Uema
UNIVERSIDADE ESTADUAL
DO MARANHÃO

UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS TECNOLÓGICAS
CURSO DE ENGENHARIA DE COMPUTAÇÃO

João Gabriel Pereira Ferreira

São João do Maranhão: Análise de Tópicos e Sentimentos em Tweets e Notícias.

São Luís - Maranhão
2025

João Gabriel Pereira Ferreira

São João do Maranhão: Análise de Tópicos e Sentimentos em Tweets e Notícias.

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Engenharia de Computação pela Universidade Estadual do Maranhão.

Universidade Estadual do Maranhão

UEMA

Orientador: Prof. Dr. Antonio Fernando Lavareda Jacob Junior

São Luís - Maranhão

2025

Ferreira, João Gabriel Pereira
São João do Maranhão: análise de tópicos e sentimentos em tweets e notícias. / João Gabriel Pereira Ferreira. – São Luis, MA, 2025.
56 f

TCC (Graduação em Engenharia de Computação) - Universidade Estadual do Maranhão, 2025.

Orientador: Prof. Dr. Antonio Fernando Lavareda Jacob Junior.

1.Análise de Sentimentos. 2.BERTopic. 3.Modelagem de Tópicos. 4.Processamento de Linguagem Natural. 5.São João. I.Titulo.

CDU: 004.43

Elaborado por Cássia Diniz - CRB 13/910

João Gabriel Pereira Ferreira

São João do Maranhão: Análise de Tópicos e Sentimentos em Tweets e Notícias.

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Engenharia de Computação pela Universidade Estadual do Maranhão.

São Luís - Maranhão, (14 de fevereiro de 2025):

Documento assinado digitalmente
gov.br ANTONIO FERNANDO LAVAREDA JACOB JUNIOI
Data: 24/02/2025 11:14:31-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Antonio Fernando Lavareda Jacob Junior
Orientador - UEMA

Documento assinado digitalmente
gov.br GUSTAVO SOARES SILVA
Data: 24/02/2025 11:24:52-0300
Verifique em https://validar.iti.gov.br

Gustavo Soares Silva
Examinador Interno - UEMA

Documento assinado digitalmente
gov.br PEDRO BRANDAO NETO
Data: 24/02/2025 18:48:08-0300
Verifique em https://validar.iti.gov.br

Prof. Me. Pedro Brandão Neto
Examinador Interno - UEMA

Dedicado à cultura do estado do Maranhão.

Agradecimentos

Agradeço primeiramente a Deus, por ter a oportunidade de desfrutar dessa experiência de aprendizado e comunhão com colegas de graduação e mestres. Agradeço profundamente a meus amigos e familiares por terem me apoiado nos momentos difíceis e por nunca terem deixado que eu desistisse. Esse trabalho é a consolidação de uma meta que, em certo momento da vida, eu ja via como inalcançável e a certeza de que a perseverança vale a pena.

"Cultura não é o que temos, mas o que somos."

(Lévi-Strauss)

Resumo

O crescente percentual da população com acesso à internet no Brasil configura uma maior utilização das mídias sociais como o X e, consequentemente, uma maior disponibilidade de dados textuais gerados por seus usuários. Dessa forma, faz-se interessante colher amostras de texto dos utilizadores e analisar opiniões acerca de determinado tema. Neste trabalho, foi realizada uma análise de sentimentos e modelagem de tópicos utilizando dados do X e do G1 sobre o São João do Maranhão durante os festejos de 2023 e 2024. O objetivo foi identificar a impressão deixada pelo evento na população maranhense. Para isso, foram aplicadas técnicas de *web scraping* para coletar os dados. Ao todo foram coletados 1756 *tweets* e 125 notícias. Após a coleta, os textos passaram por métodos de pré-processamento para permitir a realização da classificação de sentimentos e a modelagem de tópicos. Para identificar os temas abordados nas mídias sociais, utilizou-se o BERTopic, que obteve êxito em identificar os principais temas abordados pelos internautas, tais como: São João da Thay, Bumba Meu Boi, arte e folclore. Quanto à análise de sentimentos, verificou-se que a grande maioria dos *tweets* e artigos publicados pelo G1 foram de caráter positivo, confirmando a ampla satisfação dos usuários em relação ao São João do Maranhão.

Palavras-chave: Análise de Sentimentos; BERTopic; Modelagem de Tópicos; Processamento de Linguagem Natural; São João.

Abstract

The growing percentage of the population with internet access in Brazil means that social media such as X is being used more frequently and, consequently, text data generated by its users is becoming more available. Therefore, it is interesting to collect text samples from users and analyze opinions on a given topic. In this study, sentiment analysis and topic modeling were performed using data from X and G1 about São João do Maranhão during the 2023 and 2024 festivities. The objective was to identify the impression left by the event on the population of Maranhão. To this end, web scraping techniques were applied to collect data. In total, 1,756 tweets and 125 news items were collected. After collection, the texts underwent pre-processing methods to allow sentiment classification and topic modeling. To identify the topics covered in social media, BERTopic was used, which was successful in identifying the main topics covered by Internet users, such as: São João da Thay, Bumba Meu Boi, art and folklore. Regarding sentiment analysis, it was found that the vast majority of tweets and articles published by G1 were positive, confirming the broad satisfaction of users in relation to São João do Maranhão.

Keywords: BERTopic; Natural Language Processing; São João; Sentiment Analysis; Topic Modeling.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	*Application Programming Interface*
AS	Análise de Sentimentos
BERT	*Bidirectional Encoder Representations from Transformers*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
GloVe	*Global Vectors for Word Representation*
GPT	*Generative Pre-trained Transformer*
HDBSCAN	*Hierarchical Density-Based Spatial Clustering of Applications with Noise*
HTML	*HyperText Markup Language*
IP	*Internet Protocol*
LLM	*Large Language Model*
LDA	*Latent Dirichlet Allocation*
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
MLM	*Masked Language Model*
NLTK	*Natural Language Toolkit*
NSP	*Next Sentence Prediction*
PLN	*Processamento de Linguagem Natural*
RNN	*Recurrent Neural Networks*
SBERT	*Sentence-BERT*
TF-IDF	*Term Frequency-Inverse Document Frequency*
UMAP	*Uniform Manifold Approximation and Projection*
VADER	*Valence Aware Dictionary and sEntiment Reasoner*

Sumário

1 INTRODUÇÃO

Com quase dois meses de celebração, o São João do Maranhão, evento denominado como "O Maior São João do Mundo"[1], conta com milhares de atrações, incluindo apresentações de grupos do tradicional Bumba Meu Boi, forró, entre outras diversas manifestações culturais. Em 2023, as festividades receberam um fluxo de 230 mil turistas (Prefeitura de São Luís, 2023). Tal número recebeu projeções de ser superado pelo São João do Maranhão de 2024 que, de acordo com estimativas, contou mais de 250 mil desembarques de curiosos pelo evento cultural no aeroporto Marechal da Cunha Machado na capital (Radio Timbira, 2024). Diante dessas informações, percebe-se que o São João do Maranhão trata-se de um valioso patrimônio que mobiliza diversos segmentos da população, tais como empreendedores, artistas e políticos, e que agrega significativamente na movimentação da economia e turismo local, além do intercâmbio cultural que as festas juninas proporcionam aos seus participantes.

De acordo com Nascimento (2021), o ser humano é um ser de tradição e, por isso, a cultura é intermediadora de transformações sociais. Diante de tal fato, é notório que uma celebração cultural que causa comoção em larga escala, como o São João, deve ser pautada no âmbito da análise sentimental.

Uma plataforma interessante para obter feedbacks acerca da opinião ou até sentimentos por parte dos frequentadores do São João do Maranhão é o X (antigo Twitter). Em 2024, o Brasil possuía aproximadamente 24,3 milhões de usuários do X (World Population Review, 2024; Demand Sage, 2024). Esse número coloca o país como um dos maiores mercados da plataforma, atrás apenas de países como os Estados Unidos e o Japão. Esses dados são relevantes para entender o alcance e o impacto das pautas debatidas no X, considerando a grande base de usuários e o potencial de engajamento que a plataforma oferece.

Uma análise de sentimentos sobre as celebrações juninas no estado do Maranhão pode revelar as impressões deixadas por esses eventos culturais na população. Esse tipo de análise oferece uma visão em tempo real das necessidades emocionais e das reações dos participantes, no caso, os frequentadores dos festejos juninos (DEIKIS, 2024). Com base nos dados obtidos, o governo ou os organizadores das festividades podem obter valiosos *insights*, que possibilitam a adoção de estratégias de marketing mais eficazes ou a melhoria da experiência do público durante os eventos.

Este estudo visa aplicar Processamento de Linguagem Natural (PLN), mineração de texto e análise de sentimentos em bases de dados de *tweets* e notícias sobre o São João

[1] <https://www.ma.gov.br/noticias/governo-do-maranhao-inicia-programacao-2024-do-maior-sao-joao-do-mundo>

do Maranhão para descobrir a opinião predominante dos usuários sobre essa festividade e, dessa forma, contribuir para o campo de estudo da análise de sentimentos, especialmente em contextos culturais específicos, abrindo oportunidades para diversas novas pesquisas e aplicações.

1.1 Justificativa

Dada a dimensão do evento junino, aliada à crescente presença digital da população, observa-se uma maior disponibilidade de dados sobre o tema. A rede social X e portais de notícias como o G1 se destacam como canais essenciais para obter informações sobre as festividades e os tópicos relacionados ao São João do Maranhão (BOLLEN; MAO; ZENG, 2011). Ao combinar análise de sentimentos e modelagem de tópicos com foco nos eventos culturais, busca-se identificar pontos de melhoria, valorizar aspectos positivos, destacar elementos culturais que geram maior comoção no público e incorporar o contexto de transformação digital, conectando cultura e inovação por meio de tecnologias emergentes (MANOVICH, 2017).

Ademais, com este estudo, ocorre a expansão do *corpora* representativo para um diferente domínio: o cultural. Além de explorar gêneros textuais coloquiais — ao exemplo do X — e jornalísticos — ao exemplo do G1. — No que tange à bases de dados classificadas para análise de sentimentos em português, percebe-se que a disponibilidade desses tipos de objetos de estudo é escassa (BRITTO; PACíFICO, 2019). Esse fato se agrava ainda mais quando se refere às mídias sociais e à avaliação de produtos (BENITEZ, 2022). Como prova disso, percebe-se que as ferramentas de Processamento de Linguagem Natural em português frequentemente apresentam desempenho inferior em comparação com o inglês (ALUíSIO et al., 2011).

Dessa forma, o desenvolvimento desse estudo contribui diretamente com a evolução dos trabalhos em PLN, principalmente no que concerne à análise de sentimentos e modelagem de tópicos sobre eventos culturais em mídias sociais. Isso se deve pelo fato de contribuir com a disponibilidade de bases de dados classificadas para o treinamento de modelos que agregam com os estudos de Processamento de Linguagem Natural.

1.2 Objetivos

1.2.1 Objetivo Geral

Este trabalho tem como objetivo geral investigar os sentimentos expressos pelos usuários do X e pelas notícias veiculadas no G1 sobre o São João no estado do Maranhão. Além disso, busca identificar os elementos culturais do evento que geram maior comoção

e atraem mais cobertura midiática através de uma modelagem de tópicos utilizando ferramentas de aprendizado de máquina.

1.2.2 Objetivos Específicos

- Realizar a coleta de dados do X e G1 utilizando ferramentas de raspagem de dados;
- Classificar os dados coletados por meio de uma ferramenta lexical;
- Avaliar a satisfação dos usuários com base na quantidade de dados classificados como positivos;
- Identificar os temas discutidos por meio da modelagem de tópicos utilizando o BERTopic.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão explicitados os conceitos principais que norteiam o desenvolvimento deste trabalho, abordando as áreas de análise de sentimentos, PLN e redes sociais.

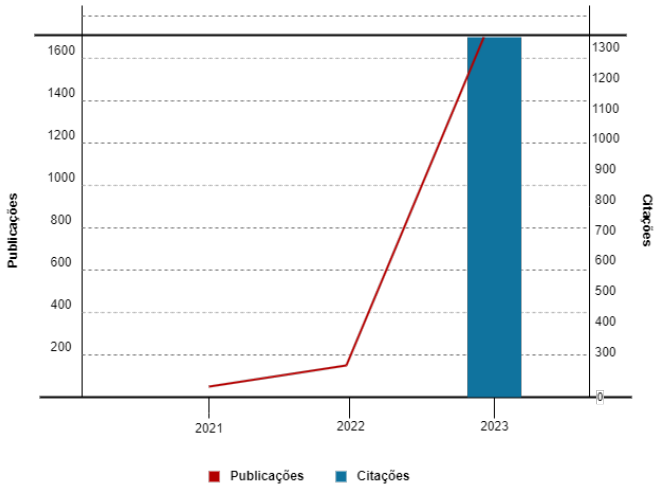
2.1 Análise de Sentimentos

A análise de sentimentos é definida como uma extensão do PLN (LIU, 2020). Essa vertente baseia-se em diversos conceitos fundamentais da linguagem, como a diferenciação entre sentimentos, emoções e a importância da subjetividade e da objetividade nos exemplos textuais. A subjetividade envolve questões pessoais e a expressão de emoções, enquanto a objetividade refere-se a declarações baseadas em fatos, sendo totalmente imparciais (PANG; LEE, 2008).

Diante dessas definições, torna-se necessário relacionar o rápido crescimento das mídias sociais, como redes sociais e blogs informativos. Esses veículos geram uma grande massa de dados textuais acerca de notícias, entretenimento e atividades cotidianas (WANKHADE; RAO; KULKARNI, 2022). Como consequência, observa-se uma massiva quantidade de informações fundamentais para fomentar estudos na área de PLN, especialmente no campo da análise de sentimentos. Essa abordagem é amplamente aplicada em setores como política, marketing e saúde, onde compreender as opiniões dos usuários é essencial para a tomada de decisões estratégicas (MEDHAT; HASSAN; KORASHY, 2014).

Uma representação da quantidade crescente de estudos na área de AS nos últimos anos é apresentada na Figura 1.

