informacoes-atualizadas-sobre-tecnologias-da-informacao-e-comunicacao.html.

[7] *Forbes*. Forbes. URL: https://forbes.com.br/forbes-tech/2023/03/brasil-e-o-terceiro-pais-que-mais-consome-redes-sociais-em-todo-o-mundo/.

[8] Roberto Franzosi. "Sociology, narrative, and the quality versus quantity debate (Goethe versus Newton): Can computer-assisted story grammars help us understand the rise of Italian fascism (1919–1922)?" Em: *Theory and Society* 39 (nov. de 2010), pp. 593–629. DOI: 10.1007/s11186-010-9131-3.

[9] Shalmoli Ghosh et al. "Stance Detection in Web and Social Media: A Comparative Study". Em: ago. de 2019, pp. 75–87. DOI: 10.1007/978-3-030-28577-7_4.

[10] Peter HEDSTRÖM, SWEDBERG e Lars UDÉHN. "Popper's Situational Analysis and Contemporary Sociology". Em: *Philosophy of the Social Sciences* 28 (1998), pp. 339–364.

[11] Serrano-Puche Javier. “La presentación de la persona en las redes sociales: una aproximación a la obra de Erving Goffman”. Em: *Anàlisi. Cuadernos de Comunicación y Cultura* 46 (2012), pp. 1–17.

[12] Karen Spark Jones. “Natural language preprocessing: a historical review”. Em: *Current Issues in Computational Linguistics: in Honour of Don Walker* (1994), pp. 3–16.

[13] Sara Klingenstein, Tim Hitchcock e Simon DeDeo. “The civilizing process in London’s Old Bailey”. Em: *Proceedings of the National Academy of Sciences* 111.26 (2014), pp. 9419–9424. DOI: 10.1073/pnas.1405984111.

[14] Jan Marco Leimeister et al. “Do virtual communities matter for the social support of patients? Antecedents and effects of virtual relationships in online communities”. Em: *Information Technology & People* 21.4 (2008), pp. 350–374.

[15] Pierre Lévy. *Cibercultura*. Trad. por Carlos Irineu da Costa. 2ª ed. São Paulo: Editora 34, 1999.

[16] Xu Liang. *GitHub - Textrank.py*. https://github.com/BrambleXu/news-graph/blob/master/textrank.py. Acesso em [data de acesso].

[17] Fabio Malini, Patrick Ciarelli e Jean Medeiros. “O sentimento político em redes sociais: big data, algoritmos e as emoções nos tweets sobre o impeachment de Dilma Rousseff”. Em: *Liinc em Revista* 13 (2017). DOI: 10.18617/liinc.v13i2.4089.

[18] Leland McInnes, John Healy e Steve Astels. “hdbscan: Hierarchical density based clustering.” Em: *J. Open Source Softw.* 2.11 (2017), p. 205.

[19] Leland McInnes, John Healy e James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. Em: 2018. URL: https://arxiv.org/abs/1802.03426.

[20] ”Rada Mihalcea e Paul Tarau”. “TextRank: Bringing Order into Text”. Em: Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 404–411. URL: https://aclanthology.org/W04-3252.

[21] John A Naslund et al. “Exploring opportunities to support mental health care using social media: A survey of social media users with mental illness”. Em: *Early intervention in psychiatry* 13.3 (2019), pp. 405–413.

[22] John A Naslund et al. “Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using YouTube”. Em: *PLOS one* 9.10 (2014), e110171.

[23] Wasseem N Ibrahem Al-Obaydy et al. “Document classification using term frequency-inverse document frequency and K-means clustering”. Em: *Indonesian Journal of Electrical Engineering and Computer Science* 27.3 (2022), pp. 1517–1524.

[24] Renan de Oliveira Rodrigues. “A sociologia de Bernard Lahire e suas críticas à sociologia de Pierre Bourdieu”. Em: *Sinais revista de ciências sociais* 22 (2018), pp. 28–47.

[25] Wisam Qader, Musa Ameen e Bilal hmed. “An Overview of Bag of Words: Importance, Implementation, Applications, and Challenges”. Em: ”Junho” de 2019, ”200–204”.

[26] *Reddit Brasil*. reddit. URL: `https://www.reddit.com/r/brasil/`.

[27] Reddit Press. *Reddit API Documentation*. URL: `https://www.reddit.com/dev/api/`.

[28] Reddit Press. *Reddit by the Numbers*. 2021. URL: `https://www.redditinc.com/press/`.

[29] Andrew Shepherd et al. “Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation”. Em: *BMC psychiatry* 15 (2015), pp. 1–9.

[30] Douglas Steinley. “K-means clustering: a half-century synthesis”. Em: *British Journal of Mathematical and Statistical Psychology* 59.1 (2006), pp. 1–34.

[31] *Subreddit Desabafos*. Reddit. URL: `https://www.reddit.com/r/desabafos/`.

[32] Amir Hossein Yazdavar et al. “Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media”. Em: *ASONAM ’17: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (jul. de 2017), pp. 1191–1198.