UNIVERSIDADE FEDERAL DO OESTE DO PARÁ

INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS

BACHARELADO INTERDISCIPLINAR EM CIÊNCIA E TECNOLOGIA

GABRIELE DE SOUSA ARAÚJO

PROCESSAMENTO DE LINGUAGEM NATURAL EM ANÁLISE DE MÍDIAS SOCIAIS: UM MAPEAMENTO SISTEMÁTICO

Santarém - PA

2023

GABRIELE DE SOUSA ARAÚJO

PROCESSAMENTO DE LINGUAGEM NATURAL EM ANÁLISE DE MÍDIAS SOCIAIS: UM MAPEAMENTO SISTEMÁTICO

Trabalho de Conclusão de Curso apresentado ao Programa de Ciência e Tecnologia para obtenção do grau de Bacharelado Interdisciplinar em Ciência e Tecnologia pela Universidade Federal do Oeste do Pará.
Orientador(a): Profº. Dr. Fábio Manoel França Lobato

Santarém - PA

2023

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema Integrado de Bibliotecas – SIBI/UFOPA

A663p Araújo, Gabriele de Sousa

Processamento de linguagem natural em análise de mídias sociais: um mapeamento sistemático./ Gabriele de Sousa Araújo. - Santarém, 2023.

124 p. : il.

Inclui bibliografias.

Orientador: Fábio Manoel França Lobato.

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Oeste do Pará, Instituto de Engenharia e Geociências, Programa de Ciência e Tecnologia, Bacharelado Interdisciplinar em Ciência e Tecnologia

1. Inteligência Artificial. 2. Processamento de Linguagem Natural. 3. Mineração de Texto. 4. Mapeamento Sistemático. 5. Mídias Sociais. I. Lobato, Fábio Manoel França, *orient.* II. Título.

CDD: 23 ed. 006.3

Bibliotecária - Documentalista: Cátia Alvarez – CRB/2 843

SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DO OESTE DO PARÁ
INSTITUTO DE ENGENHARIA E GEOCIÊNCIAS
PROGRAMA DE CIÊNCIA E TECNOLOGIA

ATA DE AVALIAÇÃO DE TCC

No dia 14 de julho do ano de dois mil e vinte e três, na sala 213, do Bloco B do Núcleo de Salas de aulas da Unidade Tapajós, da Universidade Federal do Oeste do Pará, às 16:00 horas, reuniu-se a Banca Examinadora de TCC composta pela Prof. Fábio Manoel Franca Lobato (orientador e presidente da banca), Prof. Manoel Maria Bezerra Neto e Prof. Marcelino Silva da Silva. A reunião teve por objetivo avaliar o trabalho de conclusão de curso de **Bacharelado Interdisciplinar em Ciência e Tecnologia** da discente GABRIELE DE SOUSA ARAUJO. O trabalho foi aberto pelo orientador, seguido pela apresentação realizada pela estudante. Após a apresentação, cada examinador fez perguntas à estudante durante as arguições. Após a conclusão das arguições, a Banca Examinadora procedeu ao julgamento do trabalho, chegando à conclusão de que a discente está (X) **APROVADA** ( ) **REPROVADA**, com nota <9,9 >. Nada mais havendo a tratar, foi a presente ata lavrado por mim, Fabio Manoel Franca Lobato, que vai assinada pelos membros da Banca Examinadora.

Santarém, 14 de julho de 2023.

Fábio Manoel Franca Lobato (Orientador)

Manoel Maria Bezerra Neto

Marcelino Silva da Silva

AGRADECIMENTOS

A minha família, em especial aos meus avós Rafael e Maria, aos meus tios Lene e Elson e ao meu pai Ronaldo. Eles são os pilares da pessoa que me tornei e sou eternamente grata pelo amor incondicional que recebi, pelas valiosas lições que aprendi e pelo apoio incansável ao longo desta jornada acadêmica.

Ao Prof. Dr. Fábio Lobato por ter me selecionado para fazer parte do grupo de pesquisa do LACA e ter me dado a honra de tê-lo como orientador. Sua orientação e expertise foram fundamentais para o sucesso deste trabalho e sou grata por tê-lo como guia nessa trajetória.

Aos meus amigos Domingas, Hevellym, Andreyna e Gabriel que caminharam comigo durante esse percurso.

Aos meus amigos Adrielson, Jéssica e Jonathan que estiveram ao meu lado durante a produção desse trabalho. Suas doses diárias de companheirismo, risadas e suporte tornaram essa experiência mais leve e prazerosa.

A todos que não foram mencionados acima, mas que de alguma forma contribuíram para este trabalho.

RESUMO

O número de plataformas de mídia social aumentou significativamente assim como o número de usuários ativos. Mais de 18,2 milhões de mensagens de texto são transmitidas a cada minuto nessas plataformas. Diante da quantidade de dados disponíveis, técnicas de Processamento de Linguagem Natural (PLN) têm sido utilizadas por diversos pesquisadores para analisar essa grande quantidade de dados não estruturados. Assim, é essencial entender as principais tendências e desafios da análise de mídias sociais, especialmente em eventos científicos. Nessa perspectiva, este estudo apresenta um mapeamento sistemático do uso da PLN em trabalhos publicados em cinco eventos acadêmicos: BRACIS, BraSNAM, ENIAC, STIL e PROPOR, estes eventos foram escolhidos dado a relevância dos mesmos. O estudo visa identificar as principais ferramentas e técnicas utilizadas, tarefas executadas, as fontes dos dados, e medidas de avaliação. Para tanto, 186 estudos foram analisados e selecionados criteriosamente dentre os 654 artigos publicados nesses eventos nos três anos (2020 a 2022), este período de tempo foi escolhido dada a dinamicidade da área, logo, sua rápida obsolescência. Os resultados mostram um recorte do cenário atual sobre o assunto e apontam áreas que podem ser melhoradas em pesquisas futuras utilizando técnicas para tarefas como classificação de textos, análise de sentimentos, e reconhecimento de entidades nomeadas. Assim, este trabalho pode ser útil para acadêmicos interessados em explorar o potencial dessas ferramentas e técnicas, ter uma visão clara das lacunas, desafios e oportunidades de pesquisa nessa área e analisar o cenário atual em pesquisas envolvendo PLN e mídias sociais. E ainda, guiar o setor produtivo nessa transferência de conhecimento, diminuindo a lacuna entre o estado da arte e da prática, aumentando a competitividade e inovação das ferramentas de análise de mídias sociais.

Palavras-chaves: Inteligência Artificial, Processamento de Linguagem Natural, Mineração de Texto, Mapeamento Sistemático, Mídias Sociais.

ABSTRACT

The number of social media platforms has increased significantly, as has the number of active users. More than 18.2 million text messages are transmitted every minute on these platforms. Given the amount of data available, Natural Language Processing (NLP) techniques have been used by several researchers to analyze this large amount of unstructured data. Thus, it is essential to understand social media analysis's main trends and challenges, especially in scientific events. In this perspective, this study presents a systematic mapping of PLN for social media analysis in works published in five academic events: BRACIS, BraSNAM, ENIAC, STIL, and PROPOR. These events were chosen due to their relevance. The study aims to identify the main tools and techniques used, tasks performed, data sources, and evaluation measures. For this purpose, 186 studies were analyzed and carefully selected among the 654 articles published in these events in the three years (2020 to 2022). The results show a clipping of the current scenario on the subject and point out areas that can be improved in future research using techniques such as text classification, sentiment analysis, and recognition of named entities. Thus, this work can be helpful for academics interested in exploring the potential of these tools and techniques, having a clear view of gaps, challenges, and research opportunities in this area, as well as analyzing the current scenario in research involving NLP and social media. Nevertheless, guide the productive sector in this knowledge transfer, reducing the gap between the state of the art and practice and increasing the competitiveness and innovation of social media analysis tools.

Key-words: Natural Language Processing, Text Mining, Systematic Mapping, Social media, Social networks.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACM HT	*ACM Conference on Hypertext and Hypermedia*
APIs	*Application Programming Interfaces*
ASSIN	Avaliação de Similaridade Semântica e Inferência Textual
BERT	*Bidirectional Encoder Representations form Transformers*
BiLSTM-CRF	*Bidirectional LSTM with Conditional Random Fields*
BoW	*Bag-of-words*
BRACIS	*Brazilian Conference on Intelligent Systems*
BraSNAM	*Brazilian Workshop on Social Network Analysis and Mining*
CNNs	*Convolutional Neural Networks*
CI	Critérios de Inclusão
CE	Critérios de Exclusão
CSV	*Comma-separated Values*
DL	*Deep Learning*
DTC	*Decision Tree Classifier*
ENIAC	Encontro Nacional de Inteligência Artificial e Computacional
GloVe	*Global Vectors*
IA	Inteligência Artificial
IC	Inteligência Computacional
ICSMS	*International Conference on Social Media & Society*
ICWSM	*AAAI Conference on Web and Social Media*
IEEE	*Institute of Electrical and Electronics Engineers*
LIWC	*Linguistic Inquiry and Word Count*
LNAI	*Lecture Notes in Artificial Intelligence*
LNCS	*Lecture Notes in Computer Science*

LSTM	*Long Short-term Memory*
LR	*Logistic Regression*
ML	*Machine Learning*
MLP	*Multi-layer Perceptron*
M-BERT	Multilingual BERT
NER	*Named Entity Recognition*
NLTK	*Natural Language Toolkit*
PLN	Processamento de Linguagem Natural
POS tagging	*Part-of-speech*
PROPOR	*International Conference on Computational Processing of Portuguese Language*
RF	*Random Forest*
RNNs	*Recurrent Neural Networks*
RSL	Revisão Sistemática da Literatura
SBC	Sociedade Brasileira de Computação
SOL	*SBC-OpenLib*
SNS	*Social networking service*
STIL	Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency-Inverse Document Frequency*
UGC	*User Generated Content*
WASSA	*Workshop on Computational Approaches to Subjectivity*

SUMÁRIO

1 INTRODUÇÃO

Neste Capítulo são abordadas as considerações iniciais sobre o presente trabalho, contextualizando e destacando as razões e justificativas por trás da pesquisa, a fim de compreender sua relevância e motivações. Em seguida, são apresentados os objetivos que serão alcançados ao longo do estudo, os potenciais impactos encontrados. Por fim, é descrita a sequência em que o documento está organizado, fornecendo uma visão geral da estrutura do trabalho.

1.1 Contextualização e Proposta

As mídias sociais permitem a conexão entre indivíduos e ajudam a quebrar barreiras na comunicação, proporcionando a oportunidade para que qualquer pessoa possa contar suas histórias e compartilhar suas opiniões (HOU; HAN; CAI, 2020).

Kaplan e Haenlein (2010, p.61) descrevem que:

> A mídia social é um grupo de aplicativos baseados na Internet e nos fundamentos ideológicos e tecnológicos da Web 2.0 que permitem a criação e troca de Conteúdo Gerado pelo Usuário (UGC, *User Generated Content*).

Nesse sentido, podemos pensar nas mídias sociais como as principais plataformas e suas funcionalidades, tais como *Facebook*, *Instagram* e *Twitter*. Também podemos, em termos práticos, entender as mídias sociais como um canal adicional de marketing digital em que os profissionais podem aproveitar para estabelecer comunicação com os consumidores por meio de estratégias publicitárias. Nessa perspectiva, significa que a mídia social se torna menos sobre as tecnologias ou plataformas específicas e mais sobre o compartilhamento de informações entre os usuários (APPEL *et al.*, 2020; PAN; TORRES; ZÚÑIGA, 2019).

Ao longo dos anos, o número de plataformas de mídia social e usuários ativos nessas plataformas aumentou significativamente, tornando-se uma das aplicações mais importantes da internet (AICHNER *et al.*, 2021). Este fato, consequentemente, levou ao surgimento da comunicação via texto, com mais de 18,2 milhões de mensagens de texto transmitidas a cada minuto. (BALAJI; ANNAVARAPU; BABLANI, 2021). Os dados gerados pelos usuários têm despertado o interesse acadêmico, resultando na crescente importância do campo de análise de mídias sociais, que envolve a coleta e análise de vários dados de mídias sociais e a extração de informações valiosas e ocultas (CHOI *et al.*, 2020).

De acordo com Zeng *et al.* (2010, p.14),

> A análise de mídia social se preocupa com o desenvolvimento e avaliação de ferramentas e estruturas de informática para coletar, monitorar, analisar, resumir e visualizar dados de mídia social, geralmente conduzidos por requisitos específicos de um aplicativo de destino.

Com isso, a análise de mídias sociais envolve a identificação de características comuns entre as diversas plataformas disponíveis (LIN *et al.*, 2021; ROMA; ALOINI, 2019).

Sob o mesmo ponto de vista, a grande quantidade de UGC produzido diariamente e o número expressivo de usuários ativos são motivadores para as organizações, que buscam compreender as questões e tendências emergentes a fim de identificar riscos e oportunidades na comunicação (ZHUANG *et al.*, 2023). Tanto empresas quanto organizações sem fins lucrativos utilizam ferramentas que se comunicam com as *Application Programming Interfaces* (APIs) das plataformas de mídias sociais para coletar os dados produzidos pelos usuários, buscando obter *insights* valiosos sobre a comunicação de massa (NAEEM; OZUEM, 2022). Assim, há uma busca crescente de métodos para lidar com esse volume de informações, principalmente pela disponibilidade de dados textuais nas mídias sociais.

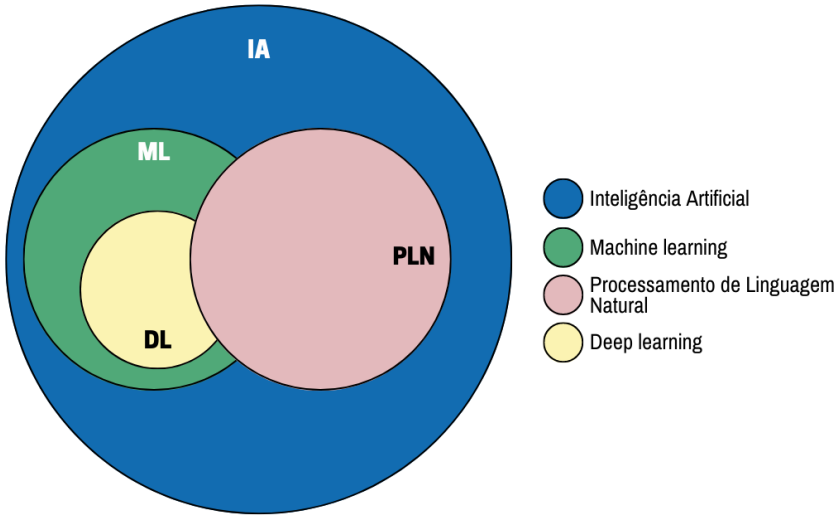
Para lidar com esses dados, técnicas avançadas de Inteligência Artificial (IA), como *Machine Learning* (ML) e *Deep Learning* (DL) são desenvolvidas e desempenham um papel fundamental (SADIKU *et al.*, 2021). A IA, sendo um campo multidisciplinar que envolve a combinação de conhecimento e métodos de ciência da computação, lógica, biologia, psicologia, filosofia e muitas outras disciplinas, cujo o objetivo é desenvolver sistemas capazes de realizar tarefas que demandam inteligência humana, como reconhecimento de padrões, tomada de decisões e Processamento de Linguagem Natural (PLN), do inglês *Natural Language Processing* (NLP) (LU, 2019; DUAN *et al.*, 2009).

Por sua vez, o ML é uma subárea da IA que se concentra no desenvolvimento de algoritmos e modelos capazes de aprender com os dados, identificar padrões e fazer previsões ou tomar decisões com base nesses padrões (MA; SUN, 2020). Já o DL é uma forma mais avançada de ML que utiliza Redes Neurais Profundas (do inglês *Deep Neural Networks* ou DNNs), uma bordagem de Rede Neural Artificial (RNA) para aprender representações mais complexas dos dados e executar tarefas sofisticadas por meio do aprendizado automático de recursos de dados. (SHARMA; SHARMA; JINDAL, 2021). Assim, o DL amplia as capacidades de aprendizado e análise dos modelos de ML.

O PLN é um subcampo da IA que usa técnicas computacionais para que os computadores possam aprender, entender e produzir conteúdo de linguagem humana, a partir da enorme quantidade de dados linguísticos disponíveis (HIRSCHBERG; MANNING, 2015), por isso também é chamada de Linguística Computacional. A área de PLN se concentra na interpretação, análise e manipulação de dados de linguagem natural para uma finalidade específica, usando diferentes algoritmos, ferramentas e métodos. Portanto, ML, DL e PLN são todos subcampos dentro da IA, e a relação entre eles é representada na Figura 6.

Dessa forma, a intersecção dessas áreas de estudo aliadas à capacidade do PLN de interpretar e analisar dados linguísticos, impulsionam o desenvolvimento de modelos, algoritmos e abordagens inovadoras na análise de mídias sociais (ZHANG; LU, 2021). Assim, o PLN tem se destacado como uma área fundamental nesse campo (LEE, 2018). Nesse contexto, existem

Figura 1 – Correlação entre IA, ML, DL e PLN.



Fonte: Adaptado de Vajjala *et al.* (2020).

diversas ferramentas de mineração de texto disponíveis, desde ferramentas de código aberto simples até bibliotecas que oferecem uma ampla gama de recursos e funcionalidades para a realização de tarefas de coleta, manipulação, limpeza e análise desses dados (BATRINCA; TRELEAVEN, 2015). Vinculada a essas ferramentas, a análise de mídia social envolve o uso de diferentes técnicas de modelagem e análise de vários campos (DERAKHSHAN; BEIGY, 2019). Essas técnicas abrangem a aplicação de algoritmos de ML e DL para tarefas como classificação de textos, análise de sentimentos, sumarização e tradução automática, extração de entidades, entre outras aplicações (BALAJI; ANNAVARAPU; BABLANI, 2021; HAYAT *et al.*, 2019; VAJJALA *et al.*, 2020).

No entanto, muitos desafios podem depender do contexto dos dados de linguagem natural, dificultando o alcance de todos os objetivos com uma única abordagem, tais como variação linguística, disponibilidade dos dados e a falta de conjuntos de dados específicos e de modelos de aprendizado em tempo real e não supervisionados (NASCIMENTO, 2019; OLIVEIRA *et al.*, 2021). Por esse motivo, o desenvolvimento de diferentes ferramentas e métodos no campo do PLN e áreas relevantes de estudo têm sido amplamente estudadas por diversos pesquisadores (KHURANA *et al.*, 2023), incluindo ferramentas e métodos específicos adaptados ao UGC (JÚNIOR *et al.*, 2020).

Com o crescimento das comunidades brasileiras de inteligência artificial, ciência de dados, análise de mídias sociais e PLN, nos perguntamos como o conhecimento nessas áreas está sendo disseminado (LOBATO; SOUSA; JR, 2021). De acordo com as pesquisas e estudos disponíveis até à elaboração deste trabalho, não há levantamento na literatura sobre métodos e técnicas de análise de mídias sociais utilizadas em eventos brasileiros nas comunidades acima citadas.

Para preencher essa lacuna na literatura, realizamos um mapeamento sistemático com o objetivo de fornecer uma visão geral da aplicação das técnicas de PLN na análise de mídias sociais, identificar os algoritmos mais usados e entender as tendências atuais no uso de PLN nesse contexto. Ao realizar esse mapeamento sistemático, é possível obter uma visão abrangente do estado da arte e da prática. Grant e Booth (2009) apontam que o mapeamento sistemático permite identificar lacunas de pesquisa na literatura existente, facilitando a identificação de áreas que necessitam de revisões adicionais e/ou novas pesquisas primárias, enriquecendo assim o conhecimento acadêmico e científico sobre o tema em questão. Para tal, foi adotada uma abordagem baseada na Revisão Sistemática da Literatura (RSL) proposta por (KITCHENHAM; CHARTERS *et al.*, 2007), visando assim criar uma base consistente para investigações futuras em áreas específicas relacionadas ao tema proposto pela autora.

Para determinar o cenário brasileiro atual, foram selecionados os cinco principais eventos científicos que publicam trabalhos na interseção da PLN e das Mídias Sociais, a saber: *Brazilian Conference on Intelligent Systems* (BRACIS), *Brazilian Workshop on Social Network Analysis and Mining* (BraSNAM), Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL) e o *International Conference on Computational Processing of Portuguese Language* (PROPOR). Consideramos o período de três anos (2020 a 2022) em nossa análise, totalizando 654 artigos listados, dos quais 186 (30%) foram examinados. Este período de tempo foi escolhido dada a dinamicidade da área, logo, sua rápida obsolescência.

Os resultados obtidos são úteis para pesquisadores e profissionais interessados em explorar o potencial dessas ferramentas e técnicas, ter uma visão clara das lacunas, desafios e oportunidades de pesquisa nessa área e analisar o cenário atual em pesquisas envolvendo PLN e mídias sociais. É fundamental ressaltar que estamos seguindo os princípios da Ciência Aberta ao disponibilizar todos os nossos dados em um repositório GitHub disponível publicamente, permitindo a reprodutibilidade dos resultados.

1.2 Objetivos

Diante do contexto apresentado, estabeleceram-se os seguintes objetivos:

1.2.1 Objetivo Geral

O objetivo deste estudo é investigar e compreender as tendências atuais no uso de PLN em análise de mídias sociais, com enfoque nos eventos acadêmicos brasileiros.

1.2.2 Objetivos Específicos

À luz do objetivo geral, destacam-se os seguintes objetivos específicos:

1. Identificar as principais ferramentas e técnicas, assim como as tarefas relacionadas de PLN utilizadas em análise de mídias sociais. Baseado nesse objetivo, uma das perguntas de pesquisa que norteia esse trabalho:

    - Quais ferramentas e técnicas de PLN são mais utilizadas nos eventos científicos selecionados?

2. Analisar as fontes e a natureza dos dados usadas para a implementação das técnicas de PLN na análise de mídias sociais. A questão de pesquisa relacionada a esse objetivo:

    - Quais são as fontes e a natureza dos dados usados na análise de mídias sociais?

3. Identificar as medidas de avaliação mais utilizadas nos estudos de PLN para medir a eficácia das técnicas aplicadas na análise de mídias sociais. A seguinte questão de pesquisa está associada a esse objetivo:

    - Quais são as medidas de avaliação mais utilizadas nos estudos de PLN?

1.3 Potenciais Impactos

Esse trabalho visa apresentar um panorama atualizado do estado da arte e da prática no que diz respeito a aplicação de PLN em análise de mídias sociais análise de mídias sociais. Espera-se que o mesmo possa ser usado como base para o avanço contínuo do conhecimento nessa área.

Para a comunidade de mineração de textos, espera-se que esse estudo identifique lacunas, desafios e oportunidades de pesquisa nessas áreas específicas. Ao destacar as áreas em que ainda há necessidade de desenvolvimento e investigação, o trabalho pode direcionar a atenção dos pesquisadores para os tópicos mais relevantes e promissores. Isso pode estimular discussões e colaborações futuras, impulsionando o progresso na mineração de textos aplicada às mídias sociais. Já para o setor produtivo, o presente estudo contribui nessa transferência de conhecimento, diminuindo a lacuna entre o estado da arte e da prática, aumentando a competitividade e inovação das ferramentas de análise de mídias sociais.

Os resultados obtidos nessa pesquisa foram submetidos na conferência BRACIS, que tem como tópico de interesse estudo voltados para IA e PLN. O artigo enviado para submissão se encontra no Apêndice A, porém devido às exigências da conferência em trabalhos que apresentem artefatos e aprimoramento de técnicas existentes, o artigo foi dado como rejeitado. O *feedback* dos revisores pode ser visualizado nos Anexos A. Uma das sugestões dos revisores é que seja submetido ao ENIAC, apontando como um evento mais aberto a esse tipo de pesquisa.

1.4 Organização do trabalho

O restante do trabalho encontra-se organizado como segue.

No Capítulo 2, são apresentados trabalhos relevantes que abordam a aplicação de técnicas de PLN na análise de texto. Essa revisão da literatura tem como objetivo fornecer um panorama dos estudos anteriores, destacando as abordagens, os desafios enfrentados e as contribuições oferecidas por essas pesquisas. Essa análise crítica da literatura permite situar o presente estudo dentro do contexto científico atual e identificar lacunas que podem ser exploradas.

No Capítulo 3, é apresentado o método de pesquisa utilizado para a realização do mapeamento sistemático. Detalhes sobre a seleção dos eventos acadêmicos brasileiros, a coleta de dados, os critérios de inclusão e exclusão dos artigos, bem como a análise dos dados, são apresentados nesse capítulo. Além disso, é descrito o protocolo de mapeamento sistemático adotado. Essa seção fornece transparência e reprodutibilidade ao processo de pesquisa adotado.

Os resultados obtidos a partir da análise exploratória dos artigos e os algoritmos mais relevantes encontrados são apresentados no Capítulo 4. Nessa seção, são apresentadas as principais descobertas, estatísticas e tendências identificadas ao analisar os estudos selecionados.

No Capítulo 5 são discutidos os resultados à luz da literatura, considerando também a amarração para com os objetivos do estudo e os achados do trabalho, destacando-se as contribuições, desafios e as limitações encontradas.

Por fim, no Capítulo 6, são apresentadas as considerações finais do estudo. Nessa seção, são feitas sínteses dos principais resultados, destacando-se as contribuições e as implicações práticas da pesquisa. Ademais, são discutidas possíveis direções para pesquisas futuras, apontando lacunas identificadas e oportunidades de aprimoramento na aplicação de técnicas de PLN na análise de mídias sociais.

2 TRABALHOS RELACIONADOS

As mídias sociais têm se tornado uma fonte importante de dados para análises de diferentes setores, incluindo negócios, governo e a indústria do lazer (HASSANI; MOSCONI, 2022; YIGITCANLAR *et al.*, 2020; MIRZAALIAN; HALPENNY, 2019). À medida que a quantidade de dados gerados diariamente cresce, as técnicas de análise de dados tornaram-se mais importantes do que nunca para fornecer *insights* valiosos (HE *et al.*, 2019).

Consequentemente, muitos pesquisadores têm explorado essa área com o objetivo de identificar a natureza dos dados e os domínios de pesquisa abordados por meio das análises realizadas. Esse fato permite uma melhor compreensão dos acontecimentos sociais. Neste sentido, Zachlod *et al.* (2022) conduziu uma revisão sistemática de 94 artigos que usaram ou discutiram a análise de dados de mídia social como o principal tópico de pesquisa entre 2017 e 2020, os autores justificam esse intervalo de tempo curto devido ao rápido avanço da área e às mudanças de acesso das grandes plataformas nos últimos anos. Desse modo, foi identificado que a maioria dos dados usados nas análises foram coletados do *Twitter*, *Facebook*, *Instagram*, *YouTube*, *TripAdvisor* e *LinkedIn*. Como esses dados são gerados por tipos muito diferentes de usuários, algumas áreas como o marketing, em particular, têm recebido muita atenção dos estudiosos. Além de que, áreas que precisam de informações instantâneas, como gerenciamento de desastres, hospitalidade e turismo, também são cobertas por esse tipo de análise. Similarmente à (ZACHLOD *et al.*, 2022), adotamos a janela de 3 anos para o processo de busca dos estudos, este período de tempo foi escolhido dada a dinamicidade da área, logo, sua rápida obsolescência.

As técnicas de PLN são comumente usadas para extrair e analisar o conteúdo criado pelos usuários (DRUS; KHALID, 2019; MADILA; DIDA; KAIJAGE, 2021). De acordo com Ghani *et al.* (2019), existem várias técnicas e métodos aplicados para a análise de dados de mídias sociais, e alguns dos principais focos são a classificação de emoções dos usuários, detecção de informações, espaço-temporais, agrupamento e avaliação de desempenho. Isso é reforçado por Choi *et al.* (2020), onde os autores realizam uma revisão sistemática de 57 estudos de mídia social focados em *Business Intelligence* (BI) entre 2014 - 2018. Três questões de pesquisa são propostas para extrair dados, metodologia e algoritmos desses trabalhos. Diante disso, diferentes plataformas e grupos são identificados, indicando que a maioria das plataformas utilizadas são do tipo *commercial review* (e.g., Amazon, Yelp e TripAdvisor) ou *social networking service* (SNS) (e.g., *Facebook*, *Twitter* e *LinkedIn*). Assim como, os algoritmos mais relevantes usados pelos pesquisadores foram análise de sentimentos, modelagem de tópicos, abordagem baseada em ML, abordagem baseada em rede e abordagem teórica para analisar os dados.

Ainda nesta perspectiva, (KHURANA *et al.*, 2023) teve como objetivo apresentar detalhadamente o estado da arte sobre tendências e desafios no campo da PLN, com trabalhos relevantes na literatura até o ano de 2022. É visto que ao longo dos anos, técnicas de ML e DL têm sido

usadas em diferentes tarefas de PLN. Por exemplo, modelos de redes neurais como *Convolutional Neural Networks* (CNNs) e as *Recurrent Neural Networks* (RNNs) são aplicados na classificação de sentenças, classificação de texto, resumo, tradução automática e recuperação de informações. Além de outras técnicas empregadas no aprendizado multitarefa como *Part-of-speech tagging* (POS tagging) e *Named Entity Recognition* (NER), em *Word Embedding* como *Global Vectors* (GloVe) e os Mecanismos de Atenção como os *Transformers*, sendo o *Bidirectional Encoder Representations from Transformers* (BERT) o mais utilizado. Os autores apontam que apesar dos avanços significativos, há desafios a serem enfrentados, um deles é a falta de modelos de linguagem abrangentes para diferentes domínios e áreas geográficas. Lidar com significados distintos de palavras e sentenças nessas áreas é problemático, devido à presença de linguagem informal, expressões idiomáticas e termos culturalmente específicos. Embora dados volumosos e treinamento regular possam melhorar os modelos, a complexidade persiste ao tratar palavras com significados diversos em diferentes regiões.

Embora tenha sido conduzida um extenso levantamento de trabalhos sobre o estado da arte e da prática no contexto brasileiro, não foi encontrado nenhum mapeamento sistemático da aplicação de técnicas de PLN, mais especificamente em análise de mídias sociais. Ao explorar trabalhos de revisão sistemática ou mapeamento sistemático que abordam esse campo, foi identificado estudos que se concentram principalmente na área geral da mineração de texto.

Por exemplo, Souza *et al.* (2018) conduziram um mapeamento sistemático de estudos relacionados à aplicação da mineração de texto para a língua portuguesa no período de 1996 a 2014. O estudo utilizou uma abordagem de busca automatizada em bibliotecas digitais (e.g., *Institute of Electrical and Electronics Engineers* (IEEE), Scopus e Scielo) e busca manual em vários anais de conferências realizadas no Brasil (e.g., PROPOR, BraSNAM e STIL). Percebeu-se que houve um aumento significativo em 2016, sendo que dos 203 estudos selecionados, 61% deles abordavam o português brasileiro. Dentre as 234 tarefas identificadas, a classificação de texto foi a mais abordada, representando 49% dos estudos. A técnica de pré-processamento mais utilizada para essa tarefa foi a remoção de *stopwords*, e os principais algoritmos empregados foram o *Support Vector Machine* (SVM) e o *Naïve Bayes*. Quanto às ferramentas, o *Natural Language Toolkit* (NLTK) foi a mais amplamente utilizada. Em relação às fontes de dados, aproximadamente 50% dos estudos basearam-se em notícias online (i.g., Folha de São Paulo e Público), enquanto o *Twitter* foi apontado como a principal fonte de mídias sociais. Como também, medidas de avaliação, como Precisão, Revocação e *F-measure*, estiveram presentes na maioria dos estudos analisados.

