UNIVERSIDADE FEDERAL DO CEARÁ

INSTITUTO UNIVERSIDADE VIRTUAL

CURSO DE GRADUAÇÃO EM SISTEMAS E MÍDIAS DIGITAIS

LUCAS FARIAS FERREIRA GOMES

IMPACTOS DA PANDEMIA DE COVID-19 EM ALUNOS DE GRADUAÇÃO: UM ESTUDO ATRAVÉS DA MODELAGEM DE TÓPICOS NEURAL EM DISCUSSÕES DO REDDIT

FORTALEZA

2022

LUCAS FARIAS FERREIRA GOMES

IMPACTOS DA PANDEMIA DE COVID-19 EM ALUNOS DE GRADUAÇÃO: UM ESTUDO ATRAVÉS DA MODELAGEM DE TÓPICOS NEURAL EM DISCUSSÕES DO REDDIT

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas e Mídias Digitais do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas e Mídias Digitais.

Orientador: Prof. Dr. José Gilvan Rodrigues Maia

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

G615i Gomes, Lucas Farias Ferreira.
Impactos da pandemia de COVID-19 em alunos de graduação : um estudo através da modelagem de tópicos neural em discussões do Reddit / Lucas Farias Ferreira Gomes. – 2022.
83 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Instituto UFC Virtual, Curso de Sistemas e Mídias Digitais, Fortaleza, 2022.
Orientação: Prof. Dr. José Gilvan Rodrigues Maia.

1. Impactos da COVID-19. 2. Alunos de graduação. 3. Modelagem de tópicos. 4. Clusterização hierárquica. I. Título.

CDD 302.23

LUCAS FARIAS FERREIRA GOMES

IMPACTOS DA PANDEMIA DE COVID-19 EM ALUNOS DE GRADUAÇÃO: UM ESTUDO ATRAVÉS DA MODELAGEM DE TÓPICOS NEURAL EM DISCUSSÕES DO REDDIT

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas e Mídias Digitais do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas e Mídias Digitais.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. José Gilvan Rodrigues Maia (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Henrique Sérgio Lima Pequeno
Universidade Federal do Ceará (UFC)

Prof. Me. José Wellington Franco da Silva
Universidade Federal do Ceará (UFC)

Prof. Dr. Paulo Antonio Leal Rego
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

À minha família, que sempre que possível priorizou minha educação, apoiando e investindo o quanto puderam ao longo dessa jornada. Sou grato pelas condições de estudo que me foram asseguradas durante a vida, incluindo escolas, cursos, livros, computadores, etc.

À Universidade Federal do Ceará (UFC), por proporcionar ensino gratuito e de qualidade para a formação de profissionais de diversas áreas. Agradeço a eficiência na tomada de decisões para redução de danos durante a pandemia. Pude reconhecer, ao finalizar este trabalho, que muitas das políticas adotadas pela universidade foram acertadas.

Aos professores e servidores do curso de Sistemas e Mídias Digitais, que sempre se empenharam para fornecer a melhor experiência educacional e integridade estrutural, mantendo uma alta qualidade mesmo durante esses anos de crises (econômica e de saúde).

Ao Prof. Dr. José Gilvan Rodrigues Maia, por ter atuado como orientador deste trabalho desde a fase de projeto, e por ter me incentivado a buscar soluções diversas para os problemas aqui abordados. Tal incentivo me ajudou a aprender muito sobre diferentes técnicas em um intervalo de tempo curto, causando um salto significativo de conhecimento durante o período em que este trabalho foi desenvolvido. Destaco essa qualidade de professor encorajador também na sala de aula, pois nesse período de pandemia ouviu sempre as necessidades dos alunos e adaptou sua metodologia de ensino sempre que possível.

Aos membros da banca examinadora – Prof. Dr. Henrique Sérgio Lima Pequeno, Prof. Me. José Wellington Franco da Silva e Prof. Dr. Paulo Antonio Leal Rego –, pela leitura atenciosa do trabalho e a contribuição significativa através de observações e sugestões de melhorias – algumas das quais me fizeram ficar mais atento a detalhes técnicos que normalmente poderiam passar despercebidos.

Ao Prof. Dr. Fernando Lincoln Carneiro Leão Mattos, por ter proporcionado meu primeiro contato com a Análise de Redes Sociais, o que aos poucos se traduziu em uma curiosidade pela mineração de textos, levando, por fim, ao estudo do Processamento de Linguagem Natural. O agradeço também por ter sido um dos primeiros professores a reconhecer em mim a capacidade de atuar algum dia como pesquisador, o que ajudou e tem ajudado a manter ativa minha auto-confiança e curiosidade para aprender mesmo em momentos difíceis.

“A informação é o único bem que pode ser dado
e conservado ao mesmo tempo.”

(Steven Pinker)

RESUMO

A pandemia de COVID-19 impactou as vidas de muitos alunos de graduação ao redor do mundo, causando mudanças significativas nas suas rotinas. A necessidade de medidas de combate à transmissão do vírus, como distanciamento social e migração para aulas online, gerou uma série de situações que afetaram os estudantes nos âmbitos sociais, educacionais e emocionais. Com o objetivo de ajudar o planejamento de políticas efetivas por instituições, buscou-se analisar esses impactos através de relatos de alunos de graduação, obtidos pela coleta de discussões publicadas no site de notícias sociais Reddit entre 1 de janeiro de 2020 a 31 de dezembro de 2021. Para essa análise, foi utilizado um modelo de tópico neural para aplicar *document embeddings* contextuais – gerados através de um modelo de linguagem pré-treinado – no agrupamento de postagens semanticamente semelhantes, resultando em *clusters* densos, que tiveram suas estruturas temáticas representadas na forma de palavras-chave. Para auxiliar a interpretação dos resultados, foram observados os documentos mais representativos de cada tópico e foi gerado um diagrama da hierarquia entre os *clusters*, ajudando a compreender a similaridade estrutural entre documentos de diferentes tópicos. Além disso, foram exploradas as postagens mais populares sobre os relatos observados, levando em conta diferentes períodos. Com a análise desses fatores e através dos tópicos de maior utilidade, obteve-se como resultado um estudo dos impactos da pandemia de COVID-19 em alunos de graduação, onde analisou-se os temas principais de cada grupo de relatos, descrevendo-os em termos compreensíveis.

Palavras-chave: Impactos da COVID-19. Alunos de graduação. Modelagem de tópicos. Clusterização hierárquica.

ABSTRACT

The COVID-19 pandemic has impacted the lives of many undergraduate students around the world, causing significant changes in their routines. The need for measures to combat the transmission of the virus, such as social distancing and migration to online classes, has caused a series of situations that affected students in the social, educational and emotional spheres. In order to help institutions plan effective policies, we sought to analyze these impacts through posts from undergraduate students, obtained by collecting discussions published on the social news site Reddit between January 1, 2020 and December 31, 2021. For this analysis, we used a neural topic model to apply contextual document embeddings – generated through a pre-trained language model – in the grouping of semantically similar posts, resulting in dense clusters, which had their thematic structures represented in the form of keywords. To help interpret the results, the most representative documents of each topic were observed and a diagram showing the hierarchy between clusters was generated, helping understand the structural similarity between documents across different topics. In addition, we explored the most popular posts about the observed themes, taking into account various periods. By analysing these factors and considering the most useful topics, the result was a study of the impacts of the COVID-19 pandemic on undergraduate students, where the main themes of each group of posts were observed and translated into understandable terms.

Keywords: Impacts of COVID-19. Undergraduate students. Topic modeling. Hierarchical clustering.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
PLN	Processamento de Linguagem Natural
AP	Aprendizagem Profunda
IA	Inteligência Artificial
FFNN	*Feedforward Neural Network*
RNN	*Recurrent Neural Network*
BERT	*Bidirectional Encoder Representations from Transformers*
NER	Reconhecimento de Entidades Nomeadas
POS	*part-of-speech*
BOW	*bag-of-words*
LDA	*Latent Dirichlet Allocation*
WE-IRBO	*Word Embedding-based Inverted Rank-Biased Overlap*
NPMI	*Normalized Pointwise Mutual Information*
KDD	*Knowledge Discovery in Databases*
API	*Application Programming Interface*
TF-IDF	*Term Frequency–Inverse Document Frequency*
JSON	*JavaScript Object Notation*
PMAW	*Pushshift Multithread* API *Wrapper*
PRAW	*Python Reddit* API *Wrapper*
c-TF-IDF	*class-based* TF-IDF
MMR	*Relevância Marginal Máxima*
UMAP	*Uniform Manifold Approximation and Projection*
HDBSCAN	*Hierarchical Density-Based Spatial Clustering of Applications with Noise*
OMS	Organização Mundial da Saúde
UTC	Tempo Universal Coordenado

SUMÁRIO

1 INTRODUÇÃO

Em 2020, com os acontecimentos da pandemia de COVID-19, as vidas de pessoas de várias partes do mundo foram impactadas, gerando mudanças significativas nas suas rotinas e hábitos. Dentre os grupos afetados, pode-se mencionar os estudantes de ensino superior, que frequentemente mostraram alterações negativas nos contextos educacionais, sociais e emocionais em diversos países (ARISTOVNIK *et al.*, 2020). A necessidade de medidas de combate à transmissão da doença refletiu ainda em mudanças na abordagem de ensino adotada pelas universidades, causando rapidamente a migração de aulas presenciais para meios online em muitos lugares do mundo (ALI, 2020).

Durante os primeiros meses de 2021, o ensino à distância ainda predominava em muitas universidades. Entretanto, com o avanço da vacinação e a redução do número de casos de coronavírus ao redor do mundo, muitas instituições voltaram a adotar o ensino presencial para o segundo semestre do ano, apesar das preocupações com a variante Delta (NADWORNY, 2021; NIETZEL, 2021). No final de 2021, porém, com a disseminação da variante Ômicron, algumas instituições anunciaram novamente a adoção de estratégias de ensino à distância para o início de 2022, tornando a situação incerta para muitos estudantes (HERNANDEZ, 2021).

No geral, a conjuntura da situação vivida durante a pandemia causou nos estudantes um aumento da sensação de isolamento, de ansiedade, de frustrações e de preocupações com o futuro profissional e acadêmico. Além disso, outros problemas enfrentados por parte dos universitários foram a dificuldade de adaptação ao ensino remoto e de estratégias de estudo individual, intensificados pela falta de motivação e a necessidade de maior autodisciplina. Porém, o apoio de professores e de setores de assistência estudantil em instituições de ensino superior tiveram um papel fundamental na redução dos impactos, mostrando que é possível criar estratégias de mitigação durante crises sanitárias, de modo que as necessidades dos estudantes sejam levadas em consideração (ARISTOVNIK *et al.*, 2020).

1.1 Problemática

Diante do exposto anteriormente, mostra-se útil a coleta e análise de uma base de dados contendo relatos de diferentes tipos de situações observadas por estudantes durante crises anteriores, podendo, possivelmente, ter aplicações no planejamento antecipado de políticas de apoio a estudantes durante eventuais crises no futuro. Para a coleta desses dados, é razoável

dirigir a atenção às redes sociais, pois é um setor que movimentou-se durante o ano de 2020 (SCHULTZ; PARIKH, 2020) e, dada a natureza desse tipo de site, é possível encontrar com facilidade relatos pessoais e discussões acerca de eventos globais.

De fato, nas redes sociais, a doença causada pelo novo coronavírus se tornou um dos assuntos mais comentados logo no início de 2020 (BRANDON, 2020) e causou, no decorrer do ano, um aumento de utilização além do esperado de sites do gênero (SCHULTZ; PARIKH, 2020). Uma dessas redes foi o Reddit, com um aumento de 44% no número de usuários diários em relação a 2019, destacando uma quantidade de mais 50 milhões de menções ao coronavírus na plataforma ao longo de 2020 (REDDIT, 2020). No ano seguinte, o site voltou a crescer, com um aumento de 19% no número de postagens e 12% no número de comentários (REDDIT, 2021b).

No Reddit, as pessoas realizam postagens anônimas em comunidades criadas por usuários acerca de interesses específicos, com regras pré-definidas e delimitação de assuntos. Dentre as comunidades universitárias, algumas das maiores são *r/College* e *r/GradSchool*, *subreddits* (como chamam-se os sub-fóruns) criados para a discussão de assuntos relacionados à vida estudantil, sendo o primeiro no contexto de graduação, e o segundo de mestrado e doutorado. Nessas comunidades é possível encontrar diversos relatos sobre diferentes aspectos de vivência, incluindo experiências com mudanças impostas pela pandemia, o que pode consistir em uma fonte de dados muito favorável, com destaque à espontaneidade de participação dos usuários nas discussões, sem a necessidade de interferência do pesquisador. Porém, devido ao grande volume de postagens e comentários existentes nesses meios, é difícil fazer uma análise adequada por métodos manuais, que podem ser extensos e cansativos.

Utilizando-se conceitos de Aprendizagem de Máquina (AM) e Processamento de Linguagem Natural (PLN), é possível lidar com os desafios supracitados de forma automatizada, reduzindo consideravelmente o tempo e o esforço necessários para analisar a vasta coleção de textos que pode ser encontrada em redes sociais. Além disso, o progresso crescente nesses campos torna acessíveis ferramentas de análise e processamento que diminuem significativamente as barreiras técnicas entre o pesquisador e o objeto de pesquisa, sendo aplicáveis em diversos âmbitos de estudo, como as humanidades, as engenharias e as ciências sociais.

De modo geral, o PLN engloba técnicas de computação para assimilar e produzir conteúdo em língua natural, visando promover a solução de tarefas diversas do mundo real (HIRSCHBERG; MANNING, 2015); enquanto a AM consiste em uma série de métodos que possibilitam detectar de modo automatizado a presença de padrões em um grupo de dados,

permitindo ainda o uso dessa experiência adquirida para realizar predições (MURPHY, 2012). Em conjunto, essas duas disciplinas viabilizam o desenvolvimento das mais variadas tecnologias, como tradução de textos, assistentes virtuais, modelagem de tópicos e até mesmo mineração de opinião (CAMBRIA; WHITE, 2014; HIRSCHBERG; MANNING, 2015).

Para o presente trabalho, a técnica citada mais relevante é a modelagem de tópicos, pois através dela é possível descobrir, a partir de uma coleção de textos de um período, uma série de tópicos – organizados em palavras-chave – que categorizam as estruturas temáticas dos documentos observados, permitindo produzir uma base de dados conveniente e acessível para a descoberta de conhecimento (BLEI *et al.*, 2003; BLEI, 2012).

Nesse sentido, essa monografia sugere contribuir para a pesquisa de relatos estudantis durante a pandemia a partir de dados coletados do Reddit. Tais dados passaram por um processo de triagem, para então serem transformados em representações vetoriais por intermédio de um modelo de linguagem baseado em Aprendizagem Profunda (AP). Através dessas representações, os relatos foram agrupados automaticamente considerando a similaridade semântica entre eles, de modo que documentos que descrevem uma mesma temática possam compartilhar uma mesma categoria. Dessa forma, foi possível extrair as palavras mais características de cada agrupamento, resultando em tópicos que puderam ser interpretados a partir dessas palavras-chave e dos relatos mais representativos de cada tema. Por fim, foram descritas as relações entre os tópicos e foram analisados os relatos mais populares referentes à pandemia de COVID-19.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral desta pesquisa é analisar, a partir da aplicação de técnicas de PLN e AM, os tópicos referentes a impactos da pandemia de COVID-19 no contexto de estudantes de graduação usuários de comunidades universitárias do Reddit.

1.2.2 Objetivos Específicos

- Identificar os tópicos referentes à pandemia;
- Analisar o nível de coerência dos tópicos;
- Identificar a relação entre tópicos; e
- Elaborar a visualização de *insights*.

2 FUNDAMENTAÇÃO TEÓRICA

Segundo Chomsky (2006), a linguagem é um atributo essencial para a humanidade, inseparável de qualquer fase crítica da existência da espécie, e que desempenha um papel fundamental no pensamento e na interação humana, atuando como uma ferramenta de livre expressão de ideias e de sentimentos. Dessa forma, linguagem expressa o que o autor chama de "a essência humana", a capacidade generativa da cognição da espécie, que permite a construção de palavras e frases novas tidas como possibilidades infinitas.

Assim, Chomsky (2002) afirma que não há limites de combinações possíveis para a composição de uma sequência gramatical, e que sempre é possível gerar uma nova frase. Porém, apesar desse aspecto criativo, Fromkin (2013) descreve que os falantes de uma língua em comum podem, através do conhecimento compartilhado da gramática, comunicar-se uns com os outros com o mínimo de distorção, sendo esse conhecimento admitido como um fator implícito, não sendo obrigatório o estudo de conceitos linguísticos para o aprendizado de uma língua.

Nesse sentido, os aspectos supracitados caracterizam alguns dos maiores desafios para o processamento de línguas naturais: a ambiguidade e a larga variabilidade, que na computação se traduzem em uma alta esparsidade de dados. Por conseguinte, é muito difícil descrever explicitamente as regras que compõem uma língua, já que as maneiras pelas quais as palavras podem ser combinadas para construir significados são praticamente infinitas, o que dificulta consideravelmente o planejamento de algoritmos aplicáveis nesse cenário, fazendo surgir a necessidade de compreender conceitos de Inteligência Artificial (IA) que permitem o aprendizado automático de estruturas e representações da linguagem: AM, AP e PLN (GOLDBERG, 2017).

No contexto de utilização da linguagem na Internet, pode-se destacar o imenso volume de informações geradas na forma de texto, potencializado em grande parte pelo uso das redes sociais, onde os usuários podem compartilhar opiniões sobre uma variedade de assuntos e atuar como produtores de conteúdo. Atualmente, na Internet, produz-se semanalmente mais dados do que havia desde os seus primórdios até 2003 – ano que precedeu a criação de um dos sites mais populares do mundo, o Facebook (CAMBRIA; WHITE, 2014).

Dessa forma, considerando o cenário de coleta e interpretação de dados em larga escala a partir de redes sociais, torna-se evidente a necessidade de uso de técnicas automatizadas para o processamento da informação textual, sendo uma das mais apropriadas para esse trabalho a modelagem de tópicos – conceito discutido na Seção 2.4.

Nesse sentido, para possibilitar o entendimento dessa monografia, são discutidos neste capítulo os principais fundamentos teóricos necessários para o desenvolvimento da pesquisa proposta: AM, AP, PLN e modelagem de tópicos. Em cada seção, os conceitos mais significativos são descritos e, quando relevante, são relacionados aos demais fundamentos abordados.

2.1 Aprendizagem de Máquina

A AM é uma vertente da IA que envolve um conjunto de métodos computacionais que utilizam informações, obtidas por meio de análises de dados, para adquirir experiência e realizar predições acerca de um problema. Os dados utilizados nesse processo podem estar na forma de grupos rotulados ou podem abranger elementos não-estruturados (MOHRI *et al.*, 2018). Quanto aos tipos de tarefas de AM, pode-se identificar diferentes categorias, especificadas na próxima subseção.

2.1.1 Tipos de Aprendizagem

A aprendizagem supervisionada envolve algoritmos que são treinados em uma série de exemplos previamente rotulados – com resultados definidos – e, a partir da experiência adquirida, geram predições para novos casos, estabelecendo rótulos conforme o que foi observado nas etapas de treinamento. É comum empregar esse tipo de tarefa em problemas de classificação e de regressão (MOHRI *et al.*, 2018).

Na aprendizagem não-supervisionada, o algoritmo recebe somente dados que não foram rotulados e, através do treinamento definido pelo modelo, aprende a reconhecer estruturas ocultas, ou seja, que não são diretamente observáveis – processo por vezes chamado de "descoberta de conhecimento" (MURPHY, 2012). No âmbito do PLN, é comum encontrar dados não-rotulados, sendo possível expandir a base de conhecimento através da aplicação de algoritmos de aprendizagem não-supervisionada (MANNING; SCHÜTZE, 1999).

Na aprendizagem semi-supervisionada, o algoritmo recebe tanto dados rotulados quanto não-rotulados e, após o treinamento, realiza predições para novos casos. Esse tipo de tarefa pode beneficiar circunstâncias onde dados rotulados requerem muito esforço para serem adquiridos, mas que haja uma ampla disponibilidade de dados não-rotulados. Além disso, o uso de dados não-rotulados pode ajudar a melhorar o desempenho de aplicações que normalmente fariam uso exclusivo de dados rotulados (MOHRI *et al.*, 2018; MURPHY, 2012).

A aprendizagem por reforço se refere a uma tarefa onde um agente de um sistema interage ativamente com o ambiente, sob a mediação de um plano de recompensas e penalidades que visam maximizar uma ação esperada, de acordo com as necessidades da aplicação. Esse tipo de tarefa tem sido empregado com sucesso em uma variedade de contextos, incluindo sistemas de telecomunicações, robótica e jogos digitais (MOHRI *et al.*, 2018).

Por fim, é importante ressaltar, para o presente trabalho, que a qualidade de algoritmos de AM levam em consideração, além dos critérios convencionais de complexidade de espaço e de tempo, a complexidade da amostra, fazendo-se necessário avaliar o tamanho mínimo desta para que um algoritmo possa aprender um conjunto de conceitos (MOHRI *et al.*, 2018). Portanto, na modelagem de tópicos a quantidade de palavras disponíveis no *corpus* influencia diretamente a qualidade dos resultados.

2.1.2 Clusterização de Dados

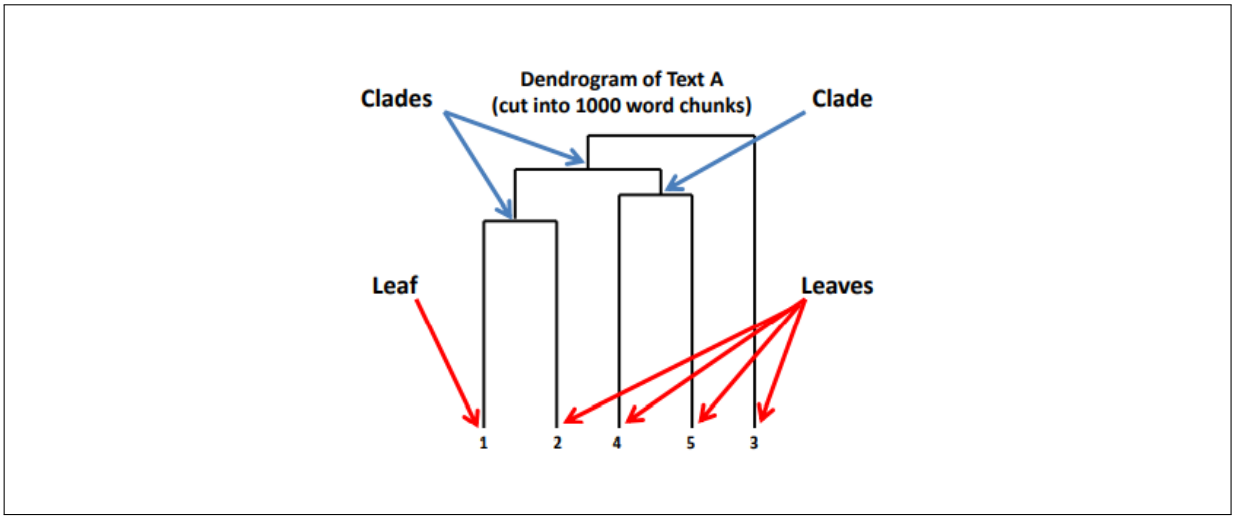
A clusterização é uma tarefa de AM não-supervisionada para o agrupamento de dados com base em um critério pré-definido, resultando em grupos denominados "*clusters*". Quanto ao formato de entrada do algoritmo, a clusterização pode ser baseada em similaridade ou baseada em *features*; já quanto ao formato de saída, ela pode ser classificada como clusterização particionada ou clusterização hierárquica (MURPHY, 2012).

Na clusterização baseada em similaridade, a entrada do algoritmo é uma matriz de dissimilaridade $N \times N$, que representa uma métrica de distância entre dois objetos em um espaço de d dimensões; enquanto na clusterização baseada em *features*, a entrada do algoritmo é uma matriz de *features* $N \times D$, onde N é a quantidade de elementos e D é a dimensionalidade de cada amostra. O agrupamento de dados com base na dissimilaridade tem como vantagem a possibilidade de aplicação de critérios de similaridade específicos do domínio que ajudam a descrever a relação entre objetos a partir dos seus *clusters*; enquanto a vantagem do agrupamento com base em *features* é a boa aplicabilidade para dados com muitos ruídos (IGUAL *et al.*, 2017).

Na clusterização particionada, o algoritmo opera para agrupar os dados em conjuntos desarticulados; enquanto na clusterização hierárquica, o algoritmo agrupa os dados a partir de uma árvore hierárquica. Esses dois tipos de clusterização distinguem-se também pelo modo que é determinado o número K de *clusters*, sendo que na clusterização particionada o valor de K deve ser previamente definido pelo pesquisador; enquanto na clusterização hierárquica esse valor é obtido automaticamente, podendo facilitar consideravelmente o processo (MURPHY, 2012).

A árvore resultante de uma clusterização hierárquica é chamada “dendrograma” (Figura 1), diagrama que pode ser usado como ferramenta de interpretação de dados, já que ele representa as relações de similaridade entre os elementos da amostra considerada (DROUT; SMITH, 2012; IGUAL *et al.*, 2017).

Figura 1 – Exemplo ilustrativo de dendrograma e seus elementos



Fonte: Drout e Smith (2012).

A Figura 1 apresenta um dendrograma ilustrativo e cada um dos elementos o compõe. A parte superior é a chamada “raiz” do dendrograma, que corresponde a um *cluster* único que contém todos os objetos da amostra. Na parte mais inferior, localizam-se as folhas (indicadas pelas setas vermelhas), que são *clusters* de uma unidade formados por cada objeto da amostra. No centro, pode-se observar os “clados” (indicados pelas setas azuis), estruturas cuja organização designam a relação de similaridade entre os objetos: as conexões descrevem a ligação entre eles, de modo que dois objetos que são diretamente conectados – como 1 e 2 – possuem maior relação; enquanto a altura descreve o grau de similaridade, sendo que quanto mais baixo for um clado mais similares são os objetos conectados por ele (DROUT; SMITH, 2012).

2.1.2.1 Métricas de Distância

Como apontado anteriormente, a matriz de dissimilaridade de uma clusterização representa uma métrica de distância entre dois objetos, e por isso é importante compreender as distâncias que podem influenciar o resultado da formação de *clusters*. Para o presente trabalho, as métricas mais relevantes são aquelas que apontam a similaridade entre dois vetores numéricos $A = [a_1, a_2, ..., a_n]$ e $B = [b_1, b_2, ..., b_n]$, descritas por Manning e Schütze (1999):

$$\textit{Produto Escalar} = a_1b_1 + a_2b_2 + ... + a_nb_n = |a||b|\cos\theta. \tag{2.1}$$

$$\textit{Similaridade do Cosseno} = \frac{a \cdot b}{|a||b|}. \tag{2.2}$$

$$\textit{Distância Euclidiana} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2}. \tag{2.3}$$

As equações acima não são proporcionais e, portanto, produzem ranqueamentos de similaridade diferentes. Além disso, a magnitude dos vetores é considerada, o que pode enviesar a clusterização de documentos com muitas palavras (GOOGLE DEVELOPERS, 2020). Porém, esses problemas podem ser facilmente enfrentados, pois uma propriedade importante da similaridade entre vetores é que, quando normalizados, o produto escalar e a distância euclidiana passam a ser proporcionais ao cosseno (MANNING; SCHÜTZE, 1999), conforme a seguir:

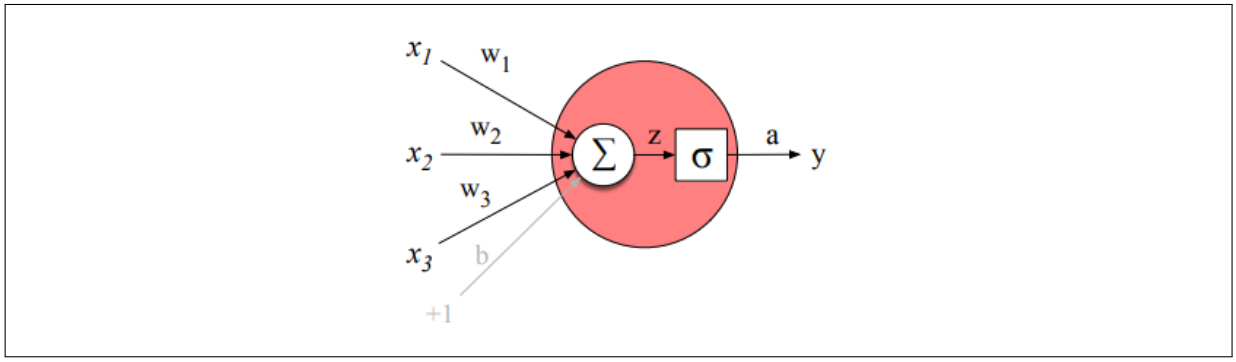
$$\textit{Produto Escalar} = \textit{Similaridade do Cosseno} = \cos\theta. \tag{2.4}$$

$$\textit{Distância Euclidiana} = \sqrt{2 - 2\cos\theta}. \tag{2.5}$$

2.2 Aprendizagem Profunda

Nos últimos anos, as redes neurais têm se tornado ferramentas cada vez mais fundamentais para aplicações de PLN. Essas redes são compostas por um conjunto de unidades computacionais – chamadas "nós" ou "neurônios" – que agem em paralelo, recebendo um vetor de valores e produzindo como saída um único valor (Figura 2) (JURAFSKY; MARTIN, 2022). Esses sistemas permitem que aplicações de AM adquiram experiência através de uma hierarquia de conceitos, captando aqueles mais simples para a definição de outros mais complexos. Assim, após uma série de treinamentos, uma aplicação desse tipo pode se tornar apta para executar tarefas que envolvem padrões cujas regras são difíceis de serem descritas por um humano, como reconhecimento de objetos e compreensão textual. Esse processo de aprendizagem ocorre ao longo de várias camadas (ou *layers*) e, por isso, o uso desse tipo sistema é chamado de AP (MURPHY, 2012; GOODFELLOW *et al.*, 2016).