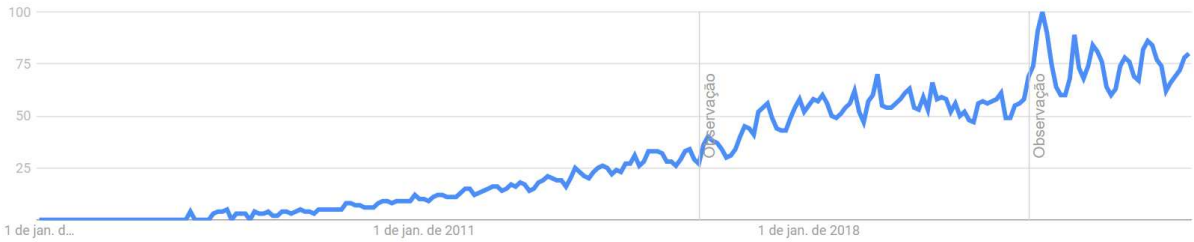
Figura 1 – Evolução da Análise de Sentimentos no âmbito acadêmico



Fonte: Adaptado de Önden et al. (2024).

A Figura 2 representa a evolução na quantidade de buscas por AS no Google de 2004 até dezembro de 2024. Nota-se uma maior incidência de pesquisas de 2022 em diante, evidenciando a atualidade e relevância do tema abordado neste estudo.

Figura 2 – Evolução da busca pelo tema de análise de sentimentos



Fonte: Google (2025).

O funcionamento da AS, é fundamentada em etapas de pré-processamento dos dados textuais, que envolvem a limpeza do texto, mais especificamente a remoção de *stopwords* (WEEDS; WEIR, 2004). *Stopwords* são palavras comuns, como artigos, preposições e conjunções, que são frequentemente removidas durante o processamento de texto, pois geralmente não contribuem para o significado semântico. Sua eliminação ajuda a reduzir o ruído e a melhorar a eficiência em tarefas de mineração de dados e recuperação de informação (MANNING; RAGHAVAN; SCHÜTZE, 2008). Como etapas complementares a essa, tem-se a tokenização e lematização, que são responsáveis por separar o texto em unidades conhecidas como "*tokens*" e reduzir os termos a sua forma base, respectivamente. Dessa forma, o processo de análise dos modelos de mineração de sentimentos processam dados relevantes e consistentes para uma classificação precisa (KOULOUMPIS; WILSON; MOORE, 2011).

Esses dados processados objetos de análise podem ser classificados através da utilização de duas abordagens distintas: uma abordagem linguística e uma abordagem de aprendizagem automática baseada em *Machine Learning* (ML) (SILVA et al., 2020; ALMEIDA; OLIVEIRA; SILVA, 2021).

Embora esses métodos sejam diferentes, todos têm o mesmo objetivo: categorizar uma frase com base em uma emoção encontrada no corpo do texto. Como resultado, a abordagem lexical utiliza léxicos de palavras que já foram classificadas de acordo com polaridades sentimentais (positiva, negativa e neutra), enquanto a abordagem baseada em aprendizado de máquina utiliza soluções matemáticas supervisionadas para treinar modelos de rotulagem de sentimento. Devido ao maior custo computacional para classificação, a abordagem baseada em ML apresenta maiores índices de precisão (SOUZA, 2021).

Para esse estudo será utilizada uma abordagem baseada em léxico com o objetivo de fazer um levantamento acerca dos sentimentos expressos pelos frequentadores do São João do Maranhão (FERNANDES; SILVA, 2019). Isso se deve ao fato de não haver uma

base de dados previamente rotulada representativa do domínio (LIU, 2020).

2.2 Jornalismo e mídias sociais: influência na percepção pública

2.2.1 X

O Twitter, renomeado recentemente como "X", configura-se como uma das principais plataformas de mídia social em contexto mundial, devido ao fato de servir como um canal para a comunicação pública em tempo real. De acordo com World Population Review (2024), o X conta com mais de 580 milhões de usuários ativos mundialmente, e, por consequência disso, dispõe de uma vasta quantidade de dados textuais curtos e altamente variados, tornando-se uma fonte consistente de material para o estudo e análise de emoções acerca de determinado tema a ser debatido (CORREIA et al., 2023). A mineração de emoções no X envolve a avaliação de opiniões e atitudes expressas pelos usuários em seus *tweets*, oferecendo *insights* valiosos para diversas áreas, incluindo marketing, política, saúde pública e previsão de tendências (KITCHIN, 2021).

A análise de sentimentos no X necessita de uma visão geral dos fundamentos teóricos inerentes tanto à plataforma quanto às técnicas de mineração de emoções. A natureza concisa dos *tweets* (textos com limite de 280 caracteres) apresenta desafios únicos, como o uso de abreviações, gírias, neologismos, sarcasmo, ironia e emojis que precisam ser devidamente tratados para uma análise precisa (GUPTA et al., 2021). Com essas devidas precauções tomadas ao usar *tweets* como objeto de estudo, tem-se uma excelente plataforma para coleta de dados para mineração de texto. Além disso, a plataforma X dispõe de ferramentas de interação em seus posts, entre elas se destaca os *retweets*, que são uma maneira de republicar textos de outros usuários, dessa forma ampliando o alcance, hashtags que consistem em palavras-chave precedidas pelo símbolo # usadas para indexação de *tweets* e simplificação da busca.

Com uma taxa de 86,6% de penetração da internet em 2024, o Brasil configura-se como um país com grande maioria da sua população com acesso à internet (Datareportal, 2024). Dessa porcentagem, tem-se que 22,9 milhões de pessoas possuem uma conta no Twitter, que representa uma parcela considerável da população brasileira, tornando o país o sexto maior em números de conta na plataforma (World Population Review, 2024). Diante de tal contexto, é notório o impacto do Twitter na difusão das informações que percorrem os diferentes cantos da nação, levando pautas regionais, como o São João do Maranhão, para curiosos a entender as nuances dessa celebração regional.

2.2.2 G1

O G1 é um portal de notícias brasileiro pertencente ao Grupo Globo, lançado em 2006, com o objetivo de oferecer informações atualizadas sobre diversos temas, como política, economia, tecnologia e entretenimento (G1, 2025). Seu modelo de funcionamento baseia-se na produção de conteúdo jornalístico profissional, garantindo credibilidade e rigor na apuração das informações (MACHADO, 2018).

O G1 segue uma estrutura editorial tradicional, publicando reportagens verificadas antes da divulgação. Diferente das redes sociais, onde qualquer usuário pode gerar conteúdo, o portal mantém um modelo hierárquico de produção de notícias (SILVA, 2020). Além disso, disponibiliza cobertura em tempo real de eventos relevantes, reforçando sua posição como fonte confiável de informações.

A mídia influencia a percepção pública por meio do conceito de "agenda-setting", segundo o qual a imprensa molda a relevância dos temas debatidos na sociedade (MCCOMBS; SHAW, 1972). O G1, devido ao seu alcance e credibilidade, desempenha um papel essencial nesse processo. Além disso, a análise de sentimentos de comentários e interações dos leitores pode revelar a recepção das notícias pelo público, permitindo compreender melhor as reações e opiniões geradas pelo portal (G1, 2025).

Com o avanço digital, o G1 passou a integrar vídeos, podcasts e transmissões ao vivo, facilitando o acesso às informações em diferentes dispositivos (G1, 2025). Embora não seja uma rede social, permite interação do público por meio de comentários e compartilhamento de conteúdos em plataformas como Facebook e X (Twitter), evidenciando a interdependência entre veículos jornalísticos e mídias sociais.

O G1 continua sendo um dos principais portais de notícias do Brasil, adaptando-se ao cenário digital sem perder sua essência jornalística. Seu papel na disseminação de informações e formação da opinião pública é essencial, apesar dos desafios impostos pelas redes sociais e novas formas de consumo de notícias. Ferramentas como análise de sentimentos e modelagem de tópicos oferecem novas perspectivas para entender a recepção do conteúdo e os temas mais relevantes na cobertura jornalística.

2.3 Representação Vetorial de Palavras (*Word Embeddings*)

Para o processo de atribuição de significado às palavras contidas em uma base de dados, faz-se importante tornar o *corpus* textual um objeto legível para as soluções algorítmicas implementadas no PLN. Com base nisso, surge a abordagem de representação vetorial de palavras, método este que transforma palavras em objetos de natureza contínua que podem carregar informações sintáticas e semânticas de maneira mais eficiente (SANDHU, 2024).

Essa abordagem de otimização do texto pode ser realizada de diferentes maneiras, que abrangem desde vetores binários que não capturam a relação semântica das palavras até soluções baseadas em ML que capturam contexto e relação entre termos (SMITH, 2019).

Como exemplo de representação, tem-se a codificação *one-hot* que destaca-se por ser o meio mais simples de representação vetorial. Esse método consiste em um vetor binário com N posições sendo cada palavra única do *corpus* representada por um índice. Nesse modelo, para representar uma palavra específica, o vetor assume valores nulos, exceto no índice representado pela palavra em questão (SUN et al., 2024; KOZINA et al., 2024). Esse método de representação não captura relações semânticas entre as palavras, além de gerar vetores esparsos em caso de vocabulários extensos. Dessa forma, sendo ideal apenas para tarefas menos complexas de PLN, pelo fato de um conjunto de vetores esparsos implicar em um alto custo computacional para o treinamento de modelos de alta complexidade. (GOODFELLOW; BENGIO; COURVILLE, 2016).

O método de representação de palavras em formato vetorial permitiu avanços em aplicações como tradução automática, análise de sentimentos, chatbots e sistemas de recomendação (WU et al., 2016; SOCHER et al., 2013; SERBAN et al., 2016). Tais avanços foram possíveis com a implementação de ferrametas geradoras de vetores densos, denominados *word embeddings*, dessa forma, capturando a relação semântica entre termos através de representações no espaço vetorial multidimensional. (ZHANG et al., 2024).

2.3.1 Representações Vetoriais Não Contextualizadas

2.3.1.1 *GloVe*

O *GloVe*, ou *Global Vectors for Word Representation*, é uma técnica que consiste na transformação de termos em vetores de valor real, facilitando tarefas de PLN através da captura de relações semânticas. Esse método permite a representação de palavras em um espaço vetorial contínuo, ou seja, termos com significados semelhantes têm representações vetoriais no espaço ficam mais próximos. Dessa forma, os procedimentos em aplicações de PLN como agrupamento, recuperação de informações e tradução automática são aprimorados (SARANYA; AMUTHA, 2024). As representações vetoriais do *GLoVe* têm a possibilidade de serem pré-treinados em *corpus* extensos, como 42 bilhões de *tokens*, ou treinados para tarefas específicas, corroborando, assim, sua versatibilidade em modelos de linguagem (MAKARENKOV; SHAPIRA; ROKACH, 2016). O *GloVe* tem sido fundamental no desenvolvimento dos métodos de incorporação textual, abrindo caminho para ferramentas mais avançadas como o BERT que incorpora representações contextualizadas (CAMACHO-COLLADOS; PILEHVAR, 2020).

2.3.1.2 *FastText*

O *FastText* é um modelo leve desenvolvido pelo Facebook capaz de gerar incorporações de termos que detecta informações de subpalavras, desenvolvendo a representação do conhecimento semântico em tarefas de PLN. Essa abordagem trata as limitações dos modelos anteriores, permitindo um melhor manuseio de linguagens morfologicamente ricas e palavras fora do vocabulário, amplicando sua aplicabilidade em vários contextos (BONANDRINI et al., 2023).

Estudos recentes incluem o desenvolvimento do *FastTextSpotter*, o qual utiliza um *back-bone Swin Transformer* para aprimorar a precisão da detecção de textos de cenas multilíngues, demonstrando desempenho superior em conjuntos de dados como ICDAR2015 (DAS et al., 2025).

2.3.2 Representações Vetoriais Contextualizadas

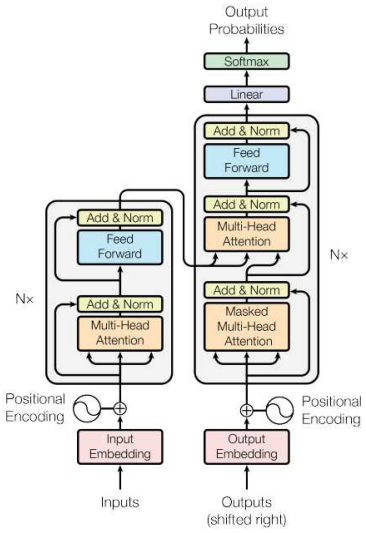
2.3.2.1 *Transformer*

A arquitetura *Transformer*, proposta por Vaswani (2017), é uma das inovações mais importantes no campo de redes neurais, especialmente no PLN. Ela revoluciona a maneira como as máquinas lidam com sequências de dados, como textos, ao abandonar o uso de Redes Neurais Recorrentes (RNNs) e Redes Convolucionais, que eram as abordagens tradicionais para esses problemas, e introduzir um novo mecanismo de atenção. Essa arquitetura se tornou a base para modelos de PLN modernos, como o BERT (*Bidirectional Encoder Representations from Transformers*) e GPT (*Generative Pre-trained Transformer*).

Antes do *Transformer*, os modelos com maior incidência de uso para o processamento de sequências eram as RNNs e as *Long Short-Term Memory* (LSTMs) (LOPYREV, 2015). Embora essas redes tivessem se mostrado eficazes em muitas tarefas de PLN, elas apresentavam dificuldades com o paralelismo durante o treinamento devido à natureza sequencial de seu processamento. Além disso, elas sofriam com o problema de gradientes que desaparecem em situações de processamento de longas sequências, dificultando o aprendizado em tarefas que envolviam longas dependências contextuais (RODRIGUES, 2023).

Ao contrário das RNNs, que processam os dados de forma sequencial, o *Transformer* permite o processamento paralelo de todos os *tokens* de uma sequência de entrada, o que acelera significativamente o treinamento, especialmente em grandes conjuntos de dados. A Figura 3 ilustra a arquitetura do modelo *Transformer*.