Em Júnior *et al.* (2020), os autores realizam um mapeamento de trabalhos publicados em conferências internacionais de grande impacto na área de análise de dados provenientes de mídias sociais. As conferências incluídas foram a *AAAI Conference on Web and Social Media* (ICWSM), o *Workshop on Computational Approaches to Subjectivity* (WASSA), a *ACM Conference on Hypertext and Hypermedia* (ACM HT) e a *International Conference on Social*

Media & Society (ICSMS). O estudo adotou uma abordagem de mapeamento sistemático, com perguntas de pesquisa direcionadas para identificar as bases de dados, ferramentas e algoritmos mais prevalentes nos estudos. Para a coleta dos dados, foi utilizado o método de *web scraping*, que permitiu obter os títulos, links, resumos e palavras-chave dos trabalhos. Ao final, um conjunto de dados foi criado para armazenar esses atributos, resultando na análise de 440 trabalhos publicados entre 2016 e 2019. Como resultado, uma análise quantitativa foi realizada, considerando a quantidade de trabalhos que utilizam cada um dos recursos extraídos. A base de dados mais amplamente utilizada foi o *Twitter*, seguida pelo *Facebook*, *Reddit* e *Wikipedia*. Em relação às ferramentas, o *Linguistic Inquiry and Word Count* (LIWC) foi identificado como a ferramenta mais utilizada, especialmente em tarefas de PLN. Outras ferramentas notáveis incluem o *Scikit-learn*, comumente utilizado em tarefas de ML, e o *Word2vec*. Quanto aos algoritmos, foram identificados o SVM, *Logistic Regression* (LR) e a *Long Short-term Memory* (LSTM) como os mais frequentes. Uma das limitações do estudo apresentada pelos autores foi a falta de uma análise relacionada a finalidade de cada algoritmo nos estudos.

No entanto, ainda há muito espaço para pesquisa e desenvolvimento nas áreas que PLN e análise de mídias sociais, principalmente com o surgimento de novas plataformas e a evolução das modelos de análise de dados, conforme discutido por Khurana *et al.* (2023). A análise de dados textuais enfrenta desafios significativos, especialmente no uso de técnicas e tarefas específicas para idiomas particulares. A limitação de algoritmos e ferramentas para esses idiomas é um obstáculo importante nesse cenário.

Assim como, poucas pesquisas têm sido direcionadas há eventos brasileiros, como no caso de Souza *et al.* (2018), cujo mapeamento abrangeu apenas até 2014, e Júnior *et al.* (2020) que se baseou em estudos de conferências internacionais. Portanto, a necessidade de um mapeamento sistemático direcionado à PLN em análise de mídias sociais advém da carência de trabalhos que evidenciam o estado da arte voltado para eventos acadêmicos brasileiros, a fim de preencher essa lacuna e fornecer uma visão abrangente do estado da arte no contexto nacional.

3 MATERIAIS E MÉTODOS

Visando alcançar os objetivos estabelecidos, neste capítulo serão descritos os procedimentos metodológicos adotados para o desenvolvimento desta pesquisa. Para isso, foi realizada uma análises dos conteúdos abordados nos trabalhos relacionados e a condução de um estudo de mapeamento sistemático usando a metodologia apresentada por Sinoara, Antunes e Rezende (2017) e Pelissari *et al.* (2022). Sua escolha deu-se ao fato destes serem baseados no percurso metodológico proposto por Kitchenham, Charters *et al.* (2007), onde é apresentado um guia útil para o planejamento e condução em estudos secundários.

Como apontado por Sinoara, Antunes e Rezende (2017), o mapeamento sistemático é uma técnica de revisão bibliográfica que, embora se diferencie de uma revisão sistemática pela profundidade e amplitude dos estudos analisados, segue um protocolo bem definido e pode ser utilizado para obter um mapeamento de publicações sobre algum assunto ou campo, identificando lacunas de pesquisa e áreas que requerem o desenvolvimento de estudos primários. Nesse sentido, a execução deste estudo consistiu de três etapas: a etapa de planejamento do estudo, em que o protocolo do mapeamento é definido, a etapa condução onde os estudos serão identificados e selecionados de acordo com os critérios estabelecidos no protocolo do mapeamento, e por fim, é realizada análise dos resultados, conforme descritos nas próximas seções.

3.1 Planejamento

Na etapa de planejamento, definiu-se o protocolo que guiou o estudo, no qual são descritas as perguntas de pesquisa, bem como o processo de pesquisa com as fontes nas quais os estudos foram mapeados, e a seleção dos estudos norteada pelos critérios de inclusão e exclusão.

3.1.1 Perguntas de Pesquisa

Com base no contexto apresentado nesta pesquisa, o problema de pesquisa abordado neste trabalho está relacionado à ausência de informações sobre os métodos e técnicas de análise de mídias sociais no contexto brasileiro. Para solucionar essa lacuna de conhecimento, foram formuladas uma série de perguntas de pesquisa (PPs) que buscaram ser respondidas por meio da aplicação do mapeamento sistemático. As PPs são apresentadas a seguir:

- PP1: Quais ferramentas e técnicas de PLN são mais utilizadas nos eventos científicos selecionados?
- PP2: Quais são as fontes e a natureza dos dados usados na análise de mídias sociais?
- PP3: Quais são as medidas de avaliação mais utilizadas nos estudos de PLN?

3.1.2 Processo de busca

O processo de pesquisa consistiu em uma busca manual dos anais dos eventos científicos, como conferências, simpósios, reuniões e *workshops* entre 2020 e 2022, listados na Tabela 1. Similarmente à (ZACHLOD *et al.*, 2022), adotamos a janela de 3 anos, dada a dinamicidade da área, logo, sua rápida obsolescência. Esses trabalhos estão disponíveis em duas fontes de pesquisa em bibliotecas digitais: o SBC-OpenLib (SOL) da Sociedade Brasileira de Computação (SBC) e o SpringerLink.

Diante disso, a escolha desses eventos é justificada por serem considerados importantes bases de pesquisa nacional em estudos relacionados às áreas de PLN, IA e Inteligência Computacional (IC), conforme apontam Lobato, Sousa e Jr (2021), Carvalho *et al.* (2022) e Pardo *et al.* (2010).

Tabela 1 – Anais de eventos selecionados.

Fonte	**Acrônimo**	**Edição**
Brazilian Conference on Intelligent Systems	BRACIS	2020–2022
Brazilian Workshop on Social Network Analysis and Mining	BraSNAM	2020–2022
Encontro Nacional de Inteligência Artificial e Computacional	ENIAC	2020–2022
International Conference on Computational Processing of the Portuguese Language	PROPOR	2020 e 2022
Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana	STIL	2021

Fonte: Elaborado pela autora.

3.1.2.1 Sobre os eventos

O *Brazilian Conference on Intelligent Systems* (BRACIS) originou-se da fusão dos dois mais importantes eventos científicos no Brasil nas áreas de IA e IC: o Simpósio Brasileiro de Inteligência Artificial e o Simpósio Brasileiro de Redes Neurais, e possui atualmente H5-index igual a 13 no *Google Metrics*[1] e estrato B1. É um evento que ocorre anualmente organizado pela SBC[2], em que busca promover teorias, aplicações que tratam do uso e análise de técnicas de IA e IC em vários campos relacionados, além pesquisas em nível internacional através do intercâmbio de ideias científicas entre pesquisadores, profissionais, cientistas e engenheiros. Seus anais podem ser encontrados na Springer em volumes da série *Lecture Notes in Computer Science* (LNCS) e *Lecture Notes in Artificial Intelligence* (LNAI) . Em conjunto a conferência, também acontece a cada ano o Encontro Nacional de Inteligência Artificial e Computacional (ENIAC) proporcionando um fórum para pesquisadores, profissionais, educadores e estudantes apresentarem e discutirem inovações e tendências em IA e IC, possui H5-index igual a 9 e

[1] `https://scholar.google.com.br/citations?view_op=top_venues`
[2] `https://www.sbc.org.br/`

estrato B2. É especialmente adequado para alunos de graduação e pós-graduação. Outro evento importante dentro do BRACIS, é o Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL), que acontece bianualmente organizado pela Comissão Especial de Processamento de Linguagem Natural[3] da SBC, tendo como objetivo fornecer um espaço de debate adequado para a integração das comunidades relacionadas à tecnologia da linguagem humana, como Linguística, Ciência da Computação, Psicolingüística, Ciência da Informação, entre outros (PARDO *et al.*, 2010). Tanto o ENIAC quanto o STIL, seus anais são publicados pela biblioteca digital SOL da SBC.

Outro evento brasileiro importante é o *Brazilian Workshop on Social Network Analysis and Mining*[4] (BraSNAM) possui H5-index igual a 7 e estrato B3. É um evento realizado em conjunto com o Congresso da Sociedade Brasileira de Computação focado em tópicos relacionados a análise de redes sociais e oferece um espaço interdisciplinar reunindo profissionais e pesquisadores de redes sociais e seus campos relacionados, para promover a colaboração e troca de ideias e práticas como, ciclo de vida de dados e informações sociais, redes sociais e de informação aspectos humanos, redes sociais e informacionais, técnicas de redes sociais e de informação. Seus anais também são publicados na biblioteca digital SOL.

O *International Conference on Computational Processing of the Portuguese Language* (PROPOR) é o principal evento na área de Processamento Computacional da Língua Portuguesa. Seu objetivo é apresentar os resultados de pesquisas acadêmicas e tecnológicas, além de promover a colaboração entre grupos de pesquisa nessa área, possui H5-index igual a 11 e estrato B2. Também é um evento bianual e ora ocorre no Brasil, ora em Portugal. Os anais do evento são publicados em volumes da série LNAI da *Springer*.

3.1.3 Seleção dos estudos

Com o intuito de identificar os estudos mais relevantes sobre as técnicas de PLN aplicadas na análise de mídias sociais dos eventos listados na Tabela 1, foram estabelecidos critérios de inclusão e exclusão, conforme descritos na Tabela 2. Os critérios de inclusão foram definidos como trabalhos publicados nos anais dos eventos, abordando técnicas, modelos e ferramentas de mineração de texto e análise textual específicas para a análise de mídias sociais.

Posteriormente, foram excluídos artigos científicos fora dos critérios de inclusão, publicações escritas em idiomas diferentes do português ou inglês, trabalhos que não sejam relevantes para o PLN e análise de mídias sociais com base no título, resumo, palavras-chave, introdução e conclusão. Essa seleção seguiu a ordem (i) título, resumo e palavras-chave; (ii) introdução e conclusão e (iii) artigo completo. Artigos que abordavam mapeamento sistemático ou revisão sistemática da literatura também foram excluídos.

[3] `http://www.nilc.icmc.usp.br/cepln/`
[4] `https://csbc.sbc.org.br/`

Tabela 2 – Critérios de Inclusão (CI) e Exclusão (CE) para a seleção de estudos relevantes.

Critério de Inclusão (CI)
CI1: Artigos que abordam técnicas, modelos, ferramentas de mineração de textos, e análise textual em análise de mídias sociais.
Critério de Exclusão (CE)
CE1: Artigos que estão fora dos critérios de inclusão
CE2: Publicações escritas em idiomas diferente do português ou inglês
CE3: Artigos que não sejam relevantes para o PLN e análise de mídias sociais com base no título, resumo, palavras-chave, introdução, e conclusão
CE4: Artigos de mapeamento sistemático ou revisão sistemática da literatura.

Fonte: Elaborado pela autora.

3.2 Condução

Com o protocolo definido, optou-se em utilizar o *Google Drive*[5] e o *Mendeley*[6] como ferramentas de gerenciamento e armazenamento dos documentos para auxiliar no processo de busca e análise dos trabalhos. Ainda mais que, a colaboração entre os pesquisadores foi facilitada pelo *Google Drive*, uma vez que era possível compartilhar os arquivos e trabalhar simultaneamente em documentos e planilhas. Iniciou-se então uma busca preliminar nas plataformas levantadas a fim de identificar os anais dos eventos definidos. Com isso, foram selecionados os anais de eventos mais relevantes na área de pesquisa em suas edições entre os anos de 2020 a 2022, e deles foram obtidos os trabalhos para a pesquisa inicial. Os eventos selecionados foram BRACIS, BraSNAM, ENIAC, PROPOR e STIL, que resultaram inicialmente em 654 artigos. Dois pesquisadores leram os artigos individualmente e avaliaram se os trabalhos atendiam aos critérios de inclusão e exclusão estabelecido na Tabela 2. Após a aplicação rigorosa desses critérios, 468 artigos foram excluídos, a maioria devido aos critérios CE3 e CE4.

A exclusão desses artigos se deve ao fato de que muitos deles tratavam de estudos que envolviam a manipulação de dados multimídia, como imagens, vídeos ou áudio, bem como a construção, descrição ou anotação de um corpus, não se enquadrando no escopo desse mapeamento sistemático, que visa avaliar estudos que abordam mineração de texto ou análise textual. Com base nos critérios de inclusão estabelecidos, 186 artigos foram selecionados para extração de dados. A Tabela 3 apresenta o número de trabalhos incluídos e excluídos em cada evento selecionado.

3.2.1 Extração de dados

Após a seleção dos trabalhos, iniciou-se a etapa de extração de dados relevantes de cada artigo. Essa etapa envolve a organização dos dados em uma planilha, selecionando atributos

[5] `https://drive.google.com/`
[6] `https://www.mendeley.com/`

Tabela 3 – Trabalhos selecionados.

Eventos	**Trabalhos de pesquisa inicial**	**Trabalhos excluídos**	**Trabalhos selecionados**
BRACIS	247	204	43
BraSNAM	67	34	33
ENIAC	206	157	49
PROPOR	83	53	30
STIL	51	20	31
TOTAL	654	468	186

Fonte: Elaborado pela autora.

relevantes, seguida pela análise dos dados, incluindo extração de informações qualitativas. Os resultados dessa análise serão apresentados no Capítulo 4.

3.2.1.1 Seleção dos atributos

Foi criada uma planilha com o intuito de registrar e organizar os dados relevantes de cada estudo. Essa planilha abrange os seguintes atributos: dados do artigo, fonte de dados (e.g., *Twitter*, *Reddit*, e *Facebook*), natureza dos dados (e.g., corpus construído/coletado, já disponível ou Não descrito), Ferramenta e Tecnologia (e.g., NLTK e spaCy), tarefas (e.g., pré-processamento, Classificação de texto e Análise de Sentimentos), ambiente de desenvolvimento (e.g., *Python* e *Jupyter Notebook*), técnicas (e.g., TF-IDF e BERT), métricas (e.g., F1-Score e Similaridade de Cosseno). Esses atributos mapeados e sua relação com as QPs estão descritos na Tabela 4.

A leitura e busca dos trabalhos relevantes ocorreram em duas etapas. Inicialmente, os títulos e resumos foram lidos para realizar uma triagem inicial. Em seguida, os trabalhos selecionados passaram por indexação dos atributos mencionados, com foco na seção de materiais e métodos. Caso as informações necessárias não estivessem disponíveis nessa seção a leitura era estendida para o trabalho completo.

3.2.1.2 Análise dos dados

Uma abordagem indutiva foi adotada para extrair e analisar as informações dos dados qualitativos coletados. Para isso, foi realizada uma análise exploratória utilizando a linguagem de programação *Python* 3.9.13 com auxílio do ambiente de programação interativa *Jupyter Notebook*. Inicialmente, a planilha criada no *Google Sheets* estava no formato de arquivo .xlsx, baseado em *Office Open XML*. No entanto, para facilitar o processo de análise dos dados, foi necessária a conversão dessa planilha para o formato *Comma-separated Values* (CSV), que é amplamente utilizado para representar dados tabulares. Essa conversão foi feita utilizando a biblioteca *Pandas*[7], usada em estudos como o de Lequertier *et al.* (2021) para analisar os dados

[7] `https://pypi.org/project/pandas/`

Tabela 4 – Mapeamento dos dados extraídos e a questão de pesquisa a que estão relacionados.

Dados	**Descrição**	**PP Relevante**
Fonte do artigo	Autor, Título, Evento, Ano, DOI	Visão geral
Fonte de dados	Fonte dos dados usada para análise no estudo	PP2
Natureza dos dados	Corpus construído ou coletado, corpus já disponível ou não descrito	PP2
Ferramenta e Tecnologia	Ferramenta ou Tecnologia usada para manipular e analisar os dados no estudo	PP1
Pré-processamento	As etapas de pré-processamento de dados realizadas no estudo	PP1
Mineração de texto/ Tarefas PLN	Tarefas PLN relacionadas com as ferramentas e técnicas que foram executadas no estudo	PP1
Técnicas usadas	Técnicas de PLN aplicadas no estudo	PP1
Métricas	medidas de avaliação usadas no estudo	PP3
Repositório	Link dos repositórios disponíveis para dados e códigos	Visão geral

Fonte: Elaborado pela autora.

extraídos de artigos.

Em seguida, essa base de dados com os atributos mapeados foi pré-processada para remover espaços e acentos desnecessários e converter as strings para minúsculas, a fim de padronizar todos os atributos. Para realizar essas operações bibliotecas como *unicodedata*[8] e o conjunto de Expressões Regulares - RegEx[9] (do inglês, *Regular Expressions*) também foram usados.

Por conseguinte, os dados qualitativos foram extraídos por meio do desenvolvimento de um dicionário utilizando a classe *Counter* da biblioteca *collections*[10]. Essa classe é especialmente útil para contar a ocorrência de elementos em uma lista, o que nos proporcionou uma maneira conveniente de contar palavras. A fim de garantir maior confiabilidade, realizamos revisões tanto no código, durante o processo de extração e manipulação dos dados com as bibliotecas mencionadas, quanto na planilha inicialmente criada a partir dos dados extraídos. Essa abordagem de dupla revisão, no código e na planilha, foi adotada para minimizar possíveis erros e assegurar a precisão dos resultados obtidos.

É importante ressaltar que todo o material produzido durante a realização deste mapeamento sistemático será deixado público em um repositório no GitHub[11].

[8] `https://docs.python.org/3/library/unicodedata.html`
[9] `https://docs.python.org/pt-br/3/library/re.html`
[10] `https://docs.python.org/3/library/collections.html`
[11] `https://github.com/GabrieleAraujo/mapeamento_sistematico_PLN`

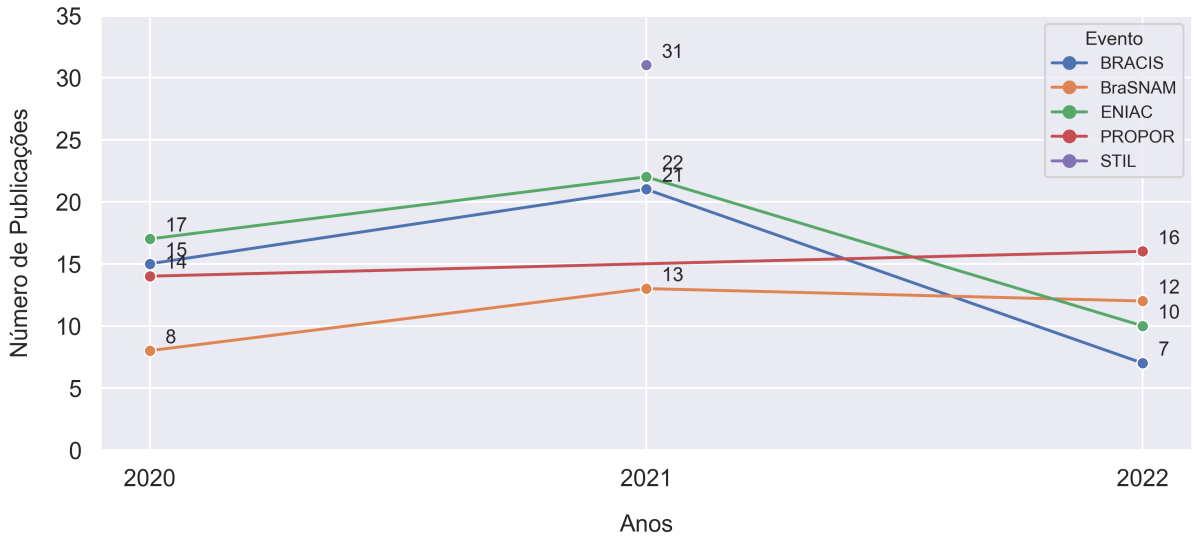
4 RESULTADOS

Este capítulo contempla os resultados do estudo, bem como as respostas às perguntas de pesquisa. Com isso, é apresentada a análise exploratória dos trabalhos selecionados, abordando as ferramentas e técnicas de PLN mais frequentes na análise de mídias sociais, seguida pelas fontes e natureza dos dados usados nessas análises. E por fim, apresentamos as medidas de avaliação identificadas nos estudos.

4.1 Análise geral

O mapeamento relatado neste trabalho foi realizado com o objetivo de fornecer uma visão geral das pesquisas desenvolvidas pela comunidade de mineração de textos e que estão relacionadas à análise de mídias sociais. Esse mapeamento é baseado em 186 estudos selecionados, correspondente a 30% dos 654 trabalhos presentes nos anais publicados entre 2020 e 2022, conforme descrito anteriormente na Seção 3.2. A distribuição temporal desses estudos por ano de publicação é apresentada na Figura 2.

Figura 2 – Distribuição anual das publicações selecionadas por evento.



Fonte: Elaborado pela autora.

A figura apresentada demonstra um aumento significativo no número de publicações no ano de 2021. Essa tendência pode ser atribuída à quantidade crescente de conteúdo gerado nas redes sociais devido à pandemia de COVID-19 (PACHUCKI; GROHS; SCHOLL-GRISSEMANN, 2022; ROSEN *et al.*, 2022). É importante destacar que alguns eventos, como o STIL e o PROPOR, possuem uma periodicidade bienal, ou seja, ocorrem a cada dois anos.

Ainda, durante a condução desta análise, percebeu-se a difusão de trabalhos em outras áreas de IA e IC nos anais dos eventos, como BRACIS e ENIAC no ano de 2022. Essa tendência de expansão e exploração de novos campos de pesquisa pode ter impactado a proporção de estudos específicos de análise textual dentro do escopo deste mapeamento.

4.2 Análise de ferramentas e técnicas

Nesta seção, realizamos uma análise exploratória visando responder às perguntas de pesquisa. Para isso, conduzimos uma análise quantitativa considerando a contagem de cada atributo único extraído em cada estudo selecionado, o mesmo se aplica para as seções seguintes. Os Apêndices B e C contém os detalhes de cada estudo primário selecionado.

PP1: Quais ferramentas e técnicas de PLN são mais utilizadas nos eventos científicos selecionados?

Dos 186 estudos analisados, identificamos um total de 135 ferramentas designadas a tarefas de PLN, o correspondente a mais de 81% dos trabalhos. A Figura 3 apresenta as 20 ferramentas mais frequentes. Dentre elas, a mais utilizada é o *Scikit-Learn*[1], aplicada em 55 dos estudos, correspondendo a 37%. No qual consiste em uma biblioteca de ML em *Python* que se destaca pela sua facilidade de uso e aplicabilidade em tarefas de PLN, como classificação, regressão e clusterização. Na maioria dos estudos foi usada para implementação de classificadores.

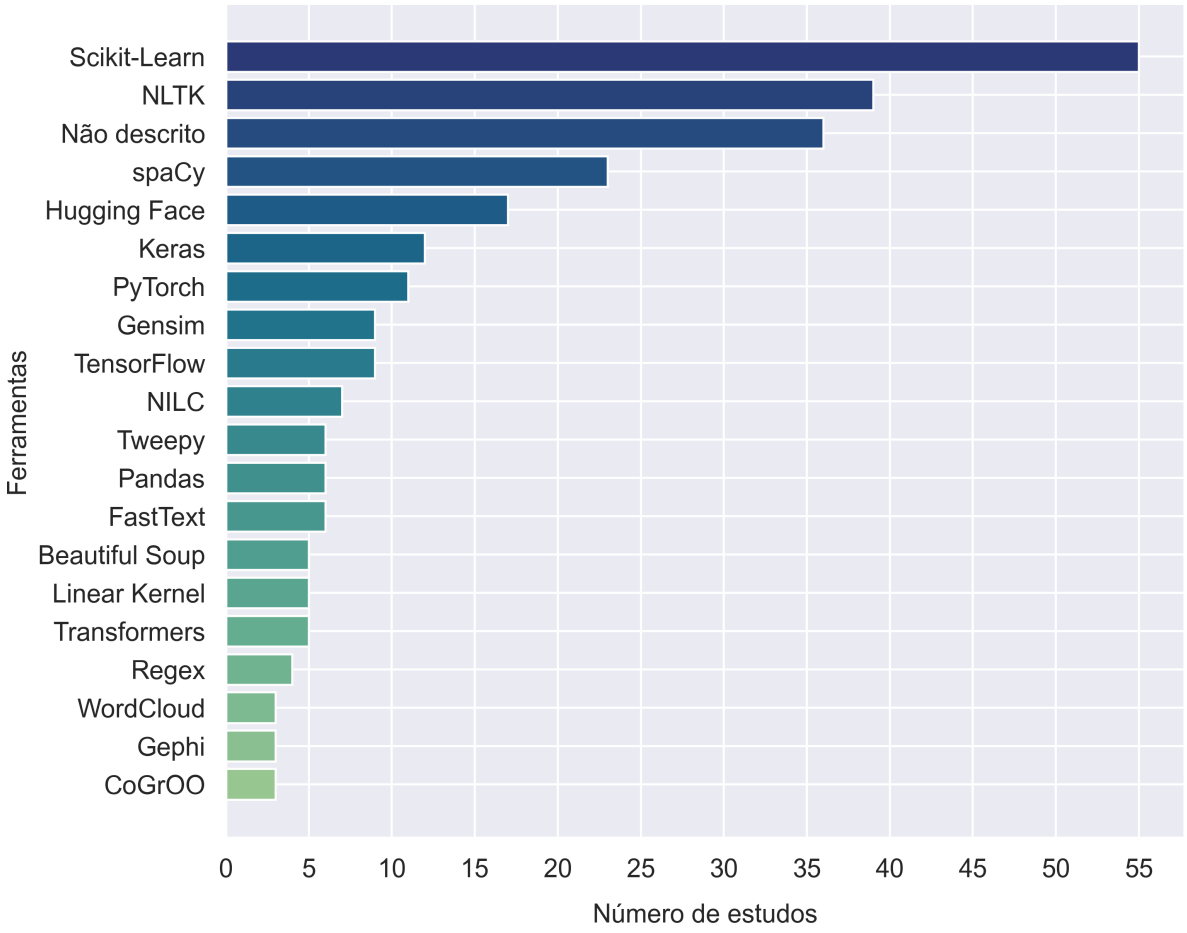
Para tarefas de pré-processamento de texto, o NLTK[2] foi uma ferramenta bastante utilizada, aplicada em 39 estudos (26%). Sendo esta uma plataforma de código aberto para PLN que oferece suporte a diversas tarefas, como tokenização, lematização, remoção de *stopwords*, POS tagging, e dentre outras. Seguida pelo spaCy[3] presente em 23 publicações (15%), por sua vez, é uma biblioteca *Python* que fornece recursos para tarefas como POS tagging, NER, análise sintática, classificação de texto, lematização e também remoção de *stopwords*.

A Figura 3 revela também um grande número de ferramentas subutilizadas, como *TensorFlow* e *Pytorch* que são *frameworks* de ML e DL, aplicáveis na criação, treinamento e implantação de modelos de linguagem. Essas ferramentas desempenham um papel crucial no tratamento e análise dos dados e estão alinhadas com as técnicas que serão descritas posteriormente. Ademais, 36 estudos (19%) não descreveram claramente as ferramentas utilizadas.

Para uma análise mais aprofundada das ferramentas, a Figura 4 proporciona uma compreensão abrangente da distribuição temporal das ferramentas predominantes nos trabalhos analisados. É evidente a prolífica utilização de diferentes ferramentas no ano de 2021, possivelmente influenciada pelo maior número de trabalhos selecionados nesse período. É notável a

[1] https://scikit-learn.org/
[2] `https://www.nltk.org/`
[3] `https://spacy.io/`

Figura 3 – As 20 ferramentas mais frequentes na análise de textos.



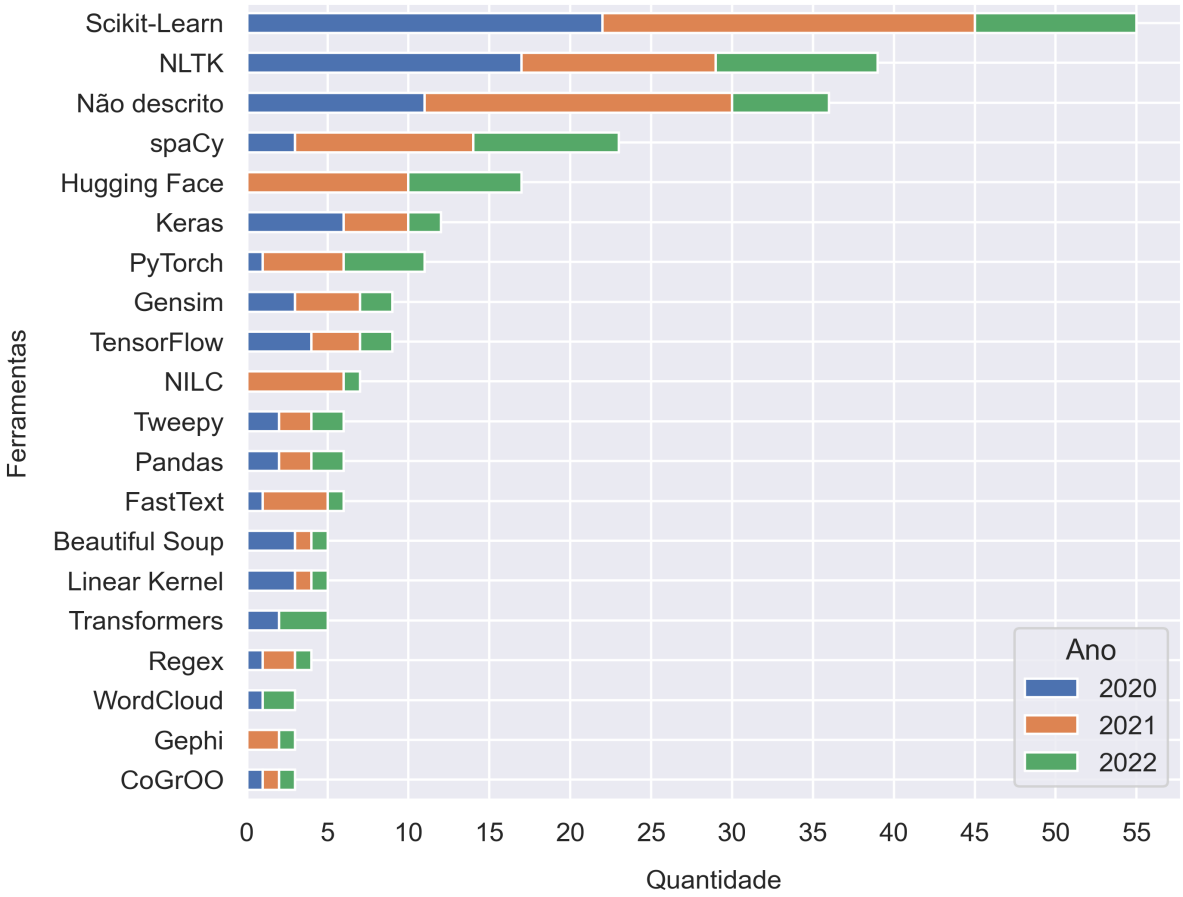
Fonte: Elaborado pela autora.

presença constante de ferramentas como Scikit-Learn e NLTK ao longo dos anos, ressaltando sua relevância contínua nas análises.

