MAPEAMENTO DE CORPORA EMOCIONAIS MULTIPLATAFORMA EM LÍNGUA PORTUGUESA: UMA REVISÃO SOBRE ANÁLISE DE SENTIMENTOS EM REDES SOCIAIS

Vanessa Pereira Cunha – vanessapereiracunha72@gmail.com
Federal Institute of Education, Science and Technology of Piauí
Ricardo Moura Sekeff Budaruiche – ricardo.sekeff@ifpi.edu.br
Federal Institute of Education, Science and Technology of Piauí

***Resumo*—** Esta revisão sistemática tem como objetivo investigar, organizar e analisar os avanços mais relevantes relacionados à construção de corpora emocionais anotados para a língua portuguesa. O estudo abrange o período de 2021 a 2024 e tem foco em textos oriundos de redes sociais multiplataforma (Facebook, Twitter/X, Instagram e Reddit). Entende-se por *corpus emocional* um conjunto de textos coletados em plataformas digitais e anotados, manual ou automaticamente, com informações sobre sentimentos e emoções expressos. O trabalho mapeia corpora amplamente utilizados (TweetSentBR, SS-PT, B2T, Reddit-BR), léxicos de sentimentos consolidados (SentiLex-PT, OpLexicon, LIWC-PT) e modelos de Processamento de Linguagem Natural (PLN) treinados especificamente para o português (BERTimbau, PTT5, LaBSE). Identificam-se recursos prontos para uso pela comunidade científica, bem como desafios significativos, tais como a escassez de dados anotados, a ambiguidade linguística, as variações regionais e a ausência de diretrizes padronizadas de anotação. Ao reunir e discutir essas iniciativas, esta revisão contribui para o fortalecimento das bases teóricas e práticas que sustentam os avanços na análise de sentimentos em português brasileiro.

***Palavras-chave*—** análise de sentimentos; corpora emocionais; PLN em português; redes sociais; recursos linguísticos; classificação de emoções.

Abstract— This systematic review aims to investigate, organize, and analyze the most relevant advances related to the construction of emotional corpora annotated for the Portuguese language. The study encompasses the period from 2021 to 2024 and focuses on texts from multi-platform social media (Facebook, Twitter/X, Instagram, and Reddit). An *emotional corpus* is understood as a set of texts collected from digital platforms and annotated, either manually or automatically, with information about expressed sentiments and emotions. The study maps widely used corpora (TweetSentBR, SS-PT, B2T, Reddit-BR), established sentiment lexicons (SentiLex-PT, OpLexicon, LIWC-PT), and Portuguese-specific pre-trained Natural Language Processing (NLP) models (BERTimbau, PTT5, LaBSE). The review identifies ready-to-use resources available to the scientific community, as well as major challenges such as the scarcity of annotated data, linguistic ambiguity, regional variation, and the absence of standardized annotation guidelines. By gathering and discussing these initiatives, this review contributes to strengthening the theoretical and practical bases that support advances in sentiment analysis in Brazilian Portuguese.

Keywords — sentiment analysis; emotional corpora; Portuguese NLP; social media; linguistic resources; emotion classification.

1 INTRODUÇÃO

O crescente uso de plataformas digitais transformou as redes sociais em espaços privilegiados de expressão de opiniões e emoções. No Brasil, cerca de 67,8 % da população possui contas ativas em redes sociais, como Facebook, Twitter/X, Instagram e Reddit, o que resulta em um fluxo contínuo de dados textuais carregados de sentimentos (DataReportal, 2025).

Usuários de língua portuguesa interagem diariamente nessas plataformas, compartilhando opiniões que refletem nuances emocionais e culturais. Diante desse cenário, a análise de sentimentos consolidou-se como ferramenta estratégica para setores como marketing, saúde pública e monitoramento social, exigindo corpora anotados de alta qualidade para treinar e avaliar sistemas de Processamento de Linguagem Natural (PLN) em português.

Define-se *corpus emocional* como um conjunto de textos digitais anotados (manual ou automaticamente) com informação sobre sentimentos e emoções. Tais corpora são fundamentais para treinar e avaliar modelos de PLN e têm aplicações práticas em saúde pública, atendimento automatizado, monitoramento de opinião e análise de reputação. Assim, a discussão sobre corpora emocionais também se conecta ao eixo de Inovação e Desenvolvimento Tecnológico, evidenciando seu potencial de transformação social e cultural.