Figura 3 – Arquitetura *Transformer*



Fonte: Vaswani (2017).

Um dos principais componentes do modelo *Transformer* é o mecanismo de *self-attention*. Esse mecanismo permite que o modelo avalie a relevância de diferentes palavras ou *tokens* dentro de uma sequência de entrada para cada posição na mesma sequência. Em vez de processar cada palavra isoladamente, o *self-attention* permite que o modelo considere o contexto completo de uma palavra em relação a todas as outras (D'AMICO; NEGRI, 2024; VASWANI, 2017). Em termos simples, cada palavra pode atender a outras palavras, ponderando sua contribuição para a representação final da palavra em questão. Esse mecanismo é calculado através da criação de três vetores — *query* (Q), *key* (K) e *value* (V) — para cada palavra, que são então combinados para gerar uma atenção ponderada. O *self-attention* torna o *Transformer* altamente eficiente, pois ele pode capturar dependências de longo alcance em dados de sequência de forma paralelizada, em contraste com as RNNs, que lidam com os dados de forma sequencial.

No caso do *Transformer*, a diferença entre *embeddings* estáticos e contextuais é de fundamental compreensão. Os *embeddings* estáticos, como os usados em modelos mais antigos (ex: *GloVe*), representam palavras de forma fixa e imutável. Ou seja, cada palavra terá sempre a mesma representação numérica, não importando o contexto em que aparece. Isso configura-se como um problema no que concerne a termos com mais de um significado possível. Já os *embeddings* contextuais, como os do *Transformer*, apresentam maior flexibilidade (VANIA; VULIć; GAšIć, 2020). Eles mudam dependendo do contexto da palavra na frase ou no texto. Por exemplo, a palavra "banco" terá uma representação diferente quando relacionada à uma instituição financeira ou a um banco de rio (ETHAYARAJH, 2019). O *Transformer* gera esses *embeddings* contextuais de forma dinâmica, ajustando-os conforme o processamento de uma sequência, permitindo uma compreensão aprimorada, precisa e adaptativa de acordo com o contexto inserido.

O *Transformer* é altamente escalável e demonstra excelente desempenho em tarefas de PLN em grandes *corpora* de texto, com a capacidade de ser adaptado para tarefas como tradução automática, sumarização e geração de textos (MELLO et al., 2024; NEMANI; VOLLALA, 2022).

2.3.2.2 BERT

O *Bidirectional Encoder Representations from Transformers* (BERT), inicialmente proposto por (DEVLIN et al., 2018), configura-se estado da arte no que concerne a projetos de PLN (RAPARTHI et al., 2021). Para isso, seu desenvolvimento é fundamentado na arquitetura *Transformer* desenvolvida por Vaswani (2017), que foca, principalmente, em pautar seu modelo em partes mais significativas de um texto, com palavras-chave para realizar suas tarefas, por exemplo. Diante disso, tem-se que a principal vantagem do BERT é a facilidade de manuseio, no qual envolve a adição de uma única camada de saída à arquitetura de rede neural para obter modelos de texto que perpassam a imprecisão existente em vários problemas de PLN (KOROTEEV, 2021).

O BERT é pré-treinado utilizando o *Masked Language Model* (MLM) e posteriormente com o *Next Sentence Prediction* (NSP), com o objetivo de entender o contexto bidirecional de um texto. Durante esse processo de treinamento do BERT com MLM, uma palavra aleatória da entrada é substituída por uma máscara [MASK]. Isso proporciona ao BERT a habilidade de prever palavras de acordo com seu contexto (YANG et al., 2024b).

Após essa etapa o NSP verifica se duas sentenças de entrada são adjacentes uma à outra (ROGERS; KOVALEVA; RUMSHISKY, 2021) com o objetivo de ensinar ao BERT as relações semânticas e coesivas entre termos.

Além disso, para possibilitar o funcionamento desse pré-treinamento, o BERT aplica o modelo *WordPiece* de tokenização para as entradas (WU et al., 2016). Essa ferramenta reduz o tamanho do vocabulário necessário e tem a habilidade de lidar com palavras raras ao representá-las como combinações de pedaços de palavras conhecidas. Essa técnica de tokenização divide o texto em subpalavras com objetivo de cobrir vocabulários grandes, unindo, dessa forma, eficiência e a capacidade de lidar com representações complexas de termos raros. Para isso, o *WordPiece* divide o texto em palavras, após isso os *tokens* são divididos em subpalavras, se a subpalavra encontra-se completa no vocabulário, ela continuará sendo um *token*. Entretanto, caso isso não aconteça, a palavra é dividida em partes menores que estão no vocabulário, como ilustrado em Wu et al. (2016):

- ***Word****: Jet makers feud over seat width with big orders at stake.*
- ***WordPieces****: _Jet et _makers _fe ud _over _seat _width _with _big _orders _at _stake*

2.4 Modelagem de Tópicos

Para Abdelrazek et al. (2023) a modelagem de tópicos é uma ferramenta poderosa para a inferência de temas que permeiam vastas coleções de documentos, dessa forma, permitindo a sumarização de grandes grupos de texto. No contexto deste estudo, a aplicação de uma modelagem de tópicos configura-se como essencial, visto que ela permite a identificação de padrões em textos grandes como notícias e, ainda que apresente desafios, em textos pequenos como *tweets* (AMORIM et al., 2022).

2.4.1 TF-IDF

A compreensão do conceito de *Term Frequency-Inverse Document Frequency* (TF-IDF) é fundamental para o entendimento do fluxo de modelagem de tópicos, visto que fundamenta-se na relevância de termos em documentos. Segundo Sun, Zuo e Wang (2023), a representação TF-IDF apresenta altas taxas de precisão, evidenciando sua utilidade prática. Para isso, esse método calcula para cada termo de um documento, um valor embasado na frequência da palavra naquele documento (*Term Frequency*) e no inverso da porcentagem de documentos em que aquele termo aparece (*Inverse Document Frequency*), dessa maneira a palavra com o maior valor obtido pelos cálculos realizado define melhor o documento (LILLEBERG; ZHU; ZHANG, 2015).

O TF-IDF é fundamentado em dois componentes principais:

TF (*Term Frequency*): feito para medir a frequência de um termo em um documento específico.

$$\mathrm{TF}\,(t,d) = \frac{f_{t,d}}{N_d} \tag{2.1}$$

O termo (t,d) representa a quantidade que um termo t aparece em um documento d. Enquanto N_d representa a quantidade de termos em um documento d.

IDF (*Inverse Document Frequency*): mensura o quanto uma palavra é rara no documento.

$$\mathrm{IDF}\,(t) = \log\left(\frac{N}{1+df\,(t)}\right) \tag{2.2}$$

O termo $df\,(t)$ representa a quantidade de documentos que contêm o termo t e N representa a quantidade de documentos.

Utilizando desses dois artifícios, o TF-IDF entrega um valor resultante na relevância de um termo para um documento:

$$\text{TF-IDF} = \mathrm{TF}\,(t,d) \times \mathrm{IDF}\,(t) \tag{2.3}$$

2.4.2 BERTopic

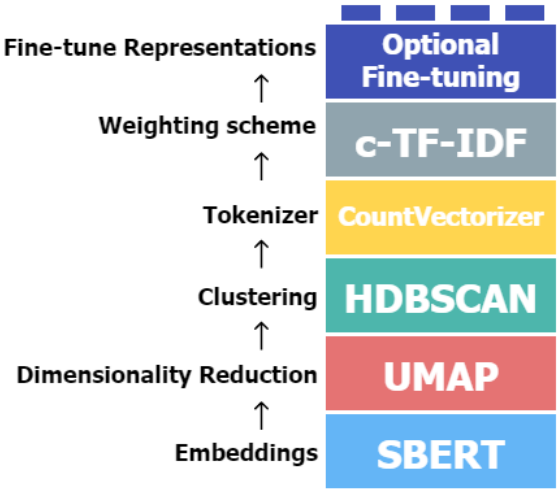
A sumarização de temas pode ser realizada utilizando algoritmos probabilísticos, que aplicam conceitos estatísticos ou até mesmo modelos de aprendizado de máquina. Essas técnicas são empregadas para identificar termos frequentemente associados e, a partir deles, sugerir tópicos. Uma abordagem clássica nesse contexto é o *Latent Dirichlet Allocation* (LDA), que fornece uma visão mais geral dos tópicos. Por outro lado, uma alternativa moderna e mais avançada baseada em aprendizado de máquina é o BERTopic (GROOTENDORST, 2022). Essa técnica combina *embeddings* de linguagem e algoritmos de *clustering* para criar tópicos de forma mais precisa e interpretável, oferecendo maior detalhamento e adaptabilidade ao conteúdo analisado (LOPES; BROTAS; MASSARANI, 2023).

Embora possuam objetivos semelhantes, as duas abordagens têm aplicações em cenários distintos. O LDA é mais indicado para *corpora* textuais de linguagem acessível, enquanto o BERTopic apresenta uma compreensão semântica mais aprofundada, sendo ideal para a análise de textos com linguagem complexa e voltados para situações específicas (SUNG-SU; HOE-CHANG, 2023).

O BERTopic foi escolhido devido à eficiência do modelo BERT em tarefas relacionadas a dados textuais, como classificação e entendimento de contexto, graças ao seu mecanismo de análise bidirecional (GROOTENDORST, 2022). Além disso, a técnica demonstra alto desempenho na análise de textos de tamanhos variados, abrangendo tanto conteúdos longos quanto curtos (AL-QURISHI, 2023; AMORIM et al., 2022). Essas características evidenciam que o BERTopic é aplicável à base de dados deste trabalho, atendendo às necessidades da tarefa proposta.

A Figura 4 tem o objetivo de ilustrar as tecnologias padrões utilizadas no BERTopic desde a entrada dos dados até a apresentação dos tópicos. As etapas detalhadas de cada processo serão apresentadas posteriormente.

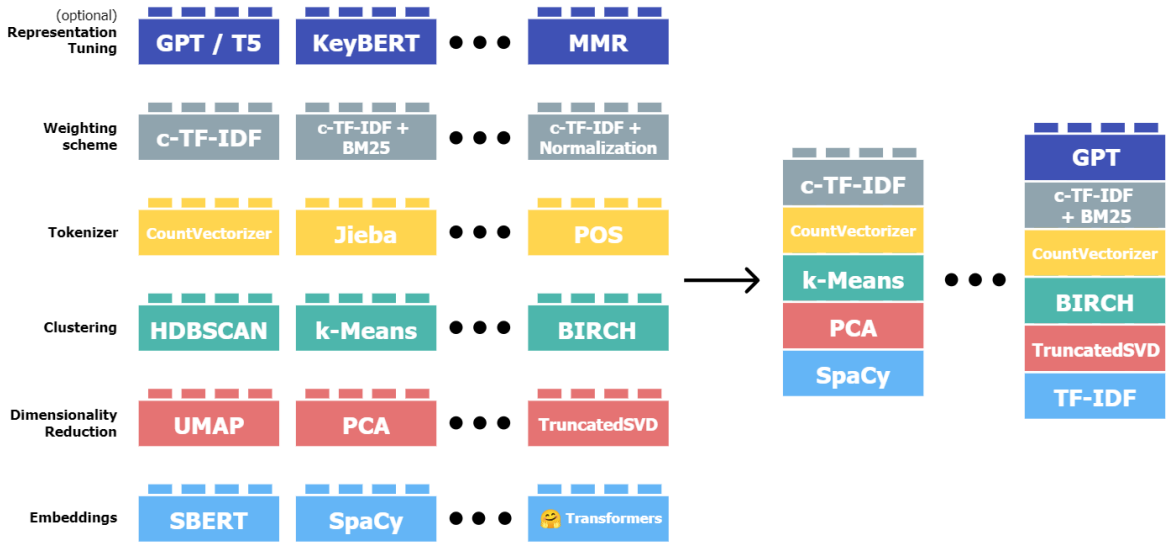
Figura 4 – Arquitetura BERTopic



Fonte: Grootendorst (2022)

Ademais, o BERTopic trata-se de uma ferramenta modular, ou seja, em cada etapa do processo de modelagem de tópicos, há a possibilidade de ajustar o modelo utilizado em cada estágio do algoritmo. A Figura 5 representa as etapas do processo de modelagem, bem como ferramentas alternativas às padrões implementadas pelo BERTopic.

Figura 5 – Modularidade BERTopic



Fonte: Grootendorst (2022)

2.4.2.1 Extração de Características

Iniciando a representação de documentos, é necessário converter os textos em uma forma vetorial que possa ser processada por algoritmos de clusterização. Para isso, utiliza-se o SBERT (Sentence-BERT), um modelo baseado em *Transformers* projetado para detectar semelhanças semânticas entre diferentes textos. Conforme demonstrado por Zhai, Wang e Zhao (2024), o SBERT é altamente eficaz na classificação de tópicos ao empregar aprendizado contrastivo, alcançando alta precisão em conjuntos de dados de benchmarks.

Devido à sua capacidade de gerar *embeddings*, a formação de tópicos pelo BERTopic utiliza o contexto de cada documento. Nesse processo, aplica-se o algoritmo *Uniform Manifold Approximation and Projection* (UMAP), que realiza a redução de dimensionalidade dos *embeddings* textuais enquanto preserva as estruturas local e global para melhor representação semântica dos dados (YANG et al., 2024a).

2.4.2.2 Agrupamento

Na etapa de clusterização, o BERTopic emprega o algoritmo *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN), um método baseado em densidade hierárquica. Segundo Stewart e Al-Khassaweneh (2022), o HDBSCAN agrupa pontos próximos ou localizados em regiões densas do espaço e constrói uma árvore hierárquica de agrupamentos. Uma característica importante do HDBSCAN é a capacidade de determinar automaticamente o número de *clusters*, eliminando a necessidade de definir esse valor como um hiperparâmetro (STROBL et al., 2021).

2.4.2.3 Representação de Tópicos

Por fim, o BERTopic utiliza o c-TF-IDF, uma versão modificada do TF-IDF adaptada para múltiplas classes. De acordo com Grootendorst (2022), essa técnica consolida os documentos de cada classe em um único texto, facilitando a identificação de palavras-chave representativas dos tópicos.