No que diz respeito às técnicas, identificou-se um total de 275 técnicas, onde apenas 3 dos estudos não descreveram claramente a técnica usada. A Figura 5 mostra que a técnica mais frequente é o BERT usada por 63 estudos (34%), em que consiste em um modelo de linguagem baseado em *Transformers* que se destacou por seu desempenho em tarefas de PLN, principalmente para pré-treinar representações de textos não rotulados (DEVLIN *et al.*, 2018). Seu uso está associado com outros modelos variantes dessa abordagem, como o caso de modelos pré-treinados para outros idiomas, como o BERTimbau presente em 32 trabalhos (17%) uma versão para o português, Multilingual BERT (M-BERT) aplicado em 32 trabalhos (17%), um modelo pré-treinado em 102 idiomas, e o modelo BERTopic usado em 6 estudos (3%) para tarefas de modelagem de tópicos.

Seguida por outras técnicas de representação textuais, o *Term Frequency-Inverse Document Frequency* (TF-IDF) presente em 58 dos estudos (37%), sendo amplamente aplicado na representação de documentos e cálculo de importância de palavras em um *corpus*. Outra muito

Figura 4 – Distribuição das 20 ferramentas mais frequentes por ano.



Fonte: Elaborado pela autora.

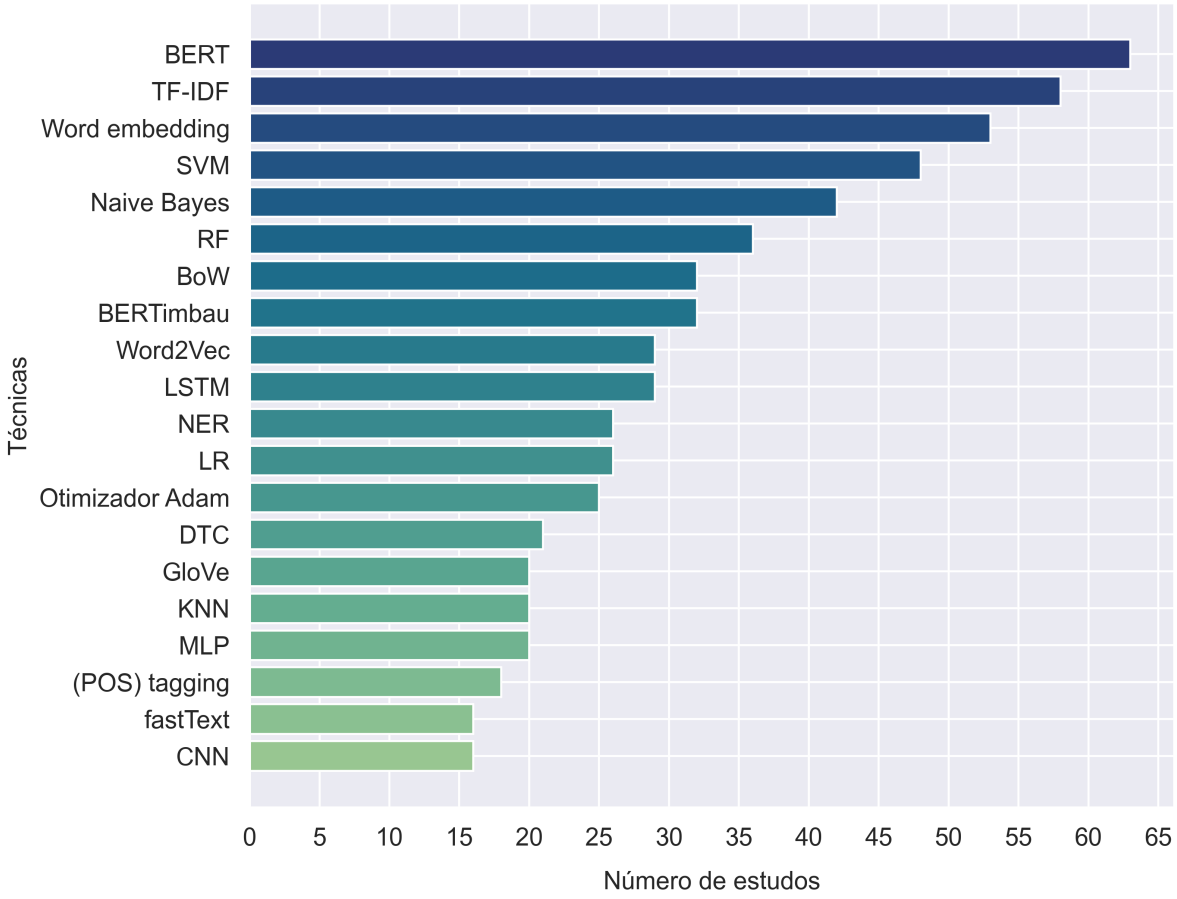
presente é o *Bag-of-words* (BoW), em 32 dos estudos (17%).

Em relação às aplicações dessas técnicas, 105 dos estudos relataram divisões de dados em tarefas de treinamento e teste. Com base nisso, muitos modelos de DL e ML foram aplicados, de tal modo que, há a presença de várias outras técnicas de DL, como *Word embedding* aplicada em 53 trabalhos (29%), Word2Vec em 29 trabalhos (16%), LSTM usado também em 29 trabalhos (16%), *Multi-layer Perceptron* (MLP) em 20 estudos (11%), GloVe mencionado em 16 trabalhos (9%), fastText aplicado também em 16 estudos (9%), CNN também em 16 (9%) e *Bidirectional LSTM with Conditional Random Fields* (BiLSTM-CRF) em 16 trabalhos (9%), aplicadas na captura informações semânticas e a modelagem do contexto linguístico.

Também são encontradas técnicas de ML que são amplamente usadas para a classificação e treinamento de dados textuais rotulados, como o SVM em 48 estudos (26%), *Naïve Bayes* usado em 42 estudos (23%), *Random Forest* (RF) presente em 36 dos estudos (20%), LR em 26 trabalhos (14%) e *Decision Tree Classifier* (DTC) em 21 trabalho (12%). Outros modelos de linguagem e representação de textos também estão presentes.

Observa-se que existe uma ampla variedade de ferramentas disponíveis para a aplicação

Figura 5 – As 20 técnicas mais frequentes na análise de textos.



Fonte: Elaborado pela autora.

de técnicas de análise textual, que variam de acordo com as tarefas de PLN que se deseja executar. A relação entre as 5 principais técnicas identificadas e as ferramentas correspondentes é descrita na Tabela 5. Pela grande quantidade de estudos mapeados, outras relações entre essas técnicas e seus específicos estudos podem ser encontradas no Apêndice B.

Outro aspecto importante é a evolução do uso dessas técnicas ao longo dos anos. Como apresentado na Figura 6, é notável que o modelo de linguagem BERT ganhou maior proeminência nos anos de 2021 e 2022, juntamente com a técnica de *Word embedding* e o BERTimbau, principalmente em 2021. Enquanto isso, as abordagens de representação de texto continuam a manter uma presença marcante ao longo dos anos, como evidenciado pelo contínuo uso do TF-IDF e BoW.

Além disso, foi realizada uma comparação das dez ferramentas mais utilizadas entre os eventos, conforme a Figura 7. Podemos identificar as preferências dos pesquisadores em relação ao uso de técnicas de PLN na análise de texto por meio desse comparativo.

Portanto, é visto que o BRACIS é o evento com maior prevalência de estudos que utilizam técnicas de DL, principalmente para a classificação de textos, com os modelos de linguagem

Tabela 5 – As 5 técnicas mais presentes e sua relação com as ferramentas a partir das tarefas associadas.

ID do artigo	Técnica	Ferramenta	Tarefas
[3, 15, 16, 29, 36, 37, 41, 42, 53, 92, 94, 95, 109, 112, 118, 161, 172, 180]	BERT	Pytorch Transformers	inferência de linguagem classificação de texto modelagem de tópicos análise de sentimentos
[7, 29, 49, 74, 160]	TF-IDF	*Scikit-Learn*	clusterização extração de características representação textual
[19, 44 53]	*Word embedding*	FastText	NER, similaridade textual extração de aspectos extração de características
[75, 114, 131, 138]	SVM	*Scikit-Learn*	classificação de texto classificação de sentimentos
[29, 77, 114, 182]	*Naïve Bayes*	*Scikit-Learn* NLTK	classificação de texto classificação de sentimentos

Fonte: Elaborado pela autora.

como o BERT, *Word embedding* e o LSTM. Esses resultados sugerem um interesse significativo na utilização de abordagens baseadas em modelos de linguagens pré-treinados, representações distribuídas de palavras e RNNs em análise de mídias sociais.

Em contraste, o ENIAC aborda mais trabalhos que usam modelos de classificação de textos, como o SVM, RF, *Naïve Bayes* e para representação de textos, como o TF-IDF e o BoW. Esses resultados evidenciam a preferência dos pesquisadores por técnicas tradicionais de ML nas análises publicadas no ENIAC.

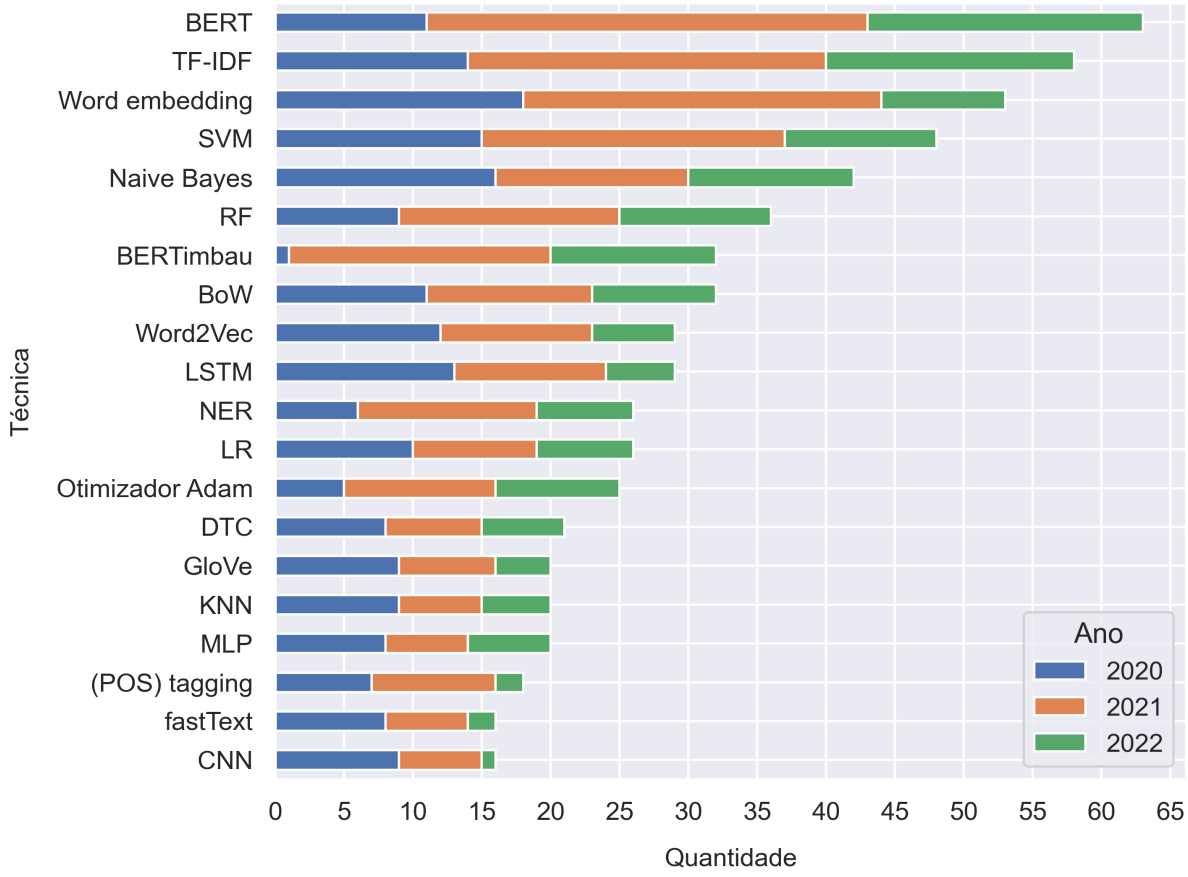
Para lidar com dados na língua portuguesa, o modelo pré-treinado BERTimbau tem sido amplamente empregado em eventos como STIL, BRACIS, ENIAC e PROPOR, respectivamente.

4.3 Fontes de dados de mídia social

PP2: Quais são as fontes e a natureza dos dados usados na análise de mídias sociais?

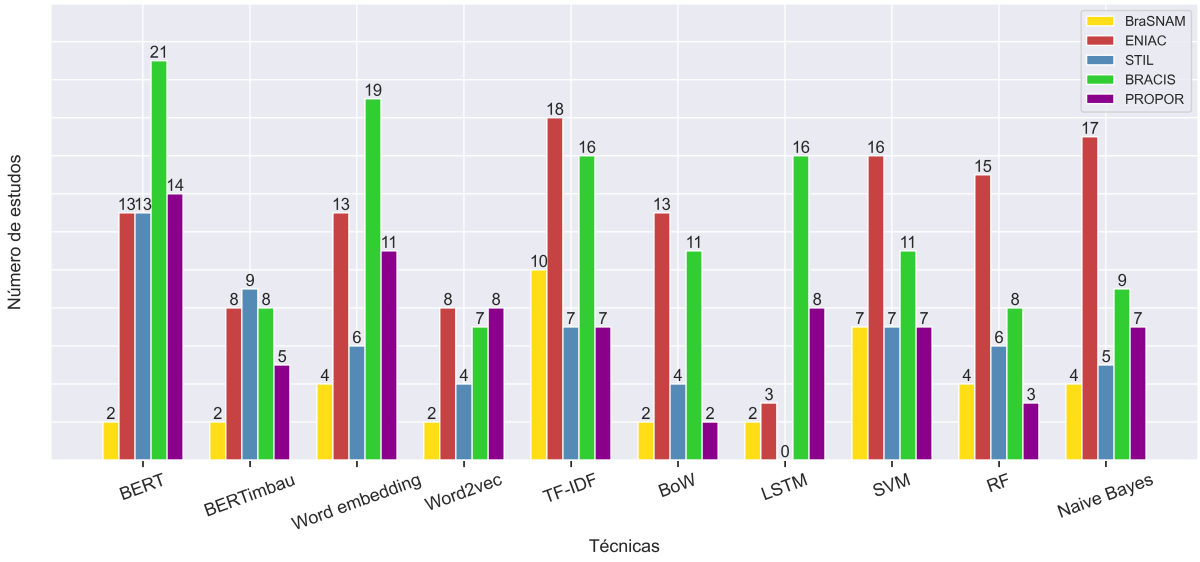
Durante a análise dos trabalhos, foram identificadas as fontes de dados mais frequentes nos estudos. Ao analisar a nuvem de palavras na Figura 8, é possível observar que o *Twitter* é a fonte de dados mais frequente, indicando que essa plataforma é amplamente explorada pelos

Figura 6 – Distribuição das 20 técnicas mais frequentes por ano.



Fonte: Elaborado pela autora.

Figura 7 – Comparação das 10 técnicas mais frequentes na análise de textos por evento.



Fonte: Elaborado pela autora.

pesquisadores. Além do *Twitter*, outra diversidade de fontes estão presentes, como o *Google News*[4], vale ressaltar que essa é a fonte de origem de um conjunto de dados amplamente utilizado nos estudos, a Avaliação de Similaridade Semântica e Inferência Textual[5] (ASSIN) que consiste em um conjunto de dados construído com 10,000 pares de sentenças de textos jornalísticos reais. Estão presentes também o *Wikipedia*, UOL[6], *TripAdvisor*[7], G1[8], Folha de São Paulo[9], *Amazon*, plataformas do Sistema Judiciário e plataformas de *e-commerce* (comércio eletrônico).

Figura 8 – Nuvem de palavras das fontes de dados em estudos.

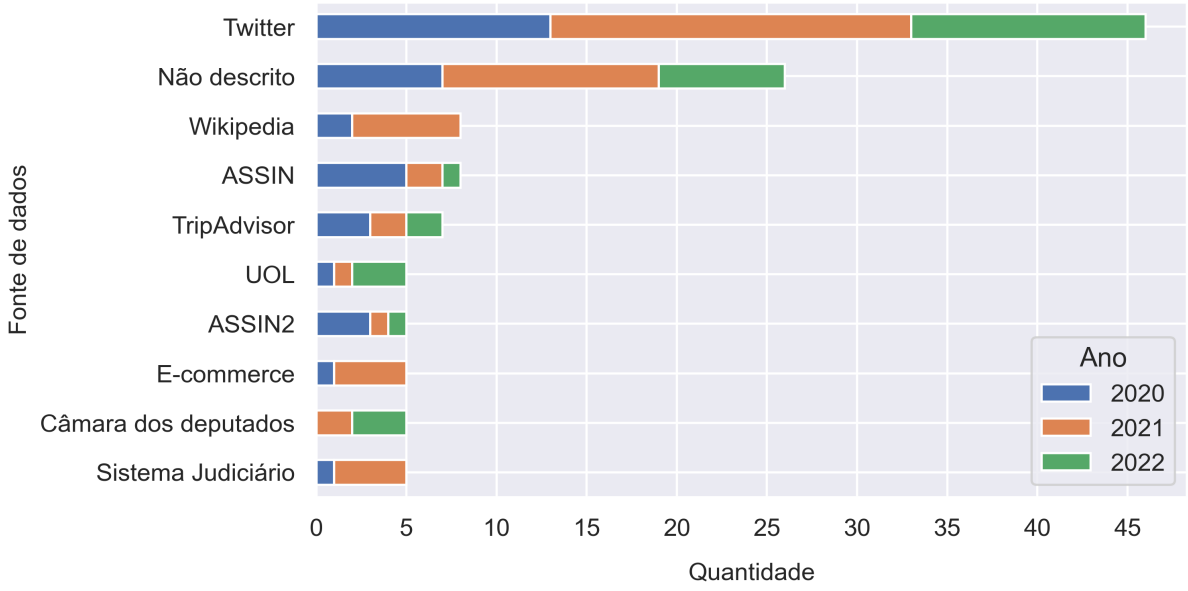
Fonte: Elaborado pela autora.

Na Figura 9 é possível visualizar melhor o uso das dez fontes de dados mais presentes nos trabalhos ao decorrer dos anos. É visto que o *Twitter* continua sendo a mídia social mais utilizada para análise de dados, sendo mais empregada em 2021. Ainda, as fontes de dados *TripAdvisor*, UOL e Câmara dos deputados também tiveram um uso significativo em 2022.

Adicionalmente, foi investigada a forma de como os dados foram obtidos. Dos estudos analisados, 110 trabalhos utilizaram *corpus* construído ou coletado, o que envolve a criação ou coleta de dados específicos para a pesquisa. Por outro lado, 75 trabalhos utilizaram *corpus* já disponíveis, ou seja, conjuntos de dados prontamente disponíveis para análise. Apenas um trabalho não descreveu claramente a natureza do *corpus* utilizado.

[4] https://news.google.com/
[5] http://nilc.icmc.usp.br/assin/
[6] https://www.uol.com.br/
[7] https://www.tripadvisor.com.br/
[8] https://g1.globo.com/
[9] https://www.folha.uol.com.br/

Figura 9 – Progressão de ano por fonte de dados.



Fonte: Elaborado pela autora.

4.4 Medidas de avaliação

Nessa seção, exploramos as métricas usadas para avaliação e validação dos dados nos estudos.

PP3: Quais são as medidas de avaliação mais utilizadas nos estudos de PLN?

Evidenciou-se uma grande variedade na metodologia de avaliação empregada em relação aos conjuntos de dados de teste ou validação. Diversas métricas foram usadas para avaliar o desempenho de modelos de classificação de dados em PLN. Entre as 101 métricas identificadas, a mais usada foi o F1-Score, presente em 108 estudos (58%), essa métrica representa a média harmônica entre Revocação e Precisão, fornecendo uma medida balanceada do desempenho do modelo. A Revocação (do inglês, *Recall*) esteve presente em 81 estudos (43%) e a Precisão (do inglês, *Precision*) em 78 estudos (42%), seguida da Acurácia em 55 estudos (30%).

Além dessas métricas, os pesquisadores também utilizam diferentes abordagens na geração de divisões de conjuntos de dados para teste e validação, incluindo método *Cross-Validation* citado em 59 estudos (32%). Denominada também de K-fold, envolve a divisão aleatória da base de dados em K subconjuntos de tamanho similar, onde K é previamente definido (BERRAR, 2019).

5 DISCUSSÕES

O interesse das pesquisas na área de análise de textos é proporcional em todos os eventos selecionados. Com ênfase no BraSNAM, PROPOR e STIL com menor número de trabalhos excluídos, mesmo apresentando um número menor de trabalhos na pesquisa inicial se comparado com o BRACIS e o ENIAC. Como também, constatou-se um crescimento no número de publicações do BraSNAM e do PROPOR nos últimos anos. Esses achados reforçam a afirmação de que esses eventos são os principais meios de divulgação de pesquisas relacionadas ao tema abordado neste estudo, corroborando com a afirmação feita por Souza *et al.* (2018).

Considerando a PP1 e a análise das ferramentas utilizadas nos estudos, destaca-se o *Scikit-Learn* como uma opção proeminente entre os pesquisadores. Essa ferramenta foi desenvolvida especificamente para a aplicação prática de técnicas de ML e pode ser utilizada em várias etapas do pipeline de PLN. Sendo amplamente empregada em tarefas de classificação de texto, em estudos como [13, 29, 41, 77, 131, 180, 182], extração de características por meio da modelagem e análise de tópicos nas publicações [65, 110, 139], e em identificação de entidades [77, 159, 167]. Assim, o *Scikit-Learn* se sobressai por sua versatilidade e pela variedade de recursos disponíveis, permitindo aos pesquisadores explorar diferentes abordagens e modelos para suas análises de texto. Sua popularidade entre os pesquisadores indica sua eficácia e utilidade na área. Como relatado em [72], a ferramenta foi usada para definir os parâmetros padrão dos modelos de linguagem e para realizar tarefas de treinamento e teste de dados. Essa aplicação específica demonstra a flexibilidade da ferramenta ao permitir a configuração dos parâmetros de acordo com as necessidades específicas de cada estudo.

Ao interpretar os resultados, observa-se que as técnicas mais mencionadas nos estudos refletem às tendências atuais no campo do PLN, conforme declarado por Khurana *et al.* (2023). Em particular, o destaque do modelo pré-treinado BERT indica o reconhecimento de sua eficácia na captura de relações semânticas e sua aplicabilidade em diversas tarefas, como na classificação de texto [5, 30, 95, 118, 164, 174], em análise de sentimentos [29, 36, 175] e identificação de entidades por meio da classificação de *tokens* [101, 114, 133, 174]. A popularidade do BERT pode ser atribuída à sua arquitetura baseada em *Transformers* e ao seu treinamento em grandes volumes de dados. Esses fatores permitem que o modelo aprenda representações contextuais das palavras, levando em consideração o contexto em que elas aparecem, diferente dos modelos clássicos de ML. Isso é especialmente importante para o PLN, pois as palavras podem ter diferentes significados dependendo do contexto em que são usadas (DEVLIN *et al.*, 2018).

No estudo feito em [180], o BERT foi utilizado em conjunto com outros modelos de representação textual variante, como o BERTimbau para rotular notícias, e é relatado que modelos de baseados em redes neurais são as melhores alternativas para cenários de avaliação de supervi-

são fraca. Em decorrência, outros estudos voltados para a análise de texto, especialmente para o português, apontam o BERTimbau como um modelo eficaz em tarefas de NER, similaridade textual de sentença e análise de sentimentos [92, 95, 174, 176].

Com base nas respostas da PP2, observamos que as fontes de dados representam diferentes contextos e conteúdos textuais, abrangendo desde opiniões de usuários, notícias, informações de produtos em comércio eletrônico, e até dados jurídicos, condizendo com Choi *et al.* (2020), Zachlod *et al.* (2022) e Júnior *et al.* (2020).

Desse modo, é visto que a análise de dados de mídia social é amplamente aplicada em várias áreas temáticas e disciplinas devido à diversidade de usuários que compartilham opiniões e experiências. O *Twitter*, em particular, recebe muita atenção dos estudiosos devido à sua natureza instantânea e à quantidade de dados textuais disponíveis. Uma vez que, os resultados obtidos corroboram com os achados de Souza *et al.* (2018), em que também foram identificados estudos que se baseiam em dados de notícias online, principalmente para serem implementados em tarefas de detecção de *fake news*, com a criação de novos conjuntos de dados voltados para o português do Brasil [20, 38, 64, 70, 88, 126]. Uma das razões da inclusão dessas fontes de dados nos estudos reflete na busca por soluções mais eficazes no combate às *fake news* e contribui para o desenvolvimento de recursos e algoritmos específicos.

Em relação a PP3, no que diz respeito aos métodos de avaliação em PLN, o destaque às métricas F1-Score, Revocação e Precisão revela uma abordagem comum para validar os modelos de classificação de textos, uma vez que, muitos estudos empregam medidas de avaliação para comparar o desempenho de diferentes modelos de DL e ML como nas publicações [41, 94, 165]. Um exemplo mencionado no artigo [19] evidencia a utilização da métrica de Revocação, que representa a porcentagem de entidades nomeadas anotadas que o modelo é capaz de detectar. Essa métrica avalia a capacidade do modelo em identificar corretamente as entidades relevantes presentes nos textos analisados. Quanto maior a Revocação, maior a capacidade do modelo em abranger um maior número de entidades nomeadas. Outra métrica relevante é a precisão, usada para avaliar a qualidade das anotações realizadas pelo modelo em relação a um conjunto de referência confiável. Quanto maior a precisão, maior a concordância entre as anotações realizadas e as anotações de referência. Assim, essas métricas permitem a avaliação do desempenho e a mensuração da qualidade das anotações realizadas pelos modelos.

Portanto, os resultados obtidos neste estudo têm o propósito de estimular os pesquisadores que têm interesse no campo da análise textual por meio do PLN, com ênfase na análise de mídias sociais. Através do mapeamento desses recursos e informações relevantes, espera-se incentivar a realização de pesquisas científicas nesse contexto específico. Ao disponibilizar esses dados e conhecimentos, esperamos que os pesquisadores possam aproveitá-los em suas aplicações e contribuir para o avanço da área. Dessa forma, o objetivo é impulsionar o desenvolvimento de novas abordagens e soluções, promovendo um maior entendimento e *insights* valiosos novas aplicações em análise de mídias sociais.

5.1 Desafios

A quantidade massiva de mensagens, comentários, *tweets* e publicações gerados pelos usuários das mídias sociais representa um desafio significativo para a análise e obtenção de *insights*. Lidar com essa enorme quantidade de dados requer um processamento eficiente para extrair informações relevantes.

Em virtude da natureza informal e das variações linguísticas utilizadas nas mídias sociais, incluindo gírias e abreviações, as tarefas se tornam ainda mais complexas. Isso se agrava ainda mais pela disseminação de notícias falsas em várias plataformas, dificultando a distinção entre informações precisas e enganosas devido à velocidade e escala das informações compartilhadas. Para realizar um processamento eficiente desses dados e identificar e classificar as informações relevantes, são necessárias técnicas avançadas de PLN.

O treinamento de modelos de linguagem capazes de lidar com tais tarefas, demanda tempo e recursos financeiros consideráveis. Coletar e rotular grandes conjuntos de dados, além de realizar ajustes e refinamentos constantes, são etapas necessárias para melhorar a precisão e o desempenho desses modelos. A disponibilidade de recursos computacionais adequados, investimentos em pesquisa e desenvolvimento são essenciais para enfrentar esse desafio e garantir uma análise eficiente e precisa dos dados das mídias sociais.

Mesmo com o surgimento de novos modelos de linguagem para idiomas específicos, como o BERTimbau, ainda existem desafios a serem superados no campo de PLN. Uma das limitações apontadas por [95] é que além da otimização dos hiperparâmetros, outros aspectos, como o desenvolvimento de conjuntos de dados anotados com uma ampla variedade de emoções e o aprimoramento da compreensão contextual, são áreas que requerem atenção para avançar em tarefa de classificação de emoções em nível de sentença em português. Essa demanda por dados anotados de alta qualidade é crucial para aprimorar os resultados e a eficácia desses modelos.

6 CONSIDERAÇÕES FINAIS

Neste artigo, realizamos um estudo de mapeamento sistemático para analisar as técnicas de PLN utilizadas na análise de mídias sociais, a fim de investigar e compreender as tendências atuais nesse campo de pesquisa. Identificamos os principais eventos científicos da área (BRACIS, BRaSNAM, ENIAC, STIL e PROPOR) e selecionamos 186 trabalhos relevantes publicados entre os anos de 2020 e 2022, dentre os 654 artigos.

Nossa pesquisa foi norteada por três Questões de Pesquisa: PP1: Quais ferramentas e técnicas de PLN são mais utilizadas nos eventos científicos selecionados? Como resultado, as ferramentas mais mencionadas nos trabalhos, foram *Scikit-Learn*, NLTK e spaCy, indicando a predominância de tarefas de pré-processamento e classificação de textos. Também identificamos 275 técnicas, dentre as quais o BERT foi a mais citada. Essas técnicas de PLN incluem tarefas de análise de sentimentos, modelagem de tópicos, bem como a classificação de textos por meio de treinamento e teste de dados, com abordagens baseadas em modelos de ML e DL. Para a segunda, PP2: Quais as fontes e natureza dos dados usados em análise de mídias sociais? Quanto às plataformas, as mais exploradas são o *Twitter*, *Wikipedia*, *TripAdvisor* e portais de notícias. Além disso, identificamos que a maioria dos estudos realizam a coleta e construção de um corpus específico para as suas análises. E por fim, a PP3: Quais são as medidas de avaliação e os desafios envolvidos em PLN? Onde foi identificado que muitos pesquisadores fizeram o uso de métricas como F1-Score, seguida por Revocação e Precisão, além do método estatístico *Cross-Validation*, para avaliar os seus modelos.

No entanto, este estudo também revelou algumas ameaças à validade do estudo. Uma delas é a limitação dos eventos selecionados, que podem representar apenas parcialmente algumas áreas de pesquisa em análise de mídias sociais. Outra ameaça à validade é a possibilidade de viés de publicação, uma vez que os artigos selecionados são aqueles disponíveis em anais de eventos específicos, uma sugestão seria realizar uma coleta automática utilizando palavras-chave bem definidas de pesquisa em outras bases de dados.