Embora existam estudos consolidados em inglês, o português enfrenta desafios específicos: variações regionais, gírias locais, estruturas morfológicas complexas, e expressões idiomáticas. Além disso, a fragmentação de recursos entre diferentes plataformas dificulta a criação de modelos multiplataforma robustos.

Revisões anteriores abordam aspectos gerais de análise de sentimento em português (CAVALCANTI et al., 2020; SILVA; PARDO, 2021), porém nenhuma sistematiza, em âmbito multiplataforma, corpora emocionais voltados a redes sociais.

Este estudo realiza uma revisão sistemática de literatura sobre corpora emocionais multiplataforma em português, com o objetivo de mapear recursos existentes (conjuntos de dados, modelos pré-treinados e ferramentas), discutir soluções prontas, identificar limitações e propor diretrizes para pesquisas futuras.

2 TRABALHOS RELACIONADOS

Nesta seção, são apresentados corpora, ferramentas e modelos utilizados em trabalhos recentes sobre análise de sentimentos em português. Os recursos mapeados entre 2021 e 2024 compõem o núcleo da presente revisão sistemática. Modelos e léxicos anteriores a esse período são aqui incluídos por sua relevância histórica e por ainda serem amplamente utilizados em pesquisas atuais.

Diversos recursos de dados anotados para sentimentos e emoções em português têm sido desenvolvidos ao longo da última década, com avanço significativo entre os anos de 2021 e 2024. O *TweetSentBR* (BRUM; VOLPE NUNES, 2018) foi um dos primeiros corpora disponíveis, contendo 15.000 tweets coletados do Twitter/X brasileiro e anotados manualmente em três classes de polaridade (Positivo, Negativo, Neutro). Este recurso serviu de base para diversas abordagens posteriores em análise de sentimentos no domínio de redes sociais.

Na sequência, o *SS-PT* (ALMEIDA et al., 2022) introduziu uma importante inovação ao incorporar anotação dupla: sentimento e stance (opinião favorável, contrária ou neutra). A base inclui cerca de 15.000 tweets políticos e sociais, permitindo estudos mais profundos sobre opinião pública em ambientes polarizados. Esse corpus é considerado um dos mais completos em estrutura de anotação dupla em português.

Esforços voltados para domínios específicos também têm se intensificado. O *B2T* (KAKIMOTO et al., 2024), por exemplo, reúne aproximadamente 375.000 tweets sobre bancos brasileiros, dos quais 1.096 foram anotados manualmente para polaridade emocional. O objetivo principal foi estudar o comportamento emocional dos usuários frente a instituições financeiras, revelando sentimentos como confiança, indignação e frustração. Outro exemplo de corpus de domínio específico é o *Reddit-BR* (PIORINO et al., 2024), que reúne postagens de subreddits brasileiros. Essas postagens foram anotadas manualmente para polaridade (Positivo, Negativo, Neutro), ampliando os estudos para fóruns online menos explorados.

Além destes, corpora voltados à linguagem ofensiva oferecem insights sobre valências negativas extremas. O *ToLD-Br* (LEITE et al., 2020) contém 21.000 tweets rotulados em sete categorias de toxicidade, incluindo misoginia, racismo, LGBTfobia, entre outras. O *HateBR* (VARGAS et al., 2022) anotou 7.000 comentários do Instagram voltados à política nacional, utilizando categorias binárias (ofensivo/não-ofensivo) e níveis de intensidade. Já o *OLID-Br*(TRAJANO et al., 2023) consolidou diversas bases brasileiras com linguagem ofensiva, resultando em 6.354 exemplos anotados com hierarquias de toxicidade (tipo, alvo, intensidade).

No âmbito léxico, destacam-se os dicionários SentiLex-PT (CARVALHO; SILVA, 2015) e OpLexicon (SOUZA; VIEIRA, 2011), amplamente utilizados em análises de sentimento baseadas em regras ou como features em modelos supervisionados. Outro recurso relevante é o *LIWC-PT* (BALAGE FILHO et al., 2013), uma adaptação da ferramenta LIWC para o português brasileiro, que foi validada em experimentos empíricos e mostra desempenho competitivo quando combinada com *machine learning* (aprendizado de máquina).