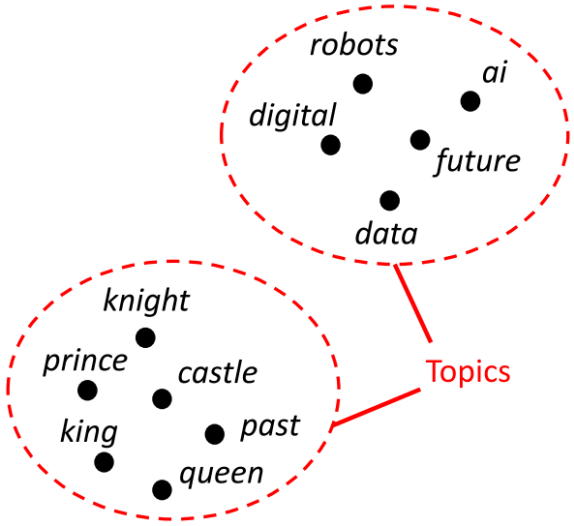
O c-TF-IDF consiste na relevância de termos em categorias em vez de documentos individuais. Ao colocar em evidência os termos que são relevantes e não partilham de categorias diferentes, o CTF-IDF colabora para tarefas de classificação de texto e agrupamento em *clusters* ao melhorar o desempenho de modelos de aprendizado de máquina.

- **CTF**: Calcula o número de vezes em que um termo aparece em documentos com uma categoria específica.

- **IDF**: Mede a relevância de um termo baseado na quantidade de categorias em que ele aparece.
- **CTF-IDF**: Combina o CTF e o ICF para atribuir um valor que reflita a importância de um termo para uma categoria em relação a todas as categorias (Hung *et al.* 2021).

Termos com maior pontuação refletem em maior importância para seu respectivo tópico, garantindo que o termo escolhido descreva o seu *cluster* correspondente. A Figura 6 ilustra a relação entre termos e tópicos.

Figura 6 – Representação de termos e tópicos no espaço vetorial



Fonte: Goodfellow, Bengio e Courville (2016)

3 TRABALHOS RELACIONADOS

Nesta seção, serão apresentados trabalhos que contribuíram para a compreensão do tema proposto por esta dissertação. A revisão de estudos relacionados permite situar a pesquisa dentro do contexto científico, apontando os métodos utilizados recentemente no contexto da mineração de sentimentos, resultados e lacunas dos trabalhos anteriores. Em específico, serão considerados os estudos de Cirqueira et al. (2017), Melo, Figueiredo et al. (2021) e Freitas, Ladeira e Caetano (2024).

Em Cirqueira et al. (2017), denota-se que o Brasil possui uma das populações mais ativas nas mídias sociais online e destaca, também, o fator comunicação como um dos mais relevantes no que concerne à população brasileira. Diante disso, os autores deste estudo realizaram uma avaliação de performance para diferentes técnicas de análise de sentimentos. A temática deste trabalho foi norteada por questões como *web bullying* e administração de instituições públicas, tópico justificado pela sua grande repercussão e alta tendência a gerar comentários e polaridade nas opiniões. Assim, para esse estudo, foram retirados dados do Facebook e do X/Twitter, duas das mídias sociais mais utilizadas no Brasil.

Para classificar tais comentários foi utilizado o sistema OpinionLabel para credibilizar ainda mais o rótulo atribuído a cada comentário (DOUGLAS et al., 2017). Por fim, as postagens foram submetidas à tradução para o inglês para que, dessa forma, os algoritmos de análise de sentimentos obtivessem melhor performance. Como resultado desse estudo, constatou-se que a desempenho dos algoritmos utilizados para análise de sentimentos em português é visivelmente menor em comparação às em inglês. Dessa forma, o trabalho concluiu que faz-se necessária a implementação de ferramentas exclusivas para trabalhar com o português visto que a tradução pode resultar em perda de informação do texto original.

No estudo realizado por Melo, Figueiredo et al. (2021) foi feita uma abordagem de modelagem de tópicos e análise de sentimentos acerca da COVID-19 em duas plataformas. Sendo estas o X e o UOL, dois dos principais canais para a disseminação de informações na web. Para essa dissertação, os dados foram coletados através de técnicas de *web scrapping* como a biblioteca TwitterScrapper para raspagem das informações na plataforma. Ao final deste procedimento, obtiveram-se cerca de 18 mil notícias e mais de 1 milhão de postagens no X acerca do tema COVID-19.

Com os dados coletados, o pré-processamento faz-se essencial para a aplicação dos algoritmos de análise sentimental e modelagem de tópicos. A biblioteca *Natural Language Toolkit* (NLTK) foi utilizada para fins de remoção de *stopwords*, menções, URL's, emotes,

pontuações e espaços extras em branco. Para a modelagem de tópicos foi utilizado o modelo de ML denominado *Machine Learning for Language Toolkit* (MALLET) que possui uma implementação de *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003). Essa técnica permite identificar grupos de termos relacionados, um total de 20 grupos com 10 termos foram criados nesse estudo para cada plataforma analisada. Como exemplo dos tópicos encontrados pela ferramenta, pode-se citar "prevenção e controle", "política" e "casos confirmados". Para a análise de sentimentos, utilizou-se a API gratuita do Google Tradutor para traduzir os dados coletados para o português e, dessa forma, aplicar a ferramenta *Valence Aware Dictionary Tool* (VADER) para calcular o grau de negatividade e positividade nos textos (HUTTO; GILBERT, 2014).

Os resultados obtidos neste trabalho demonstram que ambas as plataformas UOL e X/Twitter preocupam-se em cobrir os mesmos tópicos sobre a COVID-19. Em relação aos sentimentos expressos nas duas mídias, prevaleceu o sentimento negativo, sendo essa emoção motivada pela insatisfação perante medidas governamentais em relação à pandemia.

Em Freitas, Ladeira e Caetano (2024), objetivou-se a criação de uma ferramenta que identificasse sinais de comportamento depressivo nos usuários do X/Twitter com base nos textos das suas postagens, com foco de ampliar os estudos de análise de sentimentos na língua portuguesa utilizando fatores comportamentais e contextuais.

A metodologia de projeto utilizada foi a *Cross-Industry Standard Process for Data Mining* (CRISP-DM) e envolveu a coleta dos dados através da biblioteca Python chamada Snscrape. Para classificar os *tweets*, foram utilizados 15 atributos que incluem características como pronomes, frequência de termos depressivos, emojis e menções a medicamentos antidepressivos. As etapas de pré-processamento são semelhantes aos trabalhos já citados e, após o preparo dos dados, foram gerados vetores de características que consideram métricas estatísticas diárias, como média e variância, para cada atributo. O modelo foi treinado com dados rotulados, aplicando aprendizado supervisionado para distinguir entre grupos depressivos ou não.

Portanto, como os experimentos demonstraram, os novos atributos que os autores conjecturam contribuíram para uma precisão aprimorada do modelo para destacar os sinais de depressão. Além disso, os dados revelaram diferenças significativas no uso do X/Twitter entre os usuários antes e depois do início da pandemia, porém não encontraram evidências estatísticas significativas para afirmar que a pandemia da COVID-19 tenha ocasionado em um aumento nos casos de depressão. A pesquisa reitera o enfoque crítico por trás dos métodos automatizados que podem ajudar a encontrar condições depressivas no início em mídias sociais.

Uma síntese acerca dos trabalhos correlatos alinhados à análise de sentimentos em mídias sociais é apresentada na Tabela 1.

Tabela 1 – Resumo dos trabalhos correlatos analisados.

Autor/Ano	**Origem/Obtenção dos Dados**	**Classificação dos Dados**	**Descrição**
Cirqueira et al. (2017)	Facebook API e API REST do X/Twitter	Opinion Label	Comparação entre algoritmos de Análise de Sentimentos em suas versões em português e inglês.
Melo, Figueiredo et al. (2021)	X/Twitter e UOL (Universo Online); TwitterScraper	VADER	Metodologia para avaliar os assuntos debatidos no Twitter e no UOL acerca da COVID-19 por meio de modelagem de tópicos.
Freitas, Ladeira e Caetano (2024)	X/Twitter; Snscrape	Extração de Características e ML.	Desenvolvimento de modelo computacional para identificação de comportamento depressivo no X/Twitter com base em 15 atributos retirados dos dados.

Fonte: Elaborado pelo autor (2024)

Os estudos analisados demonstram a crescente importância da análise de sentimentos em mídias sociais para compreender a opinião pública e identificar tendências. Cirqueira et al. (2017) evidenciaram as particularidades da língua portuguesa na análise de sentimentos, enquanto Melo, Figueiredo et al. (2021) aplicaram a técnica para analisar a percepção pública sobre a COVID-19. Freitas, Ladeira e Caetano (2024) foram ainda mais além, utilizando a análise de sentimentos para identificar sinais de depressão em usuários do X/Twitter. Esses trabalhos demonstram a versatilidade da análise de sentimentos e a necessidade de adaptar as metodologias para cada contexto específico.

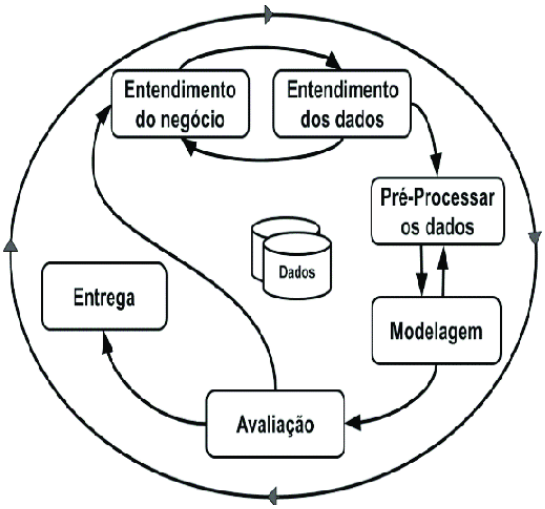
Os trabalhos citados ilustram as diversas áreas em que a análise de sentimento em mídias sociais podem alcançar. Os estudos contemplam diferentes métodos de classificação dos sentimentos expressos nos dados textuais, tais como: plataformas gamificadas como o Opinion Label ou ferramentas baseadas em léxico como o VADER.

A presente pesquisa insere-se no contexto da análise de sentimentos em mídias sociais, um campo em constante expansão. Assim como os estudos de Cirqueira et al. (2017), Melo, Figueiredo et al. (2021) e Freitas, Ladeira e Caetano (2024), este trabalho explora as potencialidades da análise de sentimentos para extrair *insights* valiosos de grandes volumes de dados textuais. Contudo, diferencia-se ao focar no nicho de eventos culturais e ao utilizar ferramentas baseadas em *Transformers*. Além disso, a pesquisa propõe uma abordagem que integra conceitos discutidos nos estudos citados, abordando um tema inédito, o que contribui para preencher lacunas existentes na literatura.

4 METODOLOGIA

Os métodos e procedimentos utilizados para conduzir a pesquisa acerca da análise de sentimentos no X sobre as festas juninas no Maranhão serão descritos nessa seção. A metodologia usada é baseada no *Cross-Industry Standard Process for Data Mining* (CRISP-DM) que se trata de um processo iterativo independente da indústria para a aplicação de projetos que envolvem extração e mineração de dados. Esse método possui 6 fases iterativas que abrangem o projeto desde a compreensão do negócio até à implementação (SCHRÖER; KRUSE; GÓMEZ, 2021). A Figura 7 ilustra as etapas da metodologia em seu fluxo circular.

Figura 7 – Fases da Metodologia CRISP-DM



Fonte: Adaptada de Wirth e Hipp (2000)

4.1 Entendimento do Negócio

A *corpora* textual com dados em português ainda é limitada quando comparada com o inglês, fato que interfere diretamente no desenvolvimento de novas ferramentas de PLN e no desempenho das existentes atualmente (OLIVEIRA et al., 2022). Esse estudo propõe bases de dados rotuladas que podem ser usadas posteriormente para treinamento de modelos de ML úteis para o desenvolvimento de diferentes funcionalidades baseadas em texto ou no âmbito cultural.

Além disso, o desenvolvimento deste estudo busca auxiliar na tomada de decisões sobre a administração dos eventos juninos, bem como identificar tendências que possam agregar valor aos festejos futuros. Isso será viabilizado pela análise de dados classificados: tópicos identificados como positivos indicarão práticas a serem mantidas nos anos seguintes,

enquanto os classificados como negativos apontarão aspectos a serem melhorados ou descontinuados.

No que concerne ao dados textuais, a presença de neologismos, gírias e abreviações, predominante na maioria dos dados coletados do X configuram-se como desafios a serem enfrentados para que este estudo atinja o seu objetivo. Visto que tais elementos presentes tendem a tornar o conjunto de dados ruidoso, além de dificultar tarefas de modelagem de tópicos e classificação de sentimentos (THELWALL, 2021).

Esse estudo visa modelar tópicos bem definidos acerca das *trends* que permeiam o São João do Maranhão e de qual forma os assuntos mais comentados se relacionam com os sentimentos expressos pelos usuários do X e pelas notícias veiculadas no portal G1.

4.2 Entendimento dos Dados

O conjunto de dados utilizado para estudo é composto de 1756 posts coletados do X e 125 artigos publicados no veículo de notícias G1 Maranhão.

Para a coleta de dados no X (antigo Twitter), foi adotada uma técnica alternativa às ferramentas que utilizam diretamente a API da plataforma, visto que os planos gratuitos não oferecem funcionalidades essenciais para este tipo de pesquisa, como a obtenção de *tweets* por meio de uma *string* de busca (X Corp., 2024). Para superar essa limitação, foi utilizada a biblioteca Python ntscraper (BOCCHI, 2024), que realiza a raspagem de dados na aplicação web conhecida como Nitter (ZEDEUS, 2019).

O Nitter é uma plataforma gratuita e de código aberto desenvolvida como alternativa ao X/Twitter, projetada para ser mais segura e responsiva aos usuários. Além disso, permite a visualização de *tweets* de forma anônima, sem a necessidade de rastreamento de IP pela plataforma original.