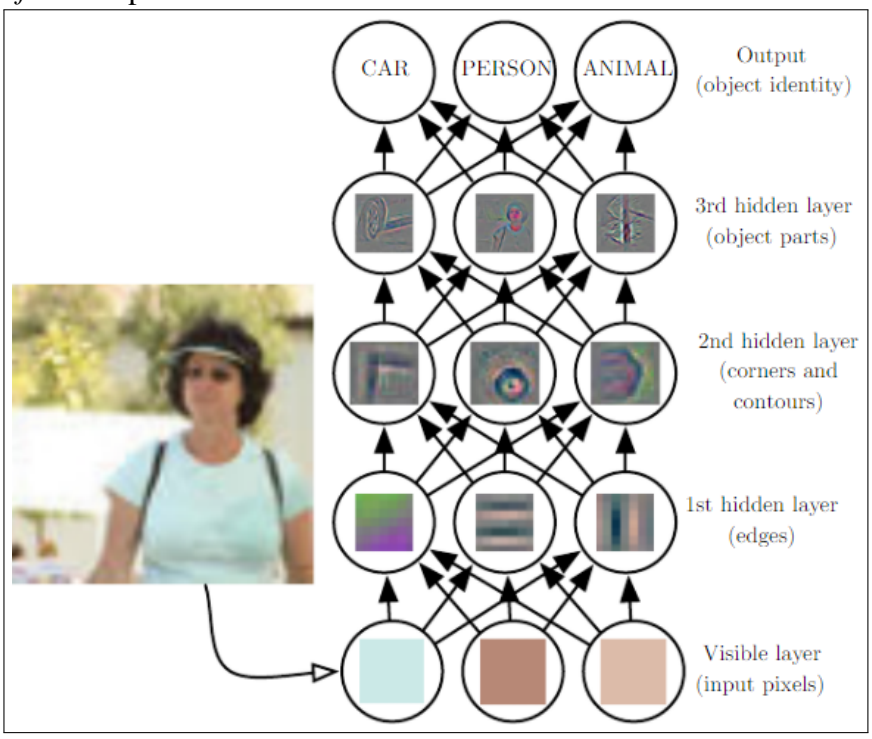
Figura 2 – Esquema de neurônio recebendo um vetor de entrada e um *bias*, e retornando um único valor de saída



Fonte: Jurafsky e Martin (2022).

A Figura 2 ilustra um neurônio recebendo três valores x_1, x_2 e x_3, e um *bias* b, que funciona como um peso para a soma ponderada dos valores de entrada, resultando em z. Esse resultado passa, então, por uma função não-linear – denominada "função de ativação"–, produzindo o valor de saída a do neurônio (JURAFSKY; MARTIN, 2022).

Figura 3 – Esquema de modelo de AP mostrando a extração de *features* para cada camada



Fonte: Goodfellow *et al.* (2016).

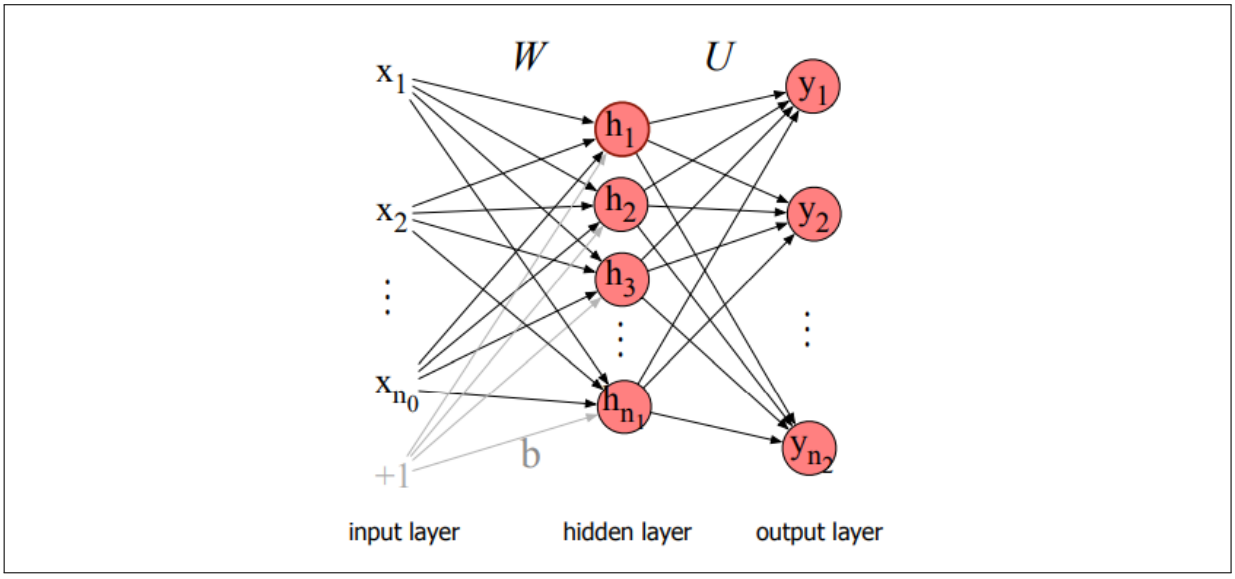
Em comparação com métodos de AM tradicionais, aplicações que fazem uso de AP tendem a ser mais eficientes, pois são capazes de aprender a representação dos dados automaticamente e expressá-la através de outras representações mais simples (Figura 3), sem exigir que o pesquisador investigue manualmente os atributos (ou *features*) significativos para a

resolução do problema. Assim, problemas complexos que normalmente exigiriam intervenção humana contínua para cada mudança, podem ser resolvidos de forma mais rápida e flexível com o uso da AP (GOODFELLOW *et al.*, 2016).

O processo que permite esses sistemas a assimilar *features* de modo automático é chamado de "aprendizado de representação", e um dos algoritmos mais básicos para sua execução é o *autoencoder*, que consiste em uma rede neural de aprendizagem não-supervisionada que é treinada para predizer os próprios dados de entrada. Esse tipo de algoritmo é formado por um *encoder*, que transforma os valores de entrada em uma nova representação, e um *decoder*, que transforma a representação obtida de volta ao estado original (GOODFELLOW *et al.*, 2016). *Autoencoders* têm sido empregados em diversas finalidades, como redução de dimensionalidade e cálculo de similaridade de sentenças (MURPHY, 2012; GOLDBERG, 2017).

Em sua forma mais simples, uma rede neural moderna é classificada como uma *Feedforward Neural Network* (FFNN) (Figura 4), que consiste em uma rede onde os neurônios de cada camada se conectam em um único sentido, passando os valores de saída para as camadas subsequentes até a camada final. Esse tipo de rede é composto por três espécies de camadas: camada de entrada, camada oculta e camada de saída. As camadas de entrada e de saída correspondem, respectivamente, aos neurônios que recebem os valores iniciais e os neurônios que determinam o resultado; enquanto a camada oculta recebe uma soma ponderada de todos os valores da camada anterior e aplica uma não linearidade, sem saber qual resultado é esperado (GOODFELLOW *et al.*, 2016; JURAFSKY; MARTIN, 2022).

Figura 4 – Esquema de FFNN com duas camadas

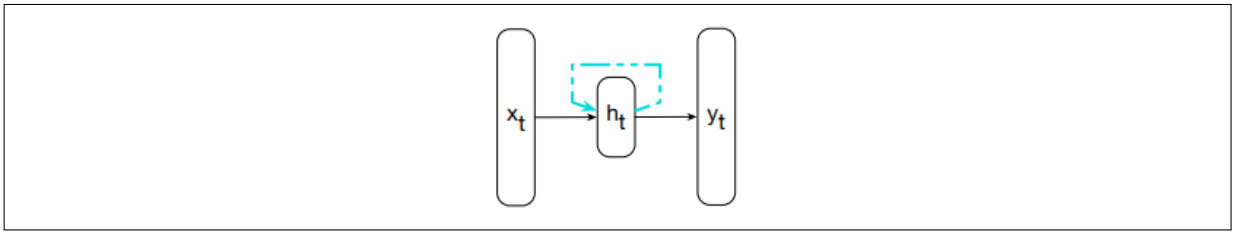


Fonte: Jurafsky e Martin (2022).

A Figura 4 ilustra uma FFNN com duas camadas[1]: uma camada de entrada, uma camada oculta e uma camada de saída. A camada de entrada recebe os valores $x_1, x_2, \ldots, x_{n_0}$, e os direcionam para a camada oculta h, onde os neurônios são transformados por uma matriz W_{ij}, que representa os parâmetros de uma soma ponderada considerando cada valor x_i, h_j e um valor único de *bias* b. O resultado dessa transformação é um vetor que caracteriza uma representação dos valores de entrada iniciais, sendo por fim passado para a camada de saída y que realiza operações de acordo com a finalidade da rede neural (JURAFSKY; MARTIN, 2022).

Quando uma FFNN é estendida para conter conexões cíclicas entre camadas, ela passa a ser chamada de *Recurrent Neural Network* (RNN), o que significa que os valores que um neurônio recebe dependem, direta ou indiretamente, dos seus próprios valores de saída anteriores (GOODFELLOW *et al.*, 2016).

Figura 5 – Esquema da *Elman Network*, um tipo de RNN



Fonte: Jurafsky e Martin (2022).

A Figura 5 ilustra uma RNN simples, conhecida como *Elman Network*, onde um vetor passa pela camada de entrada x_t e é transformado, passando para a camada oculta h_t, que calcula um resultado para a camada de saída y_t. A diferença em relação às FFNNs está na conexão recorrente, que transforma a entrada da camada oculta a partir do seu valor anterior no tempo t (JURAFSKY; MARTIN, 2022).

Tanto FFNNs quanto RNNs possuem aplicabilidade em sistemas de PLN através dos modelos de linguagem, sistemas projetados para predizer uma palavra em uma sentença a partir de um contexto, que pode ser descoberto através de representações dos significados das palavras por vetores distribuídos em um espaço semântico (MIKOLOV *et al.*, 2013a; JURAFSKY; MARTIN, 2022) – conceito melhor descrito na subseção 2.3.1.

Como a linguagem é um fenômeno que discorre com o passar do tempo através de uma sequência de palavras, FFNNs podem ser vistas como limitadas; enquanto RNNs fornecem uma melhoria no modo de representar um contexto, pois permitem que um modelo de linguagem utilize valores anteriores como referência. Entretanto, apesar da melhoria apresentada pelas

[1] Normalmente, a camada de entrada não é considerada para a contagem.

RNNs, modelos de linguagem baseados nessa rede ainda possuem limitações, como: perda de informação devido às extensas séries de transformações pelas conexões recorrentes; e baixo desempenho computacional devido à sua natureza sequencial, que dificulta cálculos em paralelo (JURAFSKY; MARTIN, 2022).

Dessa forma, modelos de linguagem mais atuais têm se baseado cada vez mais em *Transformers* (VASWANI *et al.*, 2017), redes de camadas múltiplas que são compostas por uma combinação de camadas lineares, FFNNs e camadas de auto-atenção, sendo esta última o maior diferencial dessas redes. A camada de auto-atenção fornece um método de extrair informações de contextos muito grandes, de modo que uma palavra possa ser comparada com outras para revelar sua importância naquele contexto. Além disso, essas camadas não necessitam de conexões recorrentes, o que permite a paralelização, diminuindo assim o esforço computacional em comparação às RNNs (JURAFSKY; MARTIN, 2022).

Um exemplo comum de modelo de linguagem que faz uso de *Transformers* é o *Bidirectional Encoder Representations from Transformers* (BERT) (DEVLIN *et al.*, 2019), que pode ser eficaz para a realização de diversas tarefas sensíveis ao contexto, pois, ao invés de descrever uma palavra através de uma representação vetorial estática (ou seja, um vetor para cada palavra), ele pode aprender representações vetoriais dinâmicas (onde cada palavra pode ser descrita por diferentes vetores dependendo do contexto) (JURAFSKY; MARTIN, 2022).

Recentemente, tem havido um avanço crescente no campo de PLN, fazendo surgir diversos modelos de linguagem baseados em *Transformers*, que são pré-treinados para diferentes casos de uso, sendo comumente de fácil aplicação. Um exemplo desse avanço é o MiniLM (WANG *et al.*, 2020), que utiliza um método de compressão de *Transformer* que reduz consideravelmente seu tamanho e custo de processamento, tornado-o um bom candidato para desenvolver aplicações em sistemas convencionais (como o utilizado nesse trabalho).

2.3 Processamento de Linguagem Natural

O PLN é uma vertente da IA que engloba um conjunto de técnicas computacionais, motivadas ou não por conceitos linguísticos, para a análise e representação de textos em língua natural, com o objetivo de resolver tarefas do mundo real. Dessa forma, o PLN abrange conceitos e implementações que possibilitam uma grande quantidade de aplicações em conjuntos de textos em qualquer língua, sendo comum o uso de sistemas baseados em AM e, recentemente, com avanços que estendem-se cada vez mais a sistemas baseados em AP (RAO; MCMAHAN, 2019).

Para projetar e implementar um sistema de PLN, é necessário o entendimento de um conjunto de conceitos básicos que permitem um tratamento adequado do texto. O primeiro e mais fundamental é o *token*, uma sequência de caracteres a ser tratada como uma unidade contígua. O processo de criação de *tokens* é chamado de "tokenização" (ou segmentação de palavras) e é uma tarefa essencial em PLN (BIRD *et al.*, 2009). Comumente, na língua inglesa, esse processo equivale a separar palavras e sequências numéricas de acordo com as pontuações e os espaços em branco que as envolve, porém nem sempre isso é suficiente, como nos casos onde ocorram expressões multipalavras. Nesses cenários, pode-se adotar a tokenização com base em n-gramas, que resulta em *tokens* com mais de uma palavra, ou ainda, pode-se fazer uso do Reconhecimento de Entidades Nomeadas (NER) – processo que consiste em detectar e classificar qualquer palavra que possa ser referida por um nome próprio, como locais, pessoas, organizações e entidades geo-políticas (JURAFSKY; MARTIN, 2009).

Outro conceito relevante é a lematização, que se refere à redução de uma palavra à sua forma base, também conhecida como lema. Essa técnica pode ser útil para diminuir o tamanho de um vocabulário, unificando palavras que possuem uma mesma forma que encabeça a representação no dicionário. Por exemplo, as palavras "cantam", "cantando" e "cantei", ao serem lematizadas, se reduzem à forma "cantar", já que são flexões deste mesmo verbo. Esse recurso, porém, depende da morfologia de cada palavra e por isso requer a aplicação de outro processo, o *part-of-speech* (POS) *tagging*, que consiste em classificar e rotular cada palavra de acordo com sua classe gramatical (BIRD *et al.*, 2009; JURAFSKY; MARTIN, 2009).

Por fim, os métodos supracitados fazem parte de uma etapa comum em sistemas de PLN, chamada pré-processamento textual, que consiste em um conjunto de tarefas que buscam gerar um formato padrão bem definido de unidades linguisticamente significativas, facilitando a recuperação e a extração da informação. Esse processo pode ser dividido em duas fases: triagem do documento e segmentação do texto; sendo a primeira referente à limpeza do *corpus* e a segunda aos procedimentos de tokenização e normalização (PALMER, 2010).

Após o pré-processamento, para que as palavras de um *corpus* possam enfim ser efetivamente processadas, é necessário representá-las através de formas numéricas, procedimento comumente referido como "representação vetorial para a linguagem" (VAJJALA *et al.*, 2020). Tal procedimento pode ser de diversos tipos, sendo aqueles mais relevantes para o presente trabalho o modelo de *bag-of-words* (BOW) e o modelo de *embedding vectors*, que serão brevemente discutidos na próxima subseção.

2.3.1 Representações Vetoriais para a Linguagem

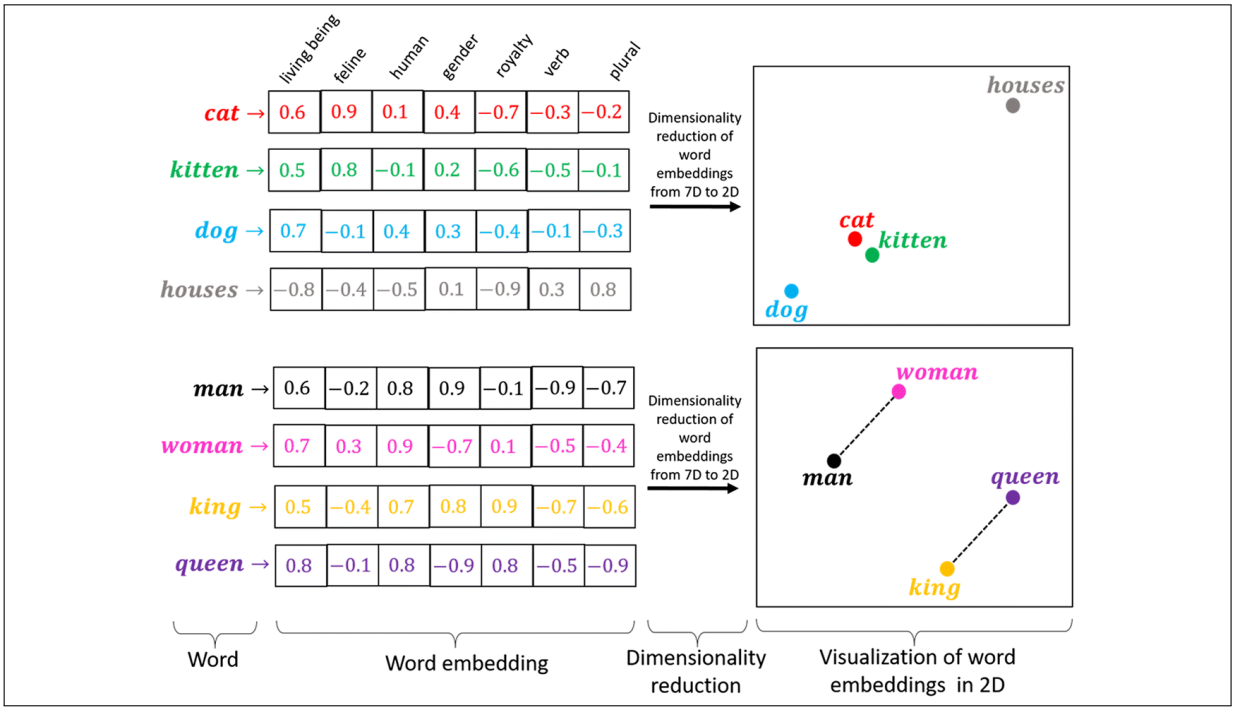
2.3.1.1 Bag-of-Words

Um dos modelos mais básicos para representação da linguagem é o BOW, que corresponde a uma contagem simples das palavras de cada documento. Esse modelo não considera a ordem em que as palavras aparecem no texto nem a relação conceitual que possa existir entre elas. O resultado da aplicação dessa técnica é uma matriz de documentos e termos, na qual cada palavra possui uma frequência entre 0 e D, onde D é a quantidade de dimensões do vetor resultante (JURAFSKY; MARTIN, 2009; VAJJALA *et al.*, 2020).

2.3.1.2 Embedding Vectors

Como antecipado na Seção 2.2, os *embedding vectors* (ou simplesmente *embeddings*) são representações vetoriais distribuídas em um espaço semântico multidimensional, onde cada dimensão compreende um aspecto conceitual da palavra – que pode ser qualquer conceito abstrato, como "realeza", "humanidade", etc. Nesse espaço, as operações vetoriais permitem determinar as relações de significado entre as palavras, de modo que vetores com direção parecida são considerados similares (MIKOLOV *et al.*, 2013a; MIKOLOV *et al.*, 2013b).

Figura 6 – Exemplo ilustrativo de representação de *embeddings*



Fonte: Rozado (2019).

A Figura 6 mostra um exemplo ilustrativo, no qual as palavras são representadas por *embeddings* de 7 dimensões, onde cada uma dessas dimensões compreendem um dos aspectos conceituais: “ser vivo”, “felino”, “humano”, “gênero”, “realeza”, “verbo” ou “plural”. Nesses *embeddings*, quanto maior o valor que compõe o vetor na dimensão d mais forte é o aspecto correspondente para a palavra considerada. Assim, na dimensão de “felino”, a palavra “gato” possui um valor maior que o da palavra “casas”; e na dimensão de “realeza”, a palavra “rainha” apresenta um valor maior que da palavra “mulher”.

Entretanto, as representações supracitadas são estáticas e, como descrito na Seção 2.2, modelos de linguagem baseados em *Transformers* têm a capacidade de gerar *embeddings* dinâmicos, que representam uma mesma palavra por diferentes vetores dependendo do contexto. Assim, palavras como “rei” apresentam diferentes *embeddings* em expressões como “rei da França” e “rei da selva” (JURAFSKY; MARTIN, 2022). Essa propriedade é muito relevante para essa monografia, já que o contexto é um aspecto que pode favorecer a identificação de temáticas.

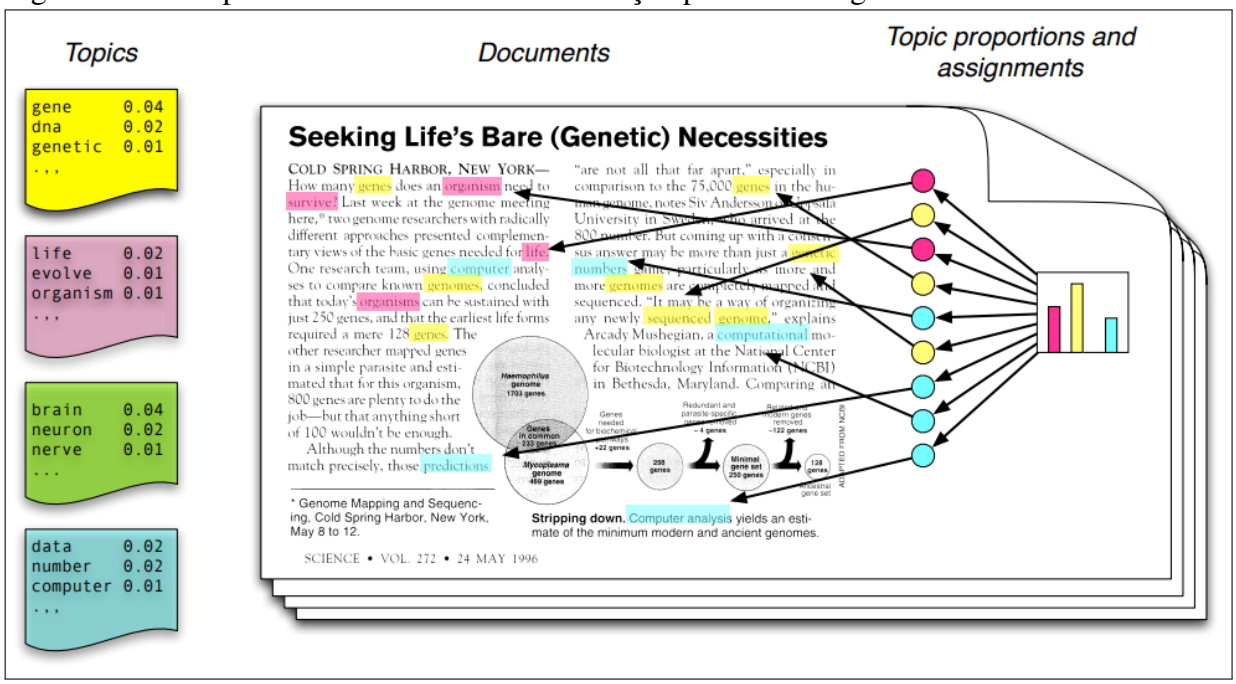
2.4 Modelagem de Tópicos

A modelagem de tópicos engloba um conjunto de algoritmos baseados em modelos estatísticos, que processam documentos com a finalidade de identificar suas estruturas temáticas em tópicos, podendo envolver ainda uma análise de como estes se relacionam uns com os outros e como se alteram em cada intervalo de tempo (BLEI, 2012).

Esses algoritmos partem da perspectiva de que um documento pode ser compreendido como uma distribuição de tópicos, enquanto estes representam uma distribuição de palavras. Mais especificamente, os tópicos consistem em estruturas latentes dentro de um documento que podem ser reveladas por modelos de AM e, assim, permitem descrever através de um conjunto de palavras as temáticas tratadas no documento analisado (BLEI *et al.*, 2003; BLEI, 2012).

Existem diversas técnicas computacionais apropriadas para a modelagem de tópicos, sendo uma das mais conhecidas a *Latent Dirichlet Allocation* (LDA), um algoritmo de AM não-supervisionada que utiliza como base um modelo probabilístico generativo para captar a permutabilidade de palavras e documentos, sem considerar a ordem que os documentos são apresentados nem das palavras que aparecem em cada documento, baseando-se, então, na representação de BOW. Assim, a LDA parte de um número pré-definido de tópicos e atribui – através de tratamentos estatísticos – cada palavra a um ou mais tópicos, e cada tópico a um ou mais documentos (BLEI *et al.*, 2003; MURPHY, 2012).

Figura 7 – Exemplo ilustrativo mostrando a intuição por trás do algoritmo de LDA



Fonte: Blei (2012).

Apesar de ser uma técnica muito utilizada, a LDA pode ser insuficiente quando a pesquisa possuir como finalidade uma análise mais detalhada das relações existentes na ocorrência de tópicos, já que ela parte do pressuposto que a ordem dos documentos e das palavras não importam, e por não observar relações além da pura modelagem de tópicos (BLEI, 2012).

Diante dessas limitações, pode-se levar em consideração algoritmos que estendem a LDA, diminuindo suas pressuposições e aumentando as variáveis de análise (BLEI, 2012). Alguns exemplos desses modelos são o *Correlated Topic Model* (BLEI; LAFFERTY, 2005), que observa a correlação entre os tópicos, e o *Dynamic Topic Model* (BLEI; LAFFERTY, 2006), que leva em consideração a ordem dos documentos e analisa as mudanças que ocorrem em cada tópico no decorrer do tempo.

Existem ainda técnicas mais complexas que vão além das modelagens exclusivamente probabilísticas citadas acima e fazem uso de *embeddings*, obtidos através de modelos de linguagem baseados em *Transformers*, para agrupar documentos com base na similaridade semântica existente entre eles, adicionando como variável o contexto que cada palavra se insere no documento. Exemplos de aplicações desse tipo são o Top2Vec (ANGELOV, 2020), o BERTopic (GROOTENDORST, 2020) e o *Combined Topic Model* (BIANCHI *et al.*, 2021). Essas técnicas neurais podem aumentar a interpretabilidade dos tópicos, já que suas palavras-chave passam a conter informações contextuais latentes – que são inexistentes na abordagem com BOW.

Para a análise desses tópicos contextuais, podem ser usadas métricas que levam em

consideração as relações conceituais entre as palavras, como a *Word Embedding-based Inverted Rank-Biased Overlap* (WE-IRBO) (BIANCHI *et al.*, 2021; TERRAGNI *et al.*, 2021b), que consiste em um método de avaliação para a diversidade dos tópicos. Tal métrica é calculada através da dissimilaridade entre os *embeddings* das palavras-chave de diferentes tópicos, o que significa que quanto mais diferentes os tópicos forem entre si maior será a pontuação WE-IRBO.

Por fim, existe ainda a *Normalized Pointwise Mutual Information* (NPMI) (LAU *et al.*, 2014), uma métrica para a avaliação da coerência de tópicos, que calcula a relação entre as palavras-chave de um tópico a partir da coocorrência observada. A otimização dos hiperparâmetros de modelos de tópicos neurais a partir dessa métrica pode prover uma boa performance, enquanto a otimização considerando somente a diversidade pode resultar em uma baixa coerência de tópicos (TERRAGNI; FERSINI, 2021). Porém, dependendo do caso observado, pode ser útil levar em conta a diversidade, pois ela permite preservar um nível adequado de variabilidade entre tópicos (BIANCHI *et al.*, 2021).