Ainda assim, a principal contribuição deste trabalho é fornecer uma visão geral da aplicação de diversas ferramentas e técnicas de PLN na análise de textos extraídos de mídias sociais. É importante destacar que estamos seguindo os Princípios da Ciência Aberta (do inglês, *Open Science*) ao disponibilizar todos os nossos dados em um repositório disponível publicamente, permitindo a reprodutibilidade dos resultados. Este trabalho pode ser útil para pesquisadores e profissionais interessados em explorar o potencial dessas ferramentas e técnicas, ter uma visão clara das lacunas, desafios e oportunidades de pesquisa nessa área e analisar o cenário atual em pesquisas envolvendo PLN e análise de mídias sociais. Para trabalhos futuros, planejamos expandir nossas fontes de dados para conferências internacionais relevantes (e.g., *AAAI Conference on*

Artificial Intelligence, *International AAAI Conference on Web and Social Media*, *etc*), com o objetivo de comparar o *status* atual dos programas de pesquisa brasileiros com os internacionais.

REFERÊNCIAS BIBLIOGRÁFICAS

HOU, Q.; HAN, M.; CAI, Z. Survey on data analysis in social media: A practical application aspect. **Big Data Mining and Analytics**, TUP, v. 3, n. 4, p. 259–279, 2020. Citado na página 17.

KAPLAN, A. M.; HAENLEIN, M. Users of the world, unite! the challenges and opportunities of social media. **Business horizons**, Elsevier, v. 53, n. 1, p. 59–68, 2010. Citado na página 17.

APPEL, G. *et al.* The future of social media in marketing. **Journal of the Academy of Marketing science**, Springer, v. 48, n. 1, p. 79–95, 2020. Citado na página 17.

PAN, Y.; TORRES, I. M.; ZÚÑIGA, M. A. Social media communications and marketing strategy: A taxonomical review of potential explanatory approaches. **Journal of Internet Commerce**, Taylor & Francis, v. 18, n. 1, p. 73–90, 2019. Citado na página 17.

AICHNER, T. *et al.* Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. **Cyberpsychology, behavior, and social networking**, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 24, n. 4, p. 215–222, 2021. Citado na página 17.

BALAJI, T.; ANNAVARAPU, C. S. R.; BABLANI, A. Machine learning algorithms for social media analysis: A survey. **Computer Science Review**, Elsevier, v. 40, p. 100395, 2021. Citado 2 vezes nas páginas 17 e 19.

CHOI, J. *et al.* Social media analytics and business intelligence research: A systematic review. **Information Processing & Management**, Elsevier, v. 57, n. 6, p. 102279, 2020. Citado 3 vezes nas páginas 17, 23 e 42.

ZENG, D. *et al.* Social media analytics and intelligence. **IEEE Intelligent Systems**, IEEE, v. 25, n. 6, p. 13–16, 2010. Citado na página 17.

LIN, M. S. *et al.* Destination image through social media analytics and survey method. **International Journal of Contemporary Hospitality Management**, Emerald Publishing Limited, v. 33, n. 6, p. 2219–2238, 2021. Citado na página 18.

ROMA, P.; ALOINI, D. How does brand-related user-generated content differ across social media? evidence reloaded. **Journal of Business Research**, Elsevier, v. 96, p. 322–339, 2019. Citado na página 18.

ZHUANG, W. *et al.* What makes user-generated content more helpful on social media platforms? insights from creator interactivity perspective. **Information Processing & Management**, Elsevier, v. 60, n. 2, p. 103201, 2023. Citado na página 18.

NAEEM, M.; OZUEM, W. Understanding the different types of ugc participants and social context for fashion brands: insights from social media platforms. **Qualitative Market Research: An International Journal**, Emerald Publishing Limited, 2022. Citado na página 18.

SADIKU, M. *et al.* Artificial intelligence in social media. **International Journal of Scientific Advances**, v. 2, n. 1, p. 15–20, 2021. Citado na página 18.

LU, Y. Artificial intelligence: a survey on evolution, models, applications and future trends. **Journal of Management Analytics**, Taylor & Francis, v. 6, n. 1, p. 1–29, 2019. Citado na página 18.

DUAN, L. *et al.* Cluster-based outlier detection. **Annals of Operations Research**, Springer, v. 168, p. 151–168, 2009. Citado na página 18.

MA, L.; SUN, B. Machine learning and ai in marketing–connecting computing power to human insights. **International Journal of Research in Marketing**, Elsevier, v. 37, n. 3, p. 481–504, 2020. Citado na página 18.

SHARMA, N.; SHARMA, R.; JINDAL, N. Machine learning and deep learning applications-a vision. **Global Transitions Proceedings**, Elsevier, v. 2, n. 1, p. 24–28, 2021. Citado na página 18.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 261–266, 2015. Citado na página 18.

VAJJALA, S. *et al.* **Practical natural language processing: a comprehensive guide to building real-world NLP systems**. [S.l.]: O'Reilly Media, 2020. Citado na página 19.

ZHANG, C.; LU, Y. Study on artificial intelligence: The state of the art and future prospects. **Journal of Industrial Information Integration**, Elsevier, v. 23, p. 100224, 2021. Citado na página 18.

LEE, I. Social media analytics for enterprises: Typology, methods, and processes. **Business Horizons**, Elsevier, v. 61, n. 2, p. 199–210, 2018. Citado na página 18.

BATRINCA, B.; TRELEAVEN, P. C. Social media analytics: a survey of techniques, tools and platforms. **Ai & Society**, Springer, v. 30, p. 89–116, 2015. Citado na página 19.

DERAKHSHAN, A.; BEIGY, H. Sentiment analysis on stock social media for stock price movement prediction. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 85, p. 569–578, 2019. Citado na página 19.

HAYAT, M. K. *et al.* Towards deep learning prospects: insights for social media analytics. **IEEE access**, IEEE, v. 7, p. 36958–36979, 2019. Citado na página 19.

NASCIMENTO, R. M. F. d. **Classificação automática de discursos de ódio em textos do Twitter**. Dissertação (B.S. thesis) — Brasil, 2019. Citado na página 19.

OLIVEIRA, N. R. de *et al.* Identifying fake news on social networks based on natural language processing: trends and challenges. **Information**, MDPI, v. 12, n. 1, p. 38, 2021. Citado na página 19.

KHURANA, D. *et al.* Natural language processing: State of the art, current trends and challenges. **Multimedia tools and applications**, Springer, v. 82, n. 3, p. 3713–3744, 2023. Citado 4 vezes nas páginas 19, 23, 25 e 41.

JÚNIOR, E. G. S. L. *et al.* Ferramentas para análise de mídias sociais: Um levantamento sistemático. **Anais do Computer on the Beach**, v. 11, n. 1, p. 389–396, 2020. Citado 4 vezes nas páginas 19, 24, 25 e 42.

LOBATO, F. M.; SOUSA, G. C. de; JR, A. F. J. Brasnam em perspectiva: uma análise da sua trajetória até os 10 anos de existência. In: SBC. **Anais do X Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2021. p. 217–228. Citado 2 vezes nas páginas 19 e 27.

GRANT, M. J.; BOOTH, A. A typology of reviews: an analysis of 14 review types and associated methodologies. **Health information & libraries journal**, Wiley Online Library, v. 26, n. 2, p. 91–108, 2009. Citado na página 20.

KITCHENHAM, B.; CHARTERS, S. *et al.* **Guidelines for performing systematic literature reviews in software engineering**. [S.l.]: UK, 2007. Citado 2 vezes nas páginas 20 e 26.

HASSANI, A.; MOSCONI, E. Social media analytics, competitive intelligence, and dynamic capabilities in manufacturing smes. **Technological Forecasting and Social Change**, Elsevier, v. 175, p. 121416, 2022. Citado na página 23.

YIGITCANLAR, T. *et al.* How can social media analytics assist authorities in pandemic-related policy decisions? insights from australian states and territories. **Health Information Science and Systems**, Springer, v. 8, p. 1–21, 2020. Citado na página 23.

MIRZAALIAN, F.; HALPENNY, E. Social media analytics in hospitality and tourism: A systematic literature review and future trends. **Journal of Hospitality and Tourism Technology**, Emerald Publishing Limited, v. 10, n. 4, p. 764–790, 2019. Citado na página 23.

HE, W. *et al.* Identifying customer knowledge on social media through data analytics. **Journal of Enterprise Information Management**, Emerald Publishing Limited, v. 32, n. 1, p. 152–169, 2019. Citado na página 23.

ZACHLOD, C. *et al.* Analytics of social media data–state of characteristics and application. **Journal of Business Research**, Elsevier, v. 144, p. 1064–1076, 2022. Citado 3 vezes nas páginas 23, 27 e 42.

DRUS, Z.; KHALID, H. Sentiment analysis in social media and its application: Systematic literature review. **Procedia Computer Science**, Elsevier, v. 161, p. 707–714, 2019. Citado na página 23.

MADILA, S.; DIDA, M.; KAIJAGE, S. A review of usage and applications of social media analytics. **Journal of Information Systems Engineering and Management**, Veritas Publications LTD, v. 6, n. 3, 2021. Citado na página 23.

GHANI, N. A. *et al.* Social media big data analytics: A survey. **Computers in Human Behavior**, Elsevier, v. 101, p. 417–428, 2019. Citado na página 23.

SOUZA, E. *et al.* Characterising text mining: a systematic mapping review of the portuguese language. **IET Software**, Wiley Online Library, v. 12, n. 2, p. 49–75, 2018. Citado 4 vezes nas páginas 24, 25, 41 e 42.

SINOARA, R. A.; ANTUNES, J.; REZENDE, S. O. Text mining and semantics: a systematic mapping study. **Journal of the Brazilian Computer Society**, Springer, v. 23, p. 1–20, 2017. Citado na página 26.

PELISSARI, R. *et al.* The use of multiple criteria decision aiding methods in recommender systems: A literature review. In: SPRINGER. **Brazilian Conference on Intelligent Systems**. [S.l.], 2022. p. 535–549. Citado na página 26.

CARVALHO, L. P. *et al.* Ethics: What is the research scenario in the brazilian conference bracis? In: SBC. **Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2022. p. 624–635. Citado na página 27.

PARDO, T. *et al.* Computational linguistics in brazil: an overview. In: **Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas**. [S.l.: s.n.], 2010. p. 1–7. Citado 2 vezes nas páginas 27 e 28.

LEQUERTIER, V. *et al.* Hospital length of stay prediction methods: a systematic review. **Medical Care**, Wolters Kluwer, v. 59, n. 10, p. 929–938, 2021. Citado na página 30.

PACHUCKI, C.; GROHS, R.; SCHOLL-GRISSEMANN, U. Is nothing like before? covid-19–evoked changes to tourism destination social media communication. **Journal of Destination Marketing & Management**, Elsevier, v. 23, p. 100692, 2022. Citado na página 32.

ROSEN, A. O. *et al.* Is social media a new type of social support? social media use in spain during the covid-19 pandemic: A mixed methods study. **International Journal of Environmental Research and Public Health**, MDPI, v. 19, n. 7, p. 3952, 2022. Citado na página 32.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Citado 2 vezes nas páginas 34 e 41.

BERRAR, D. Cross-validation. In: RANGANATHAN, S. *et al.* (Ed.). **Encyclopedia of Bioinformatics and Computational Biology**. Oxford: Academic Press, 2019. p. 542–545. Citado na página 40.

APÊNDICES

APÊNDICE A – ARTIGO SUBMETIDO AO BRACIS 2023

Natural Language Processing and Social Media: a systematic mapping on brazilian leading events

No Author Given

No Institute Given

Abstract. The number of social media platforms has increased significantly, as has the number of active users. More than 18.2 million text messages are transmitted every minute on these platforms. Given the amount of data available, Natural Language Processing (NLP) techniques have been used by several researchers to generate automated analysis. Thus, it is essential to understand social media analysis's leading trends and challenges, especially in scientific events. From this perspective, this study presents a systematic mapping of the use of NLP in works published in five academic events: BRACIS, BraSNAM, ENIAC, STIL, and PROPOR. The study aims to identify the main tools and techniques used and the development environments, tasks performed, and metrics. To this end, 186 studies were analyzed and carefully selected from the 654 papers published in these events in the three years (2020 to 2022). The results show a clipping of the current scenario on the subject and point to areas that can be improved in future research using techniques such as sentiment analysis, emotion analysis, text classification, and named entity recognition. Thus, this work can be helpful for academic researchers interested in exploring the potential of these tools and techniques, having a clear picture of gaps, challenges, and research opportunities in this area, and analyzing the current scenario in research involving NLP and social media.

Keywords: Natural Language Processing, Text Mining, Systematic Mapping · Social media · Social networks

1 Introduction

Social media facilitates the connection between individuals and helps break down communication barriers, allowing everyone to share their stories and opinions [15]. Using this definition, Kaplan and Haenlein [17] describes social media "as a group of applications based on the Internet and the ideological and technological foundations of Web 2.0 that allow the creation and exchange of User-Generated Content (UGC)". In this sense, we can think of social media as the leading platforms and their functionalities, such as Facebook, Instagram, and Twitter. In practical terms, we can also understand social media as an additional digital marketing channel that professionals can use to establish communication with consumers through advertising strategies. From this perspective, social media

becomes less about specific technologies or platforms and more about sharing information between users [2, 25].

Over the years, the number of social media platforms and active users on these platforms has increased significantly, making it one of the most important applications on the internet [1]. This fact has consequently led to the rise of communication via text, with over 18.2 million text messages transmitted every minute. [3]. The data generated by users have sparked academic interest, resulting in the increasing importance of the social media analysis field, which involves collecting and analyzing various social media data and extracting valuable and hidden information [6]. In the same direction, Natural Language Processing (NLP) has emerged as a promising approach for analyzing social media.

NLP is a subfield of computer science that uses computational techniques to learn, understand and produce human language content from the enormous amount of linguistic data available [14]. The NLP area focuses on interpreting, analyzing, and manipulating natural language data for a specific purpose, using different algorithms, tools, and methods. However, many challenges may depend on the natural language data context, making it difficult to achieve all goals with a single approach. For this reason, the development of different tools and methods in the field of NLP has been widely studied by several researchers [19], including specific tools and methods adapted to UGC [16]. With the growth of Brazilian communities of artificial intelligence, data science, social media analysis [21], and natural language processing, we wonder how knowledge in these areas is being spread. To the best of our knowledge, there is no survey in the literature on methods and techniques for analyzing social media used in Brazilian events in the communities mentioned above.

In order to fill this gap in the literature, we conducted a systematic mapping aiming to provide an overview of NLP techniques' application on social media analysis, identify the most used algorithms, and understand current trends in the use of NLP in this context. By performing this systematic mapping, it is possible to obtain a comprehensive view of the state of the art and state of practice. We have chosen the top five scientific events that publish work at the intersection of NLP and Social Media, namely: Brazilian Conference on Intelligent Systems (BRACIS), Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), *Encontro Nacional de Inteligência Artificial e Computacional* (ENIAC), Symposium in Information and Human Language Technology (STIL) and International Conference on Computational Processing of Portuguese Language (PROPOR). We considered the three-year period (2020 to 2022) in our analysis, totaling 654 papers listed, and 186 (30%) were scrutinized.

The results obtained are helpful for researchers and practitioners interested in exploring the potential of these tools and techniques, having a clear picture of gaps, challenges, and research opportunities in this area, and analyzing the current scenario in research involving NLP and social media. It is essential to point out that we are following Open Science Principles by providing all our data in a publicly available GitHub repository, allowing the reproducibility of the results.

The remainder of this paper is organized as follows. In Section 2, relevant works for applying NLP techniques in text analysis are introduced. In Section 3, the research method used for the systematic mapping is presented, along with the scope delimitation protocol adopted. The results obtained through the exploratory analysis of the papers and the most relevant algorithms found are shown in Section 4. Finally, the conclusion is presented in Section 5, including guidelines for future research.

2 Related Works

Social media has become an essential data source for analysis across different sectors, including business, government, and the leisure industry [11, 29, 23]. As the amount of data generated daily grows, data analysis techniques have become more important than ever in providing valuable insights [13].

As a result, many researchers have been exploring this area, aiming to identify the research domains addressed through the performed analyses. That fact allows a better understanding of social events, so [30] conducted a systematic review of 94 papers published between 2017 and 2020 to find the research domains used in social media data analysis. This study identified that most of this data was collected from Twitter, Facebook, Instagram, YouTube, TripAdvisor, and LinkedIn. Scholars have paid significant attention to marketing, as various users with varying interests generate social media data. In addition, areas that need instant information, such as disaster management, hospitality, and tourism, are also covered by this kind of analysis.

NLP techniques are commonly used to extract and analyze content created by users [9, 22]. In the study by [10], many techniques and methods are employed to analyze social media data. The primary focus areas include users' emotions classification, information detection, spatio-temporal analysis, clustering, and performance evaluation. This is reinforced by [6], where the author performs a systematic review of 57 social media studies focused on Business Intelligence (BI) between 2014 - 2018. Three research questions are proposed to extract these works' data, methodology, and algorithms. Given this, different platforms and groups are identified, indicating that most of the platforms used are of the commercial review type (e.g., Amazon, Yelp, and TripAdvisor) or social networking services (e.g., Facebook, Twitter, and LinkedIn). The researchers used various algorithms, including sentiment analysis, topic modeling, Machine Learning-based approach, network-based approach, and theoretical approach to analyzing the data.

Still, in this perspective, [19] aimed to present in detail the state of the art regarding trends and challenges in the field of NLP. Deep Learning and Machine Learning techniques have been used in different NLP tasks. For example, neural network models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are applied in sentence classification, text classification, summarization, machine translation, and information retrieval. In addition to other techniques employed in Multitasking Learning (e.g., Part-of-Speech tag-

ging (POS) and Named Entity Recognition (NER)), Word Embedding (e.g., GloVe), and Attention Mechanisms such as Transformers, which BERT (Bidirectional Encoder Representations from Transformers) is the most used. The authors point out that despite significant advances, there are challenges to be faced, such as the lack of language models intended for broader use in different domains or geographic areas, as it is still a problem to deal with words or sentences with different meanings between these areas.

However, there is still much room for research and development in this area, especially with the emergence of new platforms and the evolution of data analysis techniques.

3 Methodology

We carried out a systematic mapping study using the methodology presented by [28] and [26]. This choice was because they are based on the methodological path proposed by [18], where a valuable guide for planning and conducting involved in secondary studies is presented. Systematic mapping is a bibliographic review technique that, although it differs from a systematic review due to the depth and breadth of the analyzed studies, follows a well-defined protocol and can be used to obtain a mapping of publications on some subject or field, identifying research gaps and areas that require the development of primary studies [28]. In Section 4, we will present the results of our mapping.

3.1 Phase 1: Planning

In the planning phase, the protocol was defined, in which the research questions are described, as well as the research process with the sources in which the studies were mapped, the studies selection guided by the inclusion criteria and deletion.

Research Questions. The research questions (RQs) for systematic mapping are presented below:

- RQ1: Which NLP tools and techniques are used in the scientific events selected, and which are the most recurrent?
- RQ2: What are the sources and nature of the data used in social media analysis?
- RQ3: What are the most used evaluation metrics in NLP studies?

Search Process. The research process consisted of a manual search of the scientific events' proceedings, such as conferences, symposiums, meetings, and workshops between 2020 and 2022, listed in Table 1. These works are available in two digital library research sources: the SBC-OpenLib (SOL) of the Brazilian Computer Society (SBC) and SpringerLink. Given that, these events were selected because they are considered important national and international research bases in studies related to the areas of Natural Language Processing, Artificial, and Computational Intelligence, as pointed out by [21] and [5].

Table 1. Proceedings of selected events.

Source	Acronym	Edition
Brazilian Conference on Intelligent Systems	BRACIS	2020 - 2022
Brazilian Workshop on Social Network Analysis and Mining	BraSNAM	2020 - 2022
Encontro Nacional de Inteligência Artificial e Computacional	ENIAC	2020 - 2022
International Conference on Computational Processing of the Portuguese Language	PROPOR	2020 - 2022
Symposium in Information and Human Language Technology	STIL	2021

Study Selection. In order to select the most relevant works on NLP techniques applied in social media analysis of the events given in Table 1, inclusion and exclusion criteria were applied in the selection of works arranged in Table 2. Inclusion criteria were papers in the events' proceedings that address techniques, models, text mining tools, and textual analysis in social media analysis.

Subsequently, scientific papers that are outside the inclusion criteria, publications written in languages other than Portuguese or English, works that are not relevant to the NLP, and analysis of social media based on the title, abstract, keywords, introduction, and conclusion. This selection followed the order (i) title, abstract and keywords; (ii) introduction and conclusion and (iii) full paper. Papers that addressed systematic mapping or systematic literature review were also excluded.

Table 2. Inclusion (IC) and Exclusion (EC) Criteria for Selecting Relevant Studies.

Inclusion Criteria (IC)
IC1: Papers that address techniques, models, text mining tools, and textual analysis in social media analysis.
Exclusion Criteria (EC)
EC1: Papers that are outside the inclusion criteria
EC2: Publications written in languages other than Portuguese or English
EC3: Papers that are not relevant to NLP and social media analysis based on the title, abstract, keywords, introduction, and conclusion
EC4: Papers of systematic mapping or systematic literature review.

3.2 Phase 2: Conduction

The proceedings of relevant events in the research area in their editions between the years 2020 to 2022 were selected, and the works for the initial research were obtained from them. The selected events were BRACIS, BraSNAM, ENIAC, PROPOR, and STIL, which resulted in 654 papers.

Two authors read the papers individually and assessed whether the works met the inclusion and exclusion criteria in Table 2. After rigorously applying these criteria, 468 papers were excluded, most due to EC3 and EC4 criteria.

The exclusion occurs because many of them dealt with studies that involved the manipulation of multimedia data, such as images, videos, or audio, as well as the construction, description, or annotation of a corpus, not falling within the scope of this systematic mapping, which aims to evaluate studies that address text mining or text analysis. Based on the established inclusion criteria, 186 papers were selected for data extraction.

Data Extraction. The selected papers were read fully for the data extraction stage while applying the inclusion and exclusion criteria.

A spreadsheet was created to include the following data from each study: paper source, data source (e.g., Twitter, Reddit, and Facebook), nature of the data (e.g., built/collected corpus, already available or Not described), Tool and Technology (e.g., NLTK, and Spacy), tasks (e.g., pre-processing, Text Classification, and Sentiment Analysis), development environment (e.g., Python, and Jupyter Notebook), techniques (e.g., TF-IDF, and BERT), metrics (e.g., F1-Score, and Cosine Similarity). The mapped attributes and their relationship with the research questions are described in Table 3.

Table 3. Mapping the extracted data and the research question to which it is related.

Data	**Description**	**Relevant RQ**
Paper Source	Author, Title, Event, Year, DOI	Study Overview
Data Source	Data Sources used in the study	RQ2
Data Nature	Corpus built or collected, already available or not described	RQ2
Tool and Technology	Tool or technology is used to analyze the data in the study	RQ1
Preprocessing	Data pre-processing steps are performed in the study.	RQ1
Text Mining/ NLP Tasks	NLP tasks related to the tools and techniques that were performed in the study	RQ1
Techniques used	NLP techniques applied in the study	RQ1
Metrics	Evaluation measures and metrics used in study	RQ3
Repository	Link to available repositories to data and source codes	Study Overview

These information was mainly extracted from the materials and methods described in the papers, although some relevant information was also extracted

from the full text. Table 4 presents the number of papers included and excluded in each selected event.

Table 4. Selected works.

Events	Initial research papers	Deleted Papers	Selected papers
BRACIS	247	204	43
BraSNAM	67	34	33
ENIAC	206	157	49
PROPOR	83	53	30
STIL	51	20	31
TOTAL	654	468	186

An inductive approach was adopted to extract and analyze information from the qualitative data collected. For this, an exploratory analysis was carried out using the Python 3.9.13 programming language with the aid of the Jupyter Notebook interactive programming environment. Initially, the spreadsheet generated in Excel was converted from .xlsx to .csv (Comma-separated Values) using the Pandas library.

Then, this database with the mapped attributes was pre-processed to remove unnecessary spaces and accents and convert the strings to lowercase. Libraries such as unicodedata and regex were also used to perform these operations. Therefore, qualitative data were extracted by developing a dictionary for counting words using the collections library. It is essential to point out that all material produced during the conduction of this systematic mapping will be made public in a repository on GitHub[1]. These papers will be described and explored in the results section.

4 Results

In this section, the study's findings are presented, as well as the answers to the research questions. The results were organized as follows. First, we present the exploratory analysis of the selected works. Then, we present the most frequent NLP tools and techniques in social media analysis identified, followed by the sources and nature of the data used in these analyses. Finally, we discuss the evaluation metrics.

The mapping reported in this work was carried out to provide an overview of the research carried out by the Text Mining community and related to social media analysis. This mapping is based on 186 selected studies dealing with textual analysis, as described in the previous section, and the distribution of

[1] https://anonymous.4open.science/r/BRACIS-systematic-mapping-616A

these studies by year of publication is presented in Fig. 1. The graph shows that the number of publications was higher in 2021. This fact is possibly due to the amount of content generated on social media due to the COVID-19 pandemic [24, 27]. It is also worth mentioning that some events, such as STIL and PROPOR, occur every two years.

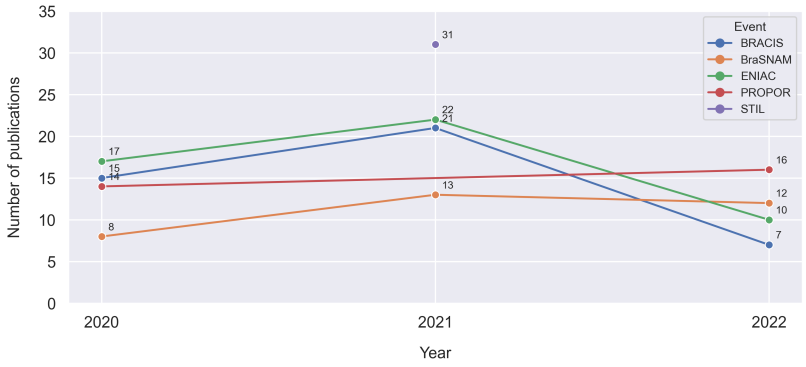


Fig. 1. Annual distribution of publications selected by the event.

During the conduction of our analysis, it was noticed the diffusion of works in other areas of Artificial Intelligence and Computational Intelligence in the events' proceedings, such as BRACIS and ENIAC in the year 2022. This tendency to expand and explore new research fields may have impacted the proportion of specific textual analysis studies within the scope of this mapping.

4.1 NLP Tools and Techniques

With the increasing availability of textual data in social media, it becomes necessary to explore adequate tools and techniques to deal with this volume of information. Thus, NLP has stood out as a whole area in this field through essential tools and techniques for analyzing and understanding these textual data [20]. Several text mining tools are available in this context, from simple open-source tools to libraries offering a wide range of resources and functionality for collecting, manipulating, cleaning, and analyzing data [4]. Linked to these tools, social media analysis involves using different modeling and analysis techniques from various fields [7]. These techniques cover the application of Machine Learning and Deep Learning algorithms for tasks such as text classification, sentiment analysis, summarization and machine translation, and entity extraction[3, 12].

State of Tool Development in NLP. We identified 135 tools for NLP tasks in the papers analyzed, the 20 most frequent being shown in Fig. 2. From this,

it was observed that the most used tool is Scikit-Learn[2], present in 55 studies of this mapping. It is essential to highlight that this tool is a Machine Learning library in Python, and it is prevalent due to its ease and applicability of use in performing NLP tasks (e.g., classification, regression, and clustering). In addition, other tools were identified as being commonly used in text pre-processing tasks, data manipulation and analysis, sentiment analysis, and topic modeling, among others. These are NLTK[3] (Natural Language Toolkit), an open-source NLP platform that supports various tasks such as tokenization, sentiment analysis, POS tagging, and analytics. of topics. Moreover, spaCy[4], in turn, is a Python library with resources for POS tagging tasks, NER, Syntactic Parsing, text classification, lemmatization, and morphological analysis, among others. Fig. 2 also reveals a large number of underused tools. While reading the papers, it was possible to observe that these tools are aligned with NLP tasks, mainly those related to pre-processing (e.g., removing stopwords, tokenization, lowercasing, etc.) and tasks related to the techniques used and described below.

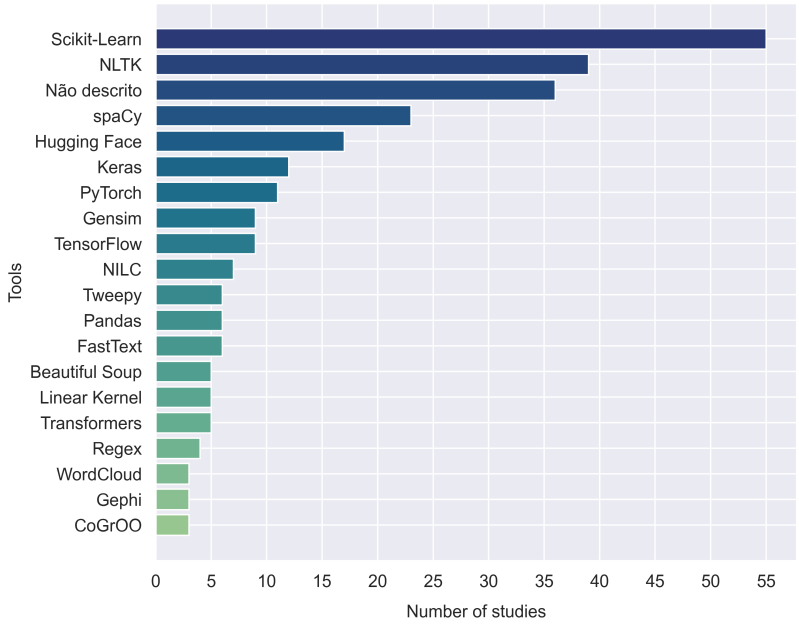


Fig. 2. The 20 most frequent tools in social media analysis.

[2] https://scikit-learn.org/
[3] https://www.nltk.org/
[4] https://spacy.io/

State of Techniques Developed in NLP. From the analysis of the selected studies, a total of 277 techniques were identified. Fig. 3 shows that the most frequent technique is BERT (used by 61 studies), a language model based on transformers that stood out for its performance in NLP [8] tasks.

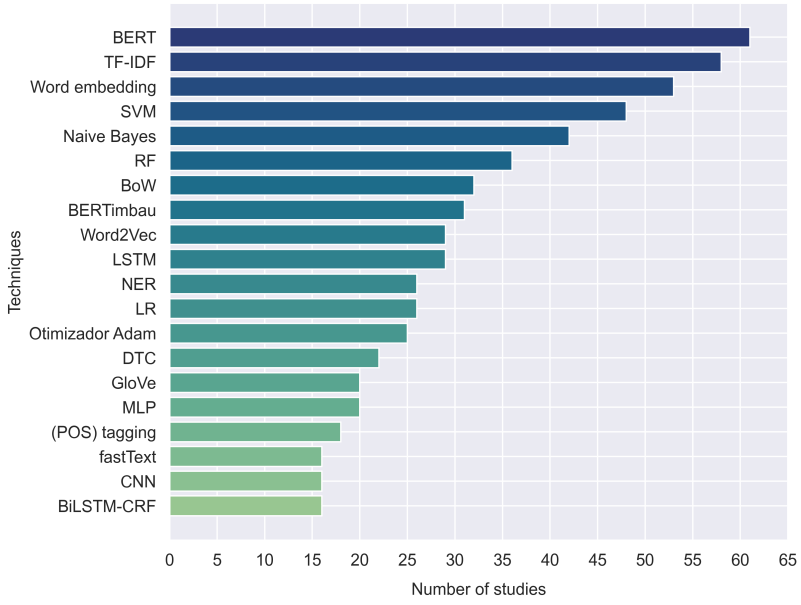


Fig. 3. The 20 most frequent techniques in social media analysis.