Os procedimentos de raspagem de dados foram realizados no Google Colab, uma plataforma que permite a execução de código Python na nuvem (Google LLC, 2024). Essa abordagem possibilitou o uso da biblioteca ntscraper com parâmetros de busca específicos para coletar os dados necessários à análise de sentimentos, por meio de técnicas de *web scraping*. Os dados obtidos abrangem o período de 1º de junho a 10 de julho dos anos de 2023 e 2024.

A coleta foi realizada utilizando queries de busca avançada, inseridas como strings no parâmetro da função get_tweets. Os operadores utilizados nas consultas de busca representam:

- **palavra1 palavra2**: busca *tweets* que contenham ambas as palavras em qualquer ordem.

- **'palavra1 palavra2'**: busca *tweets* que contenham as palavras exatamente nessa ordem.
- **lang:xx**: busca *tweets* em uma linguagem específica (código ISO 639-1, como pt para português).
- **since:YYYY-MM-DD**: busca *tweets* enviados a partir de uma data especificada.
- **until:YYYY-MM-DD**: busca *tweets* enviados até uma data especificada.

Para a coleta de dados no G1 Maranhão, foi utilizada a biblioteca BeautifulSoup do Python, por se tratar de uma ferramenta eficiente para web scraping que permite navegar pelo HTML de uma página web sem a necessidade de expressões regulares complexas (HAJBA, 2018).

Os dados de notícias coletados foram armazenados em um arquivo .csv, contendo informações sobre título, data e o corpo da notícia para análise posterior.

A Tabela 2 apresenta um exemplo de dado coletado para cada uma das fontes.

Tabela 2 – Exemplo de dados coletados de ambas as fontes

Exemplo de *tweet*	**Trecho de Notícia do G1**
o maior são joão do mundo é o do maranhão e nem adianta chorar.	94% dos turistas pretendem voltar ao Maranhão após o São João, aponta pesquisa.

Fonte: o Autor (2024)

4.3 Preparação dos Dados

Neste trabalho, a etapa de pré-processamento dos dados do X consistiu na remoção de espaços extras, links, emojis, menções e hashtags. Assim, dessa forma, facilitando o trabalho de classificação de cada *tweet*. No contexto dos dados em geral, também foram removidas *stopwords*, ou seja, palavras que não contribuem para a avaliação da sentença. A biblioteca *Natural Language Toolkit* (NLTK) possui um léxico contendo *stopwords* para língua portuguesa, facilitando a remoção de termos que dificultem a identificação de termos relevantes no contexto junino (SARICA; LUO, 2021).

Ao final da aplicação desses procedimentos, aplicou-se a tokenização que consiste na divisão do texto em unidades menores denominadas *tokens*, essa etapa facilita o processamento de modelos de PLN visto que facilita a identificação e contagem de palavras.

Por fim, aplicou-se a lematização utilizando a biblioteca spaCy do Python. Essa técnica reduz as palavras às suas formas canônicas. Por exemplo, "andando", "andou" e

"anda" podem ser reduzidos ao *lemma* "andar". Ainda que tal método não possua grande eficácia para o português brasileiro, tal redução melhora a coerência e compreensão do texto pelos modelos de PLN (MELO; FIGUEIREDO et al., 2021). Na Tabela 3 observa-se o exemplo de um *tweet* antes e depois do pré-processamento.

Tabela 3 – Exemplo de processamento dos dados

Sem pré-processamento	**Com pré-processamento**
Maranhão tem o melhor São João do mundo, e este vídeo prova isso.	maranhão melhor mundo vídeo prova

Fonte: o Autor (2024)

4.4 Modelagem

Para essa etapa, utiliza-se as bases de dados já pré-processadas para a aplicação do *Valence Aware Dictionary and Sentiment Reasoner* (VADER), que é uma ferramenta baseada em regras e léxico que utiliza um conjunto de palavras e expressões associadas a valores e sentimentos (ISNAN; ELWIREHARDJA; PARDAMEAN, 2023). Esses valores auxiliam a determinar a polaridade emocional de um texto. No caso deste trabalho, os dados serão classificados em 3 emoções: negativa, neutra e positiva. Na Tabela 4 é ilustrado um exemplo de dado para cada rótulo possível.

Tabela 4 – Exemplo de Classificação dos Dados

Tweet	**Classificação**
sentimento que arde no peito, inexplicável e nosso. te amo São João do Maranhão!	Positiva
Alô braide contrata o Jorge e Mateus pro são João do Maranhão	Neutra
eu vendo os famosos reclamando do calor daqui do maranhão no são joão da thay, aqui é o próprio inferno d tão quente	Negativa

Fonte: o Autor (2024)

Com os textos devidamente classificados, a manipulação e análise dos dados torna-se mais acessível. Dessa forma, é possível detectar padrões e nuances nos dados computados. Optou-se por organizar os *tweets* em uma base de dados `.csv` e as notícias em outra de mesmo formato, para fins de organização. Cada uma dessas bases de dados são dispostas de uma coluna para o texto e outra coluna para o sentimento expresso por ele.

A modelagem dos dados do X e do G1 foi realizada utilizando o BERTopic, que é uma técnica de modelagem de tópicos que utiliza modelos de linguagem. Apesar de sua natureza modular, que permite ao usuário da ferramenta decidir quais procedimentos utilizar em cada etapa da modelagem de tópicos, utilizou-se a configuração padrão do algoritmo neste presente estudo. Isso é justificado pela qualidade da representação semântica do S-BERT, redução de dimensionalidade do UMAP que preserva a qualidade das representações, *clusters* de diferentes tamanhos gerados pelo HDBSCAN e pela representação baseada em tópicos do cTF-IDF. As especificações de cada ferramenta configuram-se ideais para a consolidação da modelagem de tópicos baseada nos dados obtidos no presente estudo.

Para facilitar processos de comparação entre os dados e tópicos modelados para as mídias do X e do G1, foram realizadas constatações empíricas sobre a quantidade de tópicos gerados pelo BERTopic: foram testadas quantidades de tópicos variadas para avaliar o desempenho do modelo em gerar uma modelagem com a habilidade de conciliar especificidade com originalidade. Após os testes realizados alterando a função `BERTopic(language="portuguese", nr_topics=N)`, ajustando o valor de N a modo de obter uma modelagem satisfatória. Ao final desse processo, chegou-se ao número ideal de 8 tópicos para cada dataset: G1, *tweets* e *tweets* negativos. As tabelas 5 e 6 demonstram a distribuição dos dados para cada tipo de fonte.

Tabela 5 – Distribuição do dados - X

Sentimento	**Quantidade**
Positivo	966
Neutro	586
Negativo	204

Fonte: o Autor (2024)

Tabela 6 – Distribuição do dados - G1

Sentimento	**Quantidade**
Positivo	106
Neutro	19
Negativo	0

Fonte: o Autor (2024)

4.5 Avaliação

A avaliação dos tópicos gerados será realizada por uma matriz de similaridade. Esta ferramenta de avaliação representa a relação entre os temas identificados. Em cada célula da estrutura matricial há um indicador de grau de similaridade entre dois tópicos.

Essa representação destaca-se como um método eficaz para entender como os tópicos se relacionam entre si no contexto semântico.

Conforme destacado Röder, Both e Hinneburg (2015), uma boa modelagem de tópicos deve apresentar tópicos distintos o suficiente para capturar diferentes aspectos do *corpus* analisado, mas sem perder a coerência temática. Além disso, a avaliação da qualidade dos tópicos por meio da matriz de similaridade é uma abordagem complementar a outras métricas, como Mimno et al. (2011) e Lau, Newman e Baldwin (2014), garantindo uma visão mais ampla da estrutura dos tópicos.

No contexto do BERTopic, a matriz de similaridade é gerada pela função visualize_heatmap. Esse método calcula o relacionamento entre tópicos usando a medida de similaridade de cosseno, ferramenta matemática comum para comparar vetores. A similaridade de cosseno é calculada por:

$$similaridade\_cos\,(A, B) = \frac{A.B}{\|A\|\,\|B\|} \tag{4.1}$$

Onde:

- A e B são os vetores dos tópicos.
- $A.B$ é o produto escalar entre os vetores.
- $||A||$ e $||B||$ são os módulos dos vetores.

O valor da similaridade varia de -1 (totalmente diferentes) a 1 (idênticos).

4.6 Entrega

A entrega deste trabalho é consolidada pela apresentação do mesmo e dos resultados obtidos perante banca avaliadora. A base de dados de *tweets*[1] e de notícias[2] serão disponibilizadas no Kaggle.

[1] <https://www.kaggle.com/datasets/jgpfdark/so-joo-do-maranho-postagens-classificadas>
[2] <https://www.kaggle.com/datasets/jgpfdark/notcias-g1-so-joo-anlise-de-sentimentos>

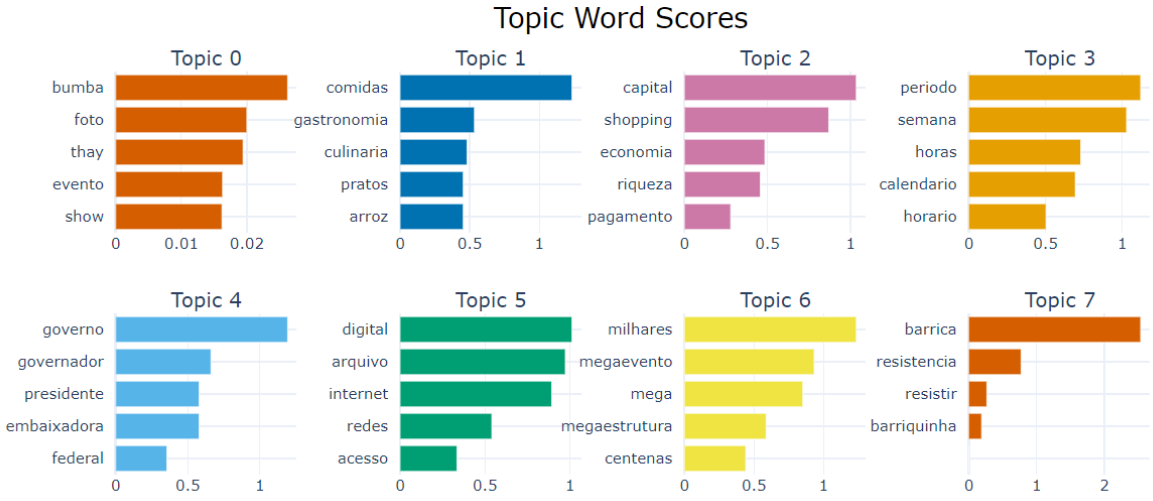
5 RESULTADOS E DISCUSSÃO

Neste capítulo, serão apresentados e debatidos os resultados provenientes dos experimentos realizados com o BERTopic, bem como a distribuição dos termos mais frequentes em cada base de dados para a comparação efetiva dos temas debatidos no G1 e no X, pelas suas semelhanças e diferenças.

A geração de gráficos pelo BERTopic foi feita pela função `visualize_barchart`. Ela resulta numa coletânea de gráficos representando cada um dos 8 tópicos gerados. O valor contido no eixo horizontal do gráfico de barras representa o valor do cTF-IDF de cada termo: o maior valor é associado à palavra que mais agrega significado naquela classe, ou seja, no tópico nela contido. Além disso, a palavra com maior valor de cTF-IDF descreve bem o conteúdo de cada tópico, visto que as palavras subsequentes possuem valor semântico semelhante.

A Figura 8 apresenta os termos mais utilizados no G1, organizados em 8 tópicos. Observa-se que temas como turismo, culinária, música e política foram abordados, sendo o mais relevante deles relacionado ao Bumba Meu Boi.

Figura 8 – Tópicos - Notícias G1

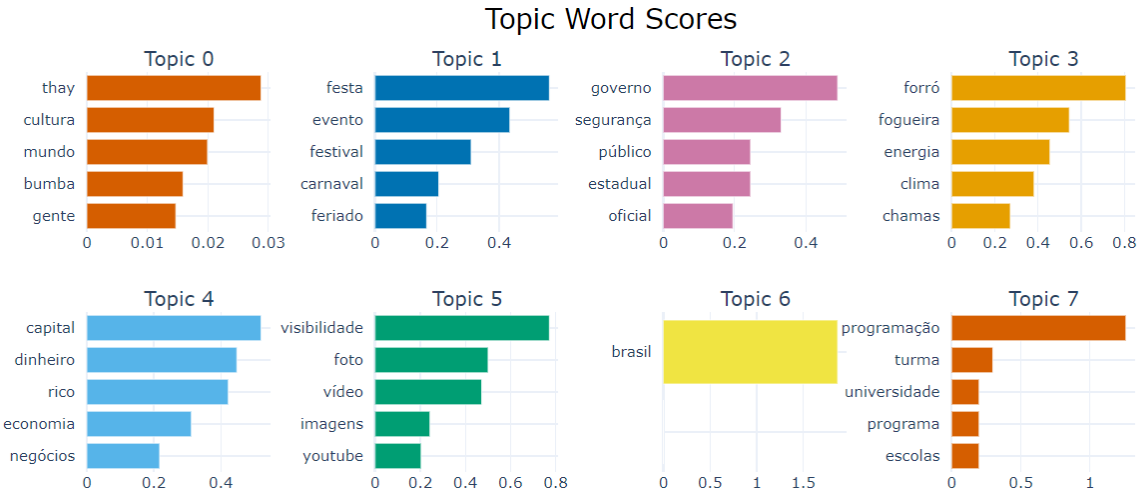


Fonte: o Autor (2024)

No caso dos *tweets*, como ilustrado na Figura 9, os tópicos mantiveram-se semelhantes ao G1. Percebe-se que o São João da Thay foi destacado mais vezes no X em relação ao G1, pois o festival gera grande repercussão nas redes sociais devido à sua forte presença midiática, à conexão com influenciadores e à divulgação de experiências e atrações exclusivas, que atraem um público engajado e participativo. O São João da Thay é um evento beneficente criado pela influenciadora Thaynara OG, realizado anualmente em São Luís do Maranhão. A festa mistura cultura junina, shows de artistas nacionais e arrecadação de fundos para projetos sociais, promovendo a tradição nordestina em grande estilo.