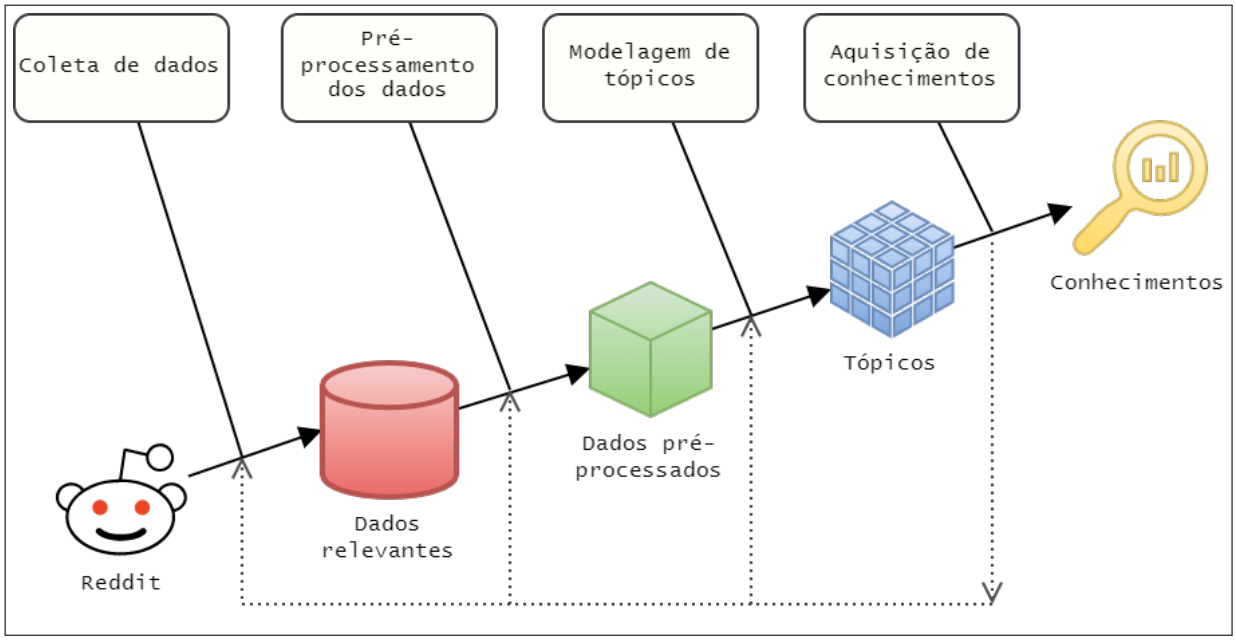
3 METODOLOGIA

No presente estudo, buscou-se compreender, através de relatos no site de notícias sociais Reddit, as particularidades enfrentadas por estudantes de graduação durante a crise sanitária global de COVID-19, levando-se em consideração os dois primeiros anos da pandemia. Dessa forma, foi realizada uma pesquisa de caráter exploratório, que segundo Gil (2002), tem como objetivo "proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a constituir hipóteses".

Quanto aos procedimentos técnicos, a pesquisa caracteriza-se como documental, pois "vale-se de materiais que não receberam ainda um tratamento analítico, ou que ainda podem ser reelaborados de acordo com os objetos da pesquisa" (GIL, 2002). Por fim, a unidade de análise consistiu nas postagens realizadas na comunidade universitária *r/College* desde o início do ano de 2020 até o fim de 2021. Portanto, os dados são inicialmente de caráter qualitativo, pois as postagens coletadas apresentam forma textual, até serem submetidas a técnicas de vetorização, como descrito na Seção 4.3, quando convertem-se em dados quantitativos, o que permite a aplicação de análises estatísticas e o processamento algorítmico do conteúdo.

3.1 Metodologia da Pesquisa

Figura 8 – Visão geral dos passos da metodologia adotada na pesquisa



Fonte: adaptada de Fayyad, Piatetsky-Shapiro e Smyth (1996).

Os passos da metodologia adotada neste estudo, conforme ilustrado na Figura 8, compreendem quatro etapas consecutivas, respectivamente: coleta de dados, pré-processamento dos dados, modelagem de tópicos e aquisição de conhecimentos. Esses passos foram adaptados da metodologia de *Knowledge Discovery in Databases* (KDD), projetada para a exploração em bases de dados de larga escala e a descoberta de padrões úteis e compreensíveis dentro deles (FAYYAD *et al.*, 1996). Assim como na KDD, todo o processo é iterativo, sendo necessárias diferentes repetições até alcançar-se o resultado final e desenvolver a descoberta de conhecimento.

A descrição de cada passo da metodologia pode ser sintetizada conforme a seguir:

1. Coleta de dados: refere-se à seleção e a obtenção do conjunto de dados através do qual o processo de descoberta será realizado;
2. Pré-processamento dos dados: corresponde à filtragem de ruídos, o tratamento de valores ausentes e a normalização da informação a ser modelada posteriormente;
3. Modelagem de tópicos: estágio onde ocorre a transformação dos dados pré-processados e a aplicação de métodos de identificação de padrões para a formação de tópicos;
4. Aquisição de conhecimentos: etapa de interpretação e visualização dos padrões extraídos, removendo informações redundantes para a pesquisa e descrevendo em termos compreensíveis aquelas consideradas úteis.

3.1.1 Coleta de Dados

A primeira etapa consistiu na seleção e coleta de dados textuais e de metadados, sendo necessário o uso da *Application Programming Interface* (API) abordada na Subseção 3.2.1. O domínio de busca foi o site de notícias sociais Reddit, sendo essa escolha determinada por tratar-se de uma rede social em constante crescimento e que favorece a criação de ambientes de discussão acerca de uma variedade de temas, de forma acessível e inclusiva para qualquer membro do site. Tais discussões ocorrem normalmente em *subreddits* públicos, criados por usuários, onde estabelecem-se regras que delimitam as possibilidades de assuntos de acordo com seu propósito, viabilizando a existência de comunidades que abrangem contextos específicos, sem a interferência de temáticas externas.

Além disso, o Reddit possui políticas que permitem o acesso público ao conteúdo de suas comunidades não-privadas e dispõe de APIs abertas (oficiais e não-oficiais), menos restritivas que de outras redes sociais – como Facebook e Twitter – e que se mostram de grande utilidade para pesquisadores (BAUMGARTNER *et al.*, 2020; REDDIT, 2021a).

Quanto ao *subreddit* cujas discussões foram coletadas, selecionou-se o *r/College*, pois é uma comunidade global ativa – com mais de 600.000 membros – voltada exclusivamente a estudantes de graduação, e que dispõe de regras específicas que restringem seu conteúdo somente a assuntos relacionados à vida universitária, proibindo memes e postagens muito particulares de uma instituição. Assim, esperou-se reduzir consideravelmente a quantidade de ruídos presentes nos dados, o que pode proporcionar um aumento da qualidade dos resultados.

Finalmente, para realizar-se a coleta, foi utilizada uma plataforma especializada no armazenamento e na disponibilização de dados do Reddit, como apresentada na Subseção 3.2.1. O modo de interação com a plataforma foi através de uma API que gera requisições à sua base de dados e retorna o conteúdo solicitado em um padrão interoperável, o que permite o tratamento acessível do texto e dos metadados, além de facilitar a conversão para outros formatos de arquivo.

3.1.2 Pré-Processamento dos Dados

Na segunda etapa, foi efetuado o pré-processamento dos dados obtidos, o que foi necessário devido à natureza não estruturada dos textos, que continham ruídos de variados tipos. Foi realizada, então, uma limpeza textual minuciosa, com o objetivo de manter o conteúdo das postagens em um formato legível em língua natural, para aumentar a eficiência dos procedimentos aplicados na etapa posterior, que requerem documentos inteligíveis para a geração de *document embeddings* que refletem apropriadamente o sentido semântico dos textos.

3.1.3 Modelagem de Tópicos

A terceira etapa consistiu em uma sequência de técnicas relativas à modelagem de tópicos aprimorada por similaridade semântica. Para sua realização, utilizou-se uma série de bibliotecas, reunidas em um pacote principal (discutido na Subseção 3.2.2), que proporciona a geração de *document embeddings*, a execução da redução de dimensionalidade e da clusterização, e a aplicação de um algoritmo de *Term Frequency–Inverse Document Frequency* (TF-IDF) baseado em classe – resultando na formação de palavras-chave equivalentes a tópicos.

Dessa forma, foi realizada de início uma testagem com valores padrões no conjunto de dados pré-processados, para então registrar-se os menores *clusters* correspondentes a relatos referentes à pandemia. Em seguida, foi feito o ajuste semi-automático dos hiperparâmetros, tomando como princípios a presença dos *clusters* observados anteriormente, além das métricas de coerência e diversidade de tópicos apresentadas na Seção 2.4.

Posteriormente, foi feito o treinamento do modelo em todo o conjunto de dados pré-processados, para então atribuir-se tópicos individualmente aos seus dois subconjuntos – correspondentes à coleção de discussões de cada ano. Ao final desse processo, obtém-se como resultado tópicos intercambiáveis entre os subconjuntos de dados, o que permite a análise comparativa dos tópicos para cada período, preservando-se as possíveis relações e hierarquias.

3.1.4 Aquisição de Conhecimentos

Finalmente, na quarta etapa, foi feita a interpretação e visualização dos tópicos relevantes para a pesquisa. Durante esse processo, foram gerados gráficos de barras para os tópicos, um dendrograma com *clusters* de tópicos, e gráficos temporais denotando a popularidade dos tópicos em cada mês. Dessa forma, foram selecionados os tópicos a serem analisados e, a partir das visualizações geradas, das palavras-chave de cada tópico e de seus documentos representativos, foi feita a interpretação dos resultados e foram elaboradas descrições em termos compreensíveis dos conhecimentos adquiridos no processo.

3.2 Aspectos de Implementação

Inicialmente, em uma fase de testes, experimentou-se utilizar a plataforma gratuita Google Colaboratory[1] para a execução das etapas, porém problemas na disponibilidade de RAM e de GPU motivaram a adoção de um sistema local, o que ocasionou um aumento significativo de performance sobre a versão então disponível da plataforma citada. Entretanto, como consequência, etapas que demandam um maior esforço computacional impossibilitaram o uso do sistema até a conclusão de suas tarefas, devido ao alto uso de recursos, gerando um custo de tempo e a indisponibilidade temporária de equipamento ao longo de vários momentos. O dispositivo utilizado foi, então, um laptop equipado conforme mostrado na Tabela 1.

Tabela 1 – Especificações do dispositivo utilizado na pesquisa

Tipo do Componente	Especificação do Componente
Sistema Operacional	Windows 10 64 bits
Processador	Intel Core i7-9750H 2,60 GHz
Placa Gráfica	Nvidia GeForce GTX 1660 Ti Max-Q 6 GB (CUDA 11.4)
Armazenamento	SSD IM2P33F3 NVMe ADATA 512 GB
Memória RAM	DDR4 2.666 MHz 8 GB

Fonte: elaborado pelo autor (2022).

[1] Disponível em <https://colab.research.google.com/>. Acesso em: 22 dez. 2021.

Além disso, todas as etapas desta pesquisa foram executadas em um ambiente virtual local da distribuição Python para computação científica Anaconda[2], e todo o processo foi registrado em documentos do Jupyter Notebook, o que possibilitou iterações mais rápidas, favoreceu a acessibilidade, a manutenção do código e até mesmo a reprodutibilidade das etapas. As versões adotadas dos instrumentos empregados foram Anaconda Individual Edition 2021.11, Python 3.9.7 e Jupyter Notebook 6.4.6.

Quanto às ferramentas, buscou-se utilizar APIs que possibilitam uma coleta rápida e consistente de dados do Reddit, levando em conta também a necessidade de seleção e obtenção de discussões realizadas em períodos específicos. Para o processo de transformação e modelagem dos dados, optou-se por adotar pacotes Python que proporcionam soluções modernas e eficientes, próximas ao estado da arte em PLN, para o cumprimento dos objetivos da pesquisa – permitindo ainda considerar múltiplos níveis de análise, como a exploração de relações entre tópicos e a geração de tópicos dinâmicos. Tais ferramentas são detalhadas nas Subseções 3.2.1 e 3.2.2, onde são descritas suas principais funcionalidades e vantagens, considerando-se as necessidades das etapas da metodologia as quais estão associadas.

3.2.1 Pushshift

A ferramenta primária utilizada na coleta de dados foi o Pushshift, uma plataforma projetada para armazenar e disponibilizar dados originados em redes sociais para pesquisadores, com a finalidade de reduzir as barreiras técnicas na aquisição de dados para pesquisas científicas. Desde 2015, essa plataforma tem coletado discussões realizadas no Reddit, e tem oferecido abertamente o download de todos os comentários e postagens transmitidos em *subreddits* públicos desde a criação do site, em junho de 2005 (BAUMGARTNER *et al.*, 2020).

Os dados supracitados fazem parte do Pushshift Reddit Dataset[3], um conjunto de dados atualizado em tempo real que é divido em dois grupos conforme o sistema de discussões do Reddit: *submissions*, que equivalem às postagens primárias em um *subreddit*, e *comments*, que referem-se aos comentários que são realizados em uma postagem. Em ambos os casos, os dados são armazenados em um arquivo *JavaScript Object Notation* (JSON) onde cada linha corresponde a uma postagem, ou um comentário, e os atributos equivalem aos valores que descrevem o objeto, como pseudônimo do autor, título, conteúdo e metadados (BAUMGARTNER *et al.*, 2020).

[2] Disponível em <https://www.anaconda.com/>. Acesso em: 24 nov. 2021.
[3] Disponível em <https://files.pushshift.io/reddit/>. Acesso em: 24 nov. 2021.

Tabela 2 – Exemplos de atributos JSON referentes a postagens do Reddit

Nome do Atributo	Descrição do Atributo	Tipo do Atributo
id	Identificador da postagem	*String*
author	Nome de usuário de quem realizou a postagem	*String*
created_utc	*Unix Timestamp* referente à data de publicação	*Integer*
selftext	Texto associado à postagem	*String*
title	Título associado à postagem	*String*
score	*Score* acumulado pela postagem	*Integer*

Fonte: adaptado de Baumgartner *et al.* (2020).

Para obter dados através da plataforma, é fornecida a Pushshift Reddit API[4], que oferece acesso dinâmico aos dados armazenados e funcionalidades como filtragem e agregação. Dessa forma, a API permite a seleção de qualquer período de tempo para a busca, em qualquer *subreddit* público e sem limites da quantidade total de dados a ser coletada, sendo necessário somente considerar o limite de requisições por minuto (BAUMGARTNER *et al.*, 2020).

Além de facilitar a aquisição, filtragem e armazenamento de dados através da sua API, o Pushshift Reddit Dataset possui um plano de gestão de dados abertos alinhados aos princípios FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) (WILKINSON *et al.*, 2016), visto que seu repositório de arquivos está disponível publicamente, podendo ser acessado por qualquer pessoa para a obtenção de dados, em um formato interoperável (JSON), e sem restrições de reutilização (BAUMGARTNER *et al.*, 2020).

Em comparação à Reddit API[5], API oficial do site, o Pushshift possui uma série de vantagens significativas, sendo uma delas a possibilidade de aquisição de dados históricos de qualquer período – funcionalidade essencial para essa pesquisa, que busca coletar discussões publicadas nos anos de 2020 e 2021. Quanto às desvantagens, nota-se que o Pushshift não mantém atualizados os metadados das discussões, o que prejudica a análise de variáveis como a popularidade (número de *upvotes*) das publicações.

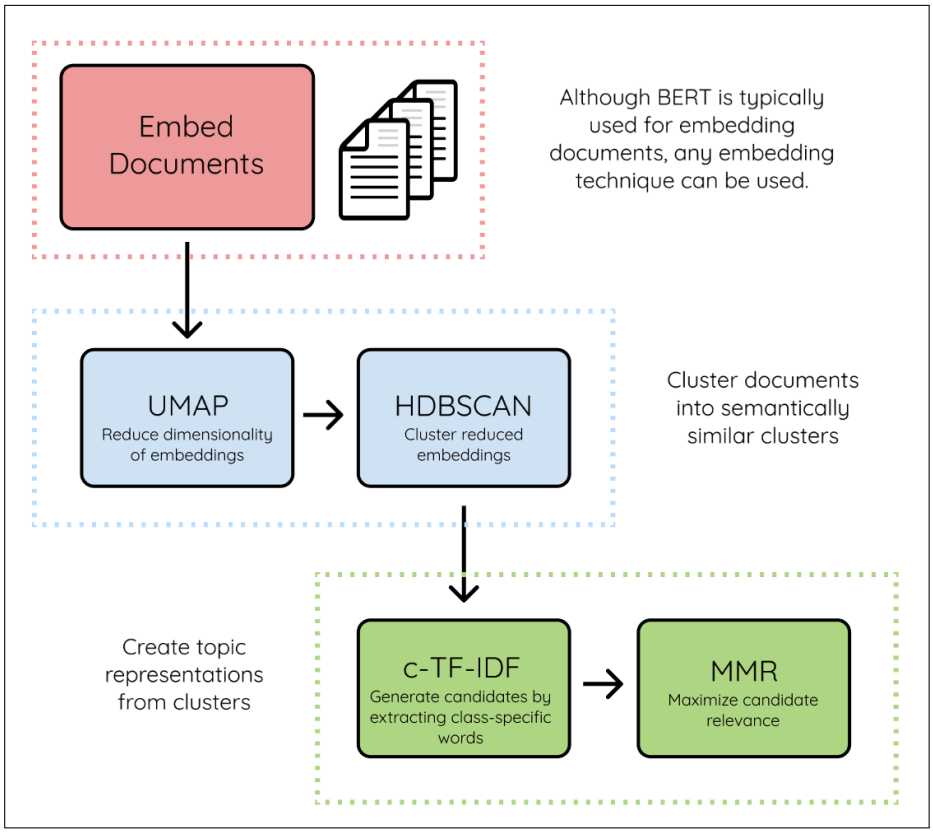
Portanto, para o presente estudo, optou-se por utilizar a combinação das APIs do Pushshift e do Reddit, sendo a primeira para a coleta dos dados no período analisado e a segunda para o enriquecimento através de metadados. Para facilitar esse processo, decidiu-se usar o *Pushshift Multithread* API *Wrapper* (PMAW)[6] e o *Python Reddit* API *Wrapper* (PRAW)[7], *wrappers* que juntos podem assegurar a realização do procedimento de coleta planejado.

[4] Disponível em <https://github.com/pushshift/api>. Acesso em: 24 nov. 2021.
[5] Disponível em <https://www.reddit.com/dev/api/>. Acesso em: 30 dez. 2021.
[6] Disponível em <https://github.com/mattpodolak/pmaw>. Acesso em: 24 nov. 2021.
[7] Disponível em <https://praw.readthedocs.io/en/stable/>. Acesso em: 30 dez. 2021.

3.2.2 BERTopic

Na terceira etapa da pesquisa, quando é realizada a modelagem de tópicos, foi utilizada a ferramenta BERTopic[8], um pacote Python que emprega modelos de *embeddings* conjuntamente a algoritmos de redução de dimensionalidade, clusterização e *class-based* TF-IDF (c-TF-IDF) para a formação de tópicos interpretáveis (GROOTENDORST, 2020).

Figura 9 – Fluxo da modelagem de tópicos com o BERTopic



Fonte: Grootendorst (2020).

Como ilustrado na Figura 9, o fluxo da modelagem de tópicos com o BERTopic pode ser sintetizado nos seguintes passos, sendo a última etapa considerada opcional:

1. Criação de *embeddings* para os documentos através de um modelo escolhido previamente;
2. Redução da dimensionalidade dos *embeddings* criados;
3. Clusterização dos *embeddings* reduzidos, resultando na geração de *clusters* densos de documentos semanticamente semelhantes;
4. Aplicação de c-TF-IDF aos *clusters*, gerando palavras-chave que equivalem a tópicos;
5. Aumento da diversidade das palavras candidatas através de uma variação do algoritmo de *Relevância Marginal Máxima* (MMR), introduzido por Carbonell e Goldstein (1998).

[8] Disponível em <https://maartengr.github.io/BERTopic/>. Acesso em: 22 dez. 2021.

Na primeira etapa do fluxo descrito acima, pode-se utilizar qualquer modelo para geração de *document embeddings*, não limitando-se unicamente a modelos baseados no BERT. Por padrão, o BERTopic faz uso da *framework* Sentence-Transformers[9], que fornece uma série de modelos pré-treinados para diferentes tipos de tarefas, como busca semântica e mineração de paráfrase (REIMERS; GUREVYCH, 2019; GROOTENDORST, 2020).

Além disso, o Sentence-Transformers dispõe de modelos projetados para casos de uso gerais (Tabela 3), sendo treinados em todo o conjunto de dados disponível na *framework* (REIMERS; GUREVYCH, 2019). Esses modelos podem ser úteis para o caso estudado no presente trabalho, pois incluem na sua base de treinamento um conjunto de dados com mais de 700 milhões de tuplas de comentários obtidos do Reddit (HENDERSON *et al.*, 2019), o que pode favorecer a criação de *embeddings* para documentos provenientes do site.

Tabela 3 – Comparação dos modelos para *document embeddings*

Nome do Modelo	Performance	Máximo de *Tokens*	Sentenças/Segundo
all-roberta-large-v1	70,23	256	800
all-mpnet-base-v1	69,98	512	2.800
all-mpnet-base-v2	69,57	384	2.800
all-MiniLM-L12-v1	68,83	256	7.500
all-distilroberta-v1	68,73	512	4.000
all-MiniLM-L12-v2	68,70	256	7.500
all-MiniLM-L6-v2	68,06	256	14.200
all-MiniLM-L6-v1	68,03	128	14.200

Fonte: Sentence-Transformers (2021).

A Tabela 3 mostra os modelos de uso geral fornecidos pelo Sentence-Transformers, e descreve os resultados das avaliações realizadas pela plataforma em um sistema equipado com uma GPU V100. A performance refere-se ao desempenho médio na codificação de sentenças em 14 tarefas diversas de diferentes domínios, e todos os modelos foram treinados em um conjunto de dados contendo mais de 1 bilhão de tuplas de treinamento (REIMERS; GUREVYCH, 2019).

Na segunda etapa do fluxo, é feita a redução da dimensionalidade dos *embeddings* então gerados – processo que consiste na projeção destes em um subespaço de dimensão inferior. Frequentemente, ao utilizar-se representações reduzidas como entrada em modelos de AM, pode-se obter predições de maior qualidade que destacam a "essência" dos dados, já que há uma espécie de filtragem de *features* não-essenciais (MURPHY, 2012). Esse processo é benéfico ainda como uma fase preparativa que antecede uma clusterização baseada em densidade, pois pode ajudar a destacar áreas mais densas no espaço vetorial (ANGELOV, 2020).

[9] Disponível em <https://www.sbert.net/>. Acesso em: 24 dez. 2021.

Para efetuar a redução da dimensionalidade, o BERTopic emprega o algoritmo *Uniform Manifold Approximation and Projection* (UMAP)[10], uma técnica que efetua a redução de forma rápida, escalando bem em termos de tamanho e dimensionalidade, e preservando de forma adequada a estrutura global do conjunto de dados (MCINNES *et al.*, 2018).

Na terceira etapa do fluxo, ocorre o agrupamento de documentos semanticamente semelhantes, processo realizado a partir da clusterização com *document embeddings* reduzidos. No BERTopic, é utilizado o *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN)[11], um algoritmo de clusterização hierárquico que encontra áreas de densidade variável e combina os pontos dessas áreas densas para a formação de *clusters*, atribuindo para as áreas mais esparsas o rótulo de ruído (MCINNES *et al.*, 2017). Como resultado, obtém-se tópicos na forma de *clusters*, que são compostos por pontos correspondentes a documentos (GROOTENDORST, 2020).

Durante o processo de clusterização com HDBSCAN, ocorre ainda o cálculo de um conjunto de pontos denominados "exemplares", que são considerados o "coração" de um *cluster*. Esses exemplares consistem nos pontos que persistem no *cluster*, sendo em última análise os pontos ao redor dos quais há a formação do agrupamento final (MCINNES *et al.*, 2017). No BERTopic, os documentos equivalentes a pontos exemplares são chamados "documentos representativos", e são considerados como aqueles que melhor caracterizam um tópico, podendo ser usados como um fator auxiliar no processo de interpretação (GROOTENDORST, 2020).

Na quarta etapa do fluxo, é efetuada a representação dos *clusters* de tópicos por meio de palavras-chave obtidas a partir do seu conteúdo. Esse processo ocorre através da aplicação do algoritmo de c-TF-IDF – uma variação do TF-IDF que, ao invés de comparar a importância das palavras entre os documentos, trata todos os documentos que compartilham uma categoria (por exemplo, um *cluster*) como um único documento, e calcula o grau de importância para palavras dentro de tal classe (GROOTENDORST, 2020). A equação de c-TF-IDF é descrita a seguir:

$$c\text{-}TF\text{-}IDF = \frac{t_i}{w_i} \times \log(\frac{m}{\sum_j^n t_j}). \tag{3.1}$$

Onde a frequência de cada palavra t é obtida para cada classe i e divida pelo número total de palavras w; e a quantidade média de palavras por classe m é divida pela frequência total da palavra t dentre todas as classes.

[10] Disponível em <https://umap-learn.readthedocs.io/en/latest/>. Acesso em: 22 dez. 2021.
[11] Disponível em <https://hdbscan.readthedocs.io/en/latest/>. Acesso em: 22 dez. 2021.

No BERTopic, o cálculo de c-TF-IDF resulta na pontuação de relevância das palavras para um *cluster*, o que é tomado como medida para a escolha das palavras-chave consideradas representativas de cada tópico. Dessa forma, um maior valor determina uma maior prioridade na representação do tópico, que normalmente é feita considerando as dez palavras de maior c-TF-IDF (GROOTENDORST, 2020).

Na última etapa do fluxo, considerada opcional, pode ser realizada a melhoria da coerência das representações c-TF-IDF, o que é feito utilizando a MMR, um algoritmo que busca maximizar a relevância de itens recuperados diminuindo o nível de similaridade entre eles (CARBONELL; GOLDSTEIN, 1998). No BERTopic, a MMR é usada para encontrar palavras-chave mais coerentes sem que haja muita sobreposição, o que pode resultar na supressão de palavras que não favorecem a representação de um tópico (GROOTENDORST, 2020).

O processo de modelagem de tópicos descrito acima é baseado no método introduzido pelo Top2Vec, que utiliza um modelo de *embeddings* para a representação conjunta de documentos e palavras no espaço vetorial e, após a redução da dimensionalidade com UMAP, os agrupa por similaridade semântica através do HDBSCAN (ANGELOV, 2020).

Entretanto, no Top2Vec, a representação dos tópicos é criada através do cálculo da centroide de cada *cluster* e a listagem das palavras nas proximidades (ANGELOV, 2020), enquanto no BERTopic os tópicos são gerados tomando como medida o grau de importância das palavras encontradas em cada *cluster* de documentos (GROOTENDORST, 2020).

Além disso, o BERTopic fornece diversas abordagens de modelagem de tópicos, similarmente às diferentes variações existentes embasadas na LDA. Porém, diferente dessas variações, o BERTopic baseia-se inteiramente no uso de *document embeddings*, ou *sentence embeddings*, e faz uso destes desde o início do processo (GROOTENDORST, 2020).

Quanto ao método de aprendizagem, os modelos de tópicos disponibilizados podem ser do tipo supervisionado, não-supervisionado ou semi-supervisionado, o que pode favorecer diversas aplicações. Em todos os casos, é possível gerar tópicos dinâmicos, que dão importância à ordem temporal dos documentos, permitindo analisar o comportamento e a frequência dos tópicos em diferentes períodos (GROOTENDORST, 2020).

Por fim, a ferramenta fornece diferentes métodos de visualização de tópicos, como gráfico de tópicos dinâmicos, dendrograma para visualizar a estrutura hierárquica dos tópicos e mapa de distância entre tópicos, sendo a última baseada na técnica de visualização de tópicos introduzida pelo LDAVis (SIEVERT; SHIRLEY, 2014).

4 ANÁLISE

Este capítulo está dividido em quatro seções, correspondentes a cada um dos passos da metodologia adotada: coleta de dados, pré-processamento dos dados, modelagem de tópicos e aquisição de conhecimentos. Para cada etapa, é detalhado o procedimento de análise e são descritos os instrumentos empregados, de modo a facilitar a compreensão. Na Tabela 4, são descritas as bibliotecas utilizadas ao longo da pesquisa, destacando-se nomes, versões e etapas correspondentes, onde os números representam cada passo na ordem descrita acima.