Next, we have the TF-IDF technique, which was used in 58 studies, and widely applied in the representation of documents and calculating the importance of words in a corpus. In addition, we observed the presence of several other Deep Learning techniques, such as Word embedding (used by 53 studies), Word2Vec, Long Short-Term Memory (LSTM), Multi-layer Perceptron (MLP), Global Vectors (GloVe), FastText, CNN and Bidirectional LSTM with Conditional Random Fields (BiLSTM-CRF), they help to capture semantic information and to model the linguistic context.

Other Machine Learning techniques are also found that are widely used for the classification and training of textual data, such as the Support Vector Machine (SVM) in 48 studies, Naive Bayes (42 studies), Random Forest (RF) in 36 studies, LR (Logistic Regression) in 26 studies and Decision Tree Classifier (DTC) in 22 studies.

In addition, a comparison of the ten most used tools between events was performed, as shown in Fig. 4. We can identify researchers' preferences regarding using NLP techniques in text analysis through this analysis.

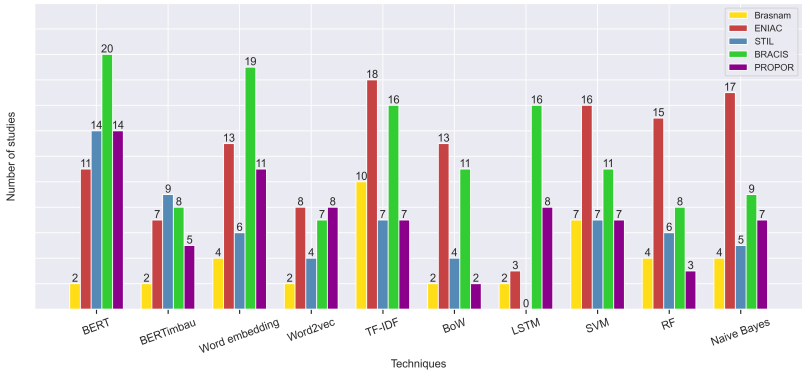


Fig. 4. Comparison of the 10 most frequent techniques in social media analysis by event.

Therefore, it is seen that BRACIS is the event with the highest prevalence of studies that use Deep Learning techniques, mainly for text classification, with language models such as BERT, Word embedding, and LSTM. These results suggest a significant interest in using approaches based on pre-trained language models, distributed word representations, and recurrent neural networks in social media analysis.

In contrast, ENIAC addresses more works that use text classification and text representation models in their analyses, such as SVM, RF, Naive Bayes, and TF-IDF. These results show the use of traditional Machine Learning techniques in social media analysis in studies published in ENIAC.

4.2 Social Media Data Sources

This section aims to discuss the sources of data and the nature of these data used in the work analyses. The most frequent data sources in the studies were identified during the analysis of the works. When analyzing the word cloud Fig. 5, it is possible to observe that Twitter is the most frequent data source, indicating that researchers widely explore this platform. In addition to Twitter, other sources, such as ASSIN (Assessment of Semantic Similarity and Textual Inference), consist of a dataset with 10,000 pairs of sentences from Google News[5]. Wikipedia, UOL, TripAdvisor, G1, Folha de São Paulo, Amazon, the Judiciary System, and E-commerce are also present.

[5] https://news.google.com/

Fig. 5. Word cloud of data sources in studies.

These data sources represent different contexts and textual content, ranging from user opinions, news, and product information in e-commerce to legal data. With this, it is possible to notice that different research domains are analyzed. Additionally, how the data were obtained was investigated. Of the analyzed studies, 110 works used corpus constructed or collected, which involves creating or collecting specific data for the research. On the other hand, 75 works used already available corpus, that is, data sets readily available for analysis. Only one piece did not describe the nature of the used corpus.

4.3 Evaluation Metrics

In this section, we explore the metrics used to evaluate and validate the data in the studies. In analyzing selected works, several evaluation metrics are used to evaluate the performance of data classification models in NLP. Among the 101 metrics identified, the most used were the F1-Score, present in 108 studies, Recall in 81 studies, Precision in 78 studies, and Accuracy in 55 studies.

The F1-Score is a metric representing the harmonic mean between Recall and Precision, providing a balanced measure of model performance. In addition to these metrics, the researchers use the Cross-validation method cited in 59 studies to evaluate and compare classification algorithms. These evaluation metrics are fundamental to assessing the quality and effectiveness of studies.

5 Conclusions

We present in this paper a systematic mapping study to analyze NLP techniques used in social media analysis. We identified the main scientific events in the area (BRACIS, BRaSNAM, ENIAC, STIL, and PROPOR) and selected 186 relevant works published between 2020 and 2022, from the 654 papers, based on our selection criteria. Our research was guided by three Research Questions: RQ1: Which NLP tools and techniques are used in scientific events, and which are the most recurrent? RQ2: What are the sources and nature of the data used in social media analysis? RQ3: What are the evaluation metrics in NLP studies?

The most mentioned tools in the works were Scikit-Learn, NLTK, and spaCy. We also identified 282 techniques, among which BERT was the most cited. These NLP techniques include sentiment analysis tasks, topic modeling, and text classification through training and testing data, with approaches based on Machine Learning and Deep Learning models. As for the platforms, the most explored are Twitter, Wikipedia, TripAdvisor, and news portals. In addition, we identified that most studies perform the collection and construction of a specific corpus for their analyses. Furthermore, many researchers have used metrics such as F1-Score, Recall and Precision, and the Cross-validation statistical method to evaluate their models.

Our research also revealed some threats to the study's validity. One is the limitation of the selected events, which may only partially represent some research areas in social media analysis. Another threat to validity is the possibility of publication bias; since the selected papers are those available in the proceedings of specific events, a suggestion would be to carry out an automatic collection using well-defined search keywords in other databases.

The main contribution of this work is to provide an overview of the application of several NLP tools and techniques in social media analytics. It is important to highlight that we are following Open Science Principles by providing all our data in a publicly available GitHub repository, allowing the reproducibility of the results. This work can be helpful for researchers and practitioners interested in exploring the potential of these tools and techniques, having a clear picture of gaps, challenges, and research opportunities in this area, and analyzing the current scenario in research involving NLP and social media. For future work, we plan to expand our data sources to relevant international conferences (e.g., AAAI Conference on Artificial Intelligence, International AAAI Conference on Web and Social Media, *etc*), aiming to compare the current status of Brazilian research programs with the international ones.

Acknowledgment.

References

1. Aichner, T., Grünfelder, M., Maurer, O., Jegeni, D.: Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. Cyberpsychology, behavior, and social networking **24**(4), 215–222 (2021)

2. Appel, G., Grewal, L., Hadi, R., Stephen, A.T.: The future of social media in marketing. Journal of the Academy of Marketing science **48**(1), 79–95 (2020)
3. Balaji, T., Annavarapu, C.S.R., Bablani, A.: Machine learning algorithms for social media analysis: A survey. Computer Science Review **40**, 100395 (2021)
4. Batrinca, B., Treleaven, P.C.: Social media analytics: a survey of techniques, tools and platforms. Ai & Society **30**, 89–116 (2015)
5. Carvalho, L.P., Murakami, L., Suzano, J.A., Oliveira, J., Revoredo, K., Santoro, F.M.: Ethics: What is the research scenario in the brazilian conference bracis? In: Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional. pp. 624–635. SBC (2022)
6. Choi, J., Yoon, J., Chung, J., Coh, B.Y., Lee, J.M.: Social media analytics and business intelligence research: A systematic review. Information Processing & Management **57**(6), 102279 (2020)
7. Derakhshan, A., Beigy, H.: Sentiment analysis on stock social media for stock price movement prediction. Engineering Applications of Artificial Intelligence **85**, 569–578 (2019)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Drus, Z., Khalid, H.: Sentiment analysis in social media and its application: Systematic literature review. Procedia Computer Science **161**, 707–714 (2019)
10. Ghani, N.A., Hamid, S., Hashem, I.A.T., Ahmed, E.: Social media big data analytics: A survey. Computers in Human Behavior **101**, 417–428 (2019)
11. Hassani, A., Mosconi, E.: Social media analytics, competitive intelligence, and dynamic capabilities in manufacturing smes. Technological Forecasting and Social Change **175**, 121416 (2022)
12. Hayat, M.K., Daud, A., Alshdadi, A.A., Banjar, A., Abbasi, R.A., Bao, Y., Dawood, H.: Towards deep learning prospects: insights for social media analytics. IEEE access **7**, 36958–36979 (2019)
13. He, W., Zhang, W., Tian, X., Tao, R., Akula, V.: Identifying customer knowledge on social media through data analytics. Journal of Enterprise Information Management **32**(1), 152–169 (2019)
14. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science **349**(6245), 261–266 (2015)
15. Hou, Q., Han, M., Cai, Z.: Survey on data analysis in social media: A practical application aspect. Big Data Mining and Analytics **3**(4), 259–279 (2020)
16. Júnior, E.G.S.L., de Sousa, G.N., Junior, A.F.L.J., Lobato, F.M.F.: Ferramentas para análise de mídias sociais: Um levantamento sistemático. Anais do Computer on the Beach **11**(1), 389–396 (2020)
17. Kaplan, A.M., Haenlein, M.: Users of the world, unite! the challenges and opportunities of social media. Business horizons **53**(1), 59–68 (2010)
18. Keele, S., et al.: Guidelines for performing systematic literature reviews in software engineering (2007)
19. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications **82**(3), 3713–3744 (2023)
20. Lee, I.: Social media analytics for enterprises: Typology, methods, and processes. Business Horizons **61**(2), 199–210 (2018)
21. Lobato, F.M., de Sousa, G.C., Jacob Jr, A.F.: Brasnam em perspectiva: uma análise da sua trajetória até os 10 anos de existência. In: Anais do X Brazilian Workshop on Social Network Analysis and Mining. pp. 217–228. SBC (2021)

22. Madila, S., Dida, M., Kaijage, S.: A review of usage and applications of social media analytics. Journal of Information Systems Engineering and Management **6**(3) (2021)
23. Mirzaalian, F., Halpenny, E.: Social media analytics in hospitality and tourism: A systematic literature review and future trends. Journal of Hospitality and Tourism Technology **10**(4), 764–790 (2019)
24. Pachucki, C., Grohs, R., Scholl-Grissemann, U.: Is nothing like before? covid-19–evoked changes to tourism destination social media communication. Journal of Destination Marketing & Management **23**, 100692 (2022)
25. Pan, Y., Torres, I.M., Zúñiga, M.A.: Social media communications and marketing strategy: A taxonomical review of potential explanatory approaches. Journal of Internet Commerce **18**(1), 73–90 (2019)
26. Pelissari, R., Alencar, P.S., Amor, S.B., Duarte, L.T.: The use of multiple criteria decision aiding methods in recommender systems: A literature review. In: Brazilian Conference on Intelligent Systems. pp. 535–549. Springer (2022)
27. Rosen, A.O., Holmes, A.L., Balluerka, N., Hidalgo, M.D., Gorostiaga, A., Gómez-Benito, J., Huedo-Medina, T.B.: Is social media a new type of social support? social media use in spain during the covid-19 pandemic: A mixed methods study. International Journal of Environmental Research and Public Health **19**(7), 3952 (2022)
28. Sinoara, R.A., Antunes, J., Rezende, S.O.: Text mining and semantics: a systematic mapping study. Journal of the Brazilian Computer Society **23**, 1–20 (2017)
29. Yigitcanlar, T., Kankanamge, N., Preston, A., Gill, P.S., Rezayee, M., Ostadnia, M., Xia, B., Ioppolo, G.: How can social media analytics assist authorities in pandemic-related policy decisions? insights from australian states and territories. Health Information Science and Systems **8**, 1–21 (2020)
30. Zachlod, C., Samuel, O., Ochsner, A., Werthmüller, S.: Analytics of social media data–state of characteristics and application. Journal of Business Research **144**, 1064–1076 (2022)

APÊNDICE B – BASE DE DADOS DOS TRABALHOS SELECIONADOS

Artigos	Fonte de dados	Natureza dos dados	Ferramentas	Pré-processamento	Tarefas	Ambiente	Técnicas	Métricas
[17]	Twitter	corpus construído/ coletado	NLTK, Spacy	Remoção (stopwords), tokenização, bigramas	Classificação (identificação de comunidades, análise de redes), Modelagem de tópicos	Python	Louvain, LDA, PageRank	Modularidade
[161]	Twitter	corpus construído/ coletado	Não descrito	Remoção (stopwords, links, emojis, hashtags, pontuações, espaços, linhas em branco, textos duplicados, menções, aspas), lowercasing, tokenização, redução de dimensionalidade	Clusterização, Modelagem de tópicos, Classificação (rotulagem de dados, detecção de posicionamento, Pontuação de valência, extração de características)	Python	UMAP, Mean Shift, BERTopic, BERT, BERTimbau, HDBSCAN, TF-IDF, c-TF-IDF, word embeddings	Similaridade de cosseno, grau de engajamento, grau de equilíbrio, Intervalo de confiança
[132]	Twitter	corpus construído/ coletado	Twint, TextBlob, Wordcloud, Syuzhet	Remoção (acentos, links, emojis, caracteres especiais), lowercasing	Visualização (nuvem de palavra, gráfico de emoções), Classificação (identificação de emoções, Análise de sentimentos, frequência de termos, identificação de polaridade)	Python, Jupyter Notebook, R	wordcloud, NRC sentiment, Syuzhet	Não descrito
[80]	UOL, Folha de São Paulo, G1, Twitter	corpus construído/ coletado	NLTK, Tweepy	Remoção (links, quebras de linha)	Classificação (rotulagem de dados, balanceamento de dados, treinamento e teste de dados), Visualização (Matriz de confusão)	Python	BERT, BERTimbau, M-BERT, word embeddings, smote, RF, SVM, otimizador AdamW	Fleiss Kappa, Acurácia, Precisão, Revocação, F1-Score, F1-Macro

[8]	Twitter	Não descrito	Angular, Flask, Pumbler, Celery, Redis, ChartJS, PrimeNG, Wordcloud, Syuzhet	Remoção (stopwords, caracteres especiais, pontuações), lowercasing, stemmization, lemmatization	Classificação (Identificação de emoções, Análise de sentimentos, identificação de polaridade), Visualização (wordcloud), teste de usabilidade	Python, R, Heroku	wordcloud, NRC sentiment, Syuzhet	Escala SUS (System Usuability Scale)
[74]	Twitter	corpus construído/ coletado	NLTK, Hatebase	Remoção (stopwords, links, emojis, imagens e outros), stemmization, lowercasing, bigramas	Clusterização, Visualização (wordcloud), Classificação (análise de léxico)	Python	TF-IDF, wordcloud, BoW	Similaridade de cosseno, Método elbow (cotovelo)
[97]	Facebook	corpus construído/ coletado	Crewdtangle	Não descrito	Classificação (Formação de Grafos, extração de backbone, identificação de comunidades, análise de engajamento, Filtragem de domínios), Visualização (boxplot), clusterização	Crowdtangle	Método de Filtro de disparidade, Louvain	Coeficiente de similaridade de Jaccard, Modularidade, distribuição cumulativa, grau médio, peso médio, densidade
[7]	TripAdvisor	corpus construído/ coletado	Selenium, BeautifulSoup	Remoção (stopwords, entidades desnecessárias), Junção de nomes próprios, PoS tagging	Modelagem de tópicos, clusterização, Classificação (Análise de sentimentos, identificação de polaridade, identificação de classes gramaticais), Visualização (wordcloud)	Python	NMF, TF-IDF, LeIA (Léxico para Inferência Adaptada), VADER, wordcloud, NER	NPMI, Root Mean Square Error (RMSE)

[43]	Twitter	corpus construído/ coletado	MongoDB, Pandas, iFeel, os_module	Remoção (stopwords, links, menções, hashtags, pontuações, quebras de linha, caracteres especiais, acentuações)	Modelagem de tópicos, Classificação (Análise de sentimentos, identificação de polaridade, rotulagem de dados)	Python	TF-IDF, NMF, Metodologia Delphi, iFeel	Não descrito
[71]	TripAdvisor, Skoob.com	corpus construído/ coletado	Spacy	Não descrito	Classificação (rotulagem de dados, identificação de classes gramaticais, análise de sentimentos, identificação de sentenças, extração de aspectos), Visualização (Diagrama de Venn)	Python	SenticNet, Algoritmo guloso, Algoritmo de seleção de regras ótimas	Precisão, Revocação, F1-Score
[129]	Twitter	corpus construído/ coletado	Gensim, networkx, Gephi	Remoção (stopwords, hashtags)	Classificação (rotulagem de dados, Identificação de comunidades, caracterização topológica, caracterização de redes), Visualização (Boxplot, wordcloud), Modelagem de tópicos	Pyhton, zenodo	LDA, Louvain, PageRank, wordcloud	Coeficiente Kappa de Cohen, In-degree, out-degree, betweenness
[182]	Twitter	corpus construído/ coletado	NLTK, Scikit-learn, Tweepy	Remoção (stopwords), unigramas, bigramas	Classificação (rotulagem de dados, análise de sentimentos, treinamento e teste de dados), Visualização (wordcloud)	Python	TF-IDF, wordcloud, NB, SVM, RF, VSM	Acurácia, F1-Score, validação cruzada

[107]	Essay-Br	corpus já disponível	NLTK, Spacy, PyHunSpell, Hunspell, UNitex-PB, CoGrOO4py, CoGrOO	Remoção (stopwords, números, pontuações, entidades desnecessárias), lowercasing, tokenização, correção de textos, hapax legonemas	Classificação (Extração de características, extração de aspectos), Análise de classes gramaticais, Visualização (matriz de confusão)	Python	LSA, word embeddings, Doc2Vec, Word2Vec, RNN, GBTD, Ridge, NER, Stanza, LSTM	Similaridade de cosseno, Kappa Quadrático Ponderado (QWK)
[29]	Buscapé	corpus construído/ coletado	Scikit-Learn, Optuna, Hugging Face	Remoção (stopwords, textos duplicados, números, pontuações, símbolos)	Classificação (rotulagem de dados, análise de sentimentos, cruzamento de domínios, extração de características, identificação de polaridade, treinamento e teste de dados), Visualização (wordcloud, matriz de confusão)	Python	Word embeddings, GloVe, TF-IDF, BERT, BERTimbau, RF, SVM, NB, LR, wordcloud	Acurácia, Precisão, Revocação, F1-Score
[180]	Notícias Agrícolas, Macrotrends, CEPEA	corpus construído/ coletado	Scikit-Learn, Hugging Face	Remoção (stopwords), normalização	Classificação (rotulagem de dados, Anotação de polaridade, identificação de polaridade, treinamento e teste de dados)	Python	BoW, TF-IDF, TF, BERT, DistilBERT, BERTimbau, KNN, MLP, NB, SVM, DTC, Friedman e Nemeyi test	Precisão, Revocação, F1-Score

[177]	Twitter	corpus construído/ coletado	MongoDB	Não descrito	Classificação (Rotulagem, avaliação automática), Refering Expression Generation (REG)	Google Colab	RNN, Bart, T5, Blenderbot, GPT2, template-based, pipeline-based, end-to-end, Sentence Aggregation	Bleu, Gleu, Rouge, Meteor
[138]	Electronic Invoices (NF-e)	corpus já disponível	DBMS, PostgreSQL	Remoção (stopwords), tokenização, stemmization	Classificação (Extração de tokens, balanceamento de dados, treinamento e teste de dados), Visualização (boxplot)	Não descrito	TF, TF-IDF, SVM, NB, undersampling e oversampling	F1-Score
[92]	ASSIN, ASSIN2	corpus já disponível	Pandas, numpy, Pytorch, scikit-learn, tqdm, Hugging Face	Tokenização	Classificação (treinamento e teste de dados)	Google Colab, Python	LDA, BERT, BERTimbau, Teste de Tukey, Analysis of variance (ANOVA), fine tuning, otimizador AdamW	Similaridade de cosseno, acurácia, validação cruzada
[118]	SNLI, COLIEE	corpus já disponível	Pytorch	Não descrito	Classificação (treinamento e teste de dados)	Google Colab	Inferência de Linguagem Natural (NLI), BERT, Analysis of variance (ANOVA), Teste de Tukey	Acurácia, validação cruzada
[139]	Senado Federal	corpus construído/ coletado	Spacy, NLTK, Gensim, MALLET	Remoção (stopwords, números, acentos, caracteres especiais, quebras de linha), lowercasing, concatenação, stemmization	Classificação (Filtragem de dados, filtragem de extremos), Modelagem de tópicos	Python	LDA	NPMI

[41]	Ouvidoria	corpus construído/ coletado	NLPAUG, scikit-learn, Pytorch, Hugging Face, nilc	Remoção (stopwords), limpeza dos dados, tokenização	Classificação (rotulagem de dados, redimensionamento, treinamento e teste de dados), Visualização (matriz de confusão)	Google Colab, Python	BERT, BERTimbau, word embeddings, Doc2Vec, Word2Vec, BoW, TF-IDF, TF, KNN, RF, SVM, NB, LR, DTC, Undersampling e Oversampling, Train-Valid-Test Split, otimizador AdamW	Acurácia, Precisão, Revocação, F1-Score
[99]	Jira (Atlassian Pty Ltd)	corpus já disponível	NLTK, RegEX, WordNetLemmatizer, Krovetz stemmization, scikit-learn, Flask	Remoção (Stopwords, símbolos, pontuações, números, ASCII), tokenização, lowercasing, lemmatization, stemmization	Classificação (treinamento e teste de dados, balanceamento de dados)	Python	BoW, TF-IDF, word embeddings, Word2Vec, SMOTE, KNN, RF, SVM, NB, LR, DTC, Mann-Whitney U Test, Friedman e Nemeyi test	AUC, validação cruzada
[125]	MorphoBr, PorGram	corpus já disponível	PyDelphin	Não descrito	Classificação (formas verbais em regulares ou irregulares), Análise morfológica	Haskell	Head-driven phrase structure grammar (HPSG), English Resource Grammar (ERG)	Não descrito
[68]	Petrolês	corpus construído/ coletado	Bosque-UD	Não descrito	Classificação (anotação morfossintática, segmentação de frases e palavras, treinamento e teste de dados), Visualização (matriz de confusão)	Python	Método Inter-Annotator Disagreement (IAD), nmod (modificador nominal do tipo adjunto adnominal), Stanza, UDPipe, Universal Dependencies (UD)	Coeficiente Kappa de Cohen, UPOS, UAS, LAS, CLAS

[57]	Pubmed	corpus construído/ coletado	CoGrOO	limpeza dos dados, tokenização, stemmization, PoS tagging	Classificação (Anotação morfossintática, identificação de classes gramaticais), Visualização (boxplot)	Não descrito	Vocabulário Teórico (VT), RSLP Stemmer, lexicalidade biomédica (LexBioMed)	DeL, DiL
[36]	Twitter	corpus já disponível	Enelvo, linear kernel, scikit-learn	Remoção (stopwords, links, emojis, retweets,hashtags, pontuações), lowercasing, normalização	Classificação (análise de sentimentos, identificação de polaridade, treinamento e teste de dados), Modelagem de tópicos, Clusterização	Python	BERT, BERTimbau, BERTopic, HDBSCAN, TF-IDF, c-tf-idf, SVM, NB, otimizador AdamW, fine tuning	F1-Score
[95]	Reddit	corpus já disponível	itranslate, google translate	Não descrito	Classificação (identificação de emoções, balanceamento de dados)	Python	BERT, BERTimbau, Class Balanced Loss (CB), fine tuning, otimizador AdamW	Precisão, revocação, F1-Score, Cross Entropy, sigmoid
[94]	Diabetes Mellitus	corpus construído/ coletado	Label Studio	Remoção (emojis)	Classificação (Anotação de polaridade, análise de sentimentos, treinamento e teste de dados), Visualização (matriz de confusão)	Python	BERT, BERTimbau, M-BERT, RF, SVM, LR, DTC, BioBERTpt, TF-IDF	Precisão, Revocação, F1-Score, Coeficiente Kappa de Cohen
[42]	Twitter	corpus construído/ coletado	Hugging Face	Remoção (hashtags)	Classificação (rotulagem de dados, Análise de sentimentos, identificação de emoções, mascaramento de dados, treinamento e teste de dados)	Python	BERT, BERTimbau	Precisão, Revocação, F1-Score

[104]	Não descrito	corpus já disponível	Spacy	Tokenização, PoS tagging	Classificação (análise de sentimentos, treinamento e teste de dados, análise de dependência, identificação de aspectos, identificação de classes gramaticais)	Python	Freq-Baseline, Word2Vec, SenticNet, Laplace precision	Precisão, Revocação, F1-Score
[77]	Reddit	corpus construído/ coletado	PRAW3, Scikit-learn	Remoção (emojis)	Classificação (rotulagem de dados, identificação de polaridade, treinamento e teste de dados, balanceamento de dados, extração de características)	Python	smote, SVM, NB	Precisão, Acurácia, Revocação, F1-Score
[15]	Americanas, e-commerce	corpus construído/ coletado	Prodigy, hugging face	Não descrito	Classificação (treinamento e teste de dados, extração de relacionamento)	Python	BERT, NER, BERTimbau, M-BERT, MTB (Multi-Task BERT)	F1-Score, Acurácia
[52]	e-commerce	corpus construído/ coletado	Keras, FastText	Remoção (stopwords, caracteres especiais, números, links), lowercasing, Concatenação	Classificação (treinamento e teste de dados)	Python	FastText, word embeddings, otimizador AdamW	Similaridade multimodal, Precisão, F1-Score, Revocação, Cross Entropy
[53]	e-commerce, Americanas	corpus já disponível	Nilc, FastText	Não descrito	Classificação (treinamento e teste de dados)	Python	Word2Vec, FastText, GloVe, word embeddings, BERT, BERTimbau, M-BERT, Nilc	similaridade de cosseno, Precisão, Revocação, F1-Score

[16]	QuintoAndar chatbot	corpus construído/ coletado	Não descrito	Remoção (acentos, stopwords), lowercasing	Classificação (treinamento e teste de dados, extração de características)	Python	BoW, LR, BERT, RF, MLP	Precisão, Acurácia
[37]	Twitter	corpus construído/ coletado	Não descrito	Limpeza, Remoção (stopwords, pontuações, números, links, menções), stemmization, lowercasing	Classificação (rotulagem de dados, extração de características, treinamento e teste de dados, Análise de sentimentos)	Python, Java	BoW, TF-IDF, BERT, RF, MLP, SVM, NB	AUC, validação cruzada
[109]	Twitter	corpus já disponível	Enelvo, LIWC, gensim, scikit-learn, nilc	normalização, redução de dimensionalidade	Classificação (rotulagem, extração de características, identificação de polaridade)	Python	rotulação fraca, LIWC, TF-IDF, bert, BERTimbau, word embeddings, NILC, SVM, MLP, LR, PCA	F1-Score, validação cruzada
[114]	Twitter	corpus construído/ coletado	NLTK, Regex, scikit-learn	Remoção (stopwords, números, pontuações, links), lowercasing, tokenização	Classificação (rotulagem de dados, treinamento e teste de dados)	Python	BoW, TF-IDF, Word2Vec, LightGBM (LGBM), RF, SVM, LR, AdaBoost (AB), NB, grid search	Acurácia, Precisão, Revocação, F1-Score, AUC, validação cruzada
[172]	Plos One	corpus já disponível	hugging face	Remoção (citações, caracteres especiais, espaços, quebra de linha, stopwords), lowercasing, stemmization	Clusterização, sumarização	Python	SumBasic, LexRank, TextRank, BERT, TF-IDF, PageRank, K-Means (KM), SciBERT	Similaridade de cosseno, Rouge

[113]	MegaLite-Es	corpus já disponível	FreeLing	bigramas	análise sintática	Python	Automatic Text Generation (ATG), RNN, Word2Vec, PERL 5.0, Canned Text method, Empty Grammatical Structures (PGSs), FreeLing, word embeddings, Language Model Analysis (Bigrams)	similaridade de cosseno
[64]	Fake.Br	corpus já disponível	Nltk, NLPNet, networkx, Scikit-Learn, MLxtend	Remoção (stopwords), Tokenização	Classificação (análise de redes, treinamento e teste de dados), clusterização	Python	SentiElection Approach (SEA), Pagerank, RF, MLP, SVM, NB, DTC, KNN, OneRule (OneR), One Class (OC), K-Means (KM)	Acurácia, Betweenness, Closeness, Eigenvector, Katz, Authorities, Cluster Coeff Avg, coeficiente de agrupamento, Correlation, Transitivity, Density, validação cruzada

[112]	Dell Accessible Learning (DAL)	corpus construído/ coletado	Não descrito	Remoção (stopwords, acentos, números, pontuações), lowercasing, lemmatization, redução de dimensionalidade	recuperação de informação, incorporação de palavras, Visualização (t-SNE)	Python	TF-IDF, Machine Reading Comprehension (MRC), Question Answering (QA), BoW, glove, word embeddings, Word Mover Distance (WMD), electra, BERT, RoBERTa, ALBERT, GLUE, PCA, t-SNE	Similaridade de cosseno, Acurácia, F1-Score
[173]	Câmara dos Deputados	corpus já disponível	Nltk	Remoção (pontuações), lowercasing, stemmization, redução de dimensionalidade	recuperação de informação	Python	NoStem, Porter, RSLP Stemmer, Savoy (UniNE), Okapi BM25, BM25L, Friedman e Nemeyi test	Revocação, reduction per document (RP), unique terms per document (UTD)
[164]	Kollemata Project	corpus já disponível	hugging face, nltk, scikit-learn	Remoção (stopwords), Tokenização, stemmization, redução de dimensionalidade	categorização multirótulo, recuperação de informação, clusterização, classificação (treinamento e teste de dados)	Python	BERT, GLUE, MBERT, BERTimbau, RSLP Stemmer, K-means (KM), otimizador AdamW	Precisão, Revocação, F1-Score, Acurácia, Cross Entropy
[111]	BDCamões	corpus já disponível	Eli5	Pos tagging	Classificação (anotação morfossintática, Classificação de gênero literário, identificação de classes gramaticais)	Python	RF, LIWC, POS tags, grid search, Universal Dependencies (UD)	validação cruzada, F1-Score, Precisão, Revocação,

[133]	Carta de Serviços do Cidadão-Ceará	corpus construído/ coletado	RASA, Hugging face	Não descrito	Classificação (detecção de intenção, geração de sentenças, identificação de entidades, treinamento e teste de dados)	Não descrito	BERT, BerTimbau, M-BERT, Dual In-tent and Entity Transformer (DIET)	BLEU, F1-Score, Acurácia, Revocação
[78]	Taobeta, TED Talks	corpus já disponível	Kaggle, sacrebleu, Nltk, Texar, PyTorch, FastText	Remoção (XML)	Classificação (treinamento e teste de dados), reduzir out-of-vocabulary (OOV), identificação de padrões de erro, multidimensional evaluation	Colab	word embeddings, Back-Translation (BT), FastText, Neural Machine Translation (NMT), Byte Pair Encoding (BPE), CEFR scale, Multiple Fisher test, otimizador AdamW	BLEU
[103]	Não descrito	corpus construído/ coletado	Não descrito	tokenização, PoS tagging	Classificação (treinamento e teste de dados, análise de triplas, extração de características, identificação de classes gramaticais), Tradução, extração de informação	Não descrito	LR, DTC, KNN	validação cruzada, Precisão, Revocação, F1-Score, Acurácia
[184]	Sensacionalista	corpus construído/ coletado	NILC-Metrix, Nilc	Não descrito	detecção automática de notícias satíricas	Não descrito	NILC	Flesch Readability Ease Score (FRES), Type-Token Ratio (TTR)
[117]	Juizados Especiais Cíveis (JECs)	corpus já disponível	AntConc, Nilc, NILC-Metrix	Não descrito	cálculo das metainformações, classificação (rotulagem de dados, identificação de classes gramaticais)	Não descrito	Nilc, AntConc	Não descrito