Em contrapartida, o Bumba Meu Boi que é uma manifestação cultural típica do Maranhão, que mistura teatro, dança e música para contar a lenda do boi encantado. Com influências indígenas, africanas e europeias, marcada por seus sotaques, ritmos e figurinos vibrantes, foi discutido com menor frequência, visto que recebe um apelo midiático menor sendo uma tradição regional. Ademais, nota-se que esse tema compartilha relevância semelhante ao termo "mundo" que faz referência ao slogan do evento "Maior São João do Mundo". Além disso, também percebe-se a incidência maior de assuntos relacionados à economia, tema que não foi abordado como os 8 principais tópicos do G1.

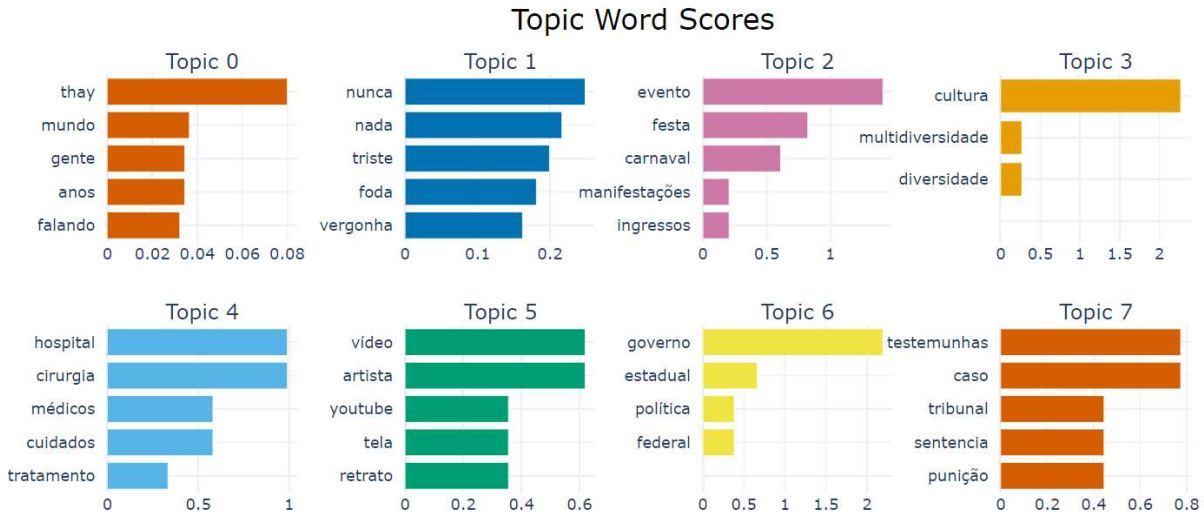
Figura 9 – Tópicos - *Tweets*



Fonte: o Autor (2024)

A Figura 10 ilustra a distribuição de *tweets* negativos. Mais uma vez, o São João da Thay é o assunto principal abordado, evidenciando a opinião heterogênea dos usuários do X sobre o evento. Observa-se, também, o termo "governo" com valor de cTF-IDF acima de 2, evidenciando um descontentamento majoritário em relação às outras palavras correlatas que compõem o tópico. Pode ser notado o mesmo acontecimento no tópico sobre cultura e diversidade, em que o termo "cultura" excede o valor de 2 no cTF-IDF, fato que corrobora a relevância da palavra para debates negativos sobre esse tema.

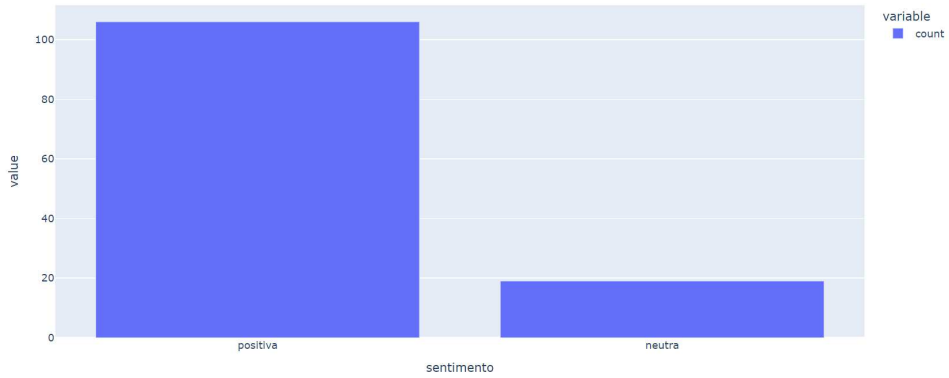
Figura 10 – Tópicos - *Tweets* rotulados como negativos



Fonte: o Autor (2024)

Ao observar na Figura 11 a distribuição de sentimentos expressos nas notícias veículadas no site G1 Maranhão, observa-se a predominância da emoção positiva nos textos com 84,6% seguido de 15,4% de textos que expressam emoção neutra em relação ao São João do Maranhão.

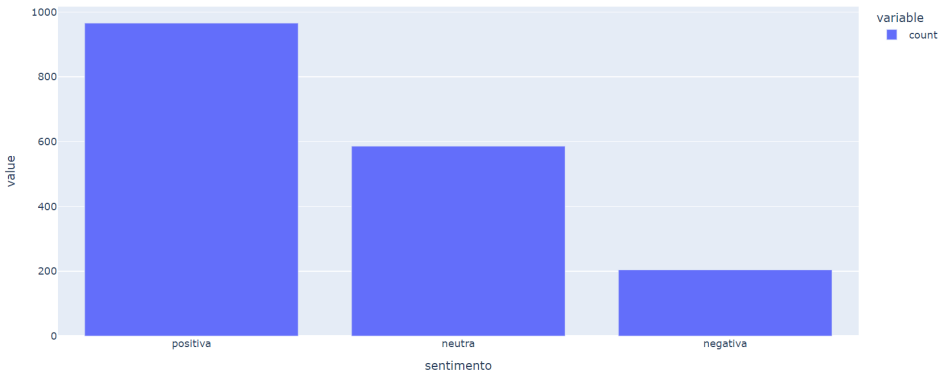
Figura 11 – Tópicos - Distribuição de Sentimentos - G1



Fonte: o Autor (2024)

Em relação ao X, dos 1756 dados coletados, 55,1% dos *tweets* foram classificados como positivos, 33,37% neutros e, por fim, 11,62% negativos. Fato que configura, no geral, o São João do Maranhão como amplamente aceito pelo público. A disposição da quantidade desses dados em cada rótulo é demonstrada na gráfico da Figura 12.

Figura 12 – Tópicos - Distribuição de Sentimentos - X



Fonte: o Autor (2024)

No que concerne aos termos mais utilizados, a nuvem de palavras obtida pela biblioteca do Python denominada WordCloud, destaca de forma proporcional o tamanho da palabra na nuvem em relação à sua frequência nas bases de dados.

No G1, os termos que receberam maior destaque, de acordo com a nuvem de palavras, fazem relação ao Bumba Meu Boi, tambor de crioula, dança portuguesa e ao Arraial do Ipem. O Arraial do Ipem é um dos principais festejos juninos de São Luís, Maranhão, reunindo apresentações de Bumba Meu Boi, quadrilhas, cacuriá e outras manifestações culturais típicas. Sendo um dos arraiais mais tradicionais e esperados do período junino, fato que justifica sua presença em grande parte dos artigos do G1. As palavras podem ser observadas na Figura 13.

Figura 13 – Nuvem de Palavras - G1

Fonte: o Autor (2024)

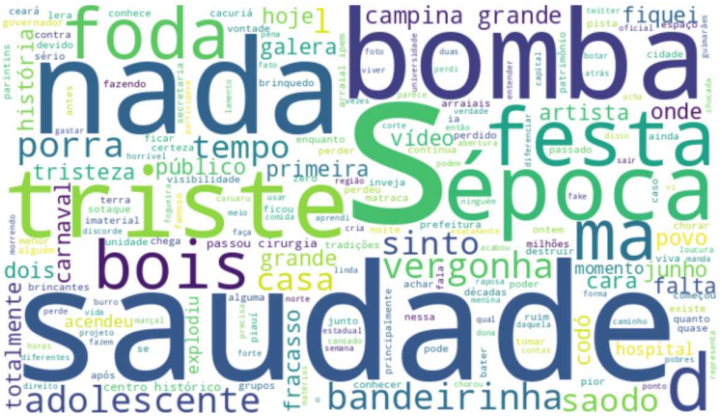
Na Figura 14, em relação ao X, considerando todos os *tweets*, os termos em destaque fazem menção ao São João da Thay, ao slogan do São João do Maranhão: "O Maior São João do Mundo" e a temas gerais como cultura, festa, evento e estado.

Figura 14 – Nuvem de Palavras - *Tweets*

Fonte: o Autor (2024)

A Figura 15 apresenta a nuvem de palavras gerada a partir dos termos mais frequentes em uma base de dados composta por *tweets* classificados como negativos. Essa visualização tem como objetivo destacar os principais tópicos e palavras recorrentes nos *tweets* analisados, facilitando a identificação de padrões linguísticos e contextuais associados a sentimentos negativos. A geração da nuvem de palavras foi realizada considerando a frequência relativa dos termos, onde palavras com maior relevância aparecem em tamanhos proporcionalmente maiores. Esse recurso é útil para sintetizar o conteúdo textual da base de dados, fornecendo *insights* iniciais sobre o vocabulário predominante em mensagens negativas.

Figura 15 – Nuvem de Palavras - *Tweets* Negativos



Fonte: o Autor (2024)

A Tabela 7 agrupa os termos citados em contextos de negatividade com um exemplo da base de dados de *tweets*.

Tabela 7 – *Tweets* negativos e termos relacionados

Tweet	**Termo**
Evento da Thaynara não representa nem o São João do Maranhão, quem dirá ser suficiente pra representar capital do Norte e Nordeste.As danças culturais daqui não tiveram nem 30 minutos de tela, tomem vergonha!	Vergonha
Um adolescente de 13 anos, perdeu quatro dedos da mão e ficou com queimaduras no rosto, após estourar uma bomba de São João	Bomba
a tristeza que todo maranhense que não está no maranhão no são joão sente	Triste
fui selecionar as fotos do São João pra postar e já me deu vontade de chorar de saudades do Maranhão	Saudades

Fonte: o Autor (2024)

Percebe-se o termo “saudade” em destaque, denotando um sentimento de saudosismo e nostalgia em relação às festividades juninas por parte dos usuários do X. No que diz respeito à palavra em destaque “triste”, tem-se duas possibilidades: o sentimento expresso pode ser tristeza pela chegada do fim da época das festividades ou por descontentamento por motivos gerais.

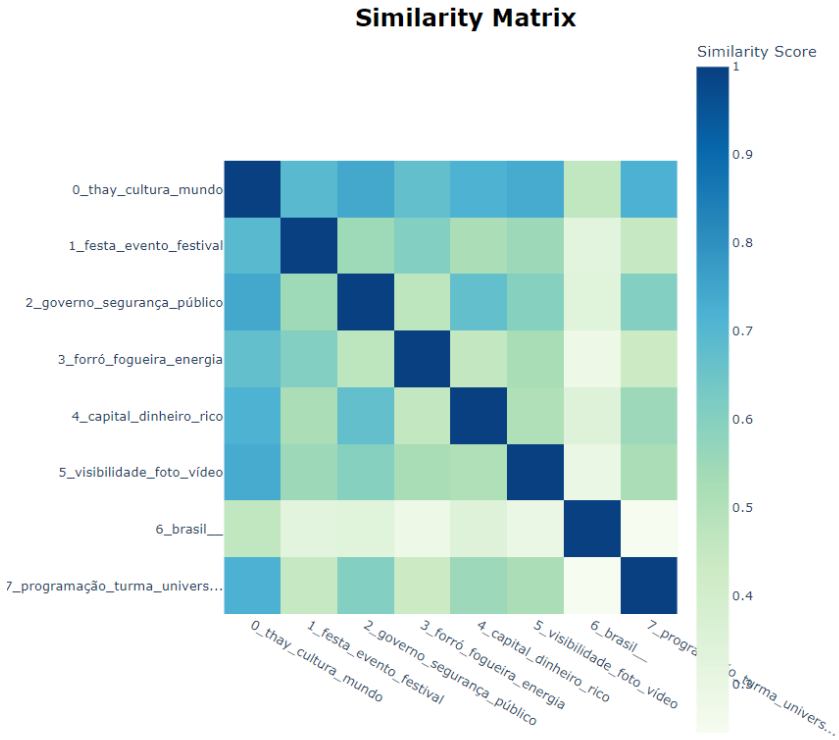
Sobre o termo “bomba” e sua associação negativa, infere-se a ocorrência de incidentes com artefatos explosivos por parte dos frequentadores das festividades.

Os *tweets* contendo o termo “vergonha”, em todas as ocorrências, denotam desapreço ocasionado pela escolha de atrações que, segundo usuários do X, não contemplam a cultura junina do Maranhão.

Por fim, serão apresentadas as matrizes de similaridade para cada conjunto de dados: *tweets* e G1.

O objetivo principal ao analisar essas estruturas é verificar a qualidade dos tópicos gerados pelo modelo. Para isso, é ideal que os tópicos não sejam redudantes, mas sim bem definidos com poucos termos em comum. A Figura 16 mostra a matriz de similaridade dos tópicos gerados para todos os *tweets*.