Tabela 4 – Bibliotecas Python utilizadas durante a análise

Nome da Biblioteca Python	Versão Utilizada	Etapas Correspondentes
pmaw	2.1.1	1
praw	7.5.0	1
pandas	1.3.5	1, 2, 3
emoji	1.6.1	2
matplotlib	3.5.0	2, 3
fasttext	0.9.2	3
gensim	4.1.2	3
umap-learn	0.5.2	3
hdbscan	0.8.27	3
scikit-learn	1.0.1	3
bertopic	0.9.4	3, 4
numpy	1.20.3	4

Fonte: elaborado pelo autor (2022).

4.1 Coleta de Dados

A primeira etapa corresponde à coleta dos dados a serem processados, que consistem em discussões realizadas no *subreddit r/College* ao longo dos dois primeiros anos da crise de COVID-19, ou seja, desde 1 de janeiro de 2020 até 31 de dezembro de 2021. Especificamente, somente em 11 de março de 2020 é que o diretor-geral da Organização Mundial da Saúde (OMS) declarou a condição de pandemia (WORLD HEALTH ORGANIZATION, 2020), porém, como a presente pesquisa considera em sua análise o período de ocorrência das discussões (através da geração de tópicos dinâmicos), pode ser útil explorar – sem prejuízo – o surgimento de tópicos relacionados à pandemia nos meses iniciais de 2020.

Quanto ao tipo de discussões a ser coletado, decidiu-se descartar os comentários e considerar somente as postagens. Essa escolha foi determinada através de uma testagem inicial com uma amostra de comentários do *r/College*, quando constatou-se que seu conteúdo gerava uma quantidade muito grande de tópicos, incluindo tópicos individuais para expressões

de agradecimento e vícios de linguagem, dificultando consideravelmente a análise.

Dessa forma, a coleta das postagens ocorreu através do PMAW, que encapsula a API do Pushshift e gera requisições à sua base de dados de forma distribuída. Esse pacote possui uma implementação de *multithreading* e limitação inteligente de solicitações, o que o torna ideal para a criação de conjuntos de dados de larga escala. Utilizou-se ainda a função do PMAW de enriquecimento de metadados via PRAW, de modo a requalificar conteúdos excluídos que tenham persistido na base de dados do Pushshift e capturar a quantidade real de *upvotes* de cada postagem, o que permite analisar a popularidade dos tópicos em diferentes períodos.

Como resultado desse processo, obteve-se um arquivo JSON com o total de 90.569 postagens publicadas no *r/College* durante o período selecionado, incluindo data de publicação, número de *upvotes* e pseudônimo dos autores. Por fim, o arquivo foi convertido para o formato CSV, para depois ser explorado mais facilmente, em forma de tabela, na etapa posterior.

4.2 Pré-Processamento dos Dados

Após a coleta das postagens, foi necessário realizar a limpeza dos dados – etapa essencial para promover a normalização dos textos obtidos, diminuindo em potencial a quantidade de ruídos e permitindo, na Seção 4.3, a criação de *document embeddings* de qualidade elevada. Conforme a abordagem de modelagem de tópicos adotada nessa pesquisa, foi necessário preservar as palavras e a ordem em que elas aparecem, o que demanda uma limpeza mais cuidadosa do que normalmente seria adotada em um tratamento com BOW. Esse processo foi efetuado com o uso intensivo de expressões regulares e componentes como *pandas*[1], para a descrição dos dados, e *Matplotlib*[2], para eventuais visualizações necessárias para a etapa posterior.

A fim de compreender a estrutura das postagens e executar uma limpeza minuciosa, utilizou-se como referência o *Guia de Formatação do Reddit*[3], que descreve entidades de HTML e a sintaxe comum de *Markdown* usada pelo site, além de padrões adicionais específicos desse domínio. Foram realizados ainda testes no seu editor de postagens, para investigar o funcionamento da hierarquia entre diferentes combinações de formatação, o que foi necessário para projetar expressões regulares que normalizam o texto de modo eficiente, além de influenciar a ordem em que essas expressões devem ser aplicadas ao texto.

[1] Disponível em <https://pandas.pydata.org/>. Acesso em: 18 jul. 2021.
[2] Disponível em <https://matplotlib.org/>. Acesso em: 18 jul. 2021.
[3] Disponível em <https://reddit.zendesk.com/hc/en-us/articles/360043033952-Formatting-Guide>. Acesso em: 20 dez. 2021.

Em uma limpeza teste, constatou-se – através da exploração e descrição dos dados – a existência de um número significativo de postagens removidas. Identificou-se também postagens recorrentes do tipo enquete, que não possuem conteúdo textual significativo, além da presença de *emojis*, *emoticons*, e-mails e links, sendo os dois últimos tanto na forma extensa quanto no padrão *Markdown* clicável.

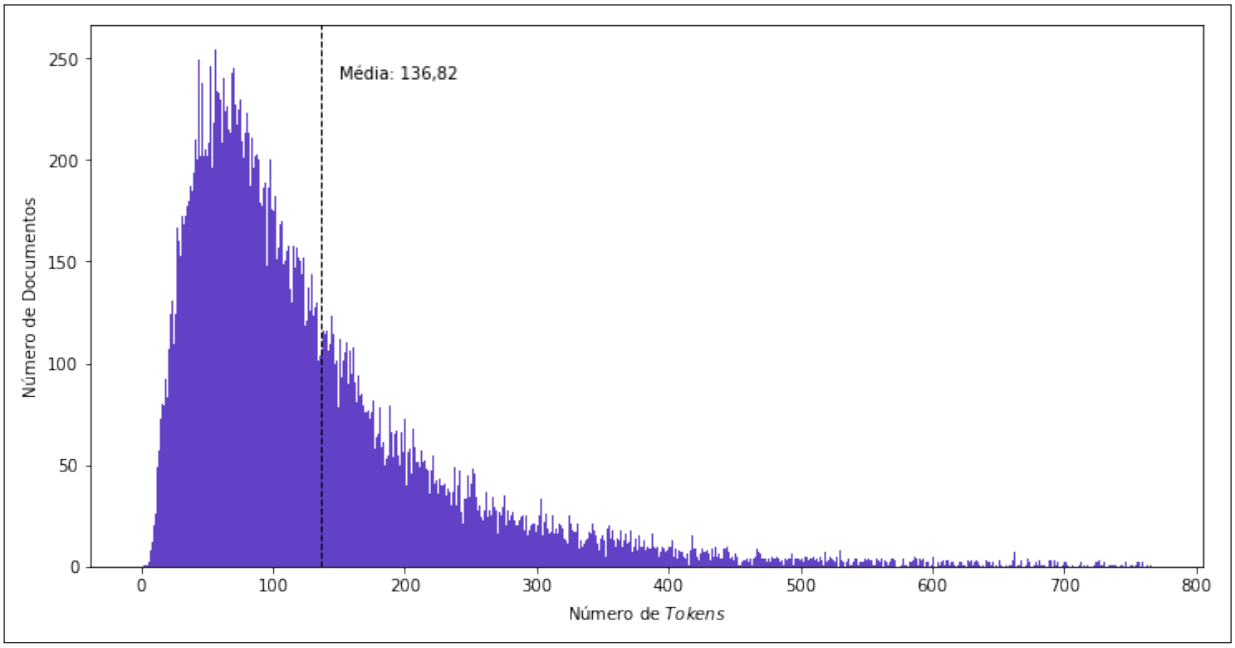
Dessa forma, foi realizado o pré-processamento dos dados nos seguintes passos:

1. Delimitar as colunas da *DataFrame*, mantendo-se somente identificadores de postagem, datas de publicação, títulos, conteúdos textuais, autores e números de *upvotes*;
2. Converter a medida de tempo, de *Unix Timestamp* a data no padrão Tempo Universal Coordenado (UTC), considerando-se somente dia, mês e ano. Destaca-se aqui que os horários de publicação das postagens não têm relevância para este estudo e, além do mais, manter uma precisão de tempo muito alta aumentaria consideravelmente o tempo de processamento de tópicos dinâmicos na Seção 4.4;
3. Limpar linhas da tabela que contenham um dos seguintes registros: valor ausente, postagem removida, postagem duplicada ou enquete.
4. Substituir entidades de HTML pelo caractere resultante;
5. Limpar, para cada ocorrência e respeitando a hierarquia, as formatações referentes a: *emojis*, citação, cabeçalho, negrito, itálico, lista pontuada, e-mail e link clicável, riscado, e-mail e link extenso, *spoiler*, bloco e linha de código, *emoticons*, sobrescrito e tabela;
6. Substituir múltiplas barras de espaço por uma individual;
7. Limpar postagens com corpo textual vazio ou espaço de largura zero do padrão *Unicode*;
8. Por fim, para cada linha da *DataFrame*, unir título e conteúdo em uma única coluna.

Com a finalização desse processo, reduziu-se o volume de dados de 90.569 a 33.199 postagens, com um total de 24.116 autores. Como passo adicional, delimitou-se ainda as colunas a data de publicação, documento – ou seja, a união de título e conteúdo – e número de *upvotes*, de modo a concluir a preparação dos dados para as etapas de modelagem de tópicos e aquisição de conhecimentos (que contempla a visualização de tópicos dinâmicos).

Observou-se ainda a quantidade de palavras em cada documento (Figura 10), a fim de auxiliar a escolha de um dos modelos pré-treinados para criação de *document embeddings*. Como na língua inglesa o método mais comum de tokenização é a decomposição do texto em palavras separadas por espaços em branco, pode-se considerar que a contagem destas equivale à quantidade de *tokens* nos documentos desta pesquisa.

Figura 10 – Gráfico do número de documentos por número de *tokens*



Fonte: elaborado pelo autor (2022).

Na Figura 10, percebe-se que a quantidade de documentos diminui conforme o número de *tokens* aumenta – sendo a média destes ao longo dos documentos aproximadamente igual 136. Normalmente, o comprimento máximo das sequências nos modelos para *document embeddings* varia entre os valores 128, 256, 384 e 512. Portanto, antes de avançar para a modelagem de tópicos, é relevante observar também a porcentagem de documentos com as quantidades correspondentes de *tokens* citadas anteriormente, conforme a Tabela 5.

Tabela 5 – Porcentagem de documentos do *corpus* com até *t tokens*

Quantidade de *Tokens* (*t*)	Porcentagem de Documentos
Até 128	62,59%
Até 256	88,75%
Até 384	95,85%
Até 512	98,14%

Fonte: elaborado pelo autor (2022).

4.3 Modelagem de Tópicos

A partir dos dados pré-processados obtidos, foi realizada a etapa correspondente à modelagem de tópicos, que incluiu três processos gerais: configuração do *CountVectorizer*, ajustagem de hiperparâmetros, e treinamento e inferência. Dessa forma, buscou-se encontrar tópicos interpretáveis com alto grau de coerência e diversidade.

Para aplicar as técnicas de modelagem de tópicos, foram utilizadas conjuntamente as ferramentas UMAP, HDBSCAN, BERTopic e Scikit-learn[4], sendo esta última para a transformação do *corpus* em uma matriz de documentos e termos, antes da obtenção da representação c-TF-IDF. Quanto às métricas para avaliação de tópicos, foram utilizadas as implementações de NPMI da biblioteca Gensim[5] (ŘEHŮŘEK; SOJKA, 2010), e de WE-IRBO disponível no repositório *topic-model-diversity*[6], juntamente ao modelo para *embeddings* na língua inglesa *cc.en.300*, da biblioteca fastText[7] (BOJANOWSKI *et al.*, 2016).

Primeiramente, foi escolhido o modelo para *document embeddings* a ser empregado, optando-se pelo modelo geral de Sentence-Transformers *all-MiniLM-L6-v2*[8], que corresponde a um modelo pré-treinado baseado no MiniLM com 6 *layers* e máximo de *tokens* igual a 256, o que significa que 88,75% dos documentos do *corpus* (Tabela 5) serão processados em sua totalidade, enquanto o restante será truncado. A escolha desse modelo se deu devido à sua alta velocidade e bom custo-benefício, como descrito na Tabela 3, além de incluir milhões de comentários do Reddit no seu conjunto de dados de treinamento, o que pode favorecer a representação vetorial para documentos originados no site.

4.3.1 Configuração do CountVectorizer

Posteriormente, foi investigada a melhor configuração de *CountVectorizer*, para o Scikit-learn, no caso estudado. Para tanto, foi necessário observar o número de tópicos encontrado automaticamente em todo o conjunto de dados, considerando valores padrões para todas as ferramentas envolvidas no processo de modelagem com o BERTopic (incluindo UMAP, HDBSCAN e Scikit-learn), o que constituiu um total de 275 tópicos – valor que foi tomado como fixo durante a configuração do *CountVectorizer*.

Em seguida, foram realizadas testagens, para o número fixo de tópicos encontrado, com diferentes opções de listas com *stop words* a serem removidas pelo *CountVectorizer*: lista vazia (ou seja, sem remoção), lista em língua inglesa do Scikit-learn, e lista básica do CoreNLP[9]. Após cada iteração, calculou-se a pontuação NPMI para então comparar-se os resultados, e ao final, foi constatado que a aplicação da lista de *stop words* do CoreNLP forneceu tópicos

[4] Disponível em <https://scikit-learn.org/stable/>. Acesso em: 18 jul. 2021.
[5] Disponível em <https://radimrehurek.com/gensim/>. Acesso em: 7 jan. 2022.
[6] Disponível em <https://github.com/silviatti/topic-model-diversity>. Acesso em: 7 jan. 2022.
[7] Disponível em <https://fasttext.cc/>. Acesso em: 7 jan. 2022.
[8] Disponível em <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>. Acesso em: 24 dez. 2021.
[9] Disponível em <https://github.com/stanfordnlp/CoreNLP>. Acesso em: 22 dez. 2021.

de maior coerência, sendo assim a configuração de *CountVectorizer* adotada para os passos subsequentes da pesquisa.

Depois, foi efetuada mais uma testagem, considerando valores padrões e a nova configuração de *CountVectorizer* obtida anteriormente, e então registrou-se os cinco menores *clusters* correspondentes a relatos referentes à pandemia, para então serem usados como critério durante a ajustagem dos hiperparâmetros – sendo um dos objetivos principais impedir a inclusão desses *clusters* em outros maiores e mais gerais, mantendo-se assim o máximo de tópicos correspondentes a relatos de problemas específicos enfrentados por estudantes.

4.3.2 Ajustagem dos Hiperparâmetros

Assim, a ajustagem dos hiperparâmetros foi efetuada de forma semi-automática e considerando todo o conjunto de dados pré-processados. Para cada hiperparâmetro a ser considerado, foram realizados 20 incrementos e a partir deles foi efetuada a modelagem de tópicos, aplicando-se ao final as métricas de NPMI e WE-IRBO.

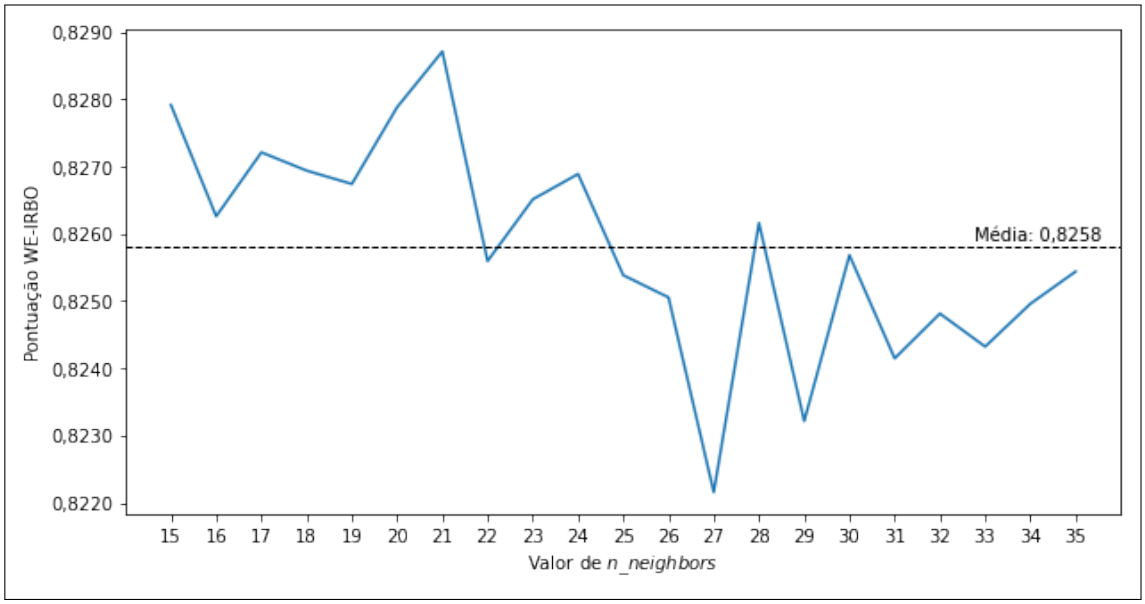
A partir dos resultados das avaliações, tomou-se como melhor candidato, através daqueles que possuem diversidade acima da média, os modelos com tópicos de maior coerência. Para cada iteração subsequente, a partir da testagem do segundo hiperparâmetro, adotou-se os melhores valores obtidos anteriormente para cada parâmetro correspondente, permitindo assim aumentar a coerência dos tópicos após cada iteração.

Dessa forma, foram considerados para essa fase os hiperparâmetros *n_neighbors* e *n_components*, do UMAP, e *min_cluster_size*, do HDBSCAN, que foram ajustados na ordem descrita. Além disso, o valor de *min_topic_size* do BERTopic, que refere-se à quantidade mínima de documentos que um *cluster* deve ter para que seja considerado um tópico, foi mantido igual a *min_cluster_size*, assim como é definido o modo de operação padrão da ferramenta.

No UMAP, o parâmetro *n_neighbors* é responsável por controlar o equilíbrio entre estruturas locais e globais dos dados, de modo que valores menores forçam o algoritmo a concentrar-se em estruturas mais localizadas, enquanto valores maiores fazem com que o foco seja voltado para estruturas globais, possivelmente causando a perda de detalhes mais precisos; enquanto o parâmetro *n_components* determina a dimensionalidade do espaço reduzido no qual os dados serão projetados (MCINNES *et al.*, 2018). E ainda, no HDBSCAN, o parâmetro *min_cluster_size* é considerado o fator primário na clusterização, determinando o menor tamanho de um agrupamento para que ele seja considerado um *cluster* (MCINNES *et al.*, 2017).

Dessa forma, realizou-se primeiro a ajustagem de *n_neighbors*, que foi testado com valores no intervalo de 15 a 35. Conforme ilustrado na Figura 11, a média da pontuação WE-IRBO foi 0,8258, sendo alcançada no geral para $n\_neighbors < 25$, com a exceção de $n\_neighbors = 22$, que esteve abaixo da média, e $n\_neighbors = 28$, que esteve acima da média.

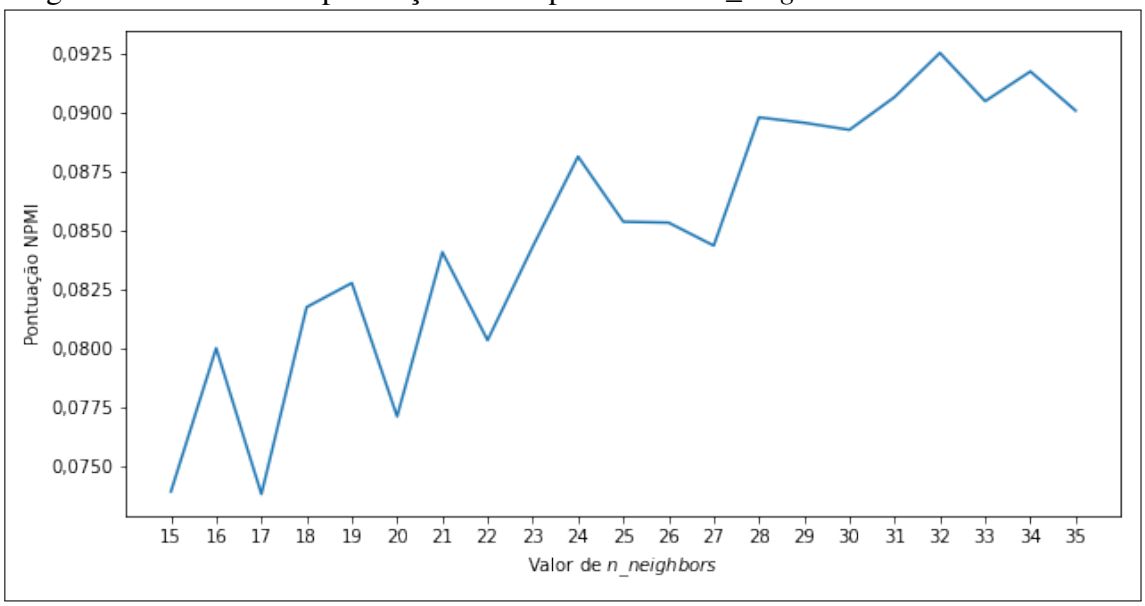
Figura 11 – Gráfico da pontuação WE-IRBO por valor de *n_neighbors*



Fonte: elaborado pelo autor (2022).

Conforme a Figura 12, dentre os valores observados, aquele que possui a maior coerência e ainda possui uma diversidade acima da média é $n\_neighbors = 28$, que é considerado então o melhor candidato, sendo tomado como o valor fixo do parâmetro na sequência.

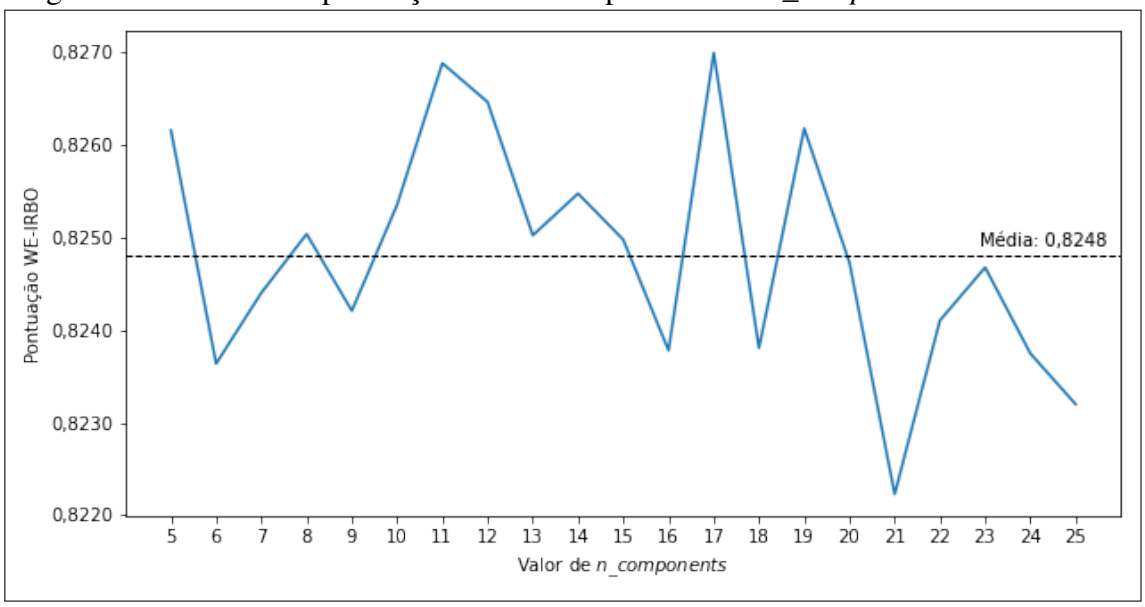
Figura 12 – Gráfico da pontuação NPMI por valor de *n_neighbors*



Fonte: elaborado pelo autor (2022).

Em seguida, o valor a ser ajustado foi *n_components* no intervalo de 5 a 25, e levando em consideração o resultado obtido de $n\_neighbors = 28$. Na Figura 13, observa-se que a média da pontuação WE-IRBO foi 0,8248, o que descarta uma série de valores para *n_components*.

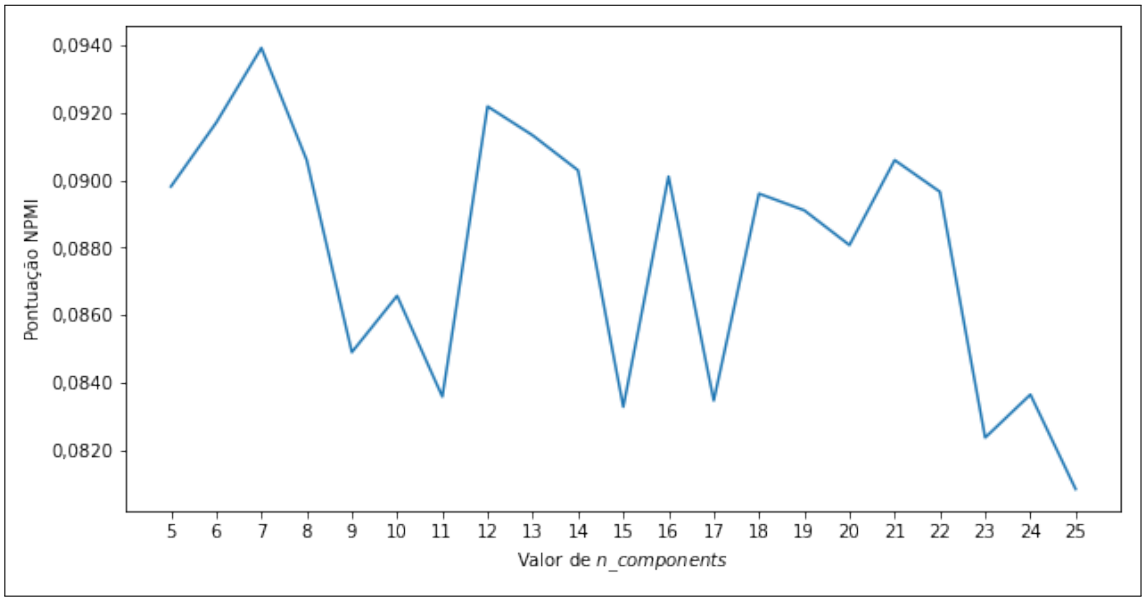
Figura 13 – Gráfico da pontuação WE-IRBO por valor de *n_components*



Fonte: elaborado pelo autor (2022).

Assim, conforme a Figura 14, o melhor candidato encontrado foi $n\_components = 12$, que possui a maior pontuação NPMI dentre os valores a serem considerados, já que para $n\_components = 7$ a medida de diversidade é abaixo da média.

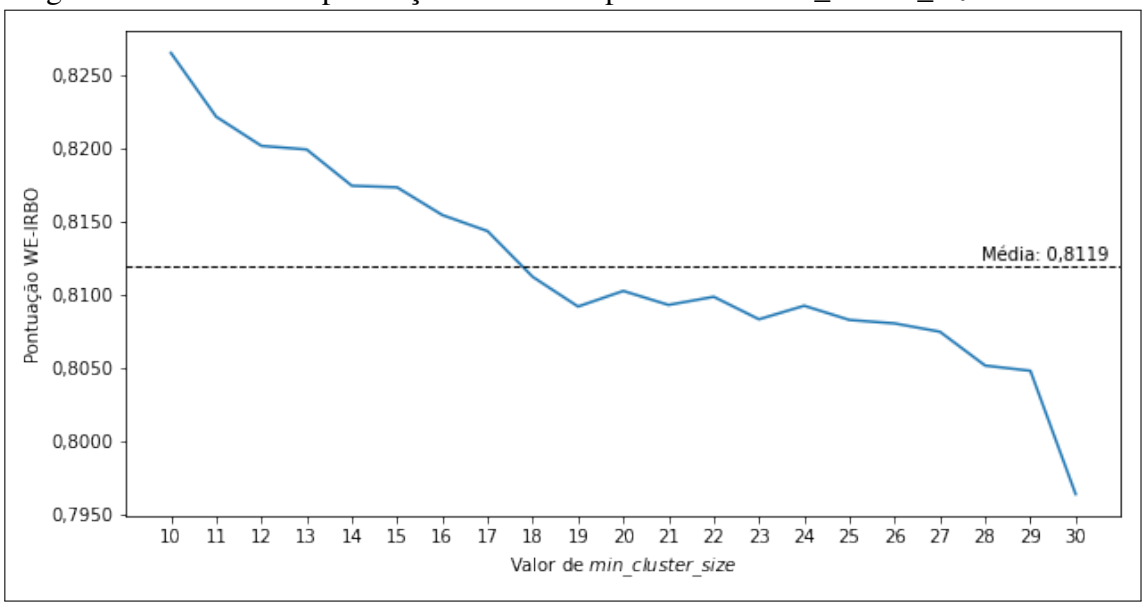
Figura 14 – Gráfico da pontuação NPMI por valor de *n_components*



Fonte: elaborado pelo autor (2022).