[6]	Língua Geral Amazônica (LGA)	corpus construído/ coletado	Python Dictionary	remoção (caracteres especiais), tokenização	classificação (rotulagem de dados, identificação de classes gramaticais, anotação morfossintática, identificação de sentenças), cálculo das metainformações	Python	POS tags	Não descrito
[178]	Não descrito	corpus construído/ coletado	WebAnno	Não descrito	classificação (rotulagem de dados, identificação de classes gramaticais, anotação morfossintática), cálculo das metainformações	WebAnno	POS tags	Coeficiente Kappa de Cohen
[148]	Twitter, Fake.br	corpus já disponível	LC-Tool, freeoffice_pt-BR, spacy	Pos tagging	cálculo das metainformações, classificação (análise de sentimentos, identificação de classes gramaticais), extração de pistas linguísticas	Python	POS tags, LeIA (Léxico para Inferência Adaptada), cálculo da emotividade	Não descrito
[67]	MIMIC-III	corpus construído/ coletado	Não descrito	Remoção (outliers), normalização	Classificação (treinamento e teste de dados), Visualização (Matriz de confusão)	Não descrito	Gradient Boosting Machine (GBM), NB, SVM, MLP, RF, AdaBoost (AB), XGBoost, LDA, Randomized Search CV, StratifiedKFold	Pearson, AUC, Desvio Padrão, validação cruzada, F1-Score, Revocação, Especificidade

[183]	Câmara dos Deputados	corpus já disponível	Savoy	Remoção (stopwords, pontuações, acentos), stemmization, unigramas, bigramas	recuperação de informação	Não descrito	Okapi BM25	validação cruzada, Revocação
[84]	Notícias Agrícolas, CEPEA	corpus construído/ coletado	Não descrito	Remoção (acentos, stopwords, pontuações, caracteres especiais), redução de dimensionalidade, unigramas, stemmization, lowercasing	Classificação (treinamento e teste de dados, rotulagem de dados)	Python	BoW, BERT, BERTimbau, M-BERT, DistilBERT, NB, SVM, MLP, KNN, Binary BoW, TF, TF-IDF, PCA	similaridade de cosseno, F1-Score, Precisão, Revocação, Acurácia
[22]	Não descrito	corpus construído/ coletado	scikit-learn, NLTK, ELI5, FastText	Remoção (acentos, stopwords, pontuações, caracteres especiais), redução de dimensionalidade, lowercasing	Classificação (treinamento e teste de dados), Visualização (nuvem de palavras)	Python	Truncated SVD, BoW, TF-IDF, word embeddings, FastText, BERT, otimizador AdamW, Gradient Boosting Machine (GBM), NB, RF, XGBoost, Extremely Randomized Trees (ERT), LightGBM (LGBM), Binary BoW, Automated Valuation Models (AVM)	coeficiente de determinação, Erro Absoluto Médio (MAE), Erro Percentual Médio Absoluto (MAPE), Root Mean Square Error (RMSE), MdAPE

[14]	arXiv	corpus construído/ coletado	Não descrito	Não descrito	Classificação (dados validados manualmente, rotulagem de dados, treinamento e teste de dados, balanceamento de dados)	Python	BoW, Gradient Boosting Machine (GBM), RF, MLP, DTC, LR, AdaBoost (AB), smote, adasyn, TF-IDF, TF	validação cruzada, AUC, Precisão, Revocação
[135]	Não descrito	corpus construído/ coletado	Rasa, Spacy	Não descrito	Classificação (treinamento e teste de dados), Visualização (matriz de confusão)	Não descrito	RASA	Acurácia, F1-Score, Precisão
[85]	PMLB	corpus já disponível	scikit-learn	Não descrito	clusterização, classificação (balanceamento de dados, treinamento e teste de dados)	Python	K-Means (KM), CBDSCV, DOBSCV, DBSCV, StratifiedKFold, LR, DTC, Support Vector Machines (SVC), RF, RBF kernel	validação cruzada, acurácia, F1-Score
[100]	SBC OpenLib (SOL), Google Scholar, IBGE, facebook, twitter	corpus construído/ coletado	BeautifulSoup, scholarly, Cytoscape, Gephi	Remoção (caracteres especiais, espaços, quebras de linha), lowercasing	classificação (identificação de comunidades, rotulagem de dados)	Python	Cytoscape, Gephi	Levenshtein, similaridade, coeficiente de agrupamento
[136]	Twitter	corpus já disponível	não descrito	não descrito	modelagem da propagação de rumores	não descrito	RNN	disseminação, persistência, Cross Entropy

[147]	Twitch	corpus construído/ coletado	Twitch-Python, Zenodo, NLTK, Scikit-Learn, PyTorch	Não descrito	modelagem de tópicos, Visualização (wordcloud), classificação (detecção de discurso de ódio)	Python	NMF, BoW, TF-IDF, wordcloud, CNN, RNN, LSTM	Não descrito
[169]	Goodreads	corpus construído/ coletado	goodreads, disparityfilter	não descrito	Classificação (rotulagem de dados, extração de backbone, identificação de comunidades)	Python, R	Louvain, Multipartite Network Modeling, Quadratic Assignment Procedure (QAP), Cross-state Cultural Analysis	similaridade de cosseno, Coeficiente Kappa de Cohen
[156]	Twitter	corpus construído/ coletado	Microsoft Azure, tweepy	não descrito	Classificação (análise de sentimentos, rotulagem de dados, identificação de polaridade, rotulagem de dados, anotação de polaridade), Visualização (Diagrama de Venn, matriz de confusão)	não descrito	POS tags, crowdsourcing, Microsoft Azure	Alfa de Krippendorff, cross entropy
[54]	IAC, imdb, reddit	corpus construído/ coletado	PRAW, NLTK, Scikit-learn	remoção (links, caracteres especiais, espaços)	Classificação (identificação de sentenças, identificação de polaridade, análise de sentimentos, treinamento e teste de dados, balanceamento de dados), Visualização (matriz de confusão)	python	otimizador SGD, SVM	Não descrito

[179]	FakeNewsNet	corpus já disponível	KNIME, mySQL, spacy	tokenização, Pos tagging	Classificação (identificação de emoções, identificação de polaridade, identificação de classes gramaticais, balanceamento de dados, treinamento e teste de dados),	Python	KNIME, POS tags, LIWC, Gradient Boosting Machine (GBM), AdaBoost (AB), NB, SVM, KNN	acurácia, validação cruzada, precisão, revocação
[75]	Twitter	corpus já disponível	SymSpellpy, NLTK, Spacy	Remoção (stopwords, acentos, pontuações, espaços), stemmization, correção de textos, lowercasing	Classificação (treinamento e teste de dados, análise de sentimentos, classificação de sentimentos, identificação de polaridade)	Python	RSLP Stemmer, Snowball, SVM, Support Vector Machines (SVC), TF-IDF, LSTM	validação cruzada, F1-Score
[61]	Twitter	corpus construído/ coletado	Twitterscraper, spacy, gensim	Remoção (links, hashtags, caracteres especiais, espaços, pontuações, stopwords)	Classificação (identificação de entidades), Modelagem de tópicos, Visualização (wordcloud)	Python	LDA, NER, wordcloud	F1-Score, Revocação, Precisão
[51]	Twitter, Youtube	corpus construído/ coletado	Netspeak, scikit-learn	Remoção (stopwords), lowercasing, tokenização	Classificação (identificação de polaridade)	Python	SVM, NB, RF, TF-IDF, TF, teste de Wilcoxon	validação cruzada, F1-Score, teste qui-quadrado
[73]	Twitter	corpus construído/ coletado	ePOCS Twitter Crawler (eTC), Gephi, IRaMuTeQ	Não descrito	Classificação (Análise de sentimentos, análise temporal, análise lexical), Modelagem de tópicos, clusterização	R	Louvain, método Reinert, IRaMuTeQ	Não descrito
[168]	twitter, facebook	corpus construído/ coletado	CrowdTangle, Botometer	Não descrito	classificação (análise temporal, rotulagem de dados)	Python	Não descrito	Coeficiente Kappa de Cohen

[153]	Wikipédia	corpus construído/ coletado	não descrito	Remoção (stopwords, acentos, pontuações), tokenização	classificação (treinamento e teste de dados), Visualização (wordcloud)	Não descrito	TF, RF, wordcloud, tf-idf, NB	validação cruzada, F1-Score, Precisão, Revocação, PMI, teste qui-quadrado
[45]	twitter	corpus construído/ coletado	Scikit-learn	Remoção (stopwords, links, acentos, pontuações), stemmization	Classificação (Análise de sentimentos, treinamento e teste de dados, identificação de polaridade, rotulagem de dados)	Python	TF-IDF, LR, tash-pt	Acurácia, Precisão, F1-Score, Revocação, Fleiss Kappa
[171]	reddit	corpus construído/ coletado	Vader, ekphrasis, nltk, gensim	Remoção (stopwords, links, acentos, pontuações, espaços, caracteres especiais), stemmization, correção de textos	classificação (análise de sentimentos, identificação de polaridade), análise do tom emocional	Python	Vader, Thread, word embeddings, Word2Vec, MLP	Mean Squared Error (MSE)
[44]	Não descrito	corpus já disponível	Enelvo	correção de textos	classificação (extração de aspectos, treinamento e teste de dados, ontologia)	Não descrito	word embeddings, POS tags, word2vec	similaridade de cosseno, Precisão, F1-Score, revocação
[25]	G1, uol	corpus construído/ coletado	scikit-learn	Não descrito	Classificação (rotulagem de dados, treinamento e teste de dados)	Python	SVM, KNN, TF, grid search	Precisão, Revocação, F1-Score, validação cruzada

[119]	Twitter	corpus construído/ coletado	WordCloud, nltk, LIWC	Remoção (stopwords, pontuações, links), lowercasing	Visualização (wordcloud), Classificação (extração de aspectos, identificação de polaridade, análise de sentimentos)	Python	NER, WordCloud, LIWC	Não descrito
[65]	e-commerce, Consumidor.gov, Reclame-Aqui	corpus construído/ coletado	scikit-learn	Remoção (stopwords, pontuações, números, acentos, caracteres especiais, links)	Classificação (extração de características), Modelagem de tópicos	Python	LDA	Flesch Readability Ease Score (FRES), PMI, coerência
[124]	Twitter, PubMed	corpus construído/ coletado	Não descrito	Remoção (emojis, links, acentos)	Modelagem de tópicos	Python	LDA	KL-Divergence
[35]	StockTwits	corpus construído/ coletado	Não descrito	Limpeza dos dados	análise estática, análise temporal, análise de correlação	Não descrito	Simulação de Monte Carlo	Spearman, indice de sharpe
[93]	Twitter	corpus já disponível	Scikit-learn	Remoção (menções, links), lowercasing, tokenização	Classificação (treinamento e teste de dados, análise de sentimentos, identificação de polaridade, balanceamento de dados)	Python	word embeddings, word2vec, LR	acurácia, F1-Score, similaridade de cosseno

[49]	Sistema Judiciário	corpus construído/ coletado	Não descrito	Remoção (pontuações, números, outliers, stopwords), lemmatization	Clusterização, Classificação (extração de características)	Python	TF-IDF, word embeddings, word2vec, PCA, K-Means (KM), Agglomerative clustering, Spectral clustering	índice de Calinski-Harabasz, índice de Davies-Bouldin, coeficiente de silhueta, Precisão, Revocação, F1-Score
[83]	opendatasus	corpus construído/ coletado	Não descrito	Não descrito	classificação (treinamento e teste de dados, balanceamento de dados)	Não descrito	DTC, NB	AUC, validação cruzada
[60]	Não descrito	corpus já disponível	scikit-learn	Remoção (links, pontuações, símbolos, referências cruzadas), bigramas, trigramas	Classificação (classificação em língua indígena, extração de características, treinamento e teste de dados), tradução, clusterização, Visualização (t-SNE)	Python	LR, Support Vector Machines (SVC), NB, K-Means (KM), t-SNE, TF-IDF, StratifiedKFold	acurácia, F1-Score, índice de Davies-Bouldin, coeficiente de silhueta, validação cruzada
[116]	Sistema Judiciário	corpus construído/ coletado	Flair, spacy, tensorflow, keras, numpy, pandas, fasttext	Não descrito	recuperação de informação, classificação (treinamento e teste de dados)	Python	NER, BiLSTM-CRF, LSTM, CNN, CRF, word embeddings, optical character recognition (OCR), flair embeddings	F1-Score, Precisão, Revocação

[10]	SEFAZ-ES	corpus já disponível	scikit-learn	redução de dimensionalidade	classificação (treinamento e teste de dados, extração de características, rotulagem de dados)	Python	KNN, RF, SVM, Neural Network (NN), grid search	validação cruzada, F1-Score, Revocação, Precisão
[19]	Sistema Judiciário	corpus construído/ coletado	Flair	Não descrito	classificação (rotulagem de dados, treinamento e teste de dados)	Python	BiLSTM-CRF, LSTM, optical character recognition (OCR), word embeddings, flair embeddings, character embedding, Pooled Contextualized Embedding, Glove, NER	Revocação, Precisão, F1-Score
[137]	Não descrito	corpus já disponível	OpenNLP, CoreNLP, spacy	Remoção (stopwords), tokenização, segmentação de sentenças	recuperação de informação, classificação (treinamento e teste de dados)	Python	NER	validação cruzada, F1-Score, Revocação, Precisão
[72]	Não descrito	corpus já disponível	scikit-learn	remoção (outliers), redução de dimensionalidade, limpeza dos dados	classificação (balanceamento de dados, treinamento e teste de dados)	Python	smote, undersampling e oversampling, KNN, RF, SVM, NB, LR, DTC, Gradient Boosting Machine (GBM), XGBoost, Extremely Randomized Trees (ERT)	F1-Score, acurácia, precisão, Revocação, Especificidade

[46]	Twitter	corpus construído/ coletado	snscrape, gensim, scikit-learn	remoção (stopwords, links, emojis), stemmization, lowercasing	classificação (rotulagem de dados, extração de características, treinamento e teste de dados), visualização (matriz de confusão)	Python	TF-IDF, word embeddings, Paragraph Vector, BoW, SVM, NB, DTC, Neural Network (NN), MLP, RSLP Stemmer	validação cruzada, especificidade, Revocação, Precisão, F1-Score
[63]	Ouvidoria Geral da União	corpus construído/ coletado	Não descrito	Não descrito	Classificação (treinamento e teste de dados)	Python	NER, BERT, TF-IDF, RF	AUC, AUPRC, Precisão, Revocação, validação cruzada
[1]	UFPR	corpus construído/ coletado	Não descrito	Remoção (links, números, textos duplicados, caracteres especiais), correção de textos	Classificação (treinamento e teste de dados, extração de caracteristicas, rotulagem de dados)	Python	Bert, word2vec, glove, SVM, RF	F1-score
[48]	Twitter	corpus já disponível	nltk, spacy, Twikenizer	Tokenização, pos tagging	classificação (rotulagem de dados, identificação de classes gramaticais, anotação morfossintática, treinamento e teste de dados)	Python	POS tags, UDPipe, Universal Dependencies (UD)	F1-score, precisão, Revocação

[20]	Twitter	corpus construído/ coletado	Tweepy, Scikit-Learn	Remoção (stopwords), stemmization, normalização	classificação (rotulagem de dados, treinamento e teste de dados, balanceamento de dados)	Python	TF-IDF, smote, undersampling e oversampling, RF, SVM, NB, DTC, XGBoost, AdaBoost (AB), Grid Search, Shapiro-Wilk, T-Student, Wilcoxon	Coeficiente Kappa de Cohen, validação cruzada, precisão, revocação, F1-Score, Mean Decrease Gini (MDG)
[160]	Twitter	corpus construído/ coletado	nltk	remoção (textos duplicados, links, emojis, hashtags, pontuações, espaços, linhas em branco, stopwords, outliers), redução de dimensionalidade, tokenização, lowercasing	detecção de posicionamento, clusterização, modelagem de tópicos, classificação (rotulagem de dados)	Python	UMAP, bertopic, word embeddings, TF-IDF, c-TF-IDF, bertimbau, bert	Não descrito
[87]	GT-RDP Brasil	corpus já disponível	Pandas, scikit-learn	Não descrito	clusterização	Python	K-Means (KM)	coeficiente de silhueta, teste qui-quadrado

[91]	Não descrito	corpus construído/ coletado	sentence-transformers	Não descrito	clusterização, classificação (treinamento e teste de dados)	Python	DistilBERT, word embeddings, bert, DBERTML, Variational Autoencoder (VAE), Neural Network (NN), one-class learning (OCL), One-Class Support Vector Machine (OCSVM), Friedman e Nemeyi test	coeficiente de silhueta, densidade, Kullback-Leibler (KL), precisão, revocação, F1-Score, AUC, acurácia
[81]	Assembleia da República	corpus construído/ coletado	tensorflow, hugging face, scikit-learn, tune	remoção (stopwords), tokenização, lemmatization	classificação (rotulagem de dados, balanceamento de dados, treinamento e teste de dados)	Python	Bag-of-N-Grams (BoNG), BoW, Bertimbau, bert, StratifiedKFold, Dataset Resampling, Data Augmentation, RF, SVM, LR, word2vec, doc2vec, Wilcoxon	validação cruzada, F1-Score, Cross Entropy, brier score
[66]	Não descrito	corpus já disponível	LIWC	Não descrito	classificação (rotulagem de dados, treinamento e teste de dados)	Python	BoW, LIWC	Alfa de Krippendorff, similaridade de cosseno, Spearman, Pearson
[128]	wikipedia	corpus construído/ coletado	hugging face	remoção (outliers)	classificação (treinamento e teste de dados), sumarização	Python, Colab	TF-IDF, PTT5, longformer, T5	rouge, precisão, revocação, f1-score

[140]	Sistema Judiciário	corpus já disponível	gensim, pytorch, hugging face, nilc, CatBoost, regex	tokenização, unigramas, bigramas, quadrigramas, lowercasing, redução de dimensionalidade	classificação (treinamento e teste de dados, mascaramento de dados)	Python	Phraser, Word2Vec, FastText, Doc2Vec, bert, bertikal, bertimbau, word embeddings, Masked Language Model (MLM), PCA, cnn, nilc	f1-score, acurácia
[176]	Blue Amazon, Wikipedia	corpus construído/ coletado	BertViz	Não descrito	tradução, Classificação (rotulagem de dados)	Python	BERTimbau, bert, bertViz, Question Answering (QA)	F1-score, rouge, exact match
[165]	Twitter	corpus construído/ coletado	nltk	remoção (stopwords, pontuações)	classificação (rotulagem de dados, treinamento e teste de dados)	Python	BoW, tf-idf, SVM, NB, LR, MLP, Analysis of variance (ANOVA), Teste de Tukey, Wilcoxon	Coeficiente Kappa de Cohen, validação cruzada, f1-score, acurácia
[158]	Não descrito	corpus já disponível	Não descrito	Não descrito	clusterização	Python	K-Means (KM), Bisecting k-Means (Bk-Means), Gaussian Mixture Models (GMMs), NMF, LDA, Label Propagation (LP), Simple Ranking (Sim. Rank.), Ranking Clustering (Rank. Clus.), KNN, BoW	cross entropy, Purity, F1-Score, Precisão, Revocação

[120]	INEP	corpus construído/ coletado	DataBricks, Scikit-learn, Pyspark-ML, Pandas	discretização, padronização, normalização	classificação (treinamento e teste de dados)	Python	DataBricks, KNN, LR, DTC, RF, K-Means (KM)	validação cruzada, AUC, precisão, revocação, F1-score
[122]	ULB Machine Learning Group	corpus construído/ coletado	scikit-learn	Não descrito	classificação (balanceamento de dados, treinamento e teste de dados)	Python, Colab	LR, K-Means (KM), oversampling e undersampling, smote, tomek, KNN, RF, SVM, NB	validação cruzada, F1-Score
[27]	Food.com	corpus já disponível	Não descrito	Remoção (tags html, stopwords, pontuações), lowercasing, unigramas, bigramas	visualização (wordcloud), classificação (extração de características)	Python	LR, RF, NB, BoW, TF-IDF, Binary BoW	validação cruzada, acurácia, Precisão, Revocação, F1-Score
[26]	Amazon	corpus construído/ coletado	nltk	Remoção (tags html, espaços, caracteres especiais, links, hashtags), lowercasing, unigramas, bigramas	classificação (anotação de polaridade, identicação de polaridade, análise de sentimentos, extração de características), visualização (wordcloud)	Python	LR, RF, NB, SVM, DTC, BoW, TF-IDF, Binary BoW	acurácia, Precisão, Revocação, F1-Score
[28]	Food.com	corpus construído/ coletado	scikit-learn	Não descrito	classificação (extração de características)	Python	Document-Term Matrix (DTM), RF, DTC, BoW	Precisão, Revocação, F1-Score, validação cruzada

[131]	TripAdvisor, Twitter, Google Play	corpus construído/ coletado	scikit-learn, spacy	Não descrito	Classificação (análise de setimentos, identificação de polaridade, identificação de sentenças, rotulagem de dados)	Python	SVM, Gradient Boosting Trees (GBT), Recursive Neural Tensor Network (RNTN), NB, SentiWordNet, EmoticonsDS	Precisão, Revocação, F1-Score
[115]	Sistema Judiciário	corpus já disponível	TesseractOCR, tensorflow, keras, numpy, pandas, fasttext	Não descrito	Page Stream Segmentation (PSS), classificação (extração de características, treinamento e teste de dados), visualização (matriz de confusão)	Python	CNN, Fasttext, VGG16, CharCNN, otimizador AdamW, optical character recognition (OCR)	acurácia, Coeficiente Kappa de Cohen
[86]	DEPEN	corpus construído/ coletado	scikit-learn, imbalanced-learn	limpeza dos dados, normalização	Classificação (extração de características, balanceamento de dados)	Python	RF, SelectFromModel (SFM), Recursive Feature Elimination (RFE), Grid Search, LabelEncoder, RandomOverSampler, StratifiedKFold	precisão, revocação, acurácia, AUC, validação cruzada, F1-Score
[13]	Não descrito	corpus já disponível	scikit-learn	Remoção (stopwords, acentos, pontuações, caracteres especiais), stemmization, lowercasing, unigramas, bigramas, redução de dimensionalidade	classificação (treinamento e teste de dados)	Python	BoW, TF-IDF, bert, DistilBERT, RoBERTa, glue, SVM, MLP, NB, KNN, VSM, DBERTML, Binary BOW, Friedman e Nemeyi test, NLM, TF	precisão, revocação, F1-Score

[141]	USP, CAPES, UNICAMP, BDTD	corpus construído/ coletado	moses, GIZA++, KenLM	Não descrito	tradução	Não descrito	GloVe, Teoria dos Grafos	meteor, bleu
[127]	Yelp	corpus construído/ coletado	nltk, langdetect	tokenização	classificação (análise de sentimentos, extração de características, treinamento e teste de dados), clusterização, detecção de idioma, sumarização	Python	TF-IDF, NB, grid search	acurácia
[110]	spotify, Lyrics Genius	corpus construído/ coletado	spotipy, Google's, WordNetLemmatizer, nltk, scikit-learn	remoção (simbolos, números, stopwords), lemmatization	Modelagem de Tópicos, detecção de idioma, visualização (grafos)	Python	LDA, GloVe, word embeddings, Kamada-Kawai, Teoria dos Grafos	Similaridade de cosseno
[96]	TripAdvisor	corpus construído/ coletado	BeautifulSoup, Selenium, nltk, Polyglot, Scikit-learn	Remoção (stopwords, pontuações, caracteres especiais, números, emojis, acentos), lowercasing	classificação (análise de sentimentos, identificação de polaridade, identificação de gênero, rotulagem de dados)	Python	NMF, DTC, BoW, TF-IDF	Não descrito
[146]	BDTD	corpus construído/ coletado	regex, Gender Guesser, spreadsheet	Não descrito	classificação (identificação de gênero), visualização (wordcloud)	Colab, Jupyter notebook, python	Não descrito	Não descrito

[162]	HuffPost Brasil, Nexo Jornal, Sensacionalista, The piaui Herald	corpus construído/ coletado	BeautifulSoup	remoção (tags html)	visualização (word-cloud), classificação (identificação de irônia)	Python	Não descrito	Não descrito
[24]	não descrito	corpus já disponível	Não descrito	Não descrito	Dados extraídos de publicações dos anais do BRASNAM, BRACIS, STIL, ENIAC, PROPOR e IBERAMIA	Não descrito	word embeddings	Coeficiente Kappa de Cohen, similaridade de cosseno, precisão, revocação, f1-score, Matthews correlation coefficient (MCC), Phi coefficient
[70]	FCN, Fake.BR, FakeNewsNet	corpus construído/ coletado	Não descrito	Não descrito	Não descrito	Python	BoW, doc2vec, Positive and Unlabeled Learning (PUL), PU-LP, KNN, TF-IDF, LPHN, GNetMine (GNM)	Katz, F1-Score, Precisão, revocação, validação cruzada
[157]	não descrito	corpus já disponível	pytorch	Não descrito	classificação (treinamento e teste de dados)	Python	BiLSTM-CRF, LSTM, NER, word embeddings, fasttext, word2vec	precisão, revocação, f1-score

[170]	Câmara dos Deputados	corpus já disponível	Não descrito	Não descrito	clusterização, modelagem de tópicos, sumarização, classificação (treinamento e teste de dados)	Python	BERTopic, word embeddings, HDBSCAN, TF-IDF, bert, bertimbau, c-tf-idf, SimCSE	NPMI, coerência
[130]	Wikipedia	corpus já disponível	Não descrito	Remoção (espaços, textos duplicados), tokenização	recuperação de informação	Python	BERT, NER, BioBERTpt	Precisão, revocação, f1-score
[18]	TripAdvisor	corpus construído/ coletado	Não descrito	tokenização, redução de dimensionalidade	classificação (extração de características, identificação de sentenças, extração de aspectos, identificação de polaridade, análise de sentimentos)	Python	BERT, glove, Aspect Sentiment Triplet Extraction (ASTE), WordPiece, BiLSTM-CRF, LSTM, Shapiro-Wilk, Friedman e Nemeyi test, word embeddings	f1-score, revocação, validação cruzada
[106]	Twitter	corpus já disponível	nltk, tweetTokenizer, scikit-learn	Remoção (simbolos, caracteres especiais), tokenização, lowercasing	classificação (extração de características)	Python	BERT, Vocabulary Graph Convolutional Network (VGCN), word embeddings, BoW, SVM, TF, TF-IDF, BiLSTM-CRF	NPMI, f1-score, precisão, revocação

[62]	TripAdvisor	corpus já disponível	spacy, Scikit-Learn	Pos tagging, tokenização	classificação (extração de características, identificação de polaridade, identificação de classes gramaticais)	Python	SVM, Gradient Boosting Trees (GBT), RF, Maximum Entropy (ME), AutoML, AutoGluon, POS tags, Universal Dependencies (UD), Recursive Feature Elimination (RFE)	precisão, revocação, f1-score, acurácia, validação cruzada
[11]	ASSIN, ASSIN2	corpus já disponível	Não descrito	Não descrito	classificação (treinamento e teste de dados)	Python	BERT, SBERT, word embeddings, bertimbau	F1-score
[121]	Diário Oficial do DF	corpus já disponível	Pytorch	Não descrito	classificação (treinamento e teste de dados)	Colab, Python	NER, CNN, LSTM, BiLSTM-CRF	Maximum Normalized Log-Probability (MNLP)
[32]	Wikipedia, folha de sao paulo, Estadao, O Globo	corpus construído/ coletado	Não descrito	remoção (caracteres especiais, quebras de linha, outliers)	classificação (treinamento e teste de dados)	Python	question answering (QA), PTT5, T5, Okapi BM25	Exact Match, F1-Score, rouge

[40]	Wikipedia, ASSIN	corpus já disponível	NILC	Não descrito	classificação (treinamento e teste de dados), concatenação, tradução	Python	word embeddings, word2vec, fasttext, NILC, Neural Network (NN), otimizador adamw, Word Relatedness, Analogy Prediction, Sentence Textual Similarity (STS), NER, TF-IDF, LSTM	Pearson, Mean squared error (MSE), Spearman, similaridade de cosseno
[144]	Estadão Economia, UOL, G1, O Globo	corpus construído/ coletado	Hugging face, Excel, spacy, TensorFlow Keras	remoção (caracteres especiais), limpeza dos dados	Classificação (identificação de entidades, treinamento e teste de dados, rotulagem de dados, identificação de sentenças)	Python	Bert, otimizador adamw	Coeficiente de similaridade de Jaccard
[98]	Amazon	corpus construído/ coletado	spacy, Tensorflow, Keras, scikit-learn	Remoção (simbolos, caracteres especiais), lemmatization, unigramas	classificação (extração de características, treinamento e teste de dados)	Python	word embeddings, biLSTM-CRF, CNN, Friedman e Nemeyi test, otimizador adamw, glove, SVM, tf-idf	validação cruzada, Cross Entropy, Mean squared error (MSE), sigmoid, F1-Score

[151]	não descrito	corpus construído/ coletado	Não descrito	tokenização	recuperação de informação, classificação (treinamento e teste de dados, rotulagem de dados, identificação de sentenças), Open Information Extraction (OIE)	Python	bertimbau, bert, word embeddings, LSTM, BiLSTM-CRF, BIO, blstme	cross entropy, Coeficiente Kappa de Cohen, Precisão, F1-Score, Revocação, validação cruzada
[167]	e-commerce	corpus construído/ coletado	scikit-learn, Prodigy	remoção (outliers, caracteres especiais, stopwords, textos duplicados), tokenização, lowercasing	classificação (extração de características, treinamento e teste de dados, identificação de entidades), clusterização, recuperação de informação	Python	NER, MITIE, BERT, tf-idf, k-means (KM), bertimbau, LSTM	f1-score, Método elbow (cotovelo)
[59]	bíblia	corpus já disponível	pylinguistic, OpenNMT	Não descrito	tradução	Não descrito	Neural Machine Translation (NMT), statistical machine translation (SMT), RNN, bert, B-RNN, otimizador adamw, word embeddings	bleu, readability ease (FRE), SARI
[12]	TJMS	corpus construído/ coletado	PDFBox, nltk, Hugging Face, scikit-learn, keras	remoção (stopwords), tokenização	classificação (identificação de sentenças, treinamento e teste de dados), visualização (matriz de confusão)	Python	Bert, Tf-idf, NB, SVM, BoW, bertimbau, MLP, otimizador adamw	F1-score