Figura 16 – Matriz de Similaridade - *Tweets*

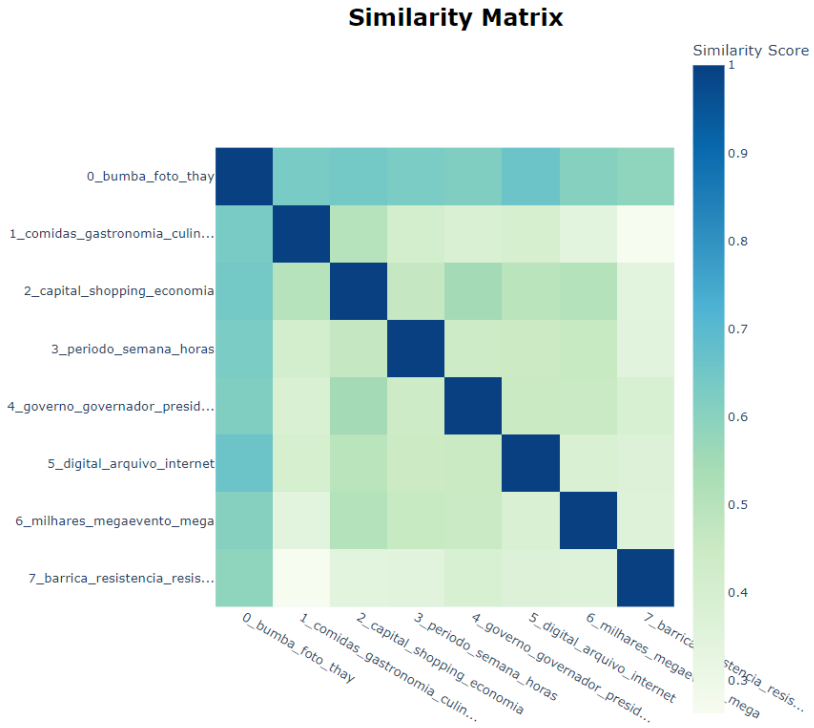


Fonte: o Autor (2024)

Ao analisar a figura, percebe-se que, os quadrados escuros indicam maior similaridade, enquanto os claros indicam menor similaridade entre os tópicos. O tópico 6 sobre “Brasil” configurou-se como o mais singular dentre os outros, apresentando similaridade abaixo de 0.3 na maioria das comparações. Enquanto o tópico 0 apresentou maior similaridade com os demais, visto que compartilha termos e conceitos com representações vetoriais semelhantes aos demais tópicos obtidos.

A Figura 17 apresenta a similaridade de tópicos para as notícias vínculadas ao G1.

Figura 17 – Matriz de Similaridade - G1



Fonte: o Autor (2024)

Ao analisar a matriz de similaridade para o G1, observa-se que os tópicos gerados pelo BERTopic apresentam menor semelhança entre si, refletindo em uma menor quantidade de termos compartilhados. Isso indica que os tópicos estão melhor definidos em comparação aos gerados para o conjunto de dados do X.

Destaca-se novamente o tópico 0, que aborda temas como o São João da Thay e o Bumba Meu Boi, apresentando maior similaridade com os demais tópicos. No entanto, os valores obtidos pela similaridade de cosseno são menores, reforçando uma separação relativa entre eles.

Além disso, o tópico 7 se destaca como o mais distinto entre os demais, fato evidenciado pela predominância de valores próximos a zero no *Similarity Score* quando comparado aos outros tópicos.

6 CONCLUSÃO E TRABALHOS FUTUROS

O uso de modelos de linguagem como o BERT para a realização de tarefas de PLN ganhou notoriedade nos últimos anos, como prova disso pode-se perceber os diversos modelos originários do BERT criados com *fine-tuning*. Apesar da ampla abordagem de diferentes contextos com a ferramenta originária da rede neural *Transformer*, ainda há a necessidade de explorar o cenário dos eventos culturais pela perspectiva dos frequentadores, proporcionando desafios ao modelo como a linguagem coloquial e gírias, por exemplo.

Dessa forma, esse estudo visou utilizar o BERTopic que faz o uso das representações vetoriais de palavras originários do BERT para compor as entrelinhas do evento cultural São João do Maranhão. Para isso, esse trabalho foi composto de técnicas de *web scraping* para coleta de dados, análise lexical para análise de sentimentos e aprendizado profundo para a modelagem de tópicos. Sendo estas, ferramentas fundamentais para o desenvolvimento de trabalhos de PLN.

Os resultados demonstraram a praticidade de extrair os temas de diferentes tópicos abordados pelo X e G1 durante as festividades. Assim, houve a possibilidade de entender os temas expostos em cada uma das duas mídias de forma rápida e eficaz, configurando uma ferramenta útil para entender os pontos positivos e corrigir os alvos de críticas negativas do evento. Além disso, com a classificação das emoções expressas nos dados textuais em positiva, negativa e neutra, houve a possibilidade de explorar e entender os problemas vivenciados durante o evento com a modelagem de tópicos e *wordclouds* exclusivamente dos dados classificados como negativos.

Diante desse contexto, a análise proposta neste estudo logrou êxito em constatar a satisfação majoritária acerca do São João do Maranhão, tendo em conta a quantidade de notícias e *tweets* de cunho positivo acerca do evento. Em relação à modelagem de tópicos, constatou-se que atrações como Bumba Meu Boi e São João da Thay foram os pontos de maior comoção em abordagens nas mídias analisadas. Além disso, ao analisar a matriz de similaridade dos tópicos criados, percebe-se uma distribuição satisfatória, criando um equilíbrio entre similaridade e disparidade. Além disso, ao realizar uma análise dos termos representativos de cada tópico no X e G1, é possível constatar que ambas as mídias cobriram temas semelhantes ao longo dos anos de 2023 e 2024.

Para trabalhos futuros, propõe-se a utlização de *Large Language Models* (LLMs) como *Generative Pre-trained Transformer* (GPT), por exemplo. Dessa forma, o estudo pode capturar nuances mais específicas acerca do evento junino. Outrossim, a utilização de

métodos para lidar com gírias, ironia e sarcasmos nos *tweets* configura-se como primordial para evoluir o estudo em questão, visto que os modelos utilizados nesse estudo ainda possuem limitações neste aspecto. Este debate acerca de ironia, gírias e sarcasmo configura-se como um desafio mesmo com modelos avançados de linguagem natural, portanto faz-se fundamental o treinamento de modelos em base de dados com altos índices de uso desses artifícios de linguagem, ou até mesmo a criação de um dicionário de gírias. Somado a isso, outra possibilidade de aprimoramento deste estudo baseia-se na utilização de outras métricas para avaliar a similaridade e qualidade dos tópicos gerados. Entre elas, pode-se destacar o `C_v` e `U_Mass`. Ademais, uma análise de sentimentos e modelagem de tópicos com levando com consideração a data de cada postagem pode ser realizada para compreender a mudança de sentimentos e de temas discutidos ao decorrer do evento. Com a aplicação dessas melhorias, a análise acerca do São João do Maranhão pode ser aprimorada.

Referências

ABDELRAZEK, A.; EID, Y.; GAWISH, E.; MEDHAT, W.; HASSAN, A. Topic modeling algorithms and applications: A survey. *Information Systems*, Elsevier, v. 112, p. 102131, 2023. Citado na página 25.

AL-QURISHI, M. Leveraging bert language model for arabic long document classification. *arXiv preprint arXiv:2305.03519*, 2023. Citado na página 26.

ALMEIDA, T.; OLIVEIRA, F.; SILVA, P. Análise de sentimentos utilizando algoritmos de aprendizado de máquina. *Revista Brasileira de Inteligência Artificial*, v. 18, n. 3, p. 100–115, 2021. Disponível em: <https://www.revistas.br/ai-analysis?utm_source=chatgpt.com>. Citado na página 18.

ALUíSIO, S. M. et al. Avaliação de ferramentas para análise linguística do português do brasil. *Revista Brasileira de Linguística Aplicada*, SciELO Brasil, v. 11, n. 2, p. 561–592, 2011. Citado na página 15.

AMORIM, A.; MURRUGARRA-LLERENA, N.; SILVA, V.; OLIVEIRA, D. de; PAES, A. Modelagem de tópicos em textos curtos: uma avaliação experimental. In: *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*. Porto Alegre, RS, Brasil: SBC, 2022. p. 254–266. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/21811>. Citado 2 vezes nas páginas 25 e 26.

BENITEZ, R. J. T. J. M. G. G. Construção de um corpus para análise de sentimentos em português. In: *Universidade do Estado de Santa Catarina (UDESC)*. [s.n.], 2022. Disponível em: <https://www.udesc.br/arquivos/udesc/id_cpmenu/15656/034_CONSTRU___O_DE_UM_CORPUS_16632622990273_15656.pdf>. Citado na página 15.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado na página 31.

BOCCHI, L. *ntscraper: A Python library for scraping content from X (formerly Twitter)*. 2024. <https://github.com/bocchilorenzo/ntscraper>. Accessed: 2024-12-16. Citado na página 35.

BOLLEN, J.; MAO, H.; ZENG, X.-J. Twitter mood predicts the stock market. *Journal of Computational Science*, v. 2, n. 1, p. 1–8, 2011. Citado na página 15.

BONANDRINI, R.; AMENTA, S.; SULPIZIO, S.; TETTAMANTI, M.; MAZZUCCHELLI, A.; MARELLI, M. Form to meaning mapping and the impact of explicit morpheme combination in novel word processing. *Cognitive Psychology*, Elsevier, v. 145, p. 101594, 2023. Citado na página 22.

BRITTO, L.; PACíFICO, L. Análise de sentimentos para revisões de aplicativos mobile em português brasileiro. In: *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre, RS, Brasil: SBC, 2019. p. 1080–1090. ISSN 2763-9061.

Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/9359>. Citado na página 15.

CAMACHO-COLLADOS, J.; PILEHVAR, M. T. Embeddings in natural language processing. In: SPECIA, L.; BECK, D. (Ed.). *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*. Barcelona, Spain (Online): International Committee for Computational Linguistics, 2020. p. 10–15. Disponível em: <https://aclanthology.org/2020.coling-tutorials.2>. Citado na página 21.

CIRQUEIRA, D.; JACOB, A.; LOBATO, F.; SANTANA, A. L. de; PINHEIRO, M. Performance evaluation of sentiment analysis methods for brazilian portuguese. In: SPRINGER. *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers 19*. [S.l.], 2017. p. 245–251. Citado 3 vezes nas páginas 30, 32 e 33.

CORREIA, G. P.; MENDONçA, R. F.; BORGES, R. M. R.; CORDEIRO, D. F. Saúde mental no twitter: análise de manifestações por meio de mineração de dados. *Novos Olhares*, Universidade de São Paulo, v. 12, n. 1, p. 1–20, 2023. Citado na página 19.

D'AMICO, F.; NEGRI, M. Self-attention as an attractor network: transient memories without backpropagation. *CoRR*, abs/2409.16112, 2024. Disponível em: <https://arxiv.org/abs/2409.16112>. Citado na página 23.

DAS, A.; BISWAS, S.; PAL, U.; LLADÓS, J.; BHATTACHARYA, S. Fasttextspotter: A high-efficiency transformer for multilingual scene text spotting. In: SPRINGER. *International Conference on Pattern Recognition*. [S.l.], 2025. p. 135–150. Citado na página 22.

Datareportal. *Digital 2024: Brazil*. 2024. Acesso em: 14 dez. 2024. Disponível em: <https://datareportal.com/reports/digital-2024-brazil>. Citado na página 19.

DEIKIS, J. *Guide to Emotional Marketing: How to Evoke Emotions in Your Audience*. 2024. Acesso em: 14 dez. 2024. Disponível em: <https://noblestudios.com/digital-marketing-services/digital-brand-strategy/guide-to-emotional-marketing/>. Citado na página 14.

Demand Sage. *Twitter Statistics 2024: Usage, Revenue, and More*. 2024. Acesso em: 14 dez. 2024. Disponível em: <https://www.demandsage.com/twitter-statistics/>. Citado na página 14.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Bidirectional encoder representations from transformers. *arXiv preprint arXiv:1810.04805*, p. 15, 2018. Citado na página 24.

DOUGLAS, C.; LUCAS, V.; MÁRCIA, P.; ANTÔNIO, J.; FÁBIO, L.; ÁDAMO, S. Opinion label: A gamified crowdsourcing system for sentiment analysis annotation. 2017. Citado na página 30.

ETHAYARAJH, K. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019. Disponível em: <https://arxiv.org/abs/1909.00512>. Citado na página 23.

FERNANDES, R.; SILVA, C. Análise de sentimento com o método vader: Um estudo de caso com textos em português. In: *Congresso de Tecnologia da Informação*. [s.n.], 2019. Disponível em: <https://cti.ufpel.edu.br/siepe/arquivos/2019/CE_04549.pdf>. Citado na página 18.

FREITAS, L. M. G.; LADEIRA, M.; CAETANO, M. F. Desenvolvimento de ferramenta de análise de sentimentos para identificação de possíveis sinais de comportamento depressivo na rede social twitter. *Revista Eletrônica de Iniciação Científica em Computação*, v. 22, n. 1, p. 91–100, 2024. Acesso em: 17 dez. 2024. Disponível em: <https://journals-sol.sbc.org.br/index.php/reic/article/view/3430>. Citado 4 vezes nas páginas 30, 31, 32 e 33.

G1. *Título do artigo ou notícia*. 2025. Disponível em: <https://g1.globo.com/ma>. Citado na página 20.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Natural language processing. In: *Deep Learning*. [S.l.]: MIT Press, 2016. cap. 12, p. 456–490. Citado 2 vezes nas páginas 21 e 29.

GOOGLE. 2025. Acessado em: 14 jan. 2025. Disponível em: <https://trends.google.com.br/trends/explore?date=all&q=sentiment%20analysis&hl=pt>. Citado na página 18.

Google LLC. *Google Search*. 2024. <https://www.google.com>. Accessed: 2024-12-16. Citado na página 35.

GROOTENDORST, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022. Citado 3 vezes nas páginas 26, 27 e 28.

GUPTA, R.; SHARMA, N.; YADAV, A. K. et al. Sentiment analysis and its applications: A survey. *Journal of Information and Data Management*, Springer, v. 12, n. 3, p. 45–56, 2021. Citado na página 19.

HAJBA, G. L. Website scraping with python. *Berkeley: Apress*, Springer, 2018. Citado na página 36.

HUTTO, C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the international AAAI conference on web and social media*. [S.l.: s.n.], 2014. v. 8, n. 1, p. 216–225. Citado na página 31.