Por fim, ajustou-se o valor de *min_cluster_size* no intervalo de 10 a 30, considerando $n\_neighbors = 28$ e $n\_components = 12$. Conforme ilustrado na Figura 15, a média de pontuação WE-IRBO foi 0,8119, o que descarta mais da metade dos candidatos possíveis.

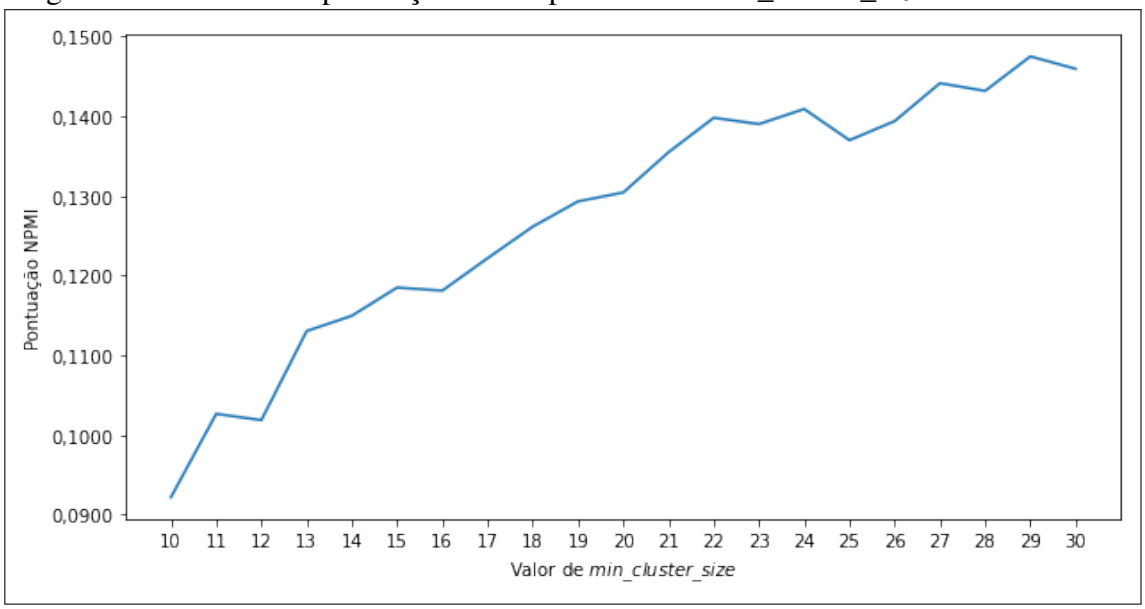
Figura 15 – Gráfico da pontuação WE-IRBO por valor de *min_cluster_size*



Fonte: elaborado pelo autor (2022).

Finalmente, conforme a Figura 16, foi obtido o melhor candidato para a opção $min\_cluster\_size = 17$, já que para $min\_cluster\_size > 17$ a medida de diversidade fica abaixo da média para as pontuações observadas, descartando esses valores como candidatos.

Figura 16 – Gráfico da pontuação NPMI por valor de *min_cluster_size*



Fonte: elaborado pelo autor (2022).

Ao final do processo de ajuste dos hiperparâmetros, que teve uma duração de aproximadamente 15 horas, obteve-se os valores $n\_neighbors = 28$, $n\_components = 12$ e $min\_cluster\_size = min\_topic\_size = 17$, com pontuações WE-IRBO e NPMI iguais a 0,8143 e 0,1222, respectivamente. Além disso, foi observada a presença de todos os *clusters* previamente registrados, sendo um deles correspondente ao menor tópico obtido, o que significa que uma redução adicional da quantidade de *clusters* poderia fazer com que este tópico fosse incorporado a outro tópico mais geral (causando sua supressão).

4.3.3 Treinamento e Inferência

Dessa forma, foi possível prosseguir com a preparação necessária para o treinamento do modelo. Primeiramente, a partir do arquivo com os dados pré-processados, foram criadas, para cada ano, três listas para o armazenamento de cada coluna da tabela – documento, data de publicação e número de *upvotes*. Durante a criação dessas listas, preservou-se a ordem dos índices para cada linha das colunas, o que é essencial para permitir a formação de tópicos associados à sua respectiva data de publicação e seu total de *upvotes*.

Em seguida, a partir de toda a estrutura obtida até então – incluindo a seleção do modelo pré-treinado *all-MiniLM-L6-v2*, a configuração do *CountVectorizer* e os valores encontrados para os hiperparâmetros – foi realizado o treinamento do modelo de tópicos em todo o conjunto de documentos, abrangendo postagens de 2020 e 2021, o que resultou em um total de 161 tópicos diversos distribuídos ao longo das postagens realizadas nos dois anos.

Finalmente, com base no modelo treinado, foi realizada a inferência de tópicos individualmente para o conjunto de postagens de cada ano, a partir das duas listas criadas anteriormente para os documentos. Assim, obteve-se como resultado representações de tópicos intercambiáveis para ambos os subconjuntos de dados, preservando toda a estrutura dos *clusters* para os dois períodos considerados, incluindo hierarquia e rótulos, o que facilita a interpretação durante o processo de aquisição de conhecimentos.

4.4 Aquisição de Conhecimentos

Após os processos relativos à modelagem de tópicos, foi realizada a etapa final para descoberta de conhecimento, que envolveu a visualização e a interpretação dos tópicos considerados úteis para a pesquisa, observando-se ainda as suas relações, com a finalidade de

descrever em termos compreensíveis os conhecimentos adquiridos.

Ao longo do processo, foram utilizadas três representações gráficas para o auxílio da interpretação dos resultados: gráficos de barras para tópicos, dendrograma com *clusters* de tópicos, e adicionalmente, gráficos temporais. Em conjunto, essas visualizações permitiram encontrar diferentes dimensões de análise relativas ao cenário dos dados, como similaridade estrutural entre tópicos e período de ocorrência (o que ajuda a entender o contexto dos tópicos).

Para a geração dessas representações gráficas, foi utilizado novamente o BERTopic, que possui implementações próprias para tais visualizações, o que facilitou todo o processo. Porém, para a criação dos gráficos de tópicos dinâmicos, foi elaborada uma modificação do código para que seja levado em consideração o número de *upvotes*, resultando em um gráfico de popularidade do tópico em função do tempo, adicionando assim mais uma dimensão de análise. Essa decisão foi tomada pois considera-se que a análise somente da frequência de um tópico não dá importância às dinâmicas de comunicação das redes sociais, onde é comum interagir através de uma postagem criada por outro usuário, sem necessariamente gerar novo conteúdo.

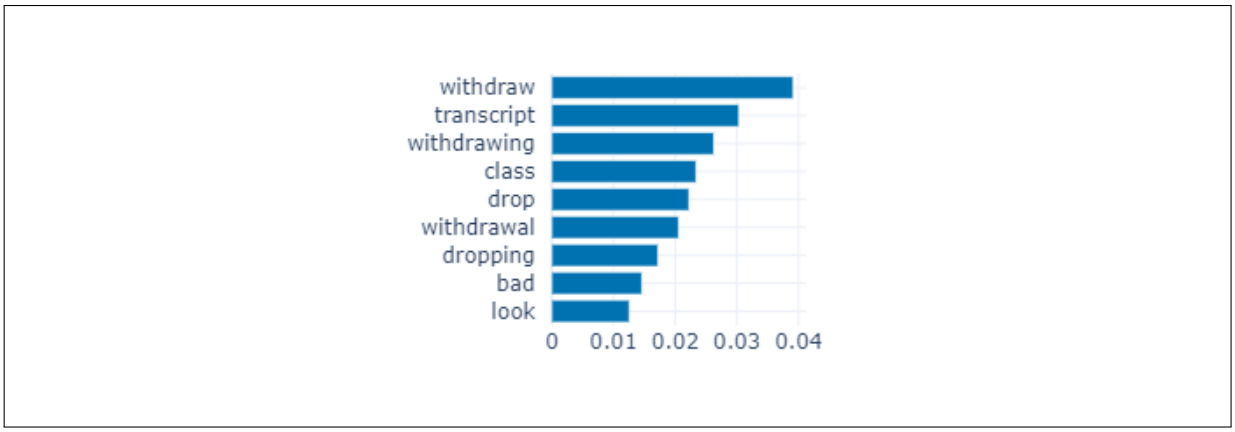
Dessa forma, o processo iniciou a partir da identificação individual dos tópicos relevantes, tomando como referência as suas palavras-chave e seus documentos representativos, sendo estes usados também como fatores auxiliares para a interpretação dos tópicos identificados. Em seguida, após a interpretação inicial, observou-se também o dendrograma dos *clusters* para visualizar a estrutura hierárquica dos tópicos, quando foi verificado quais deles são associados aos tópicos considerados relevantes para a pesquisa. Por fim, foi realizada a análise final, que leva em conta as postagens mais populares de diferentes períodos e, após a descrição dos relatos observados, foi concluído o processo de aquisição de conhecimentos.

5 RESULTADOS

Neste capítulo, são apresentados os resultados do procedimento de descoberta de conhecimento realizado na pesquisa. Em cada seção, é discutido um dos tópicos encontrados que possuem relação com impactos da pandemia de COVID-19 no contexto de estudantes de graduação. Os tópicos são ordenados por quantidade de postagens existentes sobre o mesmo e, para cada um deles, são mostradas suas palavras-chave e tópicos relacionados (Apêndice A). Além disso, foram gerados gráficos de popularidade de tópico para cada período, que podem ser visualizados no Apêndice B.

5.1 Tópico 37: Trancamento e Abandono de Disciplinas

Figura 17 – Gráfico com as palavras-chave do Tópico 37 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

O Tópico 37 abrange 115 postagens e é composto pelas palavras-chave ilustradas na Figura 17, que representam um tema relativo ao histórico escolar, ao trancamento e ao abandono de disciplinas. Os documentos representativos descrevem dois pontos principais: receio pela maior necessidade de realização de trancamentos de disciplinas durante a pandemia; e a preocupação com o histórico escolar em decorrência de trancamentos, incluindo dúvidas sobre o possível impacto destes no processo seletivo para cursos de pós-graduação.

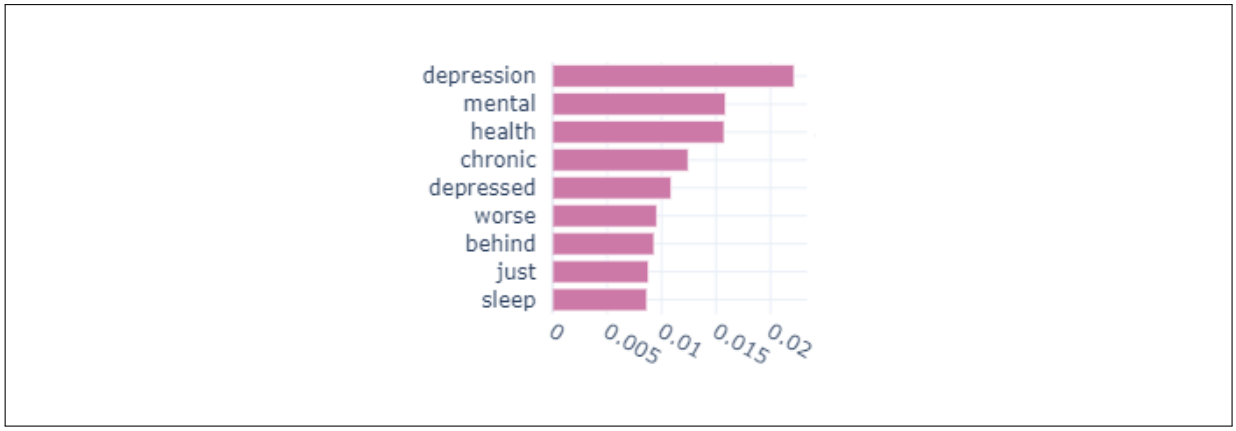
Conforme a Figura 29, o Tópico 37 possui maior relação com o Tópico 107, que compreende postagens sobre a avaliação estudantil da metodologia de ensino dos professores. Em um segundo nível de similaridade, ele está associado ao grupo dos Tópicos 74 e 135, relacionados aos temas: falta de motivação e dificuldade de assimilação do conteúdo abordado em sala de aula.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 31), o Tópico 37 foi mais votado nos meses de abril, maio e setembro; enquanto em 2021 (Figura 32), ele foi mais votado em maio, setembro e outubro. Dentre as postagens mais populares, aquelas que mencionam trancamento de disciplinas no contexto da pandemia descrevem os seguintes temas:

- Abril de 2020:
  - Sentimento de sobrecarga devido ao excesso de obrigações domésticas e familiares, em consequência de *lockdowns*;
  - Falta de motivação e dificuldade de concentração durante aulas online.
- Maio de 2021:
  - Dificuldade de adaptação com o formato online para disciplinas práticas.

5.2 Tópico 64: Depressão e Saúde Mental

Figura 18 – Gráfico com as palavras-chave do Tópico 64 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

O Tópico 64 abrange 74 postagens e é composto pelas palavras-chave ilustradas na Figura 18, que representam um tema relativo a depressão e saúde mental. Os documentos representativos descrevem três pontos principais: sentimento de solidão por estudantes novatos, motivado pela falta de oportunidades de socialização em decorrência de medidas de distanciamento social e meios de ensino à distância; piora de um quadro de depressão durante a pandemia; e declínio da performance acadêmica em decorrência do falecimento de um membro da família ou um amigo.

Conforme a Figura 29, o Tópico 64 possui maior relação com o Tópico 71, que compreende postagens sobre sentimento de cansaço e desmotivação. Em um segundo nível de similaridade, ele está associado ao grupo dos Tópicos 26 e 138, relacionados aos temas: saudade

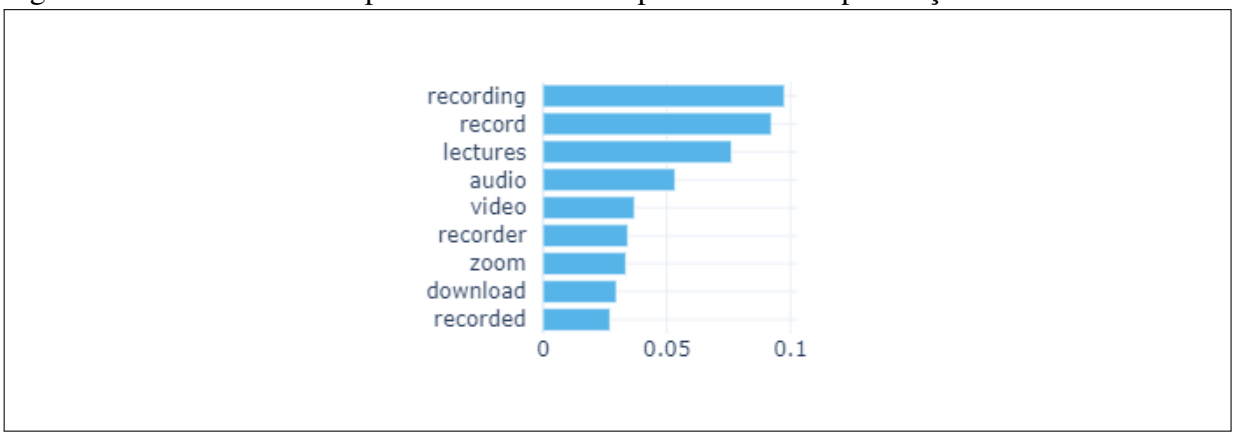
de casa, e dificuldade de relacionamento com os pais.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 33), o Tópico 64 foi mais votado no mês de setembro; enquanto em 2021 (Figura 34), ele foi mais votado em fevereiro e novembro. Dentre as postagens mais populares, aquelas que mencionam saúde mental no contexto da pandemia descrevem os seguintes temas:

- Abril de 2020:
  - Sentimento de vazio e falta de rumo durante os primeiros períodos de *lockdown*.
- Setembro de 2020:
  - Declínio da saúde mental ao longo do ano, e a busca por formas de lidar com as mudanças impostas pela pandemia;
  - Sentimento de isolamento por estudantes novatos devido à dificuldade de socialização através de aulas online.
- Fevereiro de 2021:
  - Dificuldade para acompanhar as aulas devido à piora de um quadro de depressão durante a pandemia, e a busca por estratégias para recuperar o conteúdo perdido.

5.3 Tópico 100: Necessidade de Aulas Gravadas

Figura 19 – Gráfico com as palavras-chave do Tópico 100 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

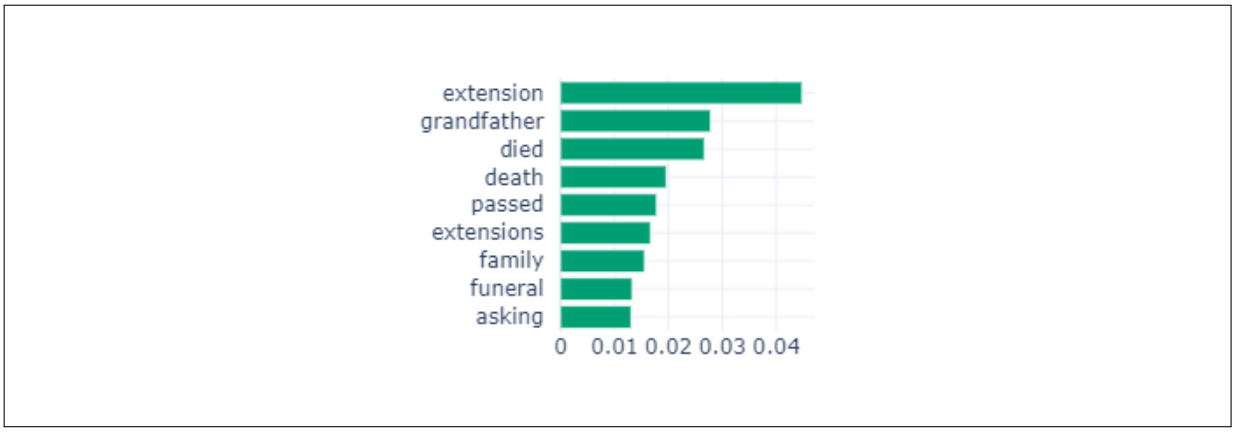
O Tópico 100 abrange 41 postagens e é composto pelas palavras-chave ilustradas na Figura 19, que representam um tema relativo a gravação de vídeo e áudio, e a videoconferências. Os documentos representativos descrevem três pontos principais: busca por recomendações de softwares para gravação de tela; relatos sobre as vantagens da disponibilidade de aulas gravadas; e a busca por formas de gravação de aulas presenciais.

Conforme a Figura 29, o Tópico 100 possui maior relação com o Tópico 19, que compreende postagens sobre plataformas de videoconferência usadas para o ensino remoto. Em um segundo nível de similaridade, ele está associado ao grupo dos Tópicos 11, 51 e 72, relacionados aos temas: busca por aparelhos móveis para realização de anotações em sala de aula; busca por aplicativos de organização; e discussões sobre plataformas para aplicação de testes online.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 35), o Tópico 100 foi mais votado no mês de agosto; enquanto em 2021 (Figura 36), ele foi mais votado em janeiro, setembro e novembro. Dentre as postagens mais populares, aquelas que mencionam gravação de aulas no contexto da pandemia descrevem um mesmo tema: busca por formas de gravação de videoconferências para consulta posterior, motivada por um favorecimento para o aprendizado e por facilitar a revisão de conteúdo.

5.4 Tópico 106: Luto e Extensão de Prazos

Figura 20 – Gráfico com as palavras-chave do Tópico 106 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

O Tópico 106 abrange 38 postagens e é composto pelas palavras-chave ilustradas na Figura 20, que representam um tema relativo ao falecimento de um familiar e a extensão de prazo para atividades em disciplinas universitárias. Os documentos representativos descrevem um ponto principal: busca por sugestões para comunicar um professor sobre o falecimento – causado por coronavírus ou não – de uma pessoa próxima, para esclarecimento da necessidade de pedido de extensão de prazos para atividades.

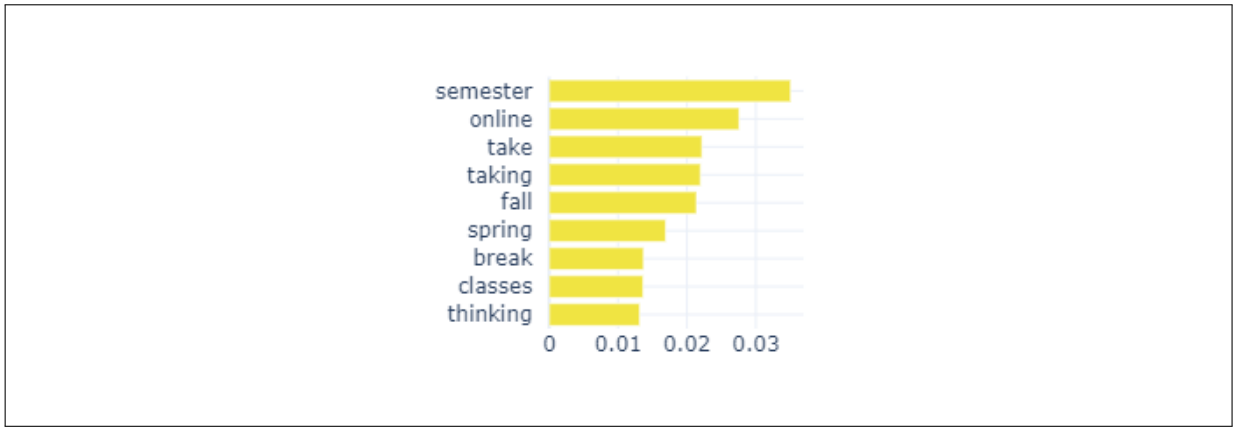
Conforme a Figura 28, o Tópico 106 possui maior relação com o Tópico 95, que compreende postagens sobre a busca de sugestões para formas apropriadas de preparar um

e-mail para contato com um professor. Em um segundo nível de similaridade, ele está associado ao grupo dos Tópicos 69 e 116, relacionados aos temas: má experiência com professores e preocupação com a opinião dos pais sobre o rendimento acadêmico.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 37), o Tópico 106 foi mais votado nos meses de abril e dezembro; enquanto em 2021 (Figura 38), ele foi mais votado em setembro e dezembro. Dentre as postagens mais populares, aquelas que mencionam luto e prazos no contexto da pandemia descrevem um mesmo tema: a necessidade de extensão de prazos devido ao falecimento de um parente ou amigo – sendo as postagens que citam mortes por coronavírus presentes em dezembro de 2020 e setembro de 2021.

5.5 Tópico 123: Trancamento do Semestre

Figura 21 – Gráfico com as palavras-chave do Tópico 123 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

O Tópico 123 abrange 29 postagens e é composto pelas palavras-chave ilustradas na Figura 21, que representam um tema relativo à contemplação da possibilidade de trancamento do semestre, cujas justificativas são representadas pelos documentos representativos, que descrevem três contextos principais: trancamento por receio de não poder focar nos estudos durante a pandemia; trancamento pela dificuldade de adaptação aos meios de ensino online; e trancamento pelas dificuldades vividas em outros momentos da pandemia, causando o declínio da performance acadêmica do aluno.

Conforme a Figura 28, o Tópico 123 não possui nenhuma relação direta com outro tópico, sendo sua maior proximidade com o *cluster* de ruídos, que corresponde a postagens muito gerais e que não são classificadas em tópicos. Em um segundo nível de similaridade, ele está associado aos Tópicos 6 e 20, relacionados aos temas: dúvidas sobre cursos não-obrigatórios

para a obtenção de créditos, e busca por dicas de como ter um bom proveito na universidade.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 39), o Tópico 123 foi mais votado no mês de julho; enquanto em 2021 (Figura 40), ele foi mais votado em janeiro. Dentre as postagens mais populares, aquelas que mencionam trancamento do semestre no contexto da pandemia descrevem os seguintes temas:

- Julho de 2020:
  - Falta de ambiente apropriado para estudo, acentuado pelo fechamento de bibliotecas;
  - Baixa motivação e dificuldade de concentração durante aulas online;
  - Declínio da saúde mental durante os primeiros meses da pandemia.
- Outubro de 2020:
  - Preocupação em contrair COVID-19 durante aulas práticas, acentuada pela existência de comorbidades para o possível agravamento da doença.
- Janeiro de 2021:
  - Falta de compatibilidade de horário devido a um novo emprego, intensificada pela carência de aulas online assíncronas.

5.6 Tópico 127: Baixa Motivação

Figura 22 – Gráfico com as palavras-chave do Tópico 127 e suas pontuações c-TF-IDF

Fonte: elaborado pelo autor (2022).

O Tópico 127 abrange 26 postagens e é composto pelas palavras-chave ilustradas na Figura 22, que representam um tema relativo ao nível de motivação de estudantes durante aulas online. Os documentos representativos descrevem dois pontos principais: dificuldade de concentração em aulas online, acentuada pela presença de distrações em casa; e dificuldade de adaptação à rotina estabelecida após a introdução de aulas majoritariamente remotas.

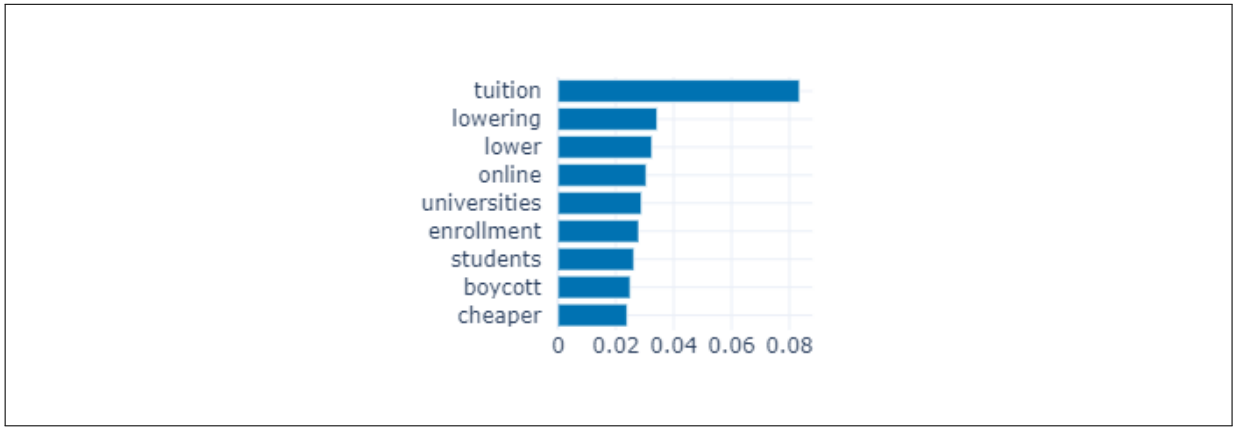
Conforme a Figura 28, o Tópico 127 possui maior relação com o Tópico 121, que compreende postagens que repercutem a utilização de meios de ensino híbrido e online. Em um segundo nível de similaridade, ele está associado aos Tópicos 99, 134 e 160, relacionados aos temas: horários de tira-dúvidas; dificuldade para entender o conteúdo ensinado; e preocupação com o fechamento temporário de bibliotecas (Seção 5.11).

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 41), o Tópico 127 foi mais votado no mês de setembro; enquanto em 2021 (Figura 42), ele foi mais votado em janeiro e maio. Dentre as postagens mais populares, aquelas que mencionam baixa motivação no contexto da pandemia descrevem os seguintes temas:

- Setembro de 2020:
  - Sentimento de isolamento e declínio da saúde mental;
  - Dificuldade de organização e de autodisciplina no contexto de aulas online.
- Janeiro e maio de 2021:
  - Dificuldade de concentração em aulas online, acentuada pela presença de distrações.

5.7 Tópico 139: Não-Redução de Mensalidades

Figura 23 – Gráfico com as palavras-chave do Tópico 139 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

O Tópico 139 abrange 23 postagens e é composto pelas palavras-chave ilustradas na Figura 23, que representam um tema relativo à mensalidade de universidades. Os documentos representativos descrevem dois pontos principais: descontentamento pela cobrança de mensalidade integral durante períodos onde o ensino foi completamente online; e desmotivação para a realização de matrícula no semestre subsequente, provocada pela não-redução do preço da mensalidade e pela preocupação com a perda de experiências proporcionadas pelo ambiente

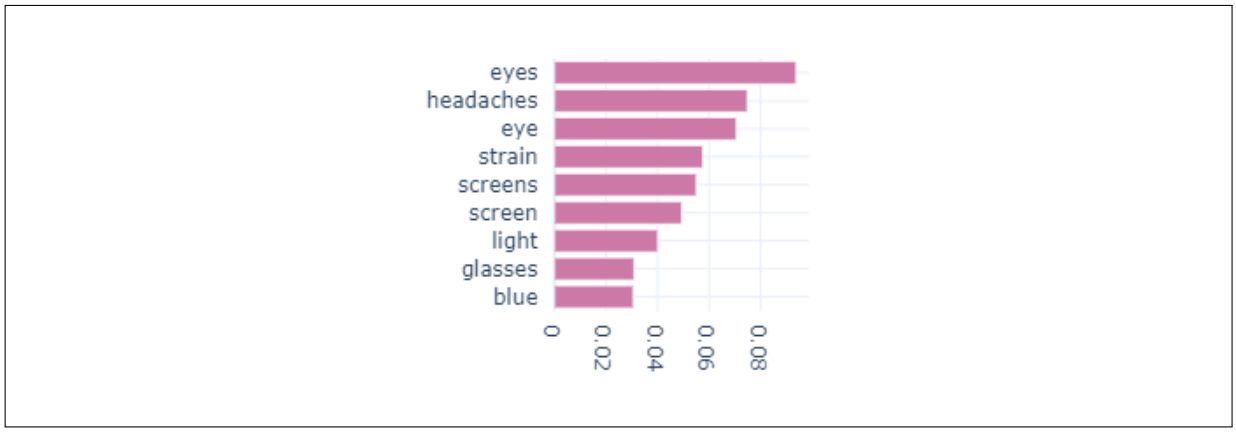
universitário, como formação de amizades e *networking*.