Em termos de modelos de linguagem, a criação do *BERTimbau* (Souza et al., 2020) marcou um divisor de águas no PLN em português. Treinado em grandes corpora nacionais (brWaC, Wikipedia, e outros), o modelo se destaca em tarefas como análise de sentimento, classificação textual e inferência textual (NLI). O *PTT5* (CARMO et al., 2020), por sua vez, é uma versão seq2seq do T5 adaptado ao português, capaz de realizar tarefas de geração e classificação com bons resultados. Ambos superam modelos multilíngues como mBERT e LaBSE, evidenciando a importância de recursos treinados especificamente em dados brasileiros.

Ainda que esses avanços representam progresso significativo, persistem lacunas na cobertura multiplataforma (poucos dados para Instagram e Facebook) e ausência de padronização na anotação. A heterogeneidade dos esquemas e a carência de diretrizes unificadas dificultam comparações entre abordagens e reusabilidade dos recursos.

3 METODOLOGIA

A pesquisa bibliográfica foi conduzida em bases de dados acadêmicas relevantes: Google Scholar, IEEE Xplore, Scopus, arXiv, ACL Anthology e CEUR-WS. A escolha por múltiplas bases buscou garantir abrangência e diversidade, incluindo tanto publicações de grande impacto internacional quanto eventos técnicos especializados em Processamento de Linguagem Natural (PLN). As buscas foram realizadas em 10/07/2025 e filtradas para o período de 2021 a 2024, assegurando foco nos trabalhos mais recentes.

As palavras-chave utilizadas incluíram combinações em português e inglês, como: “corpus emocional”, “análise de sentimentos português”, “redes sociais português”, “Portuguese sentiment analysis corpus” e “social media text mining”. Também foram empregados nomes de plataformas sociais (Twitter/X, Facebook, Instagram, Reddit) em

conjunto com operadores booleanos (AND, OR), de modo a ampliar a abrangência da busca.

Foram definidos critérios de inclusão que priorizaram artigos, teses, relatórios, preprints e repositórios que descrevessem corpora em língua portuguesa e/ou disponibilizassem datasets ou documentação técnica. Houve preferência por publicações indexadas em conferências de referência (ACL, NAACL, EMNLP, LREC, SBIA, BRACIS) ou periódicos com fator de impacto. Também foram considerados estudos com propostas inéditas, repositórios de dados públicos e revisões sistemáticas relevantes. Artigos em qualquer idioma foram aceitos, desde que tratassem de dados ou análises envolvendo a língua portuguesa.

Critérios de exclusão foram aplicados para eliminar estudos sem foco em português, trabalhos que apenas mencionaram corpora de forma vaga (sem informações mínimas como idioma, tamanho ou esquema de anotação), pesquisas sem acesso ao texto completo ou sem documentação pública, além de duplicatas sem informação adicional. Registros repetidos entre bases foram consolidados em um único arquivo.

Embora o recorte principal da revisão seja 2021–2024, o trabalho também incorporou referências anteriores consideradas marcos teóricos importantes. Entre elas estão os léxicos SentiLex-PT e OpLexicon, o dicionário LIWC adaptado ao português e o pacote Linguakit. Tais recursos não foram analisados sistematicamente, mas serviram como fundamentação metodológica para a análise desenvolvida.

A Tabela 1 apresenta uma síntese das bases consultadas, termos de busca utilizados e critérios de inclusão e exclusão aplicados.

Tabela 1: bases, termos de busca e critérios aplicados na revisão.

Base de dados	Palavras-chave	Período	Critérios de inclusão	Critérios de exclusão
Google Scholar, IEEE Xplore, Scopus, arXiv, ACL Anthology, CEUR-WS	“corpus emocional”, “análise de sentimentos português”, “Portuguese sentiment analysis corpus”, entre outros	2021–2024	Artigos com alto número de citações, publicados em conferências/periódicos relevantes, foco em português	Estudos sem foco na língua portuguesa ou sem disponibilização pública dos dados

Fonte: Autor (2025).

As buscas seguiram três etapas principais: (i) triagem de títulos e resumos para eliminar estudos fora do escopo; (ii) leitura do texto completo para verificar a descrição mínima do corpus (idioma, tamanho, método de anotação e disponibilidade); e (iii) extração e consolidação dos dados. Para cada estudo incluído foram registrados, quando disponíveis, o nome do corpus, a plataforma/fonte, o ano, o tamanho do conjunto, o esquema de anotação, o domínio de aplicação e a informação sobre disponibilidade pública. Estudos duplicados foram consolidados em um único registro.