[2]	tjce	corpus construído/ coletado	Scikit-learn	remoção (caracteres especiais, espaços)	concatenação, modelagem de tópicos, classificação (extração de características, treinamento e teste de dados), visualização (matriz de confusão)	Python	NER, optical character recognition (OCR), Teoria dos Grafos, bert, bertimbau, word embeddings, tf-idf, BoW, LDA, SVM, RF, XGBoost, otimizador adamw, Kruskal-Wallis test	F1-score
[102]	Folha de Sao Paulo	corpus já disponível	Não descrito	PoS tagging	classificação (rotulagem de dados, identificação de classes gramaticais, anotação morfossintática, treinamento e teste de dados)	Python	POS tags, UDPipe, Universal Dependencies (UD)	acurácia
[142]	Não descrito	corpus já disponível	OpenKE	Não descrito	Classificação (treinamento e teste de dados, identificação de sentenças, identificação de entidades)	Não descrito	NER, BiLSTM-CRF, BiGRU-CRF, LSTM, word embeddings, CNN, bert, glove, TransE method	f1-score, precisão, revocação
[47]	Não descrito	corpus já disponível	nltk, imbalanced-learn, scikit-learn	bigramas, PoS tagging	sumarização, classificação (extração de características, balanceamento de dados, identificação de classes gramaticais), visulaização (matriz de confusão)	Colab, Python	TextRank, TF-IDF, POS tags, RandomOverSampler, RF, NB	validação cruzada

[174]	ASSiN2	corpus já disponível	ftfy, WordPiece	remoção (tags html), tokenização	classificação (treinamento e teste de dados, identificação de entidades, identificação de classes gramaticais), concatenção	Python	Deeper transfer learning, bert, elmo, flair embeddings, bertimbau, NER, Recognizing Textual Entailment (RTE), Sentence Textual Similarity (STS), otimizador adamw, M-BERT, LSTM, BiLSTM-CRF, word embeddings	Mean Squared Error (MSE), F1-score, Pearson, acurácia, Precisão, Acurácia
[101]	twitter	corpus já disponível	transformers, keras, tensorflow	remoção (retweets), Pos tagging, lowercasing	classificação (treinamento e teste de dados, identificação de entidades, identificação de classes gramaticais)	Python	word embeddings, LSTM, Roberta, Distilbert, electra, NER, FastText, GPT2, HiCE, Comick, LSTM, BiLSTM-CRF, POS tags, bert	similaridade de cosseno, Spearman, F1-Score, acurácia
[105]	Twitter, stocktwits	corpus já disponível	Não descrito	remoção (stopwords, pontuações), tokenização, lowercasing	tradução, classificação (treinamento e teste de dados)	Python	Easy Data Augmentation (EDA), CNN, otimizador adamw, glove, word embeddings	Mean Squared Error (MSE), validação cruzada

[152]	twitter	corpus construído/ coletado	scikit-learn	remoção (números, caracteres especiais, links)	classificação (análise de sentimentos, identificação de polaridade, extração de características, treinamento e teste de dados), visualização (matriz de confusão)	Python	bert, glue, bilstm-crf, cnn, lstm, SVM, NB, KNN, tf-idf, word2vec, fasttext, M-BERT, BoW, RF, LR, Linear Regression, Support Vector Regressor (SVR), Elastic Net, Ridge, Lasso, Stochastic Gradient Descent (SGD), DTC, AdaBoost (AB), grid search	validação cruzada, Pearson, acurácia, precisão, revocação, F1-Score, mean squared error (MSE), erro absoluto medio (mae)
[149]	Wikipedia	corpus construído/ coletado	Wikiextractor, nltk	remoção (tags html, caracteres especiais)	segmentação de palavras	Python	bert, BiLSTM-CRF, cnn, lstm	f1-score, Revocação, Acurácia
[181]	Twitter	corpus construído/ coletado	tweepy, rdflib, beautifulSoup, spacy, scikit-learn	Não descrito	classificação (identificação de polaridade, análise de sentimentos, rotulagem de dados, extração de características, treinamento e teste de dados)	Python	DTC, SVM, NB, linearSVC	validação cruzada, acurácia, F1-Score, Precisão, Revocação
[55]	STF, STJ, TJSC, JusBrasil	corpus construído/ coletado	Não descrito	remoção (caracteres especiais, símbolos), redução de dimensionalidade, lowercasing, stemmization, lemmatization	classificação (extração de características, treinamento e teste de dados), clusterização, sumarização	Python	cnn, word embeddings, word2vec, glove, Optical Character Recognition (OCR), fasttext, PCA	acurácia, F1-Score, Precisão, Revocação

[58]	twitter	corpus construído/ coletado	Scikit-Learn	Remoção (pontuações), lowercasing	classificação (extração de características, rotulagem de dados)	Python	bert, lstm, word embeddings, word2vec, SVM, BoW, MLP, RNN, word2vec, StratifiedKFold, lstm, otimizador adamw	validação cruzada, acurácia, F1-Score, Precisão, Revocação
[143]	Não descrito	corpus já disponível	pytorch	Não descrito	classificação (extração de características, treinamento e teste de dados), tradução	Python	DistilGPT2, GPT2, DistilBERT, bert, cnn, lstm, NB, Easy Data Augmentation (EDA), Back-Translation (BT), Friedman e Nemeyi test	acurácia
[145]	Syrian-Lebanese Hospital (HSL)	corpus construído/ coletado	keras, tensorflow, nltk, Scikit-learn, Gensim	remoção (caracteres especiais, datas, horas, stopwords), lowercasing	classificação (extração de características, treinamento e teste de dados)	Python	cnn, rnn, LR, tf-idf, word2vec, BoW, grid search, CNN-Att, word embeddings, bilstm, gru, BiGRU	f1-score, revocação, precisão
[90]	twitter	corpus já disponível	Não descrito	remoção (caracteres especiais, tags html, acentos, menções, hashtags, pontuações), tokenização	classificação (análise de sentimentos, identificação de polaridade)	Python	Random Walk in Feature-Sample Networks (RWFSN), BoW, KNN	Coeficiente de similaridade de Jaccard, acurácia, f1-score, AUC
[166]	TJMG	corpus construído/ coletado	nltk, Scikit-learn, tensorflow	remoção (stopwords)	classificação (extração de características, treinamento e teste de dados)	Python	Glove, Neural Network (NN), Randomsearch, NB, SVM, AdaBoost (AB), tf-idf, cnn	F1-Score, Precisão, Revocação, validação cruzada

[134]	twitter	corpus já disponível	Não descrito	unigramas	classificação (rotulagem de dados, identificação de polaridade, treinamento e teste de dados)	Python	BiLSTM, LSTM, Analysis of variance (ANOVA), tf-idf	validação cruzada
[23]	Wikipedia, TCU	corpus já disponível	spacy, seqeval, Transformers, AllenNLP	remoção (tags html, links, pontuações)	classificação (identificação de entidades, treinamento e teste de dados)	Python	NER, bert, elmo, M-BERT	F1-score
[159]	ASSIN, ASSIN2	corpus já disponível	nltk, nlppyport, scikit-learn	tokenização, pos tagging, lemmatization, redução de dimensionalidade	Classificação (extração de características, identificação de classes gramaticais, identificação de entidades, treinamento e teste de dados)	não descrito	Semantic Textual Similarity (STS), POS tags, Support Vector Regressor (SVR), word embeddings, word2vec, glove, nilc, fasttext, NER	Mean Squared Error (MSE), Pearson, Coeficiente de similaridade de Jaccard, overlap, dice
[126]	Fake.br	corpus já disponível	NLTK, keras	remoção (stopwords), lowercasing, tokenização	sumarização, classificação (treinamento e teste de dados)	Python	BiGRU, GRU, MLP, word embeddings, glove, otimizador adamw	cross entropy, validação cruzada, F1-score, acurácia, precisão, revocação
[79]	E-mail, Banco do Brasil	corpus construído/ coletado	nltk, Scikit-learn, Linear Kernel	remoção (links, números, datas, tags html, caracteres especiais, stopwords, ASCII)	classificação (treinamento e teste de dados, extração de características)	Python	SVM, tf-idf, MLP, NB, KNN, grid search	F1-Score, validação cruzada, precisão, revocação

[82]	não descrito	corpus já disponível	scikit-learn, linear kernel	não descrito	classificação (identificação de similaridade)	Python	bert, word embeddings, SVM, MLP, DTC, Platt, sigmoid	F1-score, acurácia, precisão, revocação, Mean Squared Error (MSE), Pearson, Spearman
[76]	Twitter	corpus já disponível	Tweepy, scikit-learn, keras	remoção (menções, hashtags, stopwords), bigramas, trigramas, tokenização	classificação (treinamento e teste de dados, extração de características)	Python	NB, SVM, RF, TF-IDF, grid search, LSTM, BiLSTM	F1-Score, validação cruzada, precisão, revocação
[154]	LBSN, DataViva, TripAdvisor	corpus construído/ coletado	Surprise	não descrito	classificação (análise de sentimentos, identificação de polaridade)	não descrito	OpLexicon 3.0, SentiLex, Wilcoxon, KNN, SVD, grid search	validação cruzada, erro absoluto medio (mae), Root Mean Squared Error (RMSE)
[9]	ASSIN	corpus já disponível	nltk, linear kernel	tokenização, redução de dimensionalidade	classificação (extração de características, balanceamento de dados, treinamento e teste de dados)	Python	word embeddings, word2vec, fasttext, glove, Word Mover Distance (WMD), Abstract Meaning Representation (AMR), SVM, NB, DTC, Neural Network (NN), LR, SMOTE, undersampling e oversampling, baseline method	similaridade de cosseno, Smooth Inverse Frequency (SIF), F1-score, precisão, revocação

[150]	ASSIN	corpus já disponível	AllenNLP	tokenização	análise semântica, classificação (identificação de similaridade, treinamento e teste de dados), normalização	Python	elmo, fasttext, word embeddings, BiLSTM, LSTM, cnn, bert, M-BERT, word2vec	Pearson, Mean Squared Error (MSE), similaridade de cosseno
[39]	Sicredi	corpus construído/ coletado	nltk, Cogroo, ArDoq	tokenização	extração de informação, classificação (identificação de entidades, treinamento e teste de dados)	Python	NER, BiLSTM-CRF, word Embeddings, Flair Embeddings, Word2Vec, LSTM, POS tags	validação cruzada
[21]	buscapé, Amazon, Skoob	corpus construído/ coletado	nltk, scikit-learn, gensim, keras	Remoção (stopwords, pontuações, números, caracteres especiais)	classificação (rotulagem de dados, classificação subjetiva)	Python	word embeddings, Lexicon-Based Method, Graph-Based Method, PageRank, SVM, NB, Neural network (NN)	Eigenvector Centrality, Katz Index, validação cruzada, precisão, acurácia, revocação, f1-score
[123]	G1	corpus construído/ coletado	não descrito	pos tagging	classificação (treinamento e teste de dados, identificação de classes gramaticais)	não descrito	LR, POS tags, DAGGER, CRF, SEARN	validação cruzada, Precisão, REvocação, F1-Score
[56]	Não descrito	corpus já disponível	não descrito	pos tagging	extração de informação, classificação (identificação de classes gramaticais)	Python	glove, word embeddings, fasttext, Universal Dependencies (UD), BiLSTM, POS tags, MLP	UPOS, LAS, UAS, CLAS, BLEX, MLAS
[30]	ASSIN	corpus já disponível	Google Translate	não descrito	tradução, classificação (treinamento e teste de dados), identificação de inferência	Python	bert, sota	rouge, bleu, Acurácia, F1-Score

[69]	ASSIN, ASSIN2, Google News	corpus já disponível	não descrito	não descrito	classificação (identificação de similaridade, extração de características)	Python	word embeddings, Semantic Textual Similarity (STS), Siamese Neural Network (SNN), LSTM, word2vec	Similaridade de cosseno, Coeficiente de similaridade de Jaccard, Dice, Pearson, Mean Squared Error (MSE)
[5]	UOL, Folha de Sao Paulo	corpus já disponível	pytorch	não descrito	Modelagem de tópicos, Classificação (identificação de inferência)	Python	Language Models (LMs), Recognizing Textual Entailment (RTE), XLMR Transformer, RoBERTa, Bertopic, bert, M-BERT, StratifiedKFold	validação cruzada, precisão, REvocação, F1-Score
[155]	Twitter	corpus construído/ coletado	não descrito	não descrito	Modelagem de tópicos, análise semântica, classificação (extração de características)	Python	semantic role labeling (SRL), AllenNLP, bert, LDA	Toxicity Score, Identity Attack, Insult, Profanity, Threat
[163]	Domínio Público, Projecto Adamastor, BLPL	corpus já disponível	Translate tool, nltk, scikit-learn	remoção (links, emojis, stopwords, números, pontuações), tokenização, redução de dimensionalidade, lowercasing	classificação (rotulagem de dados, treinamento e teste de dados), tradução, Visualização (matriz de confusão)	Python	TF-IDF, LSA, SVD, NB, KNN, RF, Stochastic Gradient Descent (SGD), DTC, AdaBoost (AB)	Acurácia, Precisão, Revocação, F1-Score

[50]	Não descrito	corpus já disponível	FLAIR	não descrito	classificação (identificação de entidades, treinamento e teste de dados)	Python	word embeddings, word2vec, NER, Glove, Bert, elmo, Flair embeddings, BiLSTM-CRF, Allennlp	Precisão, Revocação, F1-Score
[175]	buscapé, olist, b2w	corpus já disponível	AutoTokenizer, hugging face	não descrito	classificação (análise de sentimentos)	Python	bert, bertimbau, TF-IDF, M-BERT, otimizador adamw, LR, TF-IDF	AUC
[186]	Não descrito	corpus já disponível	Transformers, seqeval	tokenização	classificação (identificação de entidades)	Python	bertimbau, bert, NER, otimizador adamw, Proof of Concept (POC), Elmo, LSTM, CRF, M-BERT	F1-Score
[3]	TJCE, CNJ	corpus construído/ coletado	spacy	remoção (Ascii), pos tagging, tokenização, lowercasing, lemmatization, redução de dimensionalidade	classificação (identificação de entidades, treinamento e teste de dados, classifying lawsuits), Modelagem de tópicos, clusterização	Python	NER, CRF, bertopic, bert, Sentence Transformer, umap, HDBSCAN, TF-IDF, c-TF-IDF, LDA, BoW, otimizador adamw, XGBoost	PWI, F1-Score

[31]	Não descrito	corpus já disponível	pytorch, spacy	remoção (textos duplicados), tokenização	extração de informação, classificação (extração de características, identificação de sentenças, treinamento e teste de dados)	Python	bert, glove, flair embeddings, POS tags, Dependency tree categories, semantic role labeling (SRL), CRF, LSTM, AllenNLP, bertimbau, XTransfomer, BiLSTM, SRU++	validação cruzada, F1-score, precisão, revocação, AUC
[33]	Federal Official Gazette	corpus construído/ coletado	não descrito	não descrito	classificação (rotulagem de dados, treinamento e teste de dados)	Python	Bert, LSTM, NB, BiLSTM, bertimbau, StratifiedKFold, T5, Pegasus, grid search	validação cruzada, MCC, Acurácia, F1-Score
[34]	Câmara dos Deputados, Twitter	corpus construído/ coletado	linear kernel	não descrito	visualização (wordcloud), classificação (extração de características, treinamento e teste de dados)	Python	wordcloud, TF-IDF, SVM, LR, MLP, grid search, BiLSTM, word2vec, word embeddings	validação cruzada, Precisão, Revocação, F1-Score, acurácia
[108]	Não descrito	corpus já disponível	keras, tensorflow, hugging face, Wordpiece	tokenização	classificação (identificação de entidades, BIO anotação, treinamento e teste de dados)	Python	bert, NER, Linguakit NER, AllenNLP, DBPedia , otimizador adamw	Precisão, Revocação, F1-Score
[4]	Câmara dos Deputados	corpus construído/ coletado	INCEpTION	pos tagging	classificação (identificação de entidades, rotulagem de dados, treinamento e teste de dados), Resolução de correferências	Python	NER, POS tags, CRF, Hidden Markov Model (HMM), Glove, BiLSTM-CRF	Coeficiente Kappa de Cohen, validação cruzada, acurácia, Precisão, Revocação, F1-Score

[38]	Boatos.org, Uol, Aos Fatos	corpus construído/ coletado	googlesearch, spacy	lowercasing	classificação (extração de características, treinamento e teste de dados, balanceamento de dados)	não descrito	word2vec, lstm	Similaridade de cosseno, validação cruzada, acurácia, Precisão, Revocação, F1-Score
[89]	Não descrito	corpus construído/ coletado	Transformers	não descrito	classificação (identificação de gênero)	Python	Target Syntactic Evaluation (TSE), M-BERT, bert, Bertinho, bertimbau	acurácia
[88]	Projeto comprova, E-farsas, Estadao Verifica, Boatos.org, AFP checamos	corpus construído/ coletado	spacy, nltk, Scikit-learn, OPFython, TensorFlow, keras	remoção (pontuações, caracteres especiais, links, stopwords), Truncation, Standardization of terms, lowercasing, lemmatization	classificação (extração de características)	Python	BoW, fasttext, NB, RF, SVM, cnn, MLP, otimizador adamw	Cross Entropy, precisão, revocação, f1-score, acurácia, validação cruzada
[185]	Twitter	corpus construído/ coletado	Transformers, hugging face	remoção (links, menções, hashtags)	classificação (Stance Detection, análise de sentimentos, rotulagem de dados, treinamento e teste de dados)	Python	Node2Vec, bert	Similaridade de cosseno, Fleiss Kappa, precisão, revocação, f1-score, acurácia

APÊNDICE C – LISTA DOS ESTUDOS PRIMÁRIOS

Lista de estudos primários

[1] Thiago Abdo and Fabiano Silva. Iterative machine learning applied to annotation of text datasets. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 374–385. SBC, 2021.

[2] André Aguiar, Raquel Silveira, Vládia Pinheiro, Vasco Furtado, and João Araújo Neto. Text classification in legal documents extracted from lawsuits in brazilian courts. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 586–600. Springer, 2021.

[3] André Aguiar, Raquel Silveira, Vasco Furtado, Vládia Pinheiro, and João A Monteiro Neto. Using topic modeling in classification of brazilian lawsuits. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 233–242. Springer, 2022.

[4] Hidelberg O Albuquerque, Rosimeire Costa, Gabriel Silvestre, Ellen Souza, Nádia FF da Silva, Douglas Vitório, Gyovana Moriyama, Lucas Martins, Luiza Soezima, Augusto Nunes, et al. Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 3–14. Springer, 2022.

[5] Alexandre Alcoforado, Thomas Palmeira Ferraz, Rodrigo Gerber, Enzo Bustos, André Seidel Oliveira, Bruno Miguel Veloso, Fabio Levy Siqueira, and Anna Helena Reali Costa. Zeroberto: Leveraging zero-shot text classification by topic modeling. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 125–136. Springer, 2022.

[6] Dominick M Alexandre, Juliana L Gurgel, and Leonel F de A Araripe. Compilação de um corpus etiquetado da língua geral amazônica. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 427–431. SBC, 2021.

[7] Antônio PS Alves, Lucas G da Silva Félix, Carlos MG Barbosa, Vinícius da Fonseca Vieira, and Carolina Ribeiro Xavier. Tiradentes no tripadvisor-o que se fala sobre essa simpática cidade histórica? In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 145–156. SBC, 2022.

[8] Otávio Alves, Taciana Pontual Falcao, George Valença, and Ermeson Andrade. Brimo: uma ferramenta para análise de sentimentos. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 97–108. SBC, 2022.

[9] Rafael Torres Anchiêta and Thiago Alexandre Salgueiro Pardo. Exploring the potentiality of semantic features for paraphrase detection. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings*, pages 228–238. Springer, 2020.

[10] Joao Paulo A Andrade, Leonardo S Paulucio, Thiago M Paixao, Rodrigo F Berriel, Teresa Cristina Janes Carneiro, Raphael V Carneiro, Alberto F De Souza, Claudine Badue, and Thiago Oliveira-Santos. A machine learning-based system for financial fraud detection. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 165–176. SBC, 2021.

[11] José E Andrade Junior, Jonathan Cardoso-Silva, and Leonardo CT Bezerra. Comparing contextual embeddings for semantic textual similarity in portuguese. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 389–404. Springer, 2021.

[12] Roberto Aragy, Eraldo Rezende Fernandes, and Edson Norberto Caceres. Rhetorical role identification for portuguese legal documents. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 557–571. Springer, 2021.

[13] Adailton Araujo, Marcos Golo, Breno Viana, Felipe Sanches, Roseli Romero, and Ricardo Marcacini. From bag-of-words to pre-trained neural language models: Improving automatic classification of app reviews for requirements engineering. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 378–389. SBC, 2020.

[14] Pedro HC Avelar, Rafael Baldasso Audibert, and Luís C Lamb. Measuring ethics in ai with ai: A methodology and dataset construction. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part I*, pages 370–384. Springer, 2022.

[15] Joao Gabriel Melo Barbirato, Livy Real, and Helena de Medeiros Caseli. Relation extraction in structured and unstructured data: a comparative investigation on smartphone titles in the e-commerce domain. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 101–110. SBC, 2021.

[16] André Barbosa and Alan Godoy. Augmenting customer support with an nlp-based receptionist. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 133–142. SBC, 2021.

[17] Carlos MG Barbosa, Lucas G da S Félix, Antônio Pedro S Alves, Carolina Ribeiro Xavier, and Vinícius da Fonseca Vieira. Uso de urls para caracterização de comunidades em redes sociais online. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 25–36. SBC, 2022.

[18] José Meléndez Barros and Glauber De Bona. A deep learning approach for aspect sentiment triplet extraction in portuguese. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 343–358. Springer, 2021.

[19] Hyan HN Batista, André CA Nascimento, Rafael Ferreira Melo, Péricles BC Miranda, Isabel WS Maldonado, and José LM Coelho Filho. A comparative analysis of text embedding approach to extract named entities in portuguese legal documents. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 221–232. SBC, 2021.

[20] Anísio Pereira Batista Filho, Débora da Conceiçao Araújo, Máverick André Dionísio Ferreira, and Paulo Salgado Gomes de Mattos Neto. Fake news detection about covid-19 in the portuguese language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 492–503. SBC, 2021.

[21] Luana Balador Belisário, Luiz Gabriel Ferreira, and Thiago Alexandre Salgueiro Pardo. Evaluating methods of different paradigms for subjectivity classification in portuguese. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 261–269. Springer, 2020.

[22] Luís Fernando Bittencourt, Otávio Parraga, Duncan D Ruiz, Isabel H Manssour, Soraia Raupp Musse, and Rodrigo C Barros. Leveraging textual descriptions for house price valuation. In *Intelligent Systems: 11th Brazilian*

Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part I, pages 355–369. Springer, 2022.

[23] Luiz Henrique Bonifacio, Paulo Arantes Vilela, Gustavo Rocha Lobato, and Eraldo Rezende Fernandes. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 648–662. Springer, 2020.

[24] Avner Dal Bosco, Renata Vieira, Bruna Zanotto, and Ana Paula Beck da Silva Etges. Ontology based classification of electronic health records to support value-based health care. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part I*, pages 359–371. Springer, 2021.

[25] M Luisa P Braga, Fabiola G Nakamura, and Eduardo F Nakamura. Criaçao e caracterizaçao de um corpus de discurso sexistas em português. *iSys-Brazilian Journal of Information Systems*, 14(2):79–95, 2021.

[26] Larissa Britto and Luciano Pacífico. Uma abordagem de classificação de sentimentos em revisões de livros em português brasileiro usando diferentes métodos de extração de características. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 116–127. SBC, 2020.

[27] Larissa Britto, Luciano Pacífico, and Teresa Ludermir. Inferência automática do nível de dificuldade em receitas culinárias usando técnicas de processamento de linguagem natural. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 104–115. SBC, 2020.

[28] Larissa Britto, Luciano Pacífico, Emilia Oliveira, and Teresa Ludermir. A cooking recipe multi-label classification approach for food restriction identification. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 246–257. SBC, 2020.

[29] Larissa FS Britto, Luis AS Pessoa, and Silvania CC Agostinho. Cross-domain sentiment analysis in portuguese using bert. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 61–72. SBC, 2022.

[30] Marco Antonio Sobrevilla Cabezudo, Marcio Inácio, Ana Carolina Rodrigues, Edresson Casanova, and Rogério Figueredo de Sousa. Natural language inference for portuguese using bert and multilingual information. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 346–356. Springer, 2020.

[31] Bruno Cabral, Marlo Souza, and Daniela Barreiro Claro. Portnoie: A neural framework for open information extraction for the portuguese language. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 243–255. Springer, 2022.

[32] Flávio Nakasato Cação, Marcos Menon José, André Seidel Oliveira, Stefano Spindola, Anna Helena Reali Costa, and Fabio Gagliardi Cozman. Deepagé: answering questions in portuguese about the brazilian environment. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 419–433. Springer, 2021.

[33] Flávio Nakasato Cação, Anna Helena Reali Costa, Natalie Unterstell, Liuca Yonaha, Taciana Stec, and Fábio Ishisaki. Tracking environmental policy changes in the brazilian federal official gazette. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 256–266. Springer, 2022.

[34] Danielle Caled and Mário J Silva. A transfer learning analysis of political leaning classification in cross-domain content. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 267–277. Springer, 2022.

[35] Ozório JS Camargos, Adriano CM Pereira, and Michele A Brandão. Em qual portfólio investir? análise e seleção a partir de dados do stocktwits. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 184–189. SBC, 2020.

[36] Leonardo Capellaro and Helena de Medeiros Caseli. Análise de polaridade e de tópicos em tweets no domínio da política no brasil. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 47–55. SBC, 2021.

[37] Vinicius Casani, Alinne C Correa Souza, Rafael G Mantovani, and Francisco Carlos M Souza. Dp-symptom-identifier: uma estratégia para classificar sintomas de depressão utilizando um conjunto de dados textuais na língua portuguesa. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 153–161. SBC, 2021.

[38] Anderson Cordeiro Charles, Livia Ruback, and Jonice Oliveira. Fakepedia corpus: A flexible fake news corpus in portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 37–45. Springer, 2022.

[39] Sandra Collovini, Patricia Nunes Gonçalves, Guilherme Cavalheiro, Joaquim Santos, and Renata Vieira. Relation extraction for competitive intelligence. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 249–258. Springer, 2020.

[40] Bernardo Scapini Consoli and Renata Vieira. Enriching portuguese word embeddings with visual information. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 434–448. Springer, 2021.

[41] Fábio Cordeiro, Ricardo de Andrade Lira Rabelo, and Raimundo Santos Moura. Classification of irregularity communications in public ombudsmen using supervised learning algorithms. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 704–715. SBC, 2022.

[42] Diogo Cortiz, Jefferson O Silva, Newton Calegari, Ana Luísa Freitas, Ana Angélica Soares, Carolina Botelho, Gabriel Gaudencio Rêgo, Waldir Sampaio, and Paulo Sergio Boggio. A weakly supervised dataset of fine-grained emotions in portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 73–81. SBC, 2021.

[43] Geandreson de S Costa, Danielle CC Couto, Antonio FL Jacob Junior, and Fábio MF Lobato. Feminismo e redes sociais online: uma análise de tweets sobre o dia internacional da mulher. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 169–180. SBC, 2022.

[44] Raul Wagner Martins Costa and Thiago Alexandre Salgueiro Pardo. Métodos baseados em léxico para extração de aspectos de opiniões em português. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 61–72. SBC, 2020.

[45] Vinicius Matheus de Medeiros Silva Coutinho and Yuri Malheiros. Detecçao de mensagens homofóbicas em português no twitter usando análise de sentimentos. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 1–12. SBC, 2020.

[46] Ramon Souza da Cruz, Gilberto Nunes Neto, and Rafael Torres Anchiêta. Detecting misinformation in tweets related to covid-19. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 280–289. SBC, 2021.

[47] Bianca da Rocha Bartolomei and Isabela Neves Drummond. Authorship attribution of brazilian literary texts through machine learning techniques. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 389–402. Springer, 2020.

[48] Emanuel Huber da Silva, Thiago Alexandre Salgueiro Pardo, Norton Trevisan Roman, and Ariani Di Fellipo. Universal dependencies for tweets in brazilian portuguese: Tokenization and part of speech tagging. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 434–445. SBC, 2021.

[49] Ingrid LA da Silva, Rafael Ferreira Mello, Péricles BC Miranda, André CA Nascimento, Isabel WS Maldonado, and José LM Coelho Filho. Assessment of text clustering approaches for legal documents. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 37–48. SBC, 2021.

[50] Messias Gomes da Silva and Hilário Tomaz Alves de Oliveira. Combining word embeddings for portuguese named entity recognition. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 198–208. Springer, 2022.

[51] Rodolpho da Silva Nascimento, Gabriel dos Santos, Flavio Carvalho, and Gustavo Guedes. Avaliando contribuições na substituição de termos informais em classificação de texto de redes sociais com netspeak-br. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 181–186. SBC, 2021.

[52] Alan da Silva Romualdo, Livy Real, and Helena de Medeiros Caseli. Classificação multimodal para detecção de produtos proibidos em uma plataforma marketplace. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 111–120. SBC, 2021.

[53] Alan da Silva Romualdo, Livy Real, and Helena de Medeiros Caseli. Measuring brazilian portuguese product titles similarity using embeddings. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 121–132. SBC, 2021.

[54] João Pedro da Silva Sousa, Rodrigo Costa Uchoa do Nascimento, Renata Mendes de Araujo, and Orlando Bisacchi Coelho. Não se perca no debate! mineração de argumentação em redes sociais. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 139–150. SBC, 2021.

[55] Thiago Raulino Dal Pont, Isabela Cristina Sabo, Jomi Fred Hübner, and Aires José Rover. Impact of text specificity and size on word embeddings performance: An empirical evaluation in brazilian legal domain. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 521–535. Springer, 2020.

[56] Juliana C Carvalho de Araújo, Cláudia Freitas, Marco Aurélio C Pacheco, and Leonardo A Forero-Mendoza. An investigation of pre-trained embeddings in dependency parsing. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 281–290. Springer, 2020.

[57] Karhyne S Padilha de Assis, Camila das Merces Silva, Janaína da Silva Leite, Wellington Araujo Nogueira, Kenji Nose Filho, André K Takahata, and Margarethe Steinberger-Elias. Lexicalidade biomédica e sua mensuração em um corpus sobre covid-19 em língua portuguesa. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 39–46. SBC, 2021.

[58] Vinícios Faustino de Carvalho, Bianca Giacon, Carlos Nascimento, and Bruno Magalhães Nogueira. Machine learning for suicidal ideation identification on twitter for the portuguese language. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 536–550. Springer, 2020.

[59] Tiago B de Lima, André CA Nascimento, George Valença, Pericles Miranda, Rafael Ferreira Mello, and Tapas Si. Portuguese neural text simplification using machine translation. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 542–556. Springer, 2021.

[60] Tiago Barbosa de Lima, André CA Nascimento, Pericles Miranda, and Rafael Ferreira Mello. Analysis of a brazilian indigenous corpus using machine learning methods. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 118–129. SBC, 2021.

[61] Tiago de Melo. Análise exploratória das duvidas sobre a covid-19 publicadas no twitter. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 175–180. SBC, 2021.

[62] Miguel de Oliveira and Tiago de Melo. An empirical study of text features for identifying subjective sentences in portuguese. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 374–388. Springer, 2021.