ISNAN, M.; ELWIREHARDJA, G. N.; PARDAMEAN, B. Sentiment analysis for tiktok review using vader sentiment and svm model. *Procedia Computer Science*, Elsevier, v. 227, p. 168–175, 2023. Citado na página 37.

KITCHIN, R. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage Publications, 2021. Citado na página 19.

KOROTEEV, M. V. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021. Citado na página 24.

KOULOUMPIS, E.; WILSON, T.; MOORE, J. D. Twitter sentiment analysis: The good the bad and the omg. *Proceedings of the Fifth International Conference on Weblogs and Social Media*, p. 538–541, 2011. Disponível em: <https://www.researchgate.net/publication/220689416_Twitter_Sentiment_Analysis_The_Good_the_Bad_and_the_OMG>. Citado na página 18.

KOZINA, A.; NADOLNY, M.; HERNES, M.; WALASZCZYK, E.; ROT, A. One hot encoding and hashing_trick transformation-performance comparision. In: IEEE. *2024 14th International Conference on Advanced Computer Information Technologies (ACIT)*. [S.l.], 2024. p. 699–704. Citado na página 21.

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*. [S.l.]: Association for Computational Linguistics, 2014. p. 530–539. Citado na página 39.

LILLEBERG, J.; ZHU, Y.; ZHANG, Y. Support vector machines and word2vec for text classification with semantic features. In: IEEE. *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. [S.l.], 2015. p. 136–140. Citado na página 25.

LIU, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press, 2020. Citado 2 vezes nas páginas 17 e 19.

LOPES, A. da S.; BROTAS, A. M. P.; MASSARANI, L. The public conversation about vaccines and vaccination against covid-19 on twitter: an infodemiological study. *Intercom: Revista Brasileira de Ciências da Comunicação*, SciELO Brasil, v. 46, p. e2023121, 2023. Citado na página 26.

LOPYREV, K. Generating news headlines with recurrent neural networks. *arXiv preprint arXiv:1512.01712*, 2015. Disponível em: <https://arxiv.org/abs/1512.01712>. Citado na página 22.

MACHADO, E. *Jornalismo digital e a transformação dos meios tradicionais*. São Paulo: Editora XYZ, 2018. Citado na página 20.

MAKARENKOV, V.; SHAPIRA, B.; ROKACH, L. Language models with pre-trained (glove) word embeddings. *arXiv preprint arXiv:1610.03759*, 2016. Citado na página 21.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008. Citado na página 18.

MANOVICH, L. Cultural analytics: The computational analysis of culture. In: _____. [S.l.]: MIT Press, 2017. Citado na página 15.

MCCOMBS, M. E.; SHAW, D. L. The agenda-setting function of mass media. *Public Opinion Quarterly*, v. 36, n. 2, p. 176–187, 1972. Citado na página 20.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Egyptian Informatics Journal*, Elsevier, v. 15, n. 4, p. 127–141, 2014. Citado na página 17.

MELLO, G. L. de; FINGER, M.; SERRAS, F.; CARPI, M. de M.; JOSE, M. M.; DOMINGUES, P. H.; CAVALIM, P. Pelle: Encoder-based language models for brazilian portuguese based on open data. *arXiv preprint arXiv:2402.19204*, 2024. Disponível em: <https://arxiv.org/abs/2402.19204>. Citado na página 24.

MELO, T. de; FIGUEIREDO, C. M. et al. Comparing news articles and tweets about covid-19 in brazil: sentiment analysis and topic modeling approach. *JMIR Public Health and Surveillance*, JMIR Publications Inc., Toronto, Canada, v. 7, n. 2, p. e24585, 2021. Citado 4 vezes nas páginas 30, 32, 33 e 37.

MIMNO, D.; WALLACH, H. M.; TALLEY, E.; LEENDERS, M.; MCCALLUM, A. Optimizing semantic coherence in topic models. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. [S.l.]: Association for Computational Linguistics, 2011. p. 262–272. Citado na página 39.

NASCIMENTO, F.-L. S. C. do. História, interculturalidade e a valorização social e educacional do festejo junino maranhense. *Educação, Ciência e Cultura*, v. 26, n. 2, p. 01–15, 2021. Citado na página 14.

NEMANI, P.; VOLLALA, S. A cognitive study on semantic similarity analysis of large corpora: A transformer-based approach. *arXiv preprint arXiv:2207.11716*, 2022. Disponível em: <https://arxiv.org/abs/2207.11716>. Citado na página 24.

OLIVEIRA, L. E. S. e.; PETERS, A. C.; SILVA, A. M. P. D.; GEBELUCA, C. P.; GUMIEL, Y. B.; CINTHO, L. M. M.; CARVALHO, D. R.; HASAN, S. A.; MORO, C. M. C. Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, Springer, v. 13, n. 1, p. 13, 2022. Citado na página 34.

ÖNDEN, A.; ALNOUR, M.; SIMIC, V.; PAMUCAR, D. The evolution of sentiment analysis across various scientific disciplines: A comprehensive review based on the bibliometric technique. *Decision Making Advances*, v. 2, n. 1, p. 222–237, 2024. Citado na página 17.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, Now Publishers Inc., v. 2, n. 1, p. 1–135, 2008. Citado na página 17.

Prefeitura de São Luís. *São Luís tem números recordes no turismo durante o São João, segundo pesquisa da SETUR*. 2023. Acessado em: 17 jan. 2025. Disponível em: <https://www.saoluis.ma.gov.br/setur/noticia/40461/sao-luis-tem-numeros-recordes-no-turismo-durante-o-sao-joao-segundo-pesquisa-da-setur>. Citado na página 14.

Radio Timbira. *Em 60 dias, o São João do Maranhão 2024 terá mais turistas do que em 2023*. 2024. Acessado em: 17 jan. 2025. Disponível em: <https://radiotimbira.ma.gov.br/em-60-dias-o-sao-joao-do-maranhao-2024-tera-mais-turistas-do-que-em-2023/?utm_source=chatgpt.com>. Citado na página 14.

RAPARTHI, M.; DODDA, S. B.; REDDY, S. R. B.; THUNKI, P.; MARUTHI, S.; RAVICHANDRAN, P. Advancements in natural language processing-a comprehensive review of ai techniques. *Journal of Bioinformatics and Artificial Intelligence*, v. 1, n. 1, p. 1–10, 2021. Citado na página 24.

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. [S.l.]: ACM, 2015. p. 399–408. Citado na página 39.

RODRIGUES, T. Redes neurais recorrentes: Rnn, lstm e gru. 2023. Disponível em: <https://rodriguesthiago.me/posts/redes_neurais_recorrentes/>. Citado na página 22.

ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 8, p. 842–866, 2021. Citado na página 24.

SANDHU, T. *Exploration of Word Embeddings with Graph-Based Context Adaptation for Enhanced Word Vectors*. Dissertação (Mestrado) — University of Windsor (Canada), 2024. Citado na página 20.

SARANYA, M.; AMUTHA, B. A survey of machine learning technique for topic modeling and word embedding. In: *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*. [S.l.: s.n.], 2024. v. 1, p. 1–6. Citado na página 21.

SARICA, S.; LUO, J. Stopwords in technical language processing. *Journal of the Association for Information Science and Technology*, Wiley Online Library, v. 72, n. 7, p. 830–841, 2021. Citado na página 36.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, Elsevier, v. 181, p. 526–534, 2021. Citado na página 34.

SERBAN, I. V.; SORDONI, A.; SHIN, J.; COURVILLE, A.; BENGIO, Y.; PINEAU, J. Building end-to-end dialogue systems using generative hierarchical neural network models. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. [S.l.: s.n.], 2016. p. 3257–3265. Citado na página 21.

SILVA, J. *Mídias sociais e o impacto na disseminação de notícias*. Rio de Janeiro: Editora ABC, 2020. Citado na página 20.

SILVA, J.; RIBEIRO, B.; PARDO, T. A. S.; OLIVEIRA, H. G.; VALE, O. Sentilex-pt: Uma base de dados lexical para análise de sentimentos em português. *1Library*, 2020. Disponível em: <https://1library.org/article/abordagem-baseada-em-l%C3%A9xico-an%C3%A1lise-de-sentimentos.yjj5vx6y>. Citado na página 18.

SMITH, N. A. Contextual word representations: A contextual introduction. *CoRR*, abs/1902.06006, 2019. Disponível em: <http://arxiv.org/abs/1902.06006>. Citado na página 21.

SOCHER, R.; PERELYGIN, A.; WU, J.; CHUANG, J.; MANNING, C. D.; NG, A. Y.; POTTS, C. Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2013. p. 1631–1642. Citado na página 21.

SOUZA, J. A. *Análise de Sentimentos em Redes Sociais: Uma Comparação entre Abordagens Baseadas em Léxico e Aprendizado de Máquina*. Tese (Doutorado) — Universidade Federal Rural do Semi-Árido, 2021. Disponível em: <https://repositorio.ufersa.edu.br/items/97383d6c-63f6-4c87-9f21-c634edf2f1a3>. Citado na página 18.

STEWART, G.; AL-KHASSAWENEH, M. An implementation of the hdbscan* clustering algorithm. *Applied Sciences*, MDPI, v. 12, n. 5, p. 2405, 2022. Citado na página 28.

STROBL, M.; SANDER, J.; CAMPELLO, R. J.; ZAÏANE, O. Model-based clustering with hdbscan. In: SPRINGER. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*. [S.l.], 2021. p. 364–379. Citado na página 28.

SUN, B.; WU, J.; GUO, C.; CHEN, K. A one-hot encoding approach for signal integrity enhancement of intra-chip interconnects. In: IEEE. *2024 25th International Conference on Electronic Packaging Technology (ICEPT)*. [S.l.], 2024. p. 01–06. Citado na página 21.

SUN, P.; ZUO, Y.; WANG, Y. Classification model for navtex navigational warning messages based on adaptive weighted tf-idf. In: *Proceedings of the 10th Multidisciplinary International Social Networks Conference*. [S.l.: s.n.], 2023. p. 133–142. Citado na página 25.

SUNG-SU, S.; HOE-CHANG, Y. A study on leadership trends from the perspective of domestic researcher's using bertopic and lda. *East Asian Journal of Business Economics (EAJBE)*, East Asia Business Economics Association, v. 11, n. 1, p. 53–71, 2023. Citado na página 26.

THELWALL, M. This! identifying new sentiment slang through orthographic pleonasm online: Yasss slay gorg queen ilysm. *IEEE Intelligent Systems*, IEEE, v. 36, n. 4, p. 114–120, 2021. Citado na página 35.

VANIA, C.; VULIć, I.; GAšIć, M. Contextual embeddings: When are they worth it? *arXiv preprint arXiv:2005.09117*, 2020. Disponível em: <https://arxiv.org/abs/2005.09117>. Citado na página 23.

VASWANI, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. Citado 3 vezes nas páginas 22, 23 e 24.

WANKHADE, M.; RAO, A. C. S.; KULKARNI, C. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, Springer, v. 55, n. 7, p. 5731–5780, 2022. Citado na página 17.

WEEDS, J.; WEIR, D. Stopword selection for text mining. *Proceedings of the 7th European Conference on Principles of Data Mining and Knowledge Discovery*, p. 7–12, 2004. Disponível em: <https://www.researchgate.net/publication/220674968_Stopword_Selection_for_Text_Mining>. Citado na página 18.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: CITESEER. *Proceedings of the 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining*. [S.l.], 2000. p. 29–39. Citado na página 34.

World Population Review. *Twitter Users by Country 2024*. 2024. Acesso em: 14 dez. 2024. Disponível em: <https://worldpopulationreview.com/country-rankings/twitter-users-by-country>. Citado 2 vezes nas páginas 14 e 19.

WU, Y.; SCHUSTER, M.; CHEN, Z.; LE, Q. V.; NOROUZI, M.; MACHEREY, W.; KRIKUN, M.; CAO, Y.; GAO, Q.; MACHEREY, K.; KLINGNER, J.; SHAH, A.; JOHNSON, M.; LIU, X.; KAISER Łukasz; GOUWS, S.; KATO, Y.; KUDO, T.;

KAZAWA, H.; STEVENS, K.; KURIAN, G.; PATIL, N.; WANG, W.; YOUNG, C.; SMITH, J.; RIESA, J.; RUDNICK, A.; VINYALS, O.; CORRADO, G.; HUGHES, M.; DEAN, J. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. Disponível em: <https://arxiv.org/abs/1609.08144>. Citado 2 vezes nas páginas 21 e 24.

X Corp. *X (formerly Twitter)*. 2024. <https://x.com>. Accessed: 2024-12-16. Citado na página 35.

YANG, D.; MA, M.; WEI, V.; LI, J.; ZHOU, J.; SONG, X.; GUAN, Z.; CHEN, X. Multiscale modelling for fatigue crack propagation of notched laminates using the umap clustering algorithm. *Thin-Walled Structures*, Elsevier, v. 199, p. 111819, 2024. Citado na página 28.

YANG, H.; ZHU, D.; HE, S.; XU, Z.; LIU, Z.; ZHANG, W.; CAI, J. Enhancing psychiatric rehabilitation outcomes through a multimodal multitask learning model based on bert and tabnet: An approach for personalized treatment and improved decision-making. *Psychiatry Research*, Elsevier, v. 336, p. 115896, 2024. Citado na página 24.

ZEDEUS. *Nitter*. [S.l.]: GitHub, 2019. <https://github.com/zedeus/nitter>. Accessed: 2024-06-28. Citado na página 35.

ZHAI, M.; WANG, X.; ZHAO, X. The importance of online customer reviews characteristics on remanufactured product sales: Evidence from the mobile phone market on amazon. com. *Journal of Retailing and Consumer Services*, Elsevier, v. 77, p. 103677, 2024. Citado na página 28.

ZHANG, C.; PENG, B.; SUN, X.; NIU, Q.; LIU, J.; CHEN, K.; LI, M.; FENG, P.; BI, Z.; LIU, M. et al. From word vectors to multimodal embeddings: Techniques, applications, and future directions for large language models. *arXiv preprint arXiv:2411.05036*, 2024. Citado na página 21.