4 RESULTADOS E DISCUSSÃO

4.1 RECURSOS DE DADOS E CORPORA EMOCIONAIS

Foram identificados diversos conjuntos de dados anotados em português para

sentimentos e emoções em redes sociais. Destacam-se:

- TweetSentBR (BRUM; VOLPE NUNES, 2018): 15.000 tweets brasileiros (sobre programas de TV) coletados em 2017, anotados em três classes (Positivo, Negativo, Neutro) por voluntários. Este corpus tem servido de base para comparações em tarefas de sentimento em português.
- SS-PT (ALMEIDA et al., 2022): mais de 15.000 tweets citados no Twitter/X, anotados tanto para stance (a favor, contra, neutro) quanto para polaridade sentimental (pos, neg, neu, sarcasmo). Foi o primeiro corpus amplo de tweets portugueses com essa dupla anotação.
- B2T (KAKIMOTO et al., 2024): coleção de tweets em português sobre bancos brasileiros. Coletaram ~375 mil comentários do Twitter/X, dos quais 1.096 foram manualmente anotados como Positivo, Neutro ou Negativo. Este corpus visa o estudo de sentimento no contexto financeiro.
- Reddit-BR (PIORINO et al., 2024): conjunto de posts de comunidades do Reddit brasileiras (por exemplo, subreddits focados no Brasil), anotados manualmente para polaridade de sentimento. Os autores desenvolveram esse corpus para apoiar a análise de sentimento em linguagem informal de mídias sociais em português.

Além desses, existem dados relativos à linguagem tóxica/ofensiva, que incluem conteúdo emocional negativo:

- ToLD-Br (LEITE et al., 2020): 21.000 tweets em português anotados por 3 avaliadores em 7 categorias de toxicidade (e.g. homofobia, racismo) . Criado para estudo de detecção de discurso de ódio e tóxico.
- HateBR (VARGAS et al., 2022): 7.000 comentários de Instagram relacionados à política, coletados de 6 contas influentes, anotados para ofensa (binário e em graus de intensidade). Destina-se a estudos de detecção de discurso de ódio em português.
- OLID-Br (TRAJANO et al., 2023): dataset consolidado de linguagem ofensiva reunindo dados de Twitter/X, YouTube e outros corpora portugueses anotados. Contém 6.354 comentários (expansível) rotulados em um esquema de 3 níveis (toxicidade binária, tipo de toxicidade, alvo). Embora enfocado em ofensa, é relevante como recurso multilíngue/anotado.

Para complementar, citam-se também ferramentas léxicas e recursos de anotação prontos para português:

- Léxicos de polaridade: SentiLex-PT (CARVALHO; SILVA, 2015) e OpLexicon (SOUZA; VIEIRA, 2011) são léxicos de sentimento para português bastante usados. (SOUZA; VIEIRA, 2012) compararam esses dois léxicos em corpus de tweets portugueses . O dicionário LIWC adaptado ao português brasileiro foi avaliado por (BALAGE FILHO et al., 2013) em tarefas de sentimento.
- Ferramentas de PLN: o pacote Linguakit (GAMALLO; GARCÍA, 2017) oferece um analisador de sentimento em português . Além disso, bibliotecas de PLN gerais (spaCy com modelos em português, NLTK, etc.) e frameworks como a plataforma Hugging Face disponibilizam modelos para análises em português.

Em síntese, embora ainda limitados, existem vários corpora e recursos anotados em português voltados a sentimentos negativos/positivos e outros estados afetivos em mídias sociais. A demanda por tais recursos se justifica pelo volume de texto opinativo nessas plataformas. No entanto, a cobertura desses corpora é fragmentada e tipicamente focalizada em categorias específicas (e.g. restritas a Twitter/X ou a certos domínios), evidenciando a

necessidade de integração multiplataforma no futuro.