[63] Eduardo de Paiva and Fernando Sola Pereira. Extraction and enrichment of features to improve complaint text classification performance. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 338–349. SBC, 2021.

[64] Roney Lira de Sales Santos and Thiago Alexandre Salgueiro Pardo. Structural characterization and graph-based detection of fake news in portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 199–208. SBC, 2021.

[65] Gustavo Nogueira de Sousa, Isabelle Guimaraes, Antonio FL Jacob Jr, Fábio MF Lobato, Sao Luıs-Maranhao-Brasil, and Oeste do Pará UFOPA Santarém-PA-Brasil. Análise comparativa das principais plataformas de reclamaç oes online: implicaç oes para análise de mıdia social em negócios. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 154–165. SBC, 2020.

[66] Rogério Figueredo De Sousa and Thiago Alexandre Salgueiro Pardo. The challenges of modeling and predicting online review helpfulness. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 727–738. SBC, 2021.

[67] Alexandre Renato Rodrigues de Souza, Fabrício Neitzke Ferreira, Rodrigo Blanke Lambrecht, Leonardo Costa Reichow, Helida Salles Santos, Renata Hax Sander Reiser, and Adenauer Correa Yamin. Mortality risk evaluation: A proposal for intensive care units patients exploring machine learning methods. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part I*, pages 1–14. Springer, 2022.

[68] Elvis de Souza, Aline Silveira, Tatiana Cavalcanti, Maria Clara Castro, and Cláudia Freitas. Petrogold–corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38. SBC, 2021.

[69] João Vitor Andrioli de Souza, Lucas Emanuel Silva E Oliveira, Yohan Bonescki Gumiel, Deborah Ribeiro Carvalho, and Claudia Maria Cabral Moro. Exploiting siamese neural networks on short text similarity tasks for multiple domains and languages. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 357–367. Springer, 2020.

[70] Mariana C de Souza, Bruno M Nogueira, Rafael G Rossi, Ricardo M Marcacini, and Solange O Rezende. A heterogeneous network-based positive and unlabeled learning approach to detect fake news. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 3–18. Springer, 2021.

[71] Vanessa de Souza Câmara and Tiago de Melo. Estudo de método de extraçao de aspectos para português do brasil baseado em regras. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 192–203. SBC, 2022.

[72] Jessica Almeida dos Santos and Lilian Berton. Applying machine learning to assist the diagnosis of covid-19 from blood and urine exams. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 256–267. SBC, 2021.

[73] Joao Guilherme Bastos dos Santos, Arthur Ituassu, Sérgio Lifschitz, Thayane Guimaraes, Diego Cerqueira, Debora Albu, Redson Fernando, Julia Hellen Ferreira, and Maria Luiza Mondelli. Das milícias digitais ao comportamento coordenado: métodos interdisciplinares de análise e identificação de bots nas eleições brasileiras. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 187–192. SBC, 2021.

[74] Vinícius S dos Santos, Felipe da R Henriques, and Gustavo Guedes. O discurso de ódio homofóbico no twitter a partir da análise de dados. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 109–120. SBC, 2022.

[75] Laerte dos Santos Cardozo and Larissa Astrogildo de Freitas. Análise de sentimentos: Avaliando o desempenho de pré-processamento e de algoritmos de aprendizagem de máquina sobre o dataset tweetsentbr. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 169–174. SBC, 2021.

[76] Luis Duarte, Luís Macedo, and Hugo Gonçalo Oliveira. Emoji prediction for portuguese. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 174–183. Springer, 2020.

[77] Carolina Eberhart, Luciano Ignaczak, and Márcio Garcia Martins. Text mining for cyberbullying detection: a brazilian portuguese evaluation. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 92–100. SBC, 2021.

[78] Arthur T Estrella and João BO Souza Filho. Tackling neural machine translation in low-resource settings: a portuguese case study. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 275–282. SBC, 2021.

[79] Rafael Faria de Azevedo, Rafael Rodrigues Pereira de Araujo, Rodrigo Guimarães Araújo, Régis Moreira Bittencourt, Rafael Ferreira Alves da Silva, Gabriel de Melo Vaz Nogueira, Thiago Marques Franca, Jair Otharan Nunes, Klailton Ralff da Silva, and Emmanuelle Regiane Cunha de Oliveira. Screening of email box in portuguese with svm at banco do brasil. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings*, pages 153–163. Springer, 2020.

[80] Marcos Paulo Fontes Feitosa, Carlos HG Ferreira, Glauber Dias Gonçalves, and Jussara Marques de Almeida. Análise da percepçao das pessoas no twitter sobre açoes policiais. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 73–84. SBC, 2022.

[81] Thomas Palmeira Ferraz, Alexandre Alcoforado, Enzo Bustos, André Oliveira, Rodrigo Gerber, Naíde Müller, André Corrêa d'Almeida, Bruno Veloso, and Anna Reali Costa. Debacer: a method for slicing moderated debates. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 667–678. Sociedade Brasileira de Computação-SBC, 2021.

[82] Pedro Fialho, Luísa Coheur, and Paulo Quaresma. Back to the feature, in entailment detection and similarity measurement for portuguese. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 164–173. Springer, 2020.

[83] José Solenir L Figuerêdo, Renata F Araújo-Calumby, and Rodrigo T Calumby. Machine learning for prognosis of patients with covid-19: An early days analysis. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 59–70. SBC, 2021.

[84] Ivan J Reis Filho, Luiz HD Martins, Antonio RS Parmezan, Ricardo M Marcacini, and Solange O Rezende. Sequential short-text classification from multiple textual representations with weak supervision. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part I*, pages 165–179. Springer, 2022.

[85] Thomas Fontanari, Tiago Comassetto Fróes, and Mariana Recamonde-Mendoza. Cross-validation strategies for balanced and imbalanced datasets. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part I*, pages 626–640. Springer, 2022.

[86] Nathan Formentin, Eduardo Borges, Giancarlo Lucca, Helida Santos, and Graçaliz Dimuro. Death registry prediction in brazilian male prisons with a random forest ensemble. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 330–341. SBC, 2020.

[87] Bruna S Freitas, Diego Bottero, Giancarlo Lucca, Eduardo N Borges, Helida Santos, and Graçaliz P Dimuro. A clustering algorithm to evaluate the attitude of brazilian researchers regarding open access research data. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 631–642. SBC, 2021.

[88] Gabriel L Garcia, Luis CS Afonso, and João P Papa. Fakerecogna: A new brazilian corpus for fake news detection. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 57–67. Springer, 2022.

[89] Marcos Garcia and Alfredo Crespo-Otero. A targeted assessment of the syntactic abilities of transformer models for galician-portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 46–56. Springer, 2022.

[90] Pedro Gengo and Filipe AN Verri. Correction to: Semi-supervised sentiment analysis of portuguese tweets with random walk in feature sample networks. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages C1–C1. Springer, 2020.

[91] Marcos PS Gôlo, Rafael G Rossi, and Ricardo M Marcacini. Triple-vae: A triple variational autoencoder to represent events in one-class event detection. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 643–654. SBC, 2021.

[92] Isabella Maria Alonso Gomes and Norton Trevisan Roman. How aspects of similar datasets can impact distributional models. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 579–590. SBC, 2022.

[93] Eliseu Guimaraes, Jonnathan Carvalho, Aline Paes, and Alexandre Plastino. Exploring model transfer strategies for sentiment analysis in twitter. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 1–12. SBC, 2021.

[94] Yohan Bonescki Gumiel, Isabela Lee, Tayane Arantes Soares, Thiago Castro Ferreira, and Adriana Pagano. Sentiment analysis in portuguese texts from online health community forums: data, model and evaluation. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 64–72. SBC, 2021.

[95] Luiz Otávio Alves Hammes and Larissa Astrogildo de Freitas. Utilizando bertimbau para a classificação de emoções em português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 56–63. SBC, 2021.

[96] Luiz F Junior, Jorge Silva Junior, and Fábio Lobato. Um olhar sobre turismo gastronômico: Um caso no tripadvisor. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 519–530. SBC, 2020.

[97] Natan Siller Laurett and Filipe Nunes Ribeiro. Caracterização das publicações e relações entre mídias alternativas polarizadas no facebook. 2022.

[98] Beatriz Lima and Tatiane Nogueira. Incorporating text specificity into a convolutional neural network for the classification of review perceived helpfulness. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 480–495. Springer, 2021.

[99] Renata F Lins, Flávia A Barros, Ricardo BC Prudêncio, and Wallace N Melo. Automatic classification of bug reports for mobile devices: An industrial case study. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 728–739. SBC, 2022.

[100] Fábio MF Lobato, Gleyce C de Sousa, and Antonio FL Jacob Jr. Brasnam em perspectiva: uma análise da sua trajetória até os 10 anos de existência. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 217–228. SBC, 2021.

[101] Johannes V Lochter, Renato M Silva, and Tiago A Almeida. Deep learning models for representing out-of-vocabulary words. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, pages 418–434. Springer, 2020.

[102] Lucelene Lopes, Magali S Duran, and Thiago AS Pardo. Universal dependencies-based pos tagging refinement through linguistic resources. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 601–615. Springer, 2021.

[103] Elian Conceiçao Luz, Camilla Rastely da Silva, and Daniela Barreiro Claro. Engenharia de features linguísticas para classificação de triplas relacionais. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 381–388. SBC, 2021.

[104] Mateus Tarcinalli Machado, Thiago Alexandre Salgueiro Pardo, Evandro Eduardo Seron Ruiz, and Ariani Di Felippo. Learning rules for automatic identification of implicit aspects in portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informaçao e da Linguagem Humana*, pages 82–91. SBC, 2021.

[105] Taynan Maier Ferreira and Anna Helena Reali Costa. Deepbt and nlp data augmentation techniques: a new proposal and a comprehensive study. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, pages 435–449. Springer, 2020.

[106] Errol Mamani-Condori and José Ochoa-Luna. Aggressive language detection using vgcn-bert for spanish texts. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 359–373, 2021.

[107] Jeziel C Marinho, Fábio Cordeiro, Rafael T Anchiêta, and Raimundo S Moura. Automated essay scoring: An approach based on enem competencies. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 49–60. SBC, 2022.

[108] Emanuel Matos, Mário Rodrigues, and António Teixeira. Named entity extractors for new domains by transfer learning with automatically annotated data. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 288–298. Springer, 2022.

[109] Augusto R Mendes, Rafael VP Passador, and Helena M Caseli. Identificando sintomas de depressao em postagens do twitter em português do brasil. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 162–171. SBC, 2021.

[110] Luan Misael, Carlos Forster, Emanuel Fontelles, Vinicius Sampaio, and Mardônio França. Temporal analysis and visualisation of music. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 507–518. SBC, 2020.

[111] Dionéia Motta Monte-Serrat, Mateus Tarcinalli Machado, and Evandro Eduardo Seron Ruiz. A machine learning approach to literary genre classification on portuguese texts: circumventing nlp's standard varieties. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 255–264. SBC, 2021.

[112] Eduardo F Montesuma, Lucas C Carneiro, Adson RP Damasceno, Joao Victor FT de Sampaio, Romulo F Férrer Filho, Paulo Henrique M Maia, and Francisco CMB Oliveira. An empirical study of information retrieval and machine reading comprehension algorithms for an online education platform. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 217–226. SBC, 2021.

[113] Luis-Gil Moreno-Jiménez, Juan-Manuel Torres-Moreno, and Roseli S Wedemann. A preliminary study for literary rhyme generation based on neuronal representation, semantics and shallow parsing. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 190–198. SBC, 2021.

[114] Ana Alice Ximenes Mota, Wellington Franco, and César Lincoln Cavalcante Mattos. Detecção de desinformação sobre covid-19 no twitter. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 172–181. SBC, 2021.

[115] Caio Mota, Andressa Lima, André Nascimento, Péricles Miranda, and Rafael de Mello. Classificação de páginas de petições iniciais utilizando redes neurais convolucionais multimodais. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 318–329. SBC, 2020.

[116] Caio CR Mota, André CA Nascimento, Péricles BC Miranda, Rafael Ferreira Mello, Isabel WS Maldonado, and José LM Coelho Filho. Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 130–140. SBC, 2021.

[117] Ester Motta and Maria José Bocorny Finatto. Constituintes frasais com função de sujeito em sentenças judiciais. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 416–424. SBC, 2021.

[118] Rubem G Nanclarez, Norton T Roman, and Fernando JV da Silva. Generalizing over data sets: a preliminary study with bert for natural language inference. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 602–611. SBC, 2022.

[119] Rodolpho Silva Nascimento, Gabriel Nascimento, Flavio Carvalho, and Gustavo Guedes. Mineração de opiniões com liwc: abordagem prática sobre sistemas judiciais eletrônicos brasileiros. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 132–141. SBC, 2020.

[120] Francisco Neto, Romero Silva, Roberta Gouveia, Maria Batista, and Igor Oliveira. Computação em nuvem e aprendizado de máquina para análise de grandes volumes de dados educacionais. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 58–69. SBC, 2020.

[121] José Reinaldo CSAVS Neto and Thiago de Paulo Faleiros. Deep active-self learning applied to named entity recognition. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 405–418. Springer, 2021.

[122] Victor Nicola, Marcelo Lauretto, and Karina Valdivia Delgado. Avaliação empírica de classificadores e métodos de balanceamento para detecção de fraudes em transações com cartões de créditos. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 70–81. SBC, 2020.

[123] Fernando AA Nóbrega, Alipio M Jorge, Pavel Brazdil, and Thiago AS Pardo. Sentence compression for portuguese. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 270–280. Springer, 2020.

[124] Diogo Nolasco and Jonice Oliveira. A study of rumor detection based on social network topic models relationship. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 166–177. SBC, 2020.

[125] Ana Luiza Nunes, Alexandre Rademaker, and Leonel Figueiredo de Alencar. Utilizando um dicionário morfológico para expandir a cobertura lexical de uma gramática do português no formalismo hpsg. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 11–18. SBC, 2021.

[126] Emerson Yoshiaki Okano, Zebin Liu, Donghong Ji, and Evandro Eduardo Seron Ruiz. Fake news detection on fake. br using hierarchical attention networks. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 143–152. Springer, 2020.

[127] André Oliveira, Anna Costa, and Eduardo Hruschka. A framework for multi-document extractive summarization of reviews with aspect-based sentiment analysis. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 471–482. SBC, 2020.

[128] André Seidel Oliveira and Anna H Reali Costa. Plsum: Generating pt-br wikipedia by summarizing multiple websites. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 751–762. SBC, 2021.

[129] Gabriel P Oliveira, Beatriz F Paiva, Ana Paula Couto da Silva, and Mirella M Moro. Characterizing the diffusion of misinformation regarding the coronavac vaccine in brazil. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 204–215. SBC, 2022.

[130] Lucas Emanuel Silva e Oliveira, Elisa Terumi Rubel Schneider, Yohan Bonescki Gumiel, Mayara Aparecida Passaura da Luz, Emerson Cabrera Paraiso, and Claudia Moro. Experiments on portuguese clinical question answering. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 133–145. Springer, 2021.

[131] Miguel Oliveira and Tiago Melo. Investigando features de sentenças para classificação de subjetividade e polaridade em português do brasil. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 270–281. SBC, 2020.

[132] Vinícius J Paes, Danilo Araújo, Kellyton Brito, and Ermeson Andrade. Análise de sentimento em tweets relacionados ao desmatamento da floresta amazônica. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 61–72. SBC, 2022.

[133] Robson T Paula, Décio G Aguiar Neto, Davi Romero, and Paulo T Guerra. Evaluation of synthetic datasets generation for intent classification tasks in portuguese. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 265–274. SBC, 2021.

[134] Matheus Camasmie Pavan, Wesley Ramos dos Santos, and Ivandré Paraboni. Twitter moral stance classification using long short-term memory networks. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 636–647. Springer, 2020.

[135] Matheus Letzov Pelozo, Marcelo Custódio, and Alison R Panisson. Answering questions about covid-19 vaccines using chatbot technologies. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part I*, pages 458–472. Springer, 2022.

[136] Fabíola SF Pereira. Caracterizaçao da propagaçao de rumores no twitter utilizando redes textuais temporais. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 25–31. SBC, 2021.

[137] Breno David Lopes Pinheiro, Ellen Polliana Ramos Souza, Douglas Vitório, and Hidelberg Oliveira Albuquerque. A comparative analysis of machine learning named entity recognition tools for the brazilian and european portuguese language variants. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 244–255. SBC, 2021.

[138] Pedro Pinheiro, Luan Siqueira, and Marcos Amaris. A four-step cascade methodology to classify mcn codes using nlp techniques. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 389–400. SBC, 2022.

[139] Victor Landim Teixeirense Pinheiro and Thiago de Paulo Faleiros. Aplicaçao de modelos de tópicos em análises automatizadas de discursos de senadores brasileiros. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 612–623. SBC, 2022.

[140] Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia, and Renato Vicente. Legalnlp-natural language processing methods for the brazilian legal language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 763–774. SBC, 2021.

[141] Lucas Porto and Evandro Ruiz. Pós-processamento de textos de tradução automática baseado em teoria de grafos. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 426–436. SBC, 2020.

[142] Pedro Ivo Monteiro Privatto and Ivan Rizzo Guilherme. When external knowledge does not aggregate in named entity recognition. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 616–627. Springer, 2021.

[143] Hugo Queiroz Abonizio and Sylvio Barbon Junior. Pre-trained data augmentation for text classification. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I*, pages 551–565. Springer, 2020.

[144] Daniel De Los Reyes, Douglas Trajano, Isabel Harb Manssour, Renata Vieira, and Rafael H Bordini. Entity relation extraction from news articles in portuguese for competitive intelligence based on bert. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 449–464. Springer, 2021.

[145] Arthur D Reys, Danilo Silva, Daniel Severo, Saulo Pedro, Marcia M de Sousa e Sá, and Guilherme AC Salgado. Predicting multiple icd-10 codes from brazilian-portuguese clinical notes. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 566–580. Springer, 2020.

[146] R Rodrigo Filho, Elismênnia Oliveira, Jordão Nunes, Marcelo Inuzuka, and Hugo do Nascimento. Computational mining on ibict bdtd's thesis and dissertation metadata for supporting social science research. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 603–614. SBC, 2020.

[147] Lucas DF Rodrigues, Luiz CCL Junior, Antonio FL Jacob Junior, and Fábio MF Lobato. Desenvolvimento de um conjunto de dados com comentários extraídos da plataforma twitch sobre o jogo league of legends. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 91–102. SBC, 2021.

[148] Rodrigo F Rodrigues and Larissa A de Freitas. Utilizando pistas linguística para detectar conteúdo enganoso em português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 447–451. SBC, 2021.

[149] Ruan Chaves Rodrigues, Acquila Santos Rocha, Marcelo Akira Inuzuka, and Hugo Alexandre Dantas do Nascimento. Domain adaptation of transformers for english word segmentation. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 483–496. Springer, 2020.

[150] Ruan Chaves Rodrigues, Jéssica Rodrigues, Pedro Vitor Quinta de Castro, Nádia Felix Felipe da Silva, and Anderson Soares. Portuguese language models and word embeddings: evaluating on semantic similarity tasks. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 239–248. Springer, 2020.

[151] Anderson da Silva Brito Sacramento and Marlo Souza. Joint event extraction with contextualized word embeddings for the portuguese language. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 496–510. Springer, 2021.

[152] Kenzo Sakiyama, Lucas de Souza Rodrigues, and Edson Takashi Matsubara. Can twitter data estimate reality show outcomes? In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 466–482. Springer, 2020.

[153] Isadora A Salles and Gisele L Pappa. Viés de gênero em biografias da wikipédia em português. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 211–216. SBC, 2021.

[154] Brenda Salenave Santana and Leandro Krug Wives. Extraction and use of structured and unstructured features for the recommendation of urban resources. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 206–214. Springer, 2020.

[155] Brenda Salenave Santana, Aline Aver Vanin, and Leandro Krug Wives. Sexist hate speech: Identifying potential online verbal violence instances. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 177–187. Springer, 2022.

[156] Jéssica S Santos, Flávia Bernardini, and Aline Paes. Measuring the degree of divergence when labeling tweets in the electoral scenario. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 127–138. SBC, 2021.

[157] Joaquim Santos, Henrique DP dos Santos, Fábio Tabalipa, and Renata Vieira. De-identification of clinical notes using contextualized language models and a token classifier. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 33–41. Springer, 2021.

[158] José Santos and Rafael Rossi. Aprendizado de máquina não supervisionado baseado em redes heterogêneas para agrupamento de textos. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 35–46. SBC, 2020.

[159] José Santos, Ana Alves, and Hugo Gonçalo Oliveira. Leveraging on semantic textual similarity for developing a portuguese dialogue system. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 131–142. Springer, 2020.

[160] Patricia D Santos and Denise H Goya. Automatic twitter stance detection on politically controversial issues: A study on covid-19's cpi. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 524–535. SBC, 2021.

[161] Patricia D Santos and Denise H Goya. Detecção de posicionamento e rotulação automática de usuários do twitter: estudo sobre o embate científico-político no contexto da cpi da covid-19. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 49–60. SBC, 2022.

[162] Gabriel Schubert and Larissa de Freitas. A construçao de um corpus para detecçao de ironia e sarcasmo em português. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 709–717. SBC, 2020.

[163] Clarisse Scofield, Mariana O Silva, Luiza de Melo-Gomes, and Mirella M Moro. Book genre classification based on reviews of portuguese-language literature. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 188–197. Springer, 2022.

[164] Felipe R Serras and Marcelo Finger. verbert: Automating brazilian case law document multi-label categorization using bert. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 237–246. SBC, 2021.

[165] Adriano Silva and Norton Roman. Hate speech detection in portuguese with naïve bayes, svm, mlp and logistic regression. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 1–12. SBC, 2020.

[166] Adriano Capanema Silva and Luiz Cláudio Gomes Maia. The use of machine learning in the classification of electronic lawsuits: An application in the court of justice of minas gerais. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 606–620. Springer, 2020.

[167] Diego F Silva, Alcides M e Silva, Bianca M Lopes, Karina M Johansson, Fernanda M Assi, Júlia TC de Jesus, Reynold N Mazo, Daniel Lucrédio, Helena M Caseli, and Livy Real. Named entity recognition for brazilian portuguese product titles. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, pages 526–541. Springer, 2021.

[168] Márcio Silva, Samuel Guimaraes, Josemar Caetano, Marcelo Araújo, Jonatas Santos, Julio CS Reis, Ana Silva, Fabrício Benevenuto, and Jussara Almeida. Propaganda eleitoral antecipada: Uma análise de postagens em mídias sociais. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 199–204. SBC, 2021.

[169] Mariana O Silva, Clarisse Scofield, Gabriel P Oliveira, Danilo B Seufitelli, and Mirella M Moro. Exploring brazilian cultural identity through reading preferences. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 115–126. SBC, 2021.

[170] Nádia FF da Silva, Marília Costa R Silva, Fabíola SF Pereira, João Pedro M Tarrega, João Vitor P Beinotti, Márcio Fonseca, Francisco Edmundo de Andrade, and André CP de LF de Carvalho. Evaluating topic models in portuguese political comments about bills from brazil's chamber of deputies. In *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II*, pages 104–120. Springer, 2021.

[171] Bárbara Silveira, Ana Paula Couto da Silva, and Fabricio Murai. Modelos de previsao do tom emocional de usuários em comunidades de saúde mental no reddit. In *Anais do IX Brazilian Workshop on Social Network Analysis and Mining*, pages 13–24. SBC, 2020.

[172] Cinthia M Souza and Renato Vimieiro. A long texts summarization approach to scientific articles. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 182–189. SBC, 2021.

[173] Ellen Souza, Gyovana Moriyama, Douglas Vitório, André CPLF de Carvalho, Nádia Félix, Hidelberg O Albuquerque, and Adriano LI Oliveira. Assessing the impact of stemming algorithms applied to brazilian legislative documents retrieval. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 227–236. SBC, 2021.

[174] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer, 2020.

[175] Frederico Dias Souza and João Baptista de Oliveira e Souza Filho. Bert for sentiment analysis: Pre-trained and fine-tuned alternatives. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 209–218. Springer, 2022.

[176] Stefano Spindola, Marcos Menon José, André Seidel Oliveira, Flávio Nakasato Cação, and Fábio Gagliardi Cozman. Interpretability of attention mechanisms in a portuguese-based question answering system about the blue amazon. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 775–786. SBC, 2021.

[177] Yan V Sym, João Gabriel M Campos, Marcos M José, and Fabio G Cozman. Comparing computational architectures for automated journalism. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 377–388. SBC, 2022.

[178] Sabrina de Fátima Barbosa Taniwaki and Jackson Wilke da Cruz Souza. Criação e anotação do corpus de resumos científicos de ciências sociais aplicadas. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 437–441. SBC, 2021.

[179] Gustavo A Testoni, Marcelo P Souza, Paulo Márcio S Freire, and Ronaldo R Goldschimidt. Um método linguístico que combina polaridade, emoção e aspectos gramaticais para detecção de fake news em inglês. In *Anais do X Brazilian Workshop on Social Network Analysis and Mining*, pages 151–162. SBC, 2021.

[180] Rodrigo Neves Trindade, Luiz HD Martins, Geraldo Nunes Correa, and Ivan José dos Reis Filho. Using a labeling function for automatic classification of agribusiness news: A weak supervisory approach. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 73–82. SBC, 2022.

[181] Francielle Alves Vargas, Rodolfo Sanches Saraiva Dos Santos, and Pedro Regattieri Rocha. Identifying fine-grained opinion and classifying polarity on coronavirus pandemic. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 511–520. Springer, 2020.

[182] Douglas Vitório, Hidelberg Oliveira Albuquerque, Ellen Polliana Ramos Souza, Adriano Lorena Inacio de Oliveira, Flávia Barros, and Ricardo BC Prudêncio. Análise do posicionamento dos usuários do twitter acerca da vacinaçao infantil contra a covid-19 no brasil. *Anais*, 2022.

[183] Douglas Vitório, Ellen Souza, Lucas Martins, Nádia FF da Silva, André Carlos Ponce de Leon Ferreira de Carvalho, and Adriano LI Oliveira. Ulysses-rfsq: A novel method to improve legal information retrieval based on relevance feedback. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part I*, pages 77–91. Springer, 2022.

[184] Gabriela Wick-Pedro and Roney LS Santos. Complexidade textual em notícias satíricas: uma análise para o português do brasil. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 409–415. SBC, 2021.

[185] Miguel Won and Jorge Fernandes. Ss-pt: A stance and sentiment data set from portuguese quoted tweets. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 110–121. Springer, 2022.

[186] Luciano Zanuz and Sandro José Rigo. Fostering judiciary applications with new fine-tuned models for legal named entity recognition in portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 219–229. Springer, 2022.

ANEXOS

ANEXO A – FEEDBACK DOS REVISORES DO BRACIS 2023

JEMS BRACIS 2023 0 Help Gabriele de Sousa Araújo

Reviews

Review 1

1) Appropriateness author:	2) Clarity author:	3) Originality author:	4) Soundness/Correctness author:	5) Meaningful comparison author:	6) Reproducibility author:	7) Overall recommendation author:
4) Probably.	2) Not so clear.	1) Poor.	2) Troublesome. The work should really have been done or evaluated differently.	1) Little awareness of related work or lacks necessary empirical comparison.	2) Only basic information about the methodology is provided.	1) Reject - Clearly below the standards of the conference.

Detailed comments:Please supply detailed comments to back up your rankings. Access strong and weak points of the paper. These comments will be forwarded to the authors of the paper. Please make suggestions to improve the quality and clarity of the paper. The more detailed you make your comments, the more useful your review will be for the committee and for the authors.:

In this paper, the authors "conducted a systematic mapping aiming to provide an overview of NLP techniques application on social media analysis, identify the most used algorithms, and understand current trends in the use of NLP in this context". So, the objective is to construct an overview of NLP techniques in some applications. Put into this perspective, it is not clear why the authors focused on these five events (bracis, brasnam, eniac, stil, and propor), while we have more focused events in the NLP fields, such as EMNLP and ACL. Moreover, why three years (2020-2022)?

As also stated by the authors: "Thus, this work can be helpful for academic researchers interested in exploring the potential of these tools and techniques, having a clear picture of gaps, challenges, and research opportunities in this area, and analyzing the current scenario in research involving NLP and social media.". A review that includes events focused on NLP (such as EMNLP and ACL) could be a better guide to analyzing the opportunities and challenges of the area.

Overall, the aim of the paper should be better stated.

Minor:
- Fig.3 shows representation strategies (e.g., tf-df) and machine learning algorithms (e.g., mlp). To show these categories separately could be more attractive to the reader.

Review 2

1) Appropriateness author:	2) Clarity author:	3) Originality author:	4) Soundness/Correctness author:	5) Meaningful comparison author:	6) Reproducibility author:	7) Overall recommendation author:
2) Probably not.	5) Very clear.	1) Poor.	3) Fairly reasonable work, but I am not entirely convinced to accept its conclusions and decisions.	3) Bibliography and comparison are somewhat helpful, but it could be hard for a reader to determine exactly how this work relates to previous work.	4) The experiments are described, but the description leaves out minor details that experts in the field can guess.	2) Weak reject - I vote for rejecting it but could be persuaded otherwise.

Detailed comments:Please supply detailed comments to back up your rankings. Access strong and weak points of the paper. These comments will be forwarded to the authors of the paper. Please make suggestions to improve the quality and clarity of the paper. The more detailed you make your comments, the more useful your review will be for the committee and for the authors.:

This work is interested in identifying the main tools and techniques, the development environments, tasks performed, and metrics used in academic studies published in Brazil.

They carried out a systematic mapping study using a methodology for bibliographic review proposed in previous works. Their finds are interesting in NLP studies perspectives. However, as a review of techniques in NLP this work does not match the main purpose of BRACIS, which encourages researchers to publish their works on new techniques or enhancement of the existing ones.

Therefore, I recommend this paper to be submitted to ENIAC.

Review 3

1) Appropriateness author:	2) Clarity author:	3) Originality author:	4) Soundness/Correctness author:	5) Meaningful comparison author:	6) Reproducibility author:	7) Overall recommendation author:
5) Certainly.	5) Very clear.	2) Just a little original	3) Fairly reasonable work, but I am not entirely convinced to accept its conclusions and decisions.	1) Little awareness of related work or lacks necessary empirical comparison.	3) The experiments are described but not in detail.	2) Weak reject - I vote for rejecting it but could be persuaded otherwise.

Detailed comments:Please supply detailed comments to back up your rankings. Access strong and weak points of the paper. These comments will be forwarded to the authors of the paper. Please make suggestions to improve the quality and clarity of the paper. The more detailed you make your comments, the more useful your review will be for the committee and for the authors.:

The work collects information about NLP works in Portuguese NLP forums doing a head count of works developed, their tools employed, and their datasets. The work lacks to mention previous work doing the same kind of analysis, which could be interesting to see the evolution further behind than the three years analyzed. The classification of the datasets is based on their names not considering the same level of origin, for example: ASSIN dataset is considered a different dataset than Google News, but ASSIN is composed by Google news sentences. A similar problem occurs while treating BERT, BERTimbau, and TensorFlow as different tools, but they actually TensorFlow approach emcompass BERT-based methods, as well as BERTimbau also is encompassed as BERT-based methods.

The work as the merit to develop this kind of survey, but given the lack of depth of analyzing the data may hinder the scientific contribution of the paper.