Conforme a Figura 30, o Tópico 139 não possui nenhuma relação direta com outro tópico, sendo sua maior proximidade com o agrupamento dos Tópicos 14 e 150, relacionados aos temas: busca pelo aumento da média de notas, e preocupação com a média necessária para a realização de transferência de instituição.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 43), o Tópico 139 foi mais votado no mês de abril; enquanto em 2021 (Figura 44), ele foi mais votado em outubro. Dentre as postagens mais populares, aquelas que mencionam mensalidades no contexto da pandemia descrevem um mesmo tema: a insatisfação com a não-redução de mensalidades, incluindo observações sobre a falta de acesso a recursos das universidades, como laboratórios, estúdios, bibliotecas e computadores – por vezes contendo licenças de software específicas.

5.8 Tópico 141: Fadiga Ocular

Figura 24 – Gráfico com as palavras-chave do Tópico 141 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

O Tópico 141 abrange 22 postagens e é composto pelas palavras-chave ilustradas na Figura 24, que representam um tema relativo a telas e desconfortos oculares. Os documentos representativos descrevem um ponto principal: queixas sobre dores de cabeça e fadiga ocular devido ao aumento da exposição a telas de aparelhos digitais durante a pandemia.

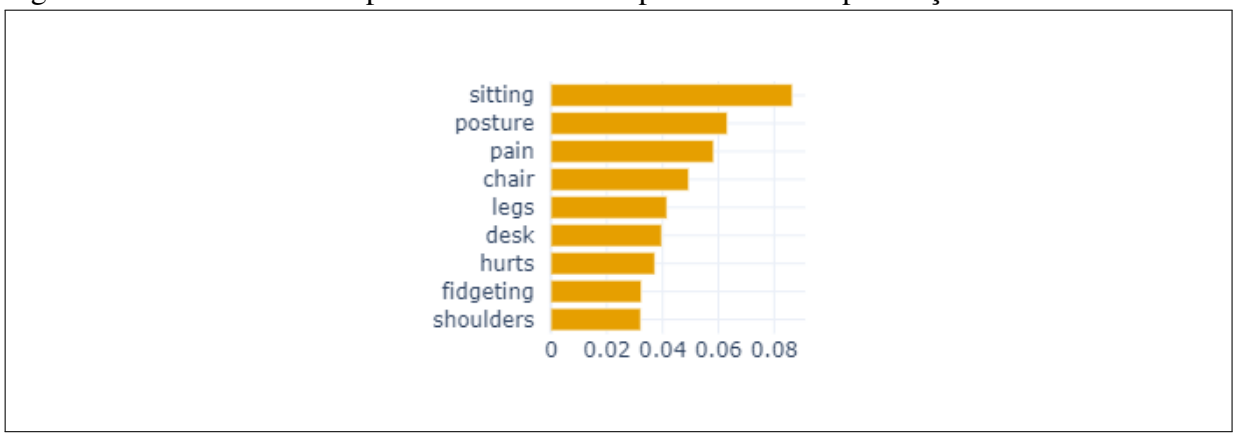
Conforme a Figura 29, o Tópico 141 não possui relação direta com outro tópico, sendo sua maior proximidade com o agrupamento dos Tópicos 58 e 82, que compreendem postagens sobre ambientes virtuais de aprendizagem e postagens sobre cursos de arte, animação e design gráfico.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 45), o Tópico 141 foi

mais votado no mês de abril; enquanto em 2021 (Figura 46), ele foi mais votado em janeiro. Dentre as postagens mais populares, aquelas que mencionam desconfortos oculares no contexto da pandemia descrevem os mesmos temas: fadiga ocular e dores de cabeça devido ao aumento do uso do computador, causando ainda a redução da produtividade e a dificuldade de realização de atividades em formato digital; aumento da frequência de dores de cabeça em relação aos semestres anteriores à pandemia; e a busca por dicas para a redução do desconforto causado pelas telas.

5.9 Tópico 156: Dores nas Costas

Figura 25 – Gráfico com as palavras-chave do Tópico 156 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

O Tópico 156 abrange 18 postagens e é composto pelas palavras-chave ilustradas na Figura 25, que representam um tema relativo a dor nas costas. Os documentos representativos descrevem dois pontos principais: preocupação com o aumento – durante a pandemia – do tempo que se passa sentado; e queixas de dores nas costas, nas pernas e nos ombros, causadas por longos períodos sentados, ocasionando ainda a desmotivação para estudar.

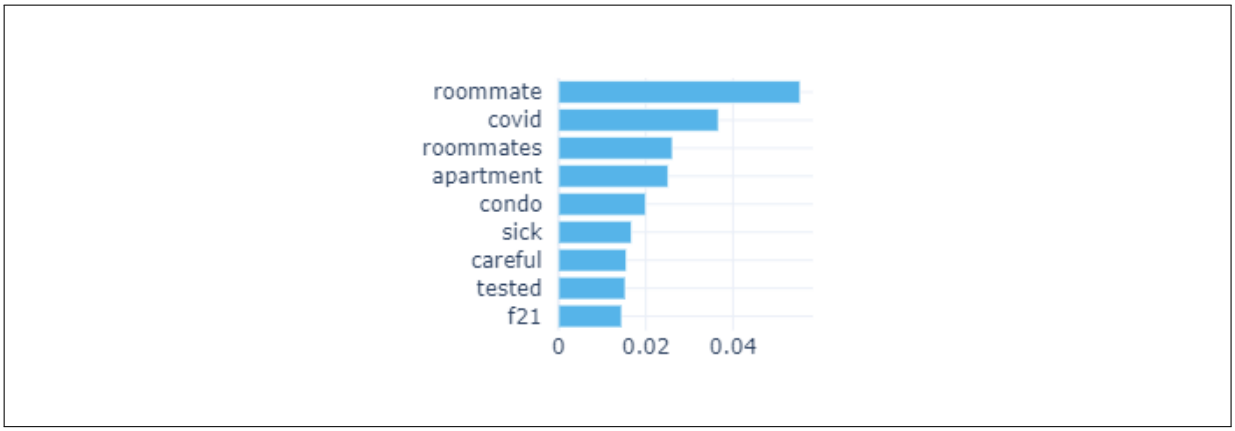
Conforme a Figura 29, o Tópico 156 possui maior relação com o Tópico 152, que compreende postagens sobre a busca de sugestões para tipos de assentos. Em um segundo nível de similaridade, ele está associado aos Tópicos 91 e 98, ambos relacionados aos tipos de estrutura dos dormitórios, como mesas, camas, cadeiras e banheiros.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 47), o Tópico 156 foi mais votado nos meses de março e setembro; enquanto em 2021 (Figura 48), ele foi mais votado em janeiro e março. Dentre as postagens mais populares, aquelas que mencionam dores nas costas no contexto da pandemia descrevem um mesmo tema: aumento de cansaço, dores nas

costas e dores nas pernas durante o período da pandemia, acentuadas pelo maior tempo que se passa sentado.

5.10 Tópico 159: Colegas de Quarto e COVID-19

Figura 26 – Gráfico com as palavras-chave do Tópico 159 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

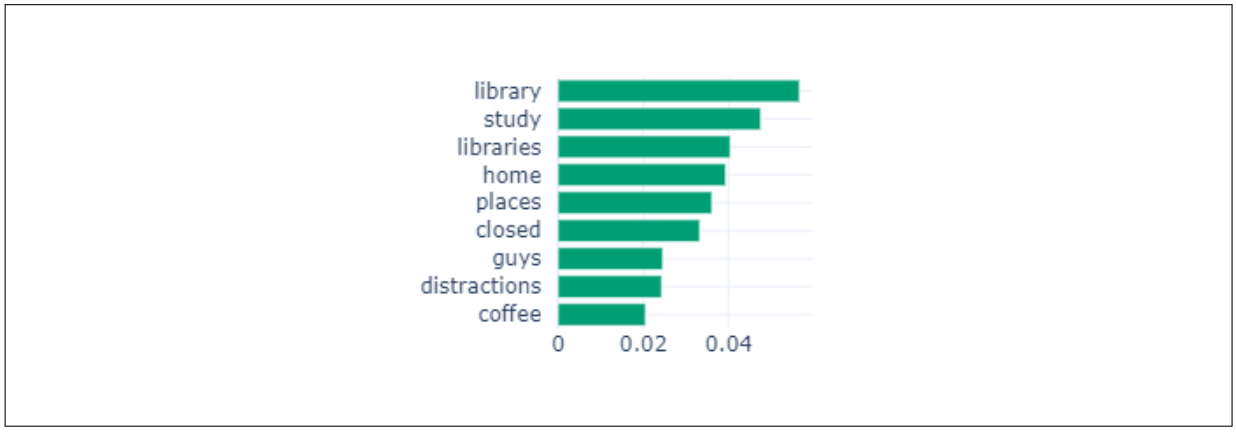
O Tópico 159 abrange 17 postagens e é composto pelas palavras-chave ilustradas na Figura 26, que representam um tema relativo a colegas de quarto e COVID. Os documentos representativos descrevem dois pontos principais: dúvidas sobre medidas de prevenção adotadas em dormitórios universitários; e preocupação quanto aos cuidados de prevenção do coronavírus adotados por colegas de quarto, causando o surgimento do receio de contração do vírus.

Conforme a Figura 29, o Tópico 159 possui maior relação com o Tópico 92, que compreende postagens sobre aluguel de apartamentos. Em um segundo nível de similaridade, ele está associado aos Tópicos 2, 33 e 155, relacionados aos temas: colegas de quarto em dormitórios universitários; meios de transporte para realização do percurso até a universidade; e busca por recomendações de tipo de apartamento ou dormitório a ser escolhido.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 49), o Tópico 159 foi mais votado no mês de junho; enquanto em 2021 (Figura 50), ele foi mais votado em janeiro e setembro. Dentre as postagens mais populares, aquelas que relacionam a pandemia com colegas de quarto descrevem um mesmo tema: preocupação com a falta de cumprimento de medidas de prevenção do coronavírus por colegas de quarto – esse receio foi mais comum em postagens que citam a existência de comorbidades para o possível agravamento da doença.

5.11 Tópico 160: Fechamento de Bibliotecas

Figura 27 – Gráfico com as palavras-chave do Tópico 160 e suas pontuações c-TF-IDF



Fonte: elaborado pelo autor (2022).

O Tópico 160 abrange 17 postagens e é composto pelas palavras-chave ilustradas na Figura 27, que representam um tema relativo a distrações, estudo e bibliotecas. Os documentos representativos descrevem dois pontos principais: preocupação com o fechamento temporário de bibliotecas de universidades, intensificada pela presença de distrações em casa ou em residência universitária; e dificuldade de execução de atividades por estudantes que costumavam utilizar bibliotecas como espaço de estudo principal.

Conforme a Figura 28, o Tópico 160 não possui nenhuma relação direta com outro tópico, sendo sua maior proximidade com o agrupamento dos Tópicos 121 e 127 (Seção 5.6), que compreendem postagens sobre a repercussão da utilização de meios de ensino híbrido e online, e postagens sobre o nível de motivação de estudantes durante aulas online. Em um segundo nível de similaridade, ele está associado aos Tópicos 99 e 134 (Seção 5.6), relacionados aos temas: horários de tira-dúvidas e dificuldade para entender o conteúdo ensinado.

Quanto à sua popularidade, ao longo do ano de 2020 (Figura 51), o Tópico 160 foi mais votado no mês de outubro; enquanto em 2021 (Figura 52), ele foi mais votado em janeiro. Dentre as postagens mais populares, aquelas que relacionam a pandemia com bibliotecas descrevem um mesmo tema: busca por novos locais de estudo devido ao fechamento temporário de bibliotecas – em períodos de *lockdown* ou por medidas de prevenção durante ondas causadas por novas variantes do coronavírus.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresenta um processo para a identificação de problemas enfrentados por estudantes de graduação durante a pandemia de COVID-19, a partir de postagens publicadas na comunidade *r/College* do Reddit. Para a realização desse procedimento, foram utilizados: conceitos básicos de KDD; uma plataforma de coleta com gestão de dados abertos; e ferramentas de código-aberto próximas ao estado da arte em PLN, incluindo um modelo de linguagem baseado em *Transformers* que é de fácil aplicação e possui baixo custo computacional. Dessa forma, o método mostrado – cujo código-fonte está disponível através do repositório do Github *covid19-rcollege-topic-model*[1] – pode ser utilizado para a obtenção de tópicos interpretáveis no Reddit mesmo em computadores de uso pessoal de potência média.

6.1 Contribuições do Trabalho

Os resultados obtidos são favoráveis para o entendimento de situações adversas enfrentadas por alunos de graduação durante a pandemia de COVID-19, possivelmente tendo aplicabilidade na preparação de políticas inclusivas ao longo desse período, ou ainda durante outras possíveis crises sanitárias de alto impacto. Apesar da análise não ter sido realizada em uma comunidade nacional do Reddit, os tópicos de problemas encontrados podem ser considerados úteis também por instituições educacionais locais, já que, no geral, esses tópicos se referem a situações globais.

Além disso, o fluxo proposto pode ser facilmente adaptado para a inferência de tópicos em outras redes sociais (como Facebook e Twitter), podendo incluir a incorporação dos metadados referentes às medidas de popularidade dessas redes, possibilitando gerar gráficos de popularidade de tópicos ao longo do tempo, conforme o Anexo B. É possível também analisar discussões realizadas em língua portuguesa, sem exigir muito esforço ou tempo, pois há disponibilidade de modelos de linguagem pré-treinados multilíngues ou em português.

6.2 Limitações

Quanto às limitações do trabalho, percebeu-se que alguns dos tópicos apresentados demonstraram ter baixa popularidade na forma de postagens, apesar de serem muito populares dentre os comentários. Um dos motivos para isso é a existência de *Megathreads* – postagens

[1] Disponível em <https://github.com/LucasFr127/covid19-rcollege-topic-model>

criadas por moderadores que são fixadas no topo do *subreddit* – para a discussão de assuntos específicos, como saúde mental e trancamento de disciplinas.

Quanto à performance do método, o uso de um modelo para *document embeddings* com número máximo de *tokens* igual a 256 pode ter desfavorecido a descoberta de alguns tópicos, já que documentos com mais *tokens* foram truncados, o que equivale ao processamento incompleto de 11,25% das postagens compreendidas pelos dados pré-processados, conforme a Tabela 5. Esse problema poderia ser enfrentado através do uso de um modelo com suporte a um maior limite máximo de *tokens*, ou ainda pela segmentação de sentenças, porém esses processos aumentariam o tempo necessário para o processamento dos documentos.

É possível também melhorar a performance do método mostrado através de uma ajustagem mais adequada para os hiperparâmetros, possivelmente através da utilização de ferramentas que fazem uso de otimização bayesiana, como Hyperopt (BERGSTRA *et al.*, 2015) e OCTIS (TERRAGNI *et al.*, 2021a).

Outro aspecto que pode beneficiar essa procedimento é investigar a viabilidade de utilização do atributo *probabilities_*, do HDBSCAN, para determinar a função de otimização, conforme demonstrado por Borrelli (2021). Tal atributo está relacionado à probabilidade de pertencimento de um ponto a um *cluster*, e estimular a formação de agrupamentos mais relacionados pode possibilitar a formação de *clusters* mais específicos (MCINNES *et al.*, 2017) – porém, em maior quantidade, o que pode aumentar o tempo de análise.

Entretanto, um fator que possivelmente facilitaria a extração de tópicos referentes a experiências negativas na pandemia, mesmo com o aumento da quantidade de tópicos citada acima, seria a aplicação prévia da análise de sentimentos, para então – considerando somente documentos, ou sentenças, de polaridade negativa – ser realizado o fluxo de modelagem de tópicos mostrado.

6.3 Trabalhos Futuros

Finalmente, todo esse processo pode ser incorporado para a modelagem de tópicos considerando discussões por meio de comentários do Reddit, em contraste às postagens, o que pode gerar a descoberta de uma quantidade ainda maior e mais variada de tópicos relacionados aos objetivos da pesquisa, já que a maior quantidade do conteúdo produzido no Reddit é por meio de comentários (REDDIT, 2020; REDDIT, 2021b).

REFERÊNCIAS

ALI, W. Online and remote learning in higher education institutes: a necessity in light of COVID-19 pandemic. **Higher Education Studies**, ERIC, v. 10, n. 3, p. 16–25, 2020. Disponível em: https://doi.org/10.5539/hes.v10n3p16. Acesso em: 4 jul. 2021.

ANGELOV, D. Top2Vec: distributed representations of topics. **CoRR**, abs/2008.09470, 2020. Disponível em: https://arxiv.org/abs/2008.09470. Acesso em: 20 jan. 2022.

ARISTOVNIK, A.; KERŽIČ, D.; RAVŠELJ, D.; TOMAŽEVIČ, N.; UMEK, L. Impacts of the COVID-19 pandemic on life of higher education students: a global perspective. **Sustainability**, v. 12, n. 20, 2020. ISSN 2071-1050. Disponível em: https://doi.org/10.3390/su12208438. Acesso em: 25 jun. 2021.

BAUMGARTNER, J.; ZANNETTOU, S.; KEEGAN, B.; SQUIRE, M.; BLACKBURN, J. The Pushshift Reddit dataset. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 14, n. 1, p. 830–839, May 2020. Disponível em: https://ojs.aaai.org/index.php/ICWSM/article/view/7347. Acesso em: 4 dez. 2021.

BERGSTRA, J.; KOMER, B.; ELIASMITH, C.; YAMINS, D.; COX, D. D. Hyperopt: a Python library for model selection and hyperparameter optimization. **Computational Science & Discovery**, IOP Publishing, v. 8, n. 1, p. 014008, jul 2015. Disponível em: https://doi.org/10.1088/1749-4699/8/1/014008. Acesso em: 22 jan. 2022.

BIANCHI, F.; TERRAGNI, S.; HOVY, D. Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In: **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**. Online: Association for Computational Linguistics, 2021. p. 759–766. Disponível em: http://dx.doi.org/10.18653/v1/2021.acl-short.96. Acesso em: 22 set. 2021.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python**: analyzing text with the Natural Language Toolkit. 1. ed. Sebastopol, CA: O'Reilly Media, 2009.

BLEI, D. M. Probabilistic topic models. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 55, n. 4, p. 77–84, apr 2012. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/2133806.2133826. Acesso em: 16 jul. 2021.

BLEI, D. M.; LAFFERTY, J. D. Correlated topic models. In: **Proceedings of the 18th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: MIT Press, 2005. (NIPS'05), p. 147–154. Disponível em: https://dl.acm.org/doi/10.5555/2976248.2976267. Acesso em: 16 jul. 2021.

BLEI, D. M.; LAFFERTY, J. D. Dynamic topic models. In: **Proceedings of the 23rd International Conference on Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2006. (ICML '06), p. 113–120. ISBN 1595933832. Disponível em: https://doi.org/10.1145/1143844.1143859. Acesso em: 16 jul. 2021.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet allocation. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 993–1022, mar 2003. ISSN 1532-4435. Disponível em: https://dl.acm.org/doi/10.5555/944919.944937. Acesso em: 25 jun. 2021.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **CoRR**, abs/1607.04606, 2016. Disponível em: http://arxiv.org/abs/1607.04606. Acesso em: 7 jan. 2022.

BORRELLI, D. **Clustering sentence embeddings to identify intents in short text**. Towards Data Science, 2021. Disponível em: https://towardsdatascience.com/clustering-sentence-embeddings-to-identify-intents-in-short-text-48d22d3bf02e. Acesso em: 22 jan. 2022.

BRANDON, J. **6.7 million people just mentioned the coronavirus on social media in one day. Here's why**. Forbes Magazine, 2020. Disponível em: https://www.forbes.com/sites/johnbbrandon/2020/03/04/67-million-people-just-mentioned-the-coronavirus-on-social-media-in-one-day-heres-why/. Acesso em: 25 jun. 2021.

CAMBRIA, E.; WHITE, B. Jumping NLP curves: a review of natural language processing research [review article]. **IEEE Computational Intelligence Magazine**, v. 9, n. 2, p. 48–57, 2014. Disponível em: https://doi.org/10.1109/MCI.2014.2307227. Acesso em: 6 ago. 2021.

CARBONELL, J.; GOLDSTEIN, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: **Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 1998. (SIGIR '98), p. 335–336. ISBN 1581130155. Disponível em: https://doi.org/10.1145/290941.291025. Acesso em: 20 jan. 2022.

CHOMSKY, N. **Syntactic structures**. 2. ed. Berlin: De Gruyter Mouton, 2002.

CHOMSKY, N. **Language and mind**. 3. ed. Cambridge, UK: Cambridge University Press, 2006.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: **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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: https://dx.doi.org/10.18653/v1/N19-1423. Acesso em: 6 ago. 2021.

DROUT, M.; SMITH, L. **How to read a dendrogram**. Wheaton College, 2012. Disponível em: https://wheatoncollege.edu/wp-content/uploads/2012/08/How-to-Read-a-Dendrogram-Web-Ready.pdf. Acesso em: 20 fev. 2022.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 39, n. 11, p. 27–34, nov 1996. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/240455.240464. Acesso em: 9 jan. 2022.

FROMKIN, V.; RODMAN, R.; HYAMS, V. **An introduction to language**. 10. ed. Boston, MA: Cengage Learning, 2013.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GOLDBERG, Y. Neural network methods for natural language processing. **Synthesis Lectures on Human Language Technologies**, v. 10, n. 1, p. 1–309, 2017. Disponível em: https://doi.org/10.2200/S00762ED1V01Y201703HLT037. Acesso em: 21 jan. 2022.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge, MA: MIT Press, 2016. Disponível em: https://www.deeplearningbook.org/. Acesso em: 21 jan. 2022.

GOOGLE DEVELOPERS. **Measuring similarity from embeddings**. Google, 2020. Disponível em: https://developers.google.com/machine-learning/clustering/similarity/measuring-similarity. Acesso em: 12 fev. 2022.

GROOTENDORST, M. **BERTopic: leveraging BERT and c-TF-IDF to create easily interpretable topics**. Zenodo, 2020. Disponível em: https://doi.org/10.5281/zenodo.4381785. Acesso em: 22 dez. 2021.

HENDERSON, M.; BUDZIANOWSKI, P.; CASANUEVA, I.; COOPE, S.; GERZ, D.; KUMAR, G.; MRKSIC, N.; SPITHOURAKIS, G.; SU, P.; VULIC, I.; WEN, T. A repository of conversational datasets. **CoRR**, abs/1904.06472, 2019. Disponível em: http://arxiv.org/abs/1904.06472. Acesso em: 24 dez. 2021.

HERNANDEZ, J. **Some colleges and universities will start the new year online as Omicron spreads**. NPR, 2021. Disponível em: https://www.npr.org/2021/12/22/1066788973/colleges-universities-remote-distance-learning-omicron. Acesso em: 22 jan. 2022.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. **Science**, v. 349, n. 6245, p. 261–266, 2015. Disponível em: https://doi.org/10.1126/science.aaa8685. Acesso em: 18 jan. 2022.

IGUAL, L.; SEGU, S.; VITRI, J.; PUERTAS, E.; RADEVA, P.; PUJOL, O.; ESCALERA, S.; DANT, F.; GARRIDO, L. **Introduction to data science**: a Python approach to concepts, techniques and applications. 1. ed. Cham: Springer Publishing Company, Incorporated, 2017. ISBN 3319500163.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**: an introduction to natural language processing, computational linguistics, and speech recognition. 2. ed. Upper Saddle River, NJ: Prentice-Hall, 2009.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**: an introduction to natural language processing, computational linguistics, and speech recognition. 3. ed. [*S. n.*], 2022. Draft of January 12, 2022. Disponível em: https://web.stanford.edu/~jurafsky/slp3/. Acesso em: 21 jan. 2022.

LAU, J. H.; NEWMAN, D.; BALDWIN, T. Machine reading tea leaves: automatically evaluating topic coherence and topic model quality. In: **Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics**. Gothenburg, Sweden: Association for Computational Linguistics, 2014. p. 530–539. Disponível em: http://dx.doi.org/10.3115/v1/E14-1056. Acesso em: 6 ago. 2021.

MANNING, C.; SCHÜTZE, H. **Foundations of statistical natural language processing**. 2. ed. Cambridge, MA: MIT Press, 1999.

MCINNES, L.; HEALY, J.; ASTELS, S. HDBSCAN: Hierarchical Density Based Clustering. **Journal of Open Source Software**, The Open Journal, v. 2, n. 11, p. 205, 2017. Disponível em: https://doi.org/10.21105/joss.00205. Acesso em: 22 dez. 2021.

MCINNES, L.; HEALY, J.; SAUL, N.; GROßBERGER, L. UMAP: Uniform Manifold Approximation and Projection. **Journal of Open Source Software**, The Open Journal, v. 3, n. 29, p. 861, 2018. Disponível em: https://doi.org/10.21105/joss.00861. Acesso em: 22 dez. 2021.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. In: BENGIO, Y.; LECUN, Y. (Ed.). **1st International Conference on Learning Representations, Workshop Track Proceedings**. Scottsdale, AZ: [*S. n.*], 2013. Disponível em: http://arxiv.org/abs/1301.3781. Acesso em: 6 ago. 2021.

MIKOLOV, T.; YIH, W.-T.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Atlanta, Georgia: Association for Computational Linguistics, 2013. p. 746–751. Disponível em: https://aclanthology.org/N13-1090. Acesso em: 6 ago. 2021.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. 2. ed. Cambridge, MA: MIT Press, 2018.

MURPHY, K. P. **Machine learning**: a probabilistic perspective. Cambridge, MA: MIT Press, 2012.

NADWORNY, E. **College move-in was supposed to mark a return to normal. Then came the Delta variant**. NPR, 2021. Disponível em: https://www.npr.org/sections/back-to-school-live-updates/2021/08/18/1028802323/colleges-begin-the-fall-semester-in-person-with-covid-worries-abound. Acesso em: 22 jan. 2022.

NIETZEL, M. T. **Scores of colleges announce plans for near-normal Fall semesters**. Forbes Magazine, 2021. Disponível em: https://www.forbes.com/sites/michaeltnietzel/2021/03/04/scores-of-colleges-announce-plans-for-near-normal-fall-semesters/?sh=691f2978779f. Acesso em: 22 jan. 2022.

PALMER, D. D. Text preprocessing. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of natural language processing**. 2. ed. Boca Raton, FL: Chapman and Hall/CRC, 2010. p. 9–30.

RAO, D.; MCMAHAN, B. **Natural language processing with PyTorch**: build intelligent language applications using deep learning. 1. ed. Sebastopol, CA: O'Reilly Media, 2019.

REDDIT. **Reddit's 2020 year in review**. 2020. Disponível em: https://redditblog.com/2020/12/08/reddits-2020-year-in-review/. Acesso em: 25 jun. 2021.

REDDIT. **Reddit privacy policy**. 2021. Disponível em: https://www.redditinc.com/policies/privacy-policy. Acesso em: 24 nov. 2021.

REDDIT. **Reddit recap 2021**. 2021. Disponível em: https://www.redditinc.com/blog/reddit-recap-2021. Acesso em: 9 jan. 2022.

ŘEHŮŘEK, R.; SOJKA, P. Software framework for topic modelling with large corpora. In: **Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks**. Valletta, Malta: University of Malta, 2010. p. 46–50. ISBN 2-9517408-6-7. Disponível em: https://is.muni.cz/publication/884893/en. Acesso em: 7 jan. 2022.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: sentence embeddings using siamese BERT-networks. **CoRR**, abs/1908.10084, 2019. Disponível em: http://arxiv.org/abs/1908.10084. Acesso em: 24 dez. 2021.

ROZADO, D. Using word embeddings to analyze how universities conceptualize "diversity" in their online institutional presence. **Society**, Springer, v. 56, n. 3, p. 256–266, 2019. Disponível em: https://doi.org/10.1007/s12115-019-00362-9. Acesso em: 12 fev. 2022.