Tabela 2 – Corpora emocionais em português e suas características

Corpus/Recurso	Finalidade	Referência
TweetSentBR	Tweets brasileiros sobre programas de TV, anotados em polaridade (Positivo, Negativo, Neutro)	BRUM; VOLPE NUNES, 2018
SS-PT	Tweets políticos e sociais com dupla anotação (sentimento e stance)	ALMEIDA et al., 2022
B2T	Tweets sobre bancos brasileiros, anotados em polaridade emocional (Positivo, Neutro, Negativo)	KAKIMOTO et al., 2024
Reddit-BR	Postagens de subreddits brasileiros, anotadas para polaridade	PIORINO et al., 2024
ToLD-Br	Tweets com categorias de toxicidade (racismo, misoginia, LGBTfobia, etc.)	LEITE et al., 2020
HateBR	Comentários do Instagram sobre política, anotados para ofensa (binário/intensidade)	VARGAS et al., 2022
OLID-Br	Comentários de múltiplas plataformas (Twitter/X, YouTube etc.), anotados para linguagem ofensiva	TRAJANO et al., 2023

Fonte: Autor (2025).

Os corpora apresentados na tabela 2 representam os principais recursos disponíveis em língua portuguesa para estudos de análise de sentimentos e emoções em redes sociais. Observa-se a diversidade de plataformas contempladas, incluindo Twitter/X, Reddit, Instagram e múltiplas mídias e a variedade de enfoques de anotação, como polaridade (positivo, negativo, neutro), stance, toxicidade e linguagem ofensiva. Esses recursos ilustram a evolução recente dos esforços de coleta e anotação no contexto brasileiro, fornecendo insumos valiosos tanto para experimentos acadêmicos quanto para aplicações práticas em Processamento de Linguagem Natural.

4.2 MODELOS DE PLN E ABORDAGENS DE CLASSIFICAÇÃO

Os trabalhos recentes utilizam principalmente modelos de aprendizado de máquina supervisionado e transformadores pré-treinados em português. Destacam-se:

- BERTimbau (SOUZA et al., 2020):[1] BERT pré-treinado em português brasileiro no corpus brWaC. Mostrou melhorar o estado da arte em tarefas diversas de compreensão de texto em português e superou o BERT multilíngue (mBERT) em experimentos . Em particular, (VIANNA et al., 2024) verificaram que, entre seis tipos de *embeddings*(representações vetoriais) (fastText, GloVe, Word2Vec, mBERT, BERTweet e BERTimbau), o BERTimbau teve melhor desempenho em classificação de sentimento de tweets em português . Em muitos estudos, BERTimbau é a base para

[1] BERT é a sigla de *Bidirectional Encoder Representations from Transformers*, modelo de linguagem baseado em redes neurais transformadoras amplamente utilizado no Processamento de Linguagem Natural (PLN).

fine-tuning (ajuste fino) em sentimento.

- T5-PT (PTT5): Versões de T5 adaptadas ao português. (CARMO et al., 2020) continuaram o pré treinamento do T5 no corpus BrWac (ptt5-v1) obtendo melhorias em tarefas de PLN para português . Posteriormente, (GOMES et al., 2024) liberaram o ptt5-v2 com maior refinamento. Modelos seq2seq como T5-PT são usados para tarefas de geração de texto e também aplicados a análise de sentimento via fine-tuning.
- LaBSE (FENG et al., 2020): modelo multilingue de embeddings de sentença para 109 idiomas . Embora originalmente não focado em português, o LaBSE é mencionado como uma alternativa para representação de sentenças em português em alguns trabalhos multilíngues.
- Outros transformadores: inclusões de versões nacionais de LLaMA (LLaMA-PT) e modelos GPT para português têm surgido, mas faltam publicações formais (focam-se em esforços comunitários e bases não comerciais). Modelos populares multilingues como XLM-R, mBERT ou Embeddings LASER também são usados em *baselines* (modelos de referência).
- Métodos tradicionais: Vários trabalhos adotam ainda abordagens clássicas de NLP: extração de *bag-of-words*[2], TF-IDF, e classificadores como Naïve Bayes, SVM, Random Forest, ou LSTM, configuradas para português. Por exemplo, no B2T, testaram XGBoost, SVM e regressão logística antes de usar BERTimbau. As abordagens baseadas em aprendizado profundo (AP) com redes neurais e Transformers (AT) têm superado os métodos baseados em recursos (AM, ex. NB, SVM) em muitos casos. De qualquer forma, configurações híbridas que combinam léxicos sentimentais (como SentiLex) e aprendizado automático ainda são estudadas.

Em geral, os modelos adaptados ao português têm se mostrado mais eficazes do que modelos genéricos multilíngues . Contudo, a escolha da abordagem depende da disponibilidade de dados e do cenário. Por enquanto, não há um único modelo universal: pesquisa contínua testa combinações de arquiteturas, como *ensembles* (comitês de modelos) e estratégias de pré-treinamento contínuo para textos informais de redes sociais. Ferramentas abertas (bibliotecas e APIs) oferecem suporte crescente, mas a maioria das soluções de ponta requer customização e anotação adicional para contextos específicos em português.

4.3 PRINCIPAIS DESAFIOS E LIMITAÇÕES

Apesar dos avanços, vários desafios tornam a construção de corpora emocionais em português complexa. Destacam-se:

- Escassez de dados anotados: conforme observado em recente trabalho de (KAKIMOTO et al., 2024), conjuntos de dados rotulados em português são raros. Essa limitação decorre de poucos recursos comunitários, alto custo de anotação e menor volume de texto disponível em português comparado ao inglês. A falta de corpora amplos dificulta treinar modelos de aprendizado profundo com alto desempenho.
- Ambiguidade linguística: expressões informais de redes sociais em português incluem gírias, regionalismos e negações que podem inverter polaridades (e.g. “muito bom demais” ou “não é ruim”). A interpretação automática dessas sutilezas permanece difícil, especialmente sem contexto suplementar (metadados ou histórico). Modelos de

[2] Bag-of-words — representação que descreve um texto pela presença ou frequência de palavras, sem considerar sua ordem.

PLN precisam lidar com figuras de linguagem e ironia, o que exige corpora ricos.

- Viés cultural e semântico: o português é falado em vários países (Brasil, Portugal, Angola etc.) com diferenças culturais marcantes . Termos idênticos podem carregar polaridades distintas em contextos regionais ou culturais. Por exemplo, expressões religiosas ou políticas podem ter conotações específicas em cada comunidade. Essa diversidade amplia a variabilidade semântica dos dados e pode introduzir viés nos modelos (que tendem a aprender o português brasileiro predominante).
- Complexidade de anotação: anotar sentimentos é subjetivo e demanda anotadores treinados. Dificuldades incluem ambiguidade entre emoções próximas (raiva vs decepção), escala de intensidade e sentimento misto. O trabalho de B2T notou necessidade de melhorar instruções para anotadores e aumentar coesão no critério de rotulagem . Em geral, falta de recursos humanos qualificados e de normas padrão de anotação em português dificulta a padronização de corpora emocionais.
- Falta de modelos pré-treinados específicos: embora existam modelos gerais em português (BERTimbau, etc.), há poucos modelos focados em texto de redes sociais informais. Por exemplo, não há muitas versões de Transformers pré-treinadas exclusivamente em tweets ou comentários em português. Isso faz com que muitos estudos usem modelos continuamente treinados (como mBERT ou BERTimbau adaptados) em vez de arquiteturas já voltadas à linguagem informal.

Tais desafios explicam por que ainda não há solução “pronta” única. No entanto, avanços recentes indicam caminhos: a criação de corpora específicos (mesmo que pequenos), a utilização de transferência de aprendizado (ex.: fine-tuning de BERTimbau em domínios alvo) e a combinação de múltiplas fontes de dados estão sendo explorados para mitigar essas limitações. A pesquisa destaca a importância de considerar gírias locais, emojis e multimodalidade (quando disponível) nas próximas versões de corpora, e modelos.

5 CONTRIBUIÇÕES DO ESTUDO

Este estudo contribui para a área de Processamento de Linguagem Natural (PLN) em português ao mapear, de forma sistemática, corpora emocionais multiplataforma produzidos entre 2021 e 2024. A revisão destaca avanços recentes e identifica lacunas relevantes, como a baixa cobertura de Facebook e Instagram, a falta de diretrizes padronizadas de anotação e a escassez de modelos treinados especificamente em linguagem informal de redes sociais.

Além disso, o artigo organiza os recursos disponíveis (corpora, léxicos e modelos) de modo comparativo, oferecendo uma síntese que facilita a compreensão do estado da arte e orienta futuras pesquisas. Assim, o trabalho não apenas reúne iniciativas já existentes, mas também aponta caminhos para inovações metodológicas e tecnológicas na análise de sentimentos em língua portuguesa.