SCHULTZ, A.; PARIKH, J. **Keeping our services stable and reliable during the COVID-19 outbreak**. 2020. Disponível em: https://about.fb.com/news/2020/03/keeping-our-apps-stable-during-covid-19/. Acesso em: 25 jun. 2021.

SENTENCE-TRANSFORMERS. **Pretrained models**. 2021. Disponível em: https://www.sbert.net/docs/pretrained_models.html. Acesso em: 24 dez. 2021.

SIEVERT, C.; SHIRLEY, K. LDAvis: a method for visualizing and interpreting topics. In: **Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces**. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. p. 63–70. Disponível em: https://dx.doi.org/10.3115/v1/W14-3110. Acesso em: 9 ago. 2021.

TERRAGNI, S.; FERSINI, E. An empirical analysis of topic models: uncovering the relationships between hyperparameters, document length and performance measures. In: **Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)**. Held Online: INCOMA Ltd., 2021. p. 1408–1416. Disponível em: https://aclanthology.org/2021.ranlp-1.157. Acesso em: 7 jan. 2022.

TERRAGNI, S.; FERSINI, E.; GALUZZI, B. G.; TROPEANO, P.; CANDELIERI, A. OCTIS: comparing and optimizing topic models is simple! In: **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations**. Online: Association for Computational Linguistics, 2021. p. 263–270. Disponível em: http://dx.doi.org/10.18653/v1/2021.eacl-demos.31. Acesso em: 22 jan. 2022.

TERRAGNI, S.; FERSINI, E.; MESSINA, E. Word embedding-based topic similarity measures. In: MÉTAIS, E.; MEZIANE, F.; HORACEK, H.; KAPETANIOS, E. (Ed.). **Natural Language Processing and Information Systems**. Cham: Springer International Publishing, 2021. p. 33–45. ISBN 978-3-030-80599-9. Disponível em: https://doi.org/10.1007/978-3-030-80599-9_4. Acesso em: 7 jan. 2022.

VAJJALA, S.; MAJUMDER, B.; GUPTA, A.; SURANA, H. **Practical natural language processing**: a comprehensive guide to building real-world NLP systems. 1. ed. Sebastopol, CA: O'Reilly Media, 2020.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Acesso em: 6 ago. 2021.

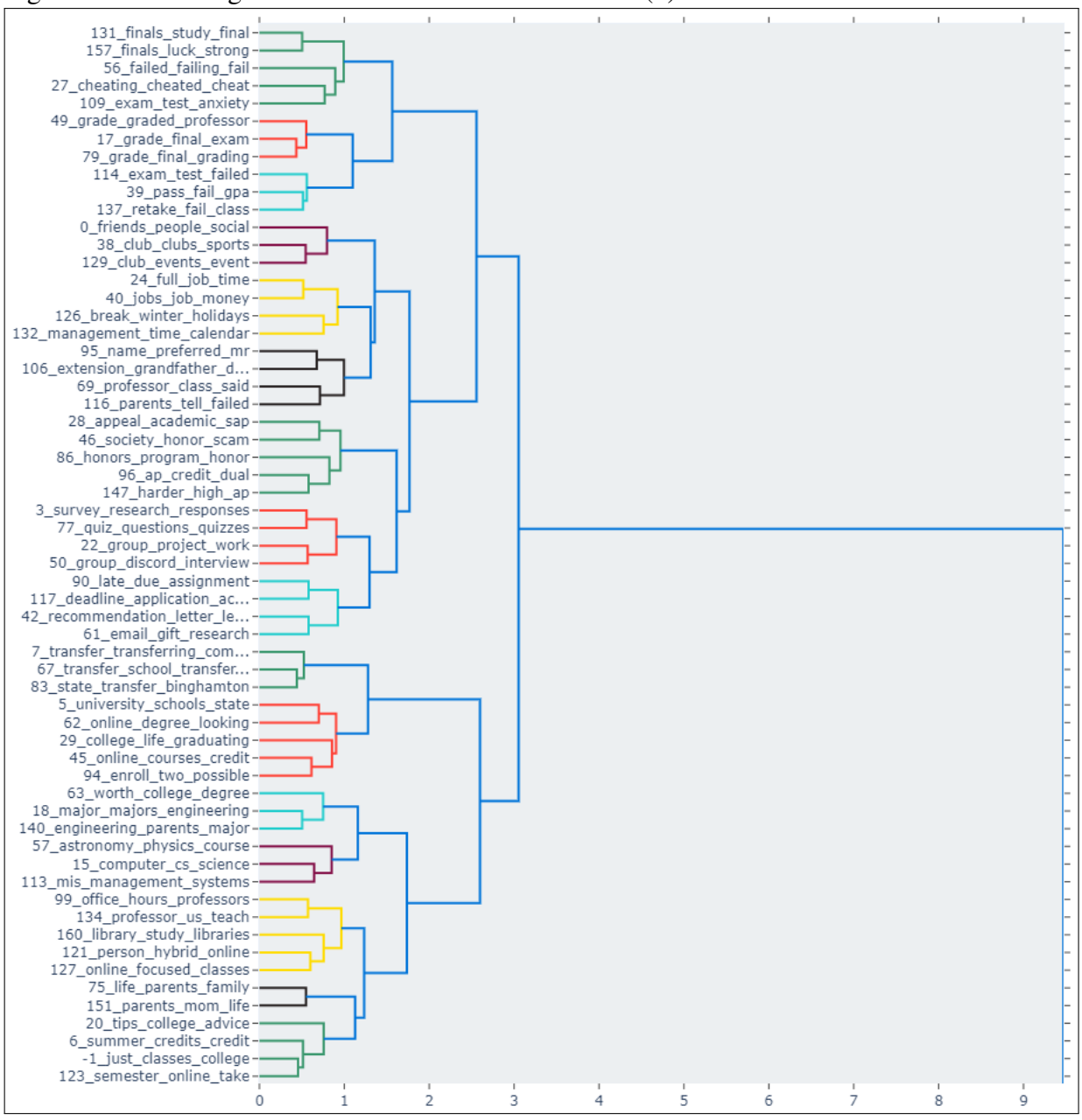
WANG, W.; WEI, F.; DONG, L.; BAO, H.; YANG, N.; ZHOU, M. MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained Transformers. **CoRR**, abs/2002.10957, 2020. Disponível em: https://arxiv.org/abs/2002.10957. Acesso em: 24 dez. 2021.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J.-W.; SANTOS, L. B. da S.; BOURNE, P. E. *et al.* The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, Nature Publishing Group, v. 3, 2016. Disponível em: https://doi.org/10.1038/sdata.2016.18. Acesso em: 24 nov. 2021.

WORLD HEALTH ORGANIZATION. **WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020**. World Health Organization, 2020. Disponível em: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Acesso em: 1 ago. 2021.

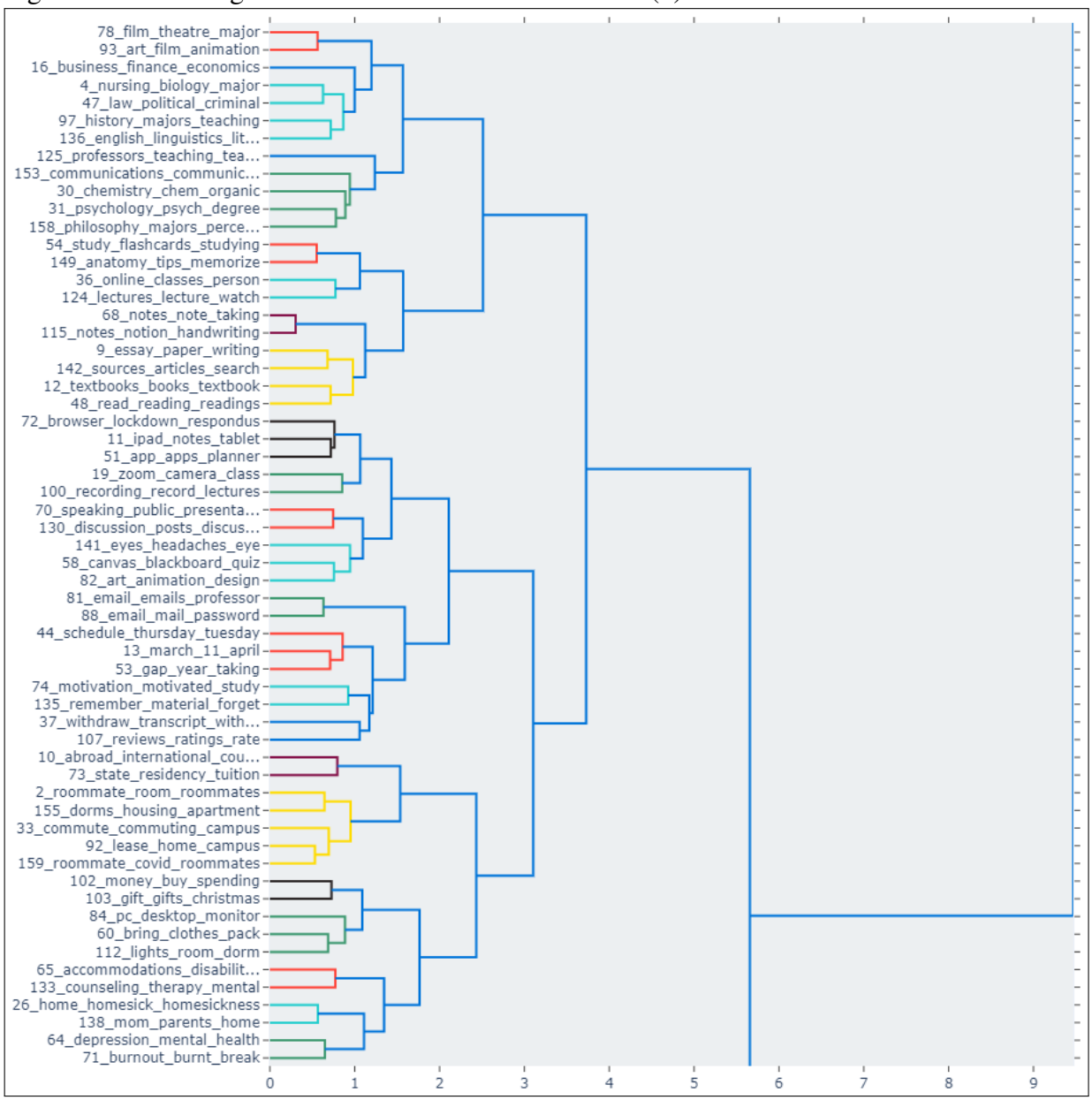
APÊNDICE A – DENDROGRAMAS RESULTANTES DOS CLUSTERS

Figura 28 – Dendrograma mostrando os *clusters* obtidos (1)



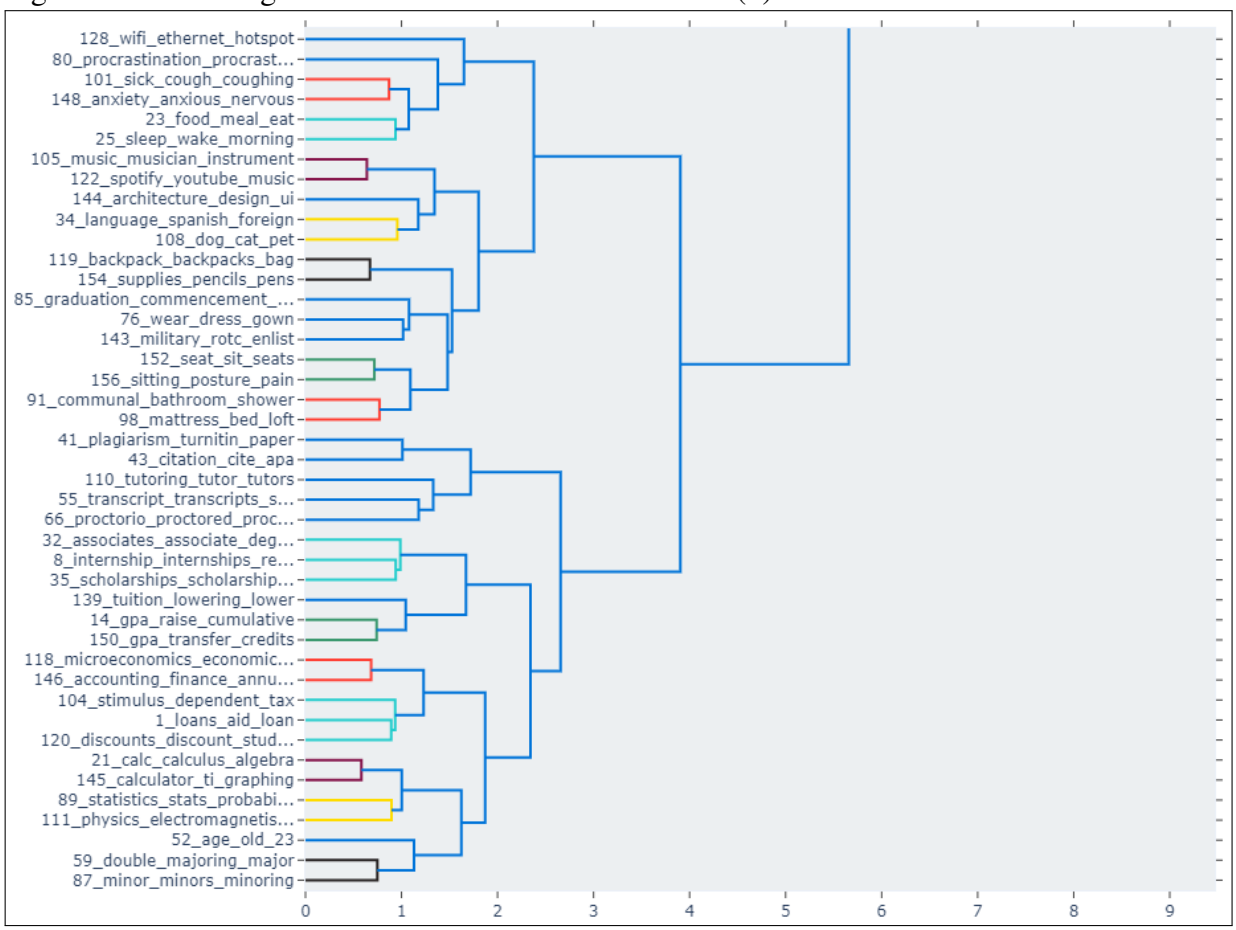
Fonte: elaborado pelo autor (2022).

Figura 29 – Dendrograma mostrando os *clusters* obtidos (2)



Fonte: elaborado pelo autor (2022).

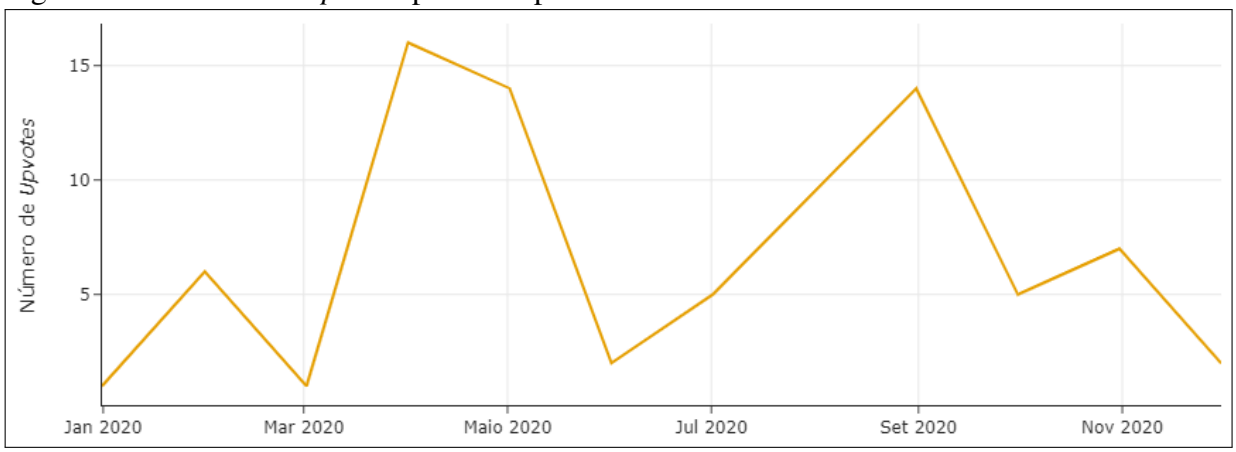
Figura 30 – Dendrograma mostrando os *clusters* obtidos (3)



Fonte: elaborado pelo autor (2022).

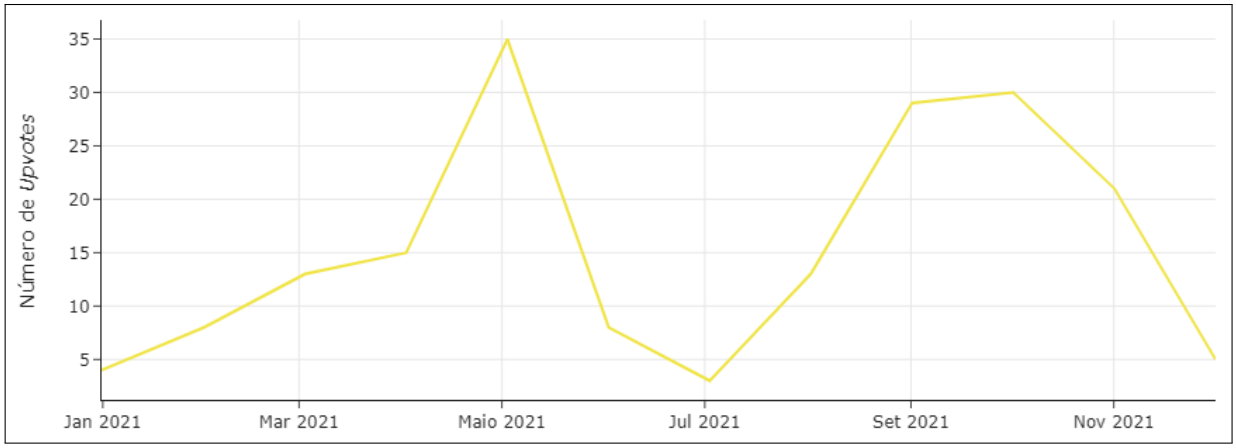
APÊNDICE B – GRÁFICOS DE TÓPICOS AO LONGO DO TEMPO

Figura 31 – Gráfico de *upvotes* para o Tópico 37 durante o ano de 2020



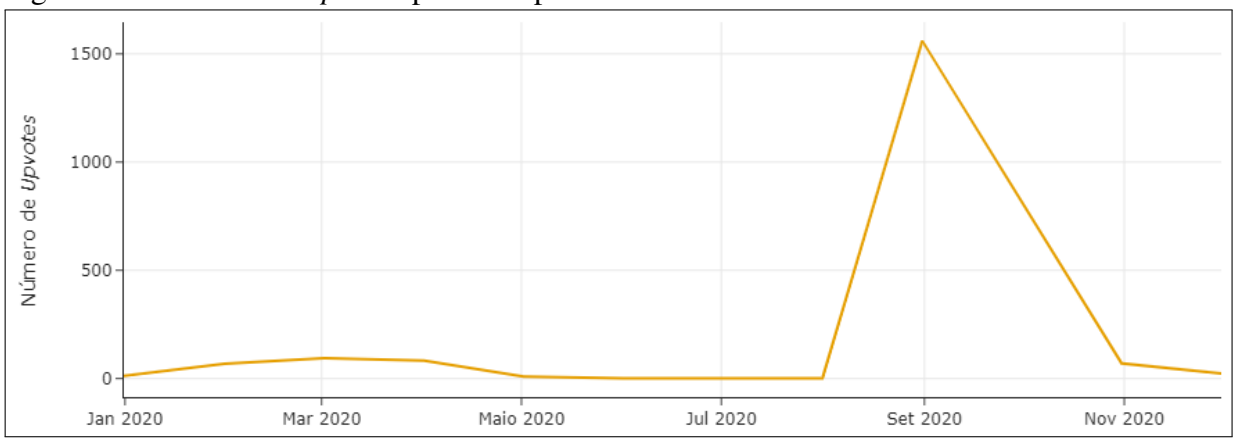
Fonte: elaborado pelo autor (2022).

Figura 32 – Gráfico de *upvotes* para o Tópico 37 durante o ano de 2021



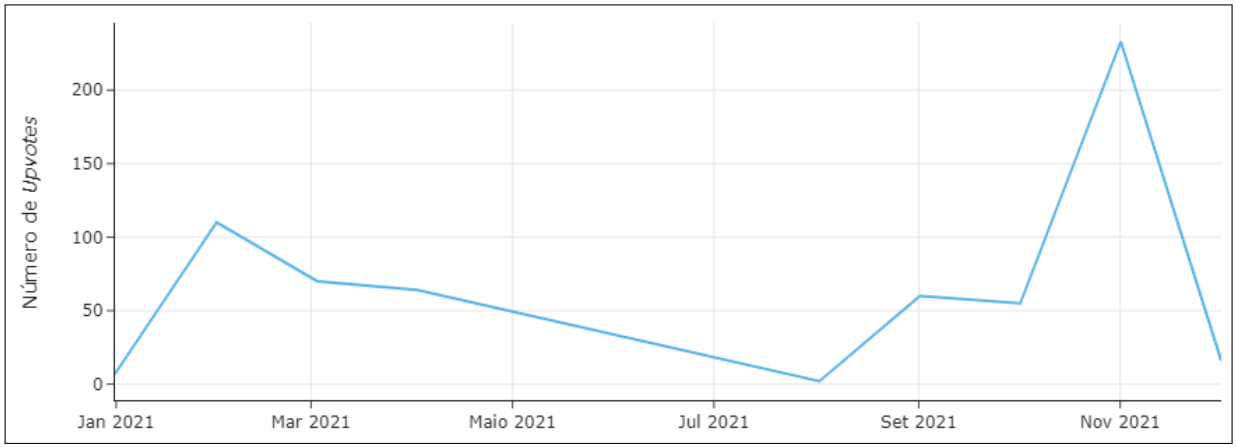
Fonte: elaborado pelo autor (2022).

Figura 33 – Gráfico de *upvotes* para o Tópico 64 durante o ano de 2020



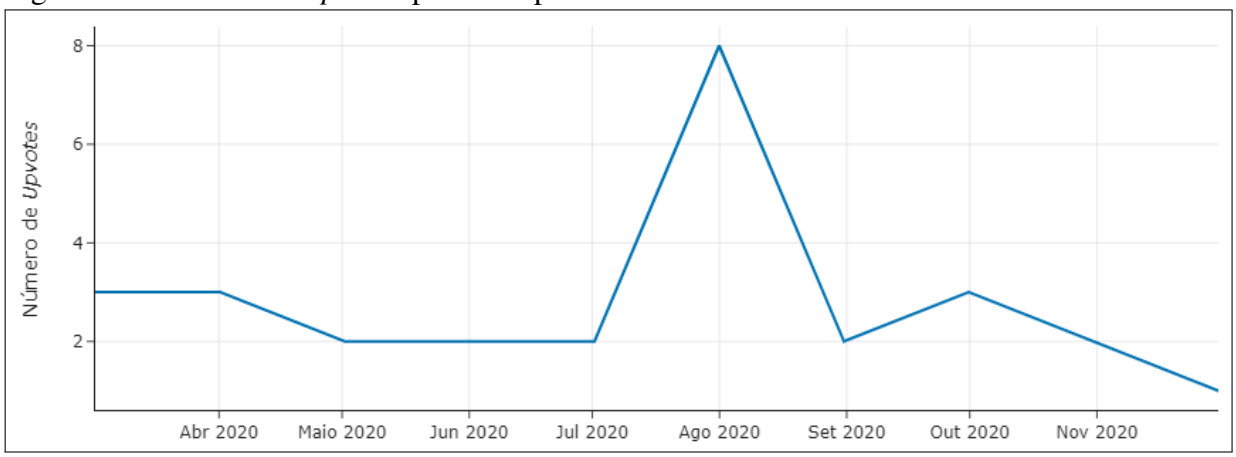
Fonte: elaborado pelo autor (2022).

Figura 34 – Gráfico de *upvotes* para o Tópico 64 durante o ano de 2021



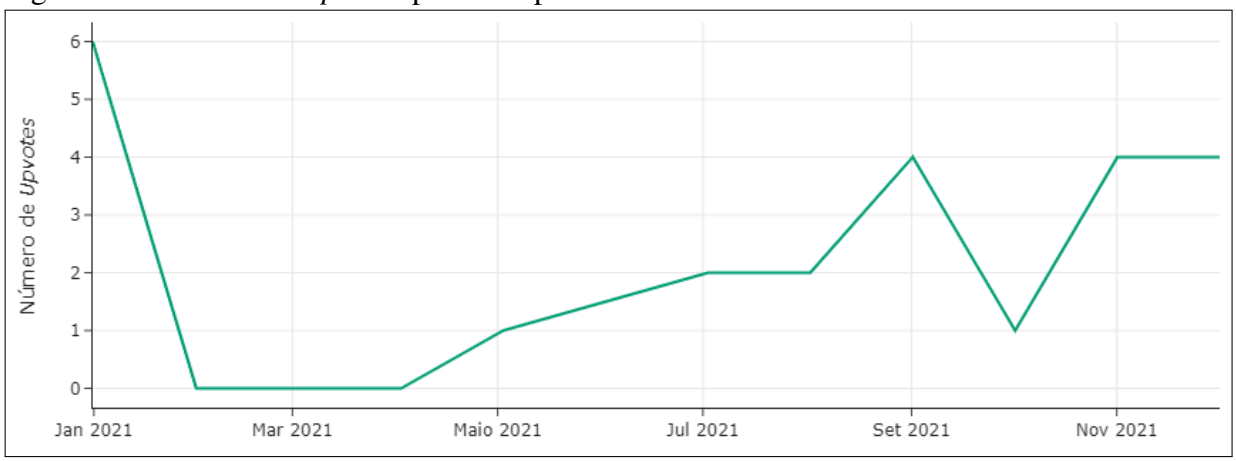
Fonte: elaborado pelo autor (2022).

Figura 35 – Gráfico de *upvotes* para o Tópico 100 durante o ano de 2020



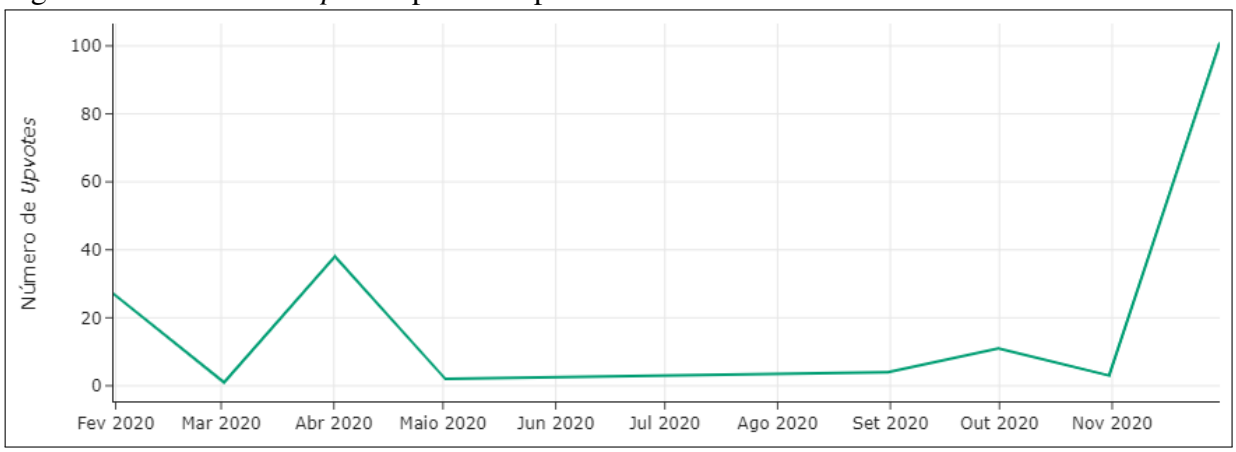
Fonte: elaborado pelo autor (2022).