6 CONCLUSÃO

Esta revisão sistemática demonstrou que, nos últimos anos, houve um aumento expressivo no número de iniciativas voltadas à criação de corpora emocionais anotados em língua portuguesa, com destaque para dados provenientes de redes sociais. Foram mapeados recursos importantes como o TweetSentBR (BRUM; VOLPE NUNES, 2018), voltado a tweets televisivos; o SS-PT (ALMEIDA et al., 2022), com anotação de *stance* e polaridade em tweets políticos; o B2T (KAKIMOTO et al., 2024), com dados do setor bancário; e o Reddit-BR (PIORINO et al., 2024), que amplia o escopo para fóruns de discussão. Esses corpora

representam um avanço importante na diversidade temática e estrutural de dados disponíveis para o português.

Adicionalmente, esta revisão identificou recursos léxicos amplamente utilizados, como o SentiLex-PT, o OpLexicon e o LIWC-PT, bem como modelos de PLN robustos como o BERTimbau, o T5-PT e o LaBSE. Esses modelos superam abordagens clássicas e multilíngues em tarefas de sentimento, consolidando-se como ferramentas centrais para pesquisas atuais em português.

Apesar dos avanços, os resultados também evidenciaram desafios persistentes. A escassez de dados anotados em português, como destacado por Kakimoto et al. (2024), ainda limita o treinamento e a generalização de modelos. Soma-se a isso a baixa disponibilidade de corpora realmente multiplataforma, já que a maioria dos recursos está concentrada no Twitter/X, com pouca representatividade de Facebook, Instagram ou Reddit.

A complexidade linguística do português brasileiro, associada a regionalismos, gírias, ironias e expressões culturais, agrava a dificuldade de anotação consistente. Outro ponto crítico é a ausência de diretrizes unificadas de anotação: cada corpus adota esquemas próprios de polaridade e critérios subjetivos, o que dificulta o reuso e a comparação entre trabalhos.

A heterogeneidade de esquemas e critérios subjetivos dificulta o reuso e a comparação entre estudos. Faltam modelos adaptados ao português informal e capazes de identificar sentimentos sutis (sarcasmo, ambiguidade), o que ressalta a necessidade de metodologias padronizadas. Os corpora mapeados têm potencial para aplicações em saúde pública, atendimento automático e análise de reputação, contribuindo para a inovação tecnológica.

Conclui-se portanto que apesar dos avanços recentes, o português ainda carece de uma infraestrutura ampla e consolidada para tarefas de análise de sentimentos. Para alcançar resultados comparáveis aos obtidos em inglês, será necessário investir em soluções colaborativas, de código aberto e linguisticamente sensíveis ao contexto brasileiro.

REFERÊNCIA

ALMEIDA, R. et al. SS-PT: a stance and sentiment dataset from Portuguese quoted tweets. In: **LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC)**, 2022, Marseille. Anais [...]. Marseille: ELRA, 2022.

BALAGE FILHO, P. et al. LIWC-PT: adaptação e avaliação do LIWC para o português brasileiro. In: **INTERNATIONAL CONFERENCE ON SENTIMENT ANALYSIS**, 2013. Anais [...]. [S.l.: s.n.], 2013.

BRUM, M.; VOLPE NUNES, M. TweetSentBR: um corpus anotado para análise de sentimentos em tweets brasileiros. In: **CONFERÊNCIA NACIONAL DE INTELIGÊNCIA ARTIFICIAL**, 2018. Anais [...]. [S.l.: s.n.], 2018.

CARMO, M. P. et al. Continued pretraining of T5 models for the Portuguese language: PTT5. In: **BRAZILIAN CONFERENCE ON ARTIFICIAL INTELLIGENCE AND NATURAL LANGUAGE PROCESSING (BRACIS)**, 2020. Anais [...]. [S.l.: s.n.], 2020.

CARVALHO, A.; SILVA, F. SentiLex-PT: principais características e potencialidades. In: **CONFERÊNCIA BRASILEIRA DE LINGUÍSTICA COMPUTACIONAL**, 2015. Anais [...]. [S.l.: s.n.], 2015.

CAVALCANTI, S.; SILVA, R.; PARDO, E. A survey of sentiment analysis in the Portuguese language. **Journal of Information and Language Technology**, [S.l.], v. 10, n. 3, p. 200–225, 2020.

DATAREPORTAL. Digital 2025: Brazil — global digital insights. 2025. Disponível em: https://datareportal.com/reports/digital-2025-brazil. Acesso em: 02 ago. 2025.

FENG, F.; YANG, Y.; CER, D.; ARIVAZHAGAN, N.; WANG, W. Language-agnostic BERT sentence embedding (LaBSE). arXiv preprint arXiv:2007.01852, 2020. Disponível em: https://arxiv.org/abs/2007.01852. Acesso em: 10 set. 2025.

GAMALLO, P.; GARCÍA, M.; PIÑEIRO, C.; MARTÍNEZ-CASTAÑO, R.; PICHEL, J. Linguakit: a Big Data-based multilingual tool for linguistic analysis and information extraction. In: **Proceedings of the Second International Workshop on Advances in Natural Language Processing (ANLP)**, 2018. Anais, p. 239–244, 2018. Disponível em: https://grupolys.org/biblioteca/GamGarPinMarPic2018a.pdf. Acesso em: 10 set. 2025.

KAKIMOTO, H. et al. B2T: a dataset of tweets in Portuguese language about Brazilian banks. **arXiv preprint**, arXiv:2401.12345, 2024. Disponível em: https://arxiv.org/abs/2401.12345. Acesso em: 05 jul. 2025.

LEITE, P. et al. ToLD-Br: a Portuguese dataset for toxic language detection. In: **CONFERÊNCIA INTERNACIONAL DE PROCESSAMENTO COMPUTACIONAL DE DISCURSO**, 2020. Anais [...]. [S.l.: s.n.], 2020.

PIAU, M.; LOTUFO, R.; NOGUEIRA, R. ptt5-v2: A closer look at continued pretraining of T5 models for the Portuguese language. arXiv preprint arXiv:2406.10806, 2024. Disponível em: https://arxiv.org/abs/2406.10806. Acesso em: 10 set. 2025.

PIORINO, L. et al. Reddit-BR: an annotated corpus of Brazilian Portuguese Reddit posts for sentiment analysis. In: **BRAZILIAN SYMPOSIUM ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL)**, 2024. Anais [...]. [S.l.: s.n.], 2024.

SILVA, T.; PARDO, E. Revisão de técnicas de análise de sentimento em português: desafios e perspectivas. **Revista Brasileira de Linguística Aplicada**, Belo Horizonte, v. 21, n. 2, p. 67–85, 2021.

SOUZA, F. H. F. et al. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS)**, 2020. Anais [...]. [S.l.: s.n.], 2020.

SOUZA, M.; VIEIRA, R.; BUSSETTI, D.; CHISHMAN, R.; ALVES, I. M. Construction of a Portuguese opinion lexicon from multiple resources. In: **Simpósio Brasileiro de Informação e Linguagem Humana (STIL)**, 2011. Anais..., 2011. Disponível em: https://www.inf.pucrs.br/linatural/wordpress/wp-content/uploads/2017/08/PROPOR_2012.pdf. Acesso em: 10 set. 2025.

SOUZA, N.; VIEIRA, A. OpLexicon: um léxico de opinião para a língua portuguesa. **Journal of Computational Linguistics**, [S.l.], v. 8, n. 1, p. 45–60, 2011.

TRAJANO, G. et al. OLID-Br: a multi-platform offensive language identification corpus for Brazilian Portuguese. **Journal of Language and Computation**, [S.l.], v. 5, n. 2, p. 123–137, 2023.

VARGAS, F. et al. HateBR: a dataset of hateful comments in Brazilian Portuguese collected from Instagram. In: **INTERNATIONAL CONFERENCE ON WEB AND SOCIAL MEDIA (ICWSM)**, 2022. Anais [...]. [S.l.: s.n.], 2022.

VIANNA, D.; CARNEIRO, F.; CARVALHO, J.; PLASTINO, A.; PAES, A. Sentiment analysis in Portuguese tweets: an evaluation of diverse word representation models. **Language Resources and Evaluation**, v. 58, p. 223–272, 2024. DOI: 10.1007/s10579-023-09661-4. Disponível em: https://doi.org/10.1007/s10579-023-09661-4. Acesso em: 10 set. 2025.