Figura 36 – Gráfico de *upvotes* para o Tópico 100 durante o ano de 2021



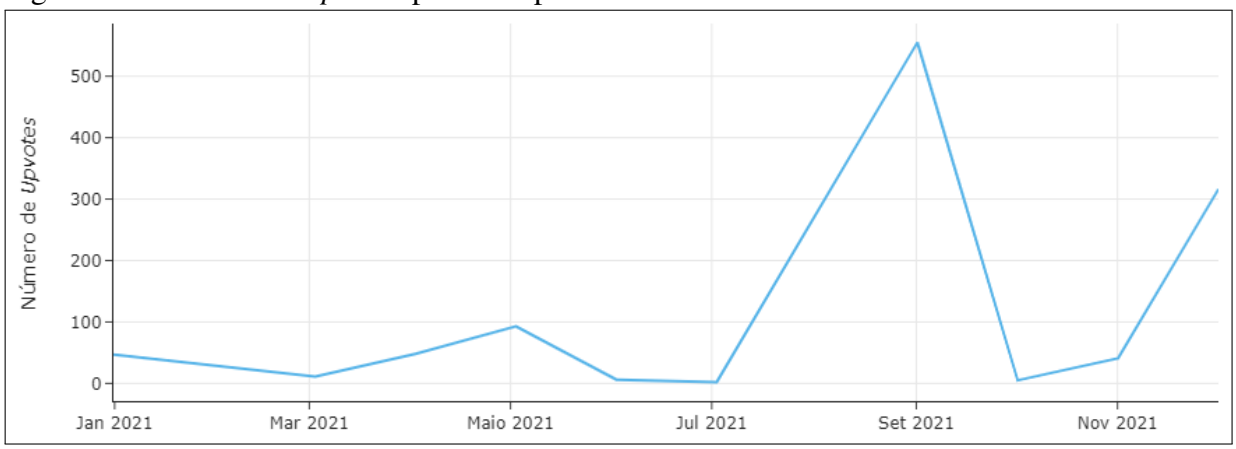
Fonte: elaborado pelo autor (2022).

Figura 37 – Gráfico de *upvotes* para o Tópico 106 durante o ano de 2020



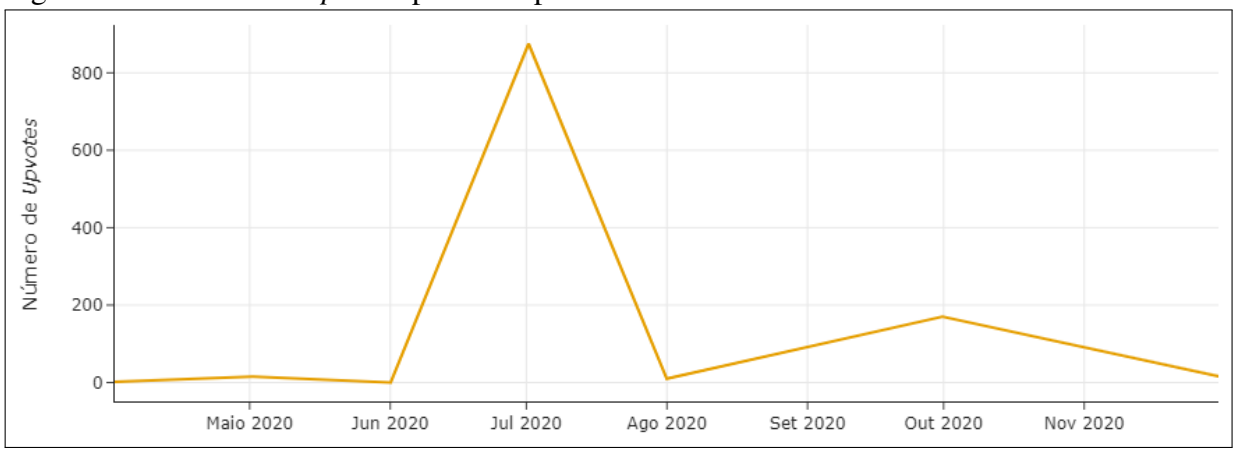
Fonte: elaborado pelo autor (2022).

Figura 38 – Gráfico de *upvotes* para o Tópico 106 durante o ano de 2021



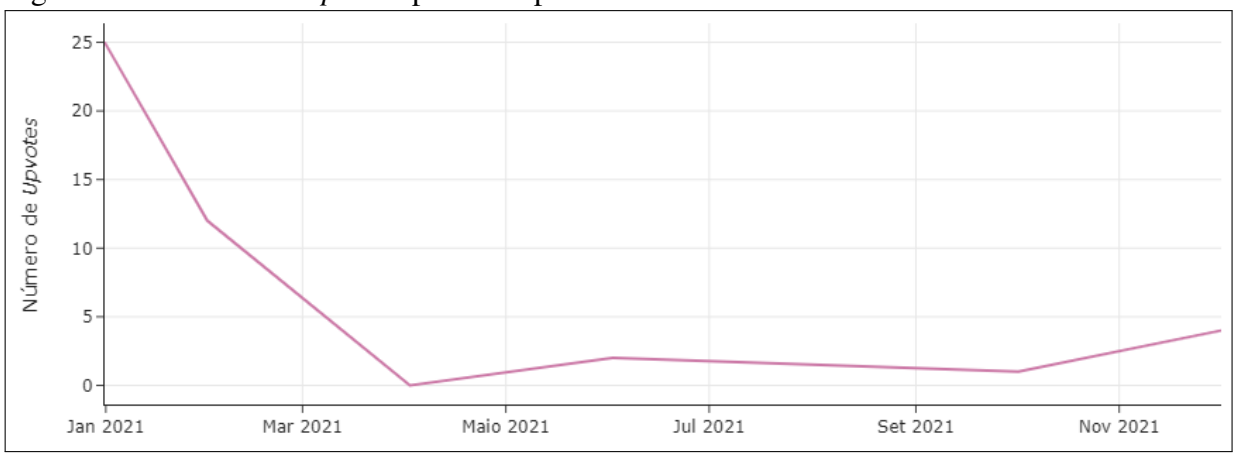
Fonte: elaborado pelo autor (2022).

Figura 39 – Gráfico de *upvotes* para o Tópico 123 durante o ano de 2020



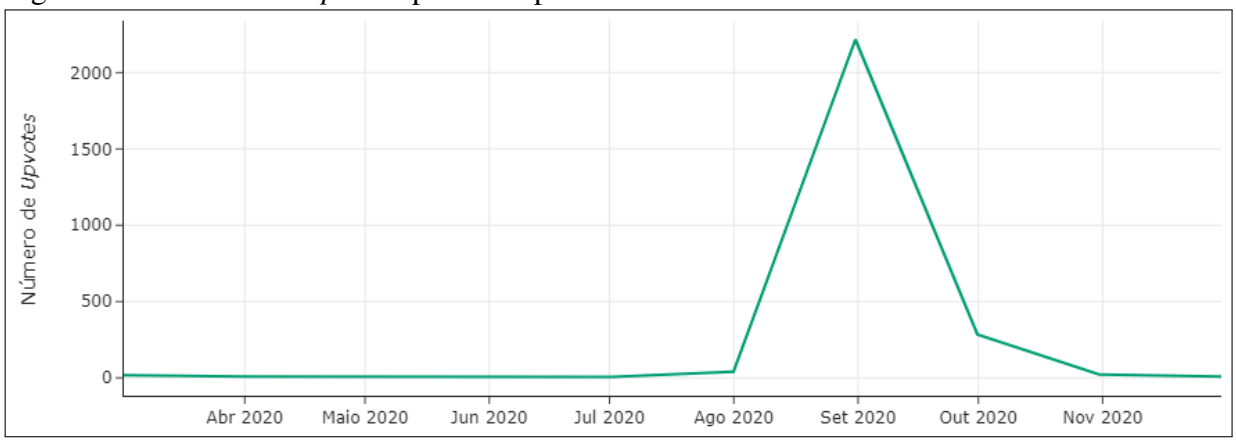
Fonte: elaborado pelo autor (2022).

Figura 40 – Gráfico de *upvotes* para o Tópico 123 durante o ano de 2021



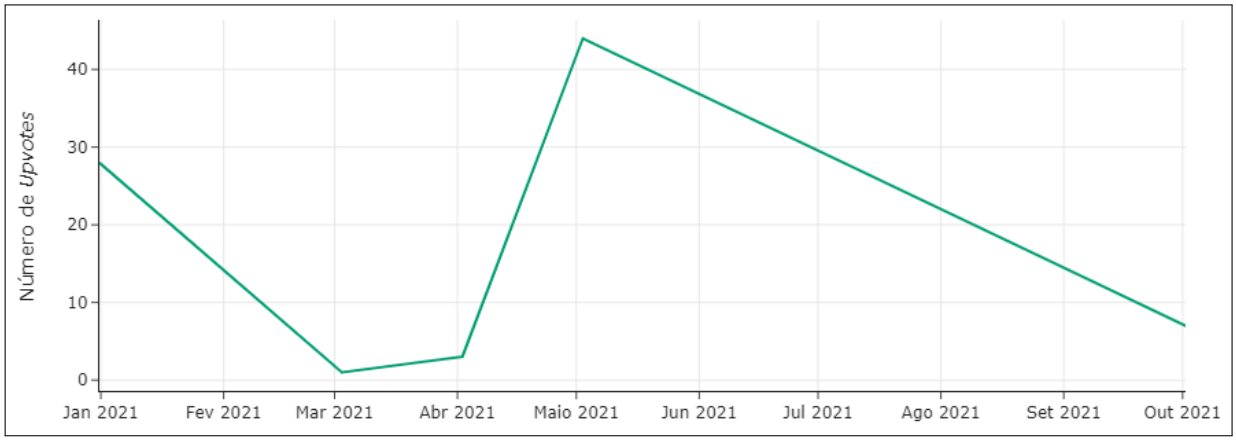
Fonte: elaborado pelo autor (2022).

Figura 41 – Gráfico de *upvotes* para o Tópico 127 durante o ano de 2020



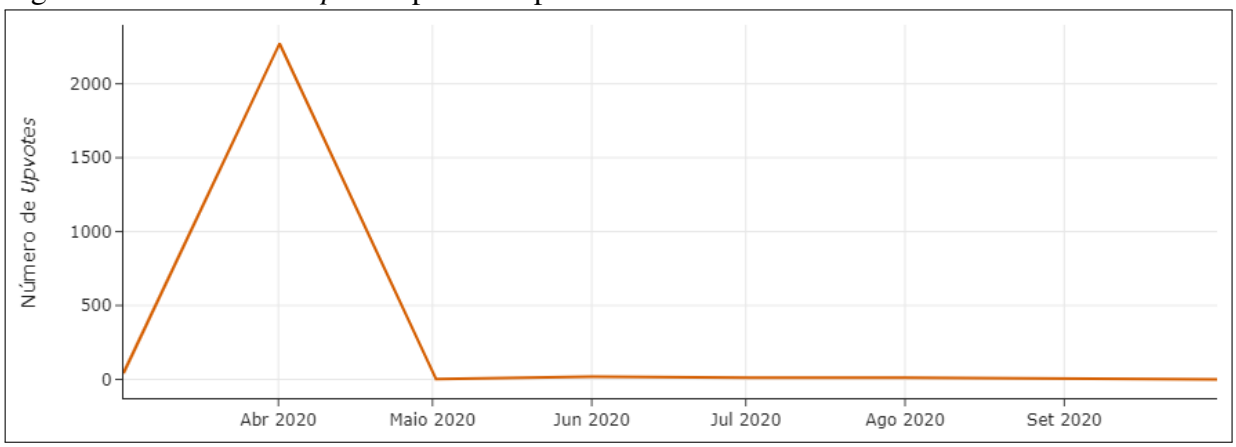
Fonte: elaborado pelo autor (2022).

Figura 42 – Gráfico de *upvotes* para o Tópico 127 durante o ano de 2021



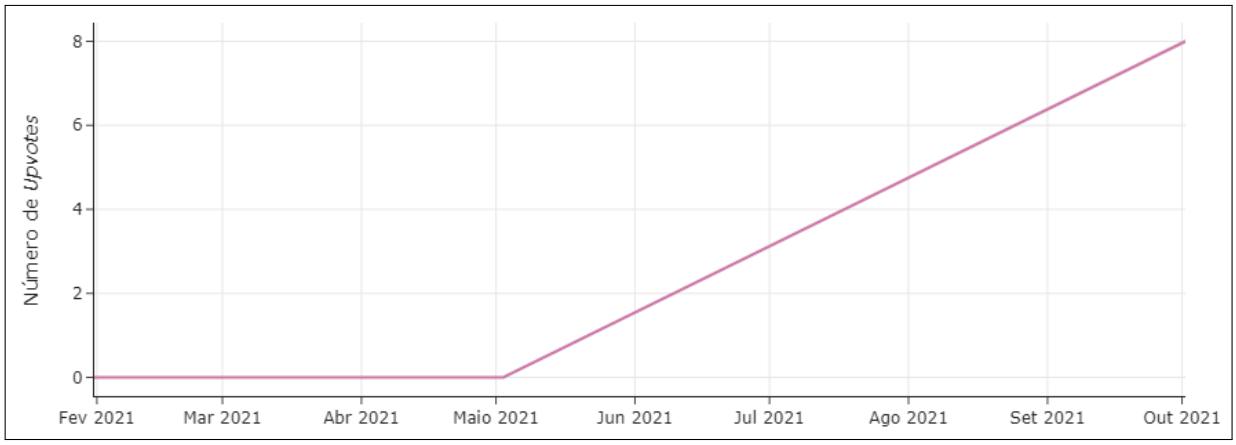
Fonte: elaborado pelo autor (2022).

Figura 43 – Gráfico de *upvotes* para o Tópico 139 durante o ano de 2020



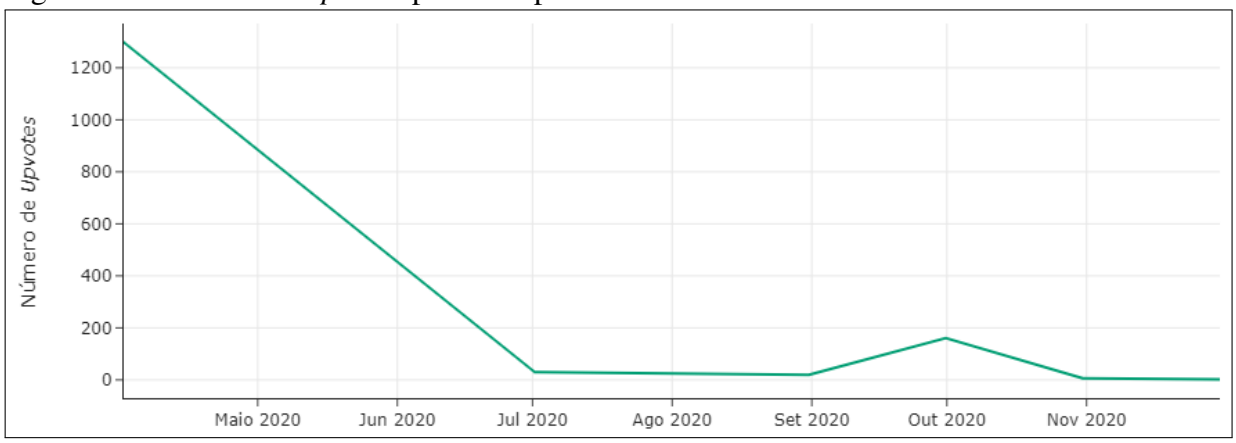
Fonte: elaborado pelo autor (2022).

Figura 44 – Gráfico de *upvotes* para o Tópico 139 durante o ano de 2021



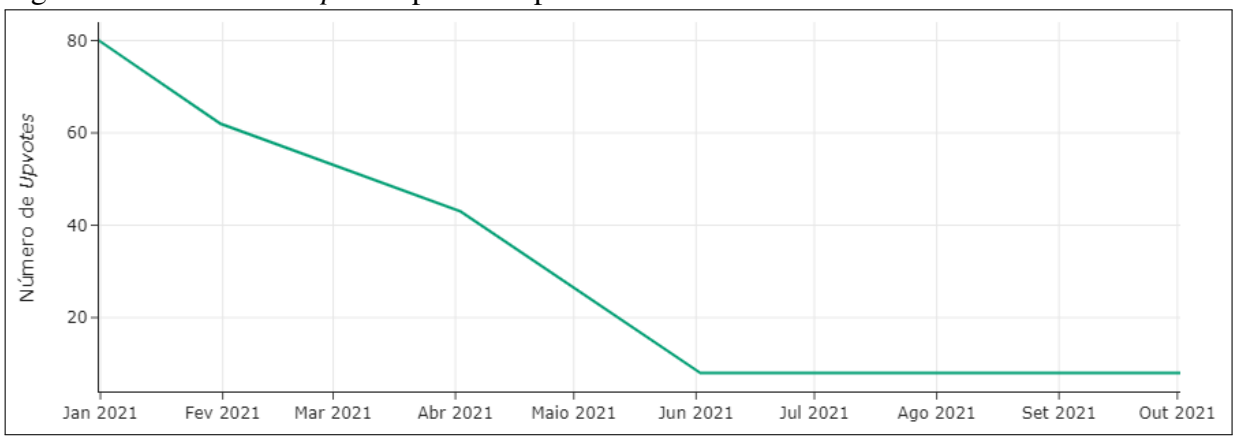
Fonte: elaborado pelo autor (2022).

Figura 45 – Gráfico de *upvotes* para o Tópico 141 durante o ano de 2020



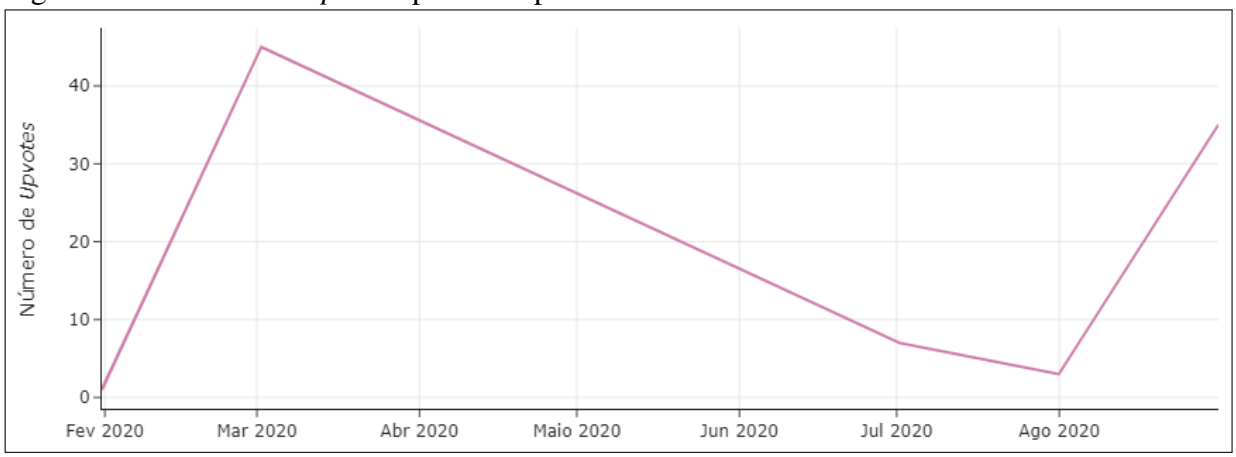
Fonte: elaborado pelo autor (2022).

Figura 46 – Gráfico de *upvotes* para o Tópico 141 durante o ano de 2021



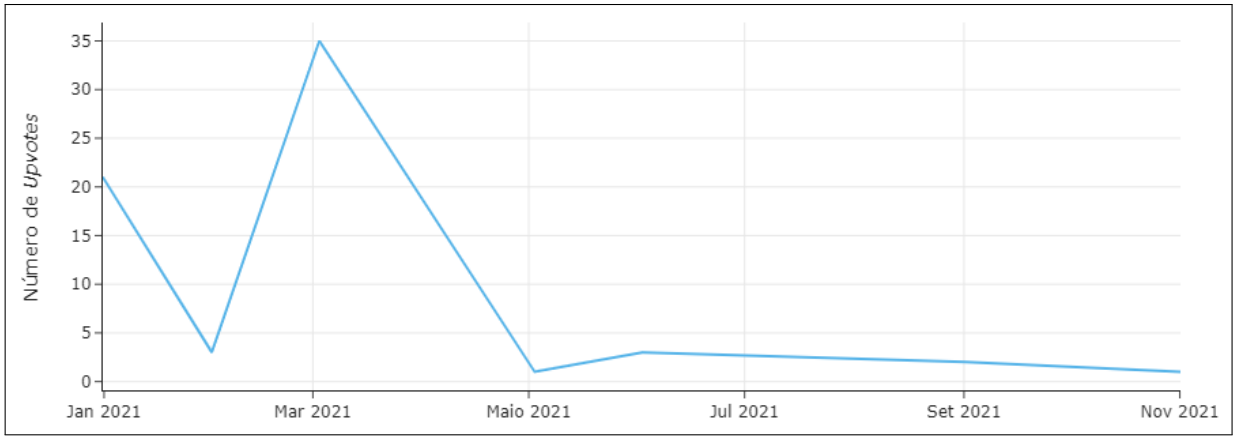
Fonte: elaborado pelo autor (2022).

Figura 47 – Gráfico de *upvotes* para o Tópico 156 durante o ano de 2020



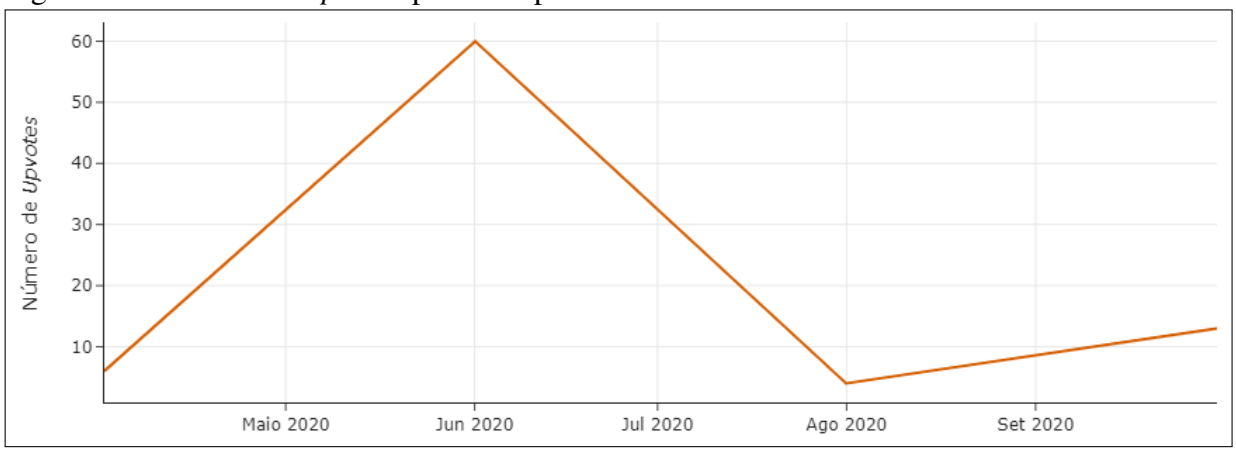
Fonte: elaborado pelo autor (2022).

Figura 48 – Gráfico de *upvotes* para o Tópico 156 durante o ano de 2021



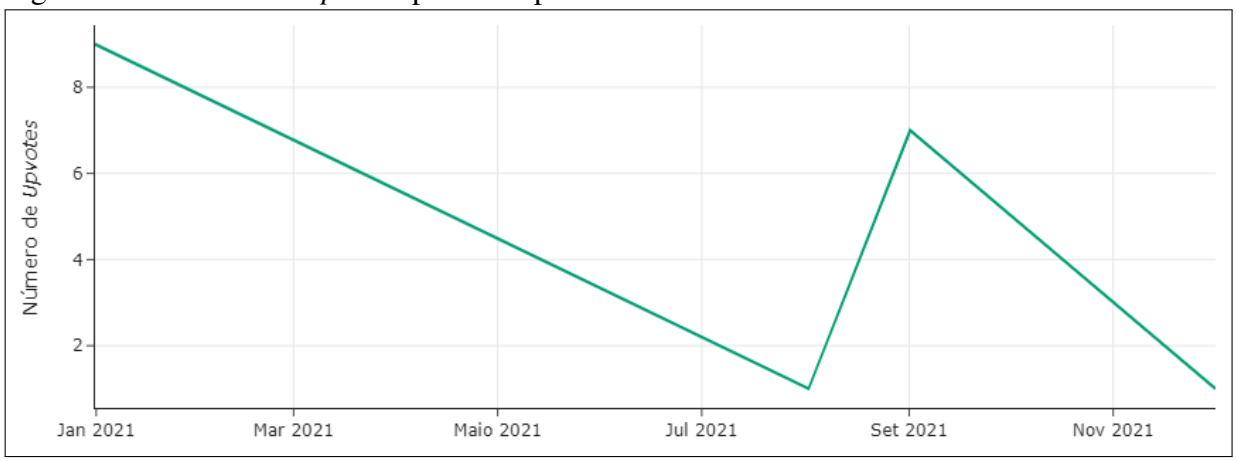
Fonte: elaborado pelo autor (2022).

Figura 49 – Gráfico de *upvotes* para o Tópico 159 durante o ano de 2020



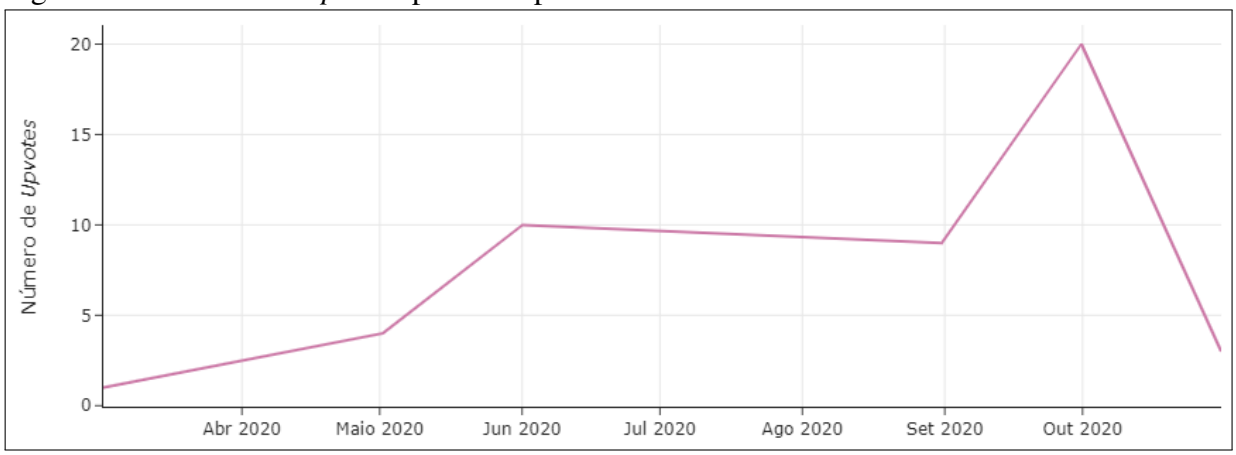
Fonte: elaborado pelo autor (2022).

Figura 50 – Gráfico de *upvotes* para o Tópico 159 durante o ano de 2021



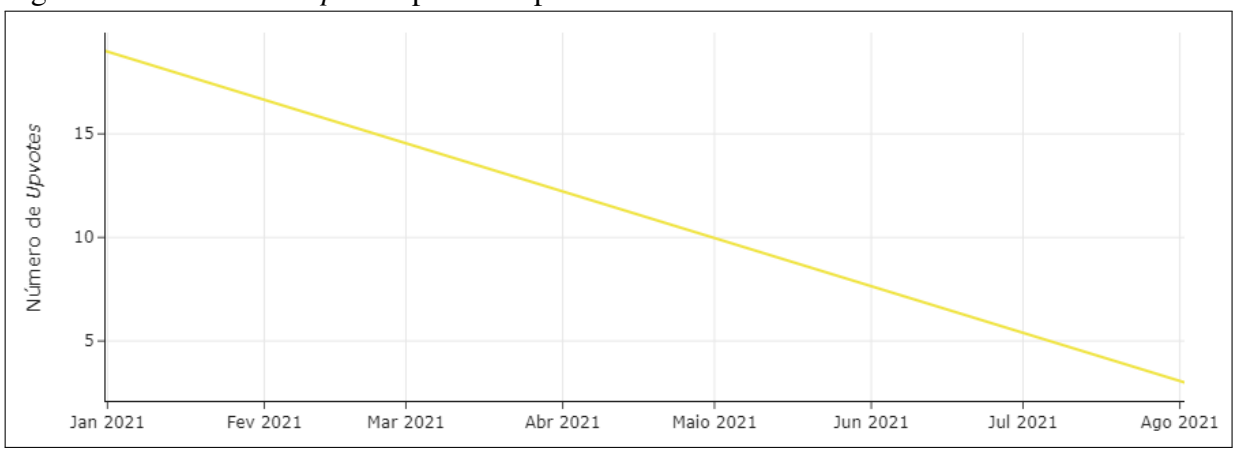
Fonte: elaborado pelo autor (2022).

Figura 51 – Gráfico de *upvotes* para o Tópico 160 durante o ano de 2020



Fonte: elaborado pelo autor (2022).

Figura 52 – Gráfico de *upvotes* para o Tópico 160 durante o ano de 2021



Fonte: elaborado pelo autor (2022).