# **Uma Análise Aprofundada do Portuguese Hate Speech Expanded Dataset (TuPyE) e seu Posicionamento no Cenário da Deteção de Discurso de Ódio**

**1. Introdução**

* **Contexto:** O discurso de ódio online tornou-se um problema crescente e pervasivo, manifestando-se de forma proeminente em plataformas de redes sociais.1 A deteção e mitigação deste fenómeno representam um desafio significativo, especialmente na língua portuguesa. As particularidades linguísticas do português, incluindo seu rico vocabulário (mais de 600 mil palavras), gramática complexa e variações regionais, podem ocultar o discurso de ódio ou conferir significados ofensivos a expressões que, em outros contextos, seriam inócuas.2 Além disso, as nuances culturais inerentes aos países lusófonos adicionam outra camada de complexidade à tarefa.2
* **Problemática:** O desenvolvimento de modelos eficazes de aprendizagem de máquina (ML) para a deteção automática de discurso de ódio depende criticamente da disponibilidade de conjuntos de dados anotados, em larga escala e de alta qualidade.5 No entanto, historicamente, tem havido uma escassez relativa desses recursos para a língua portuguesa em comparação com o inglês 2, embora esse viés venha diminuindo nos últimos anos.14 A criação manual de corpora anotados é um processo caro e demorado, exigindo conhecimento linguístico e pragmático considerável.5
* **Objetivo do Relatório:** Este relatório visa fornecer uma análise abrangente do Portuguese Hate Speech Expanded Dataset (TuPyE).20 O estudo examinará detalhadamente sua criação, estrutura e conteúdo. Além disso, comparará o TuPyE com outros conjuntos de dados relevantes, tanto em português quanto em outras línguas (com foco em metodologias significativas), identificará suas características distintivas e avaliará criticamente seus pontos fortes e limitações no contexto do panorama atual de recursos para deteção de discurso de ódio.
* **Âmbito:** O âmbito deste relatório abrange a análise da metodologia de criação de datasets, esquemas de anotação, fontes de dados, tamanho dos corpora e uma análise comparativa focada em aspetos relevantes para a investigação em Processamento de Linguagem Natural (PLN) e ML.
* **Público-Alvo:** Este documento destina-se a investigadores, profissionais e estudantes das áreas de PLN, ML e computação social, que possuam interesse na deteção de discurso de ódio e na análise de conjuntos de dados linguísticos.

**2. O Portuguese Hate Speech Expanded Dataset (TuPyE): Uma Análise Detalhada**

* **2.1. Origens e Objetivos:**
  + **Criadores:** O dataset TuPyE foi introduzido por Oliveira et al..20 O projeto tem suas raízes no trabalho de tese de Felipe Oliveira e resulta de uma colaboração envolvendo a Universidade Federal do Rio de Janeiro (UFRJ) e o Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE).27
  + **Motivação:** A principal motivação para a criação do TuPyE foi suprir a lacuna existente de corpora anotados em larga escala para a deteção de discurso de ódio em português.13 O objetivo era facilitar o desenvolvimento e o avanço de modelos de deteção mais sofisticados, utilizando técnicas de ML e PLN.20 O TuPyE foi concebido com a ambição de se tornar o maior corpus público e anotado em português para esta tarefa específica.13
  + **Disponibilidade:** O dataset é disponibilizado publicamente através de um repositório no GitHub 25, com uma abordagem de código aberto para incentivar a colaboração e o avanço da investigação na área pela comunidade científica.13
* **2.2. Metodologia: Consolidação e Expansão:**
  + **Definição de "Expandido":** O termo "Expandido" no nome do dataset refere-se ao processo de aumento e consolidação de recursos existentes. A metodologia central envolveu a integração de vários conjuntos de dados previamente publicados com um novo conjunto de dados, resultando num recurso unificado de maior escala e âmbito.20 Trata-se, portanto, de uma expansão alcançada pela fusão de corpora.
  + **Datasets Consolidados:** O TuPyE integra documentos rotulados de três estudos proeminentes anteriores sobre discurso de ódio em português 13:
    - **Fortuna et al. (2019):** Um conjunto de 5.670 tweets (abrangendo português europeu e brasileiro), originalmente anotados por três avaliadores quanto à presença ou ausência de discurso de ódio. Para a integração no TuPyE, foi aplicado um mecanismo de votação por maioria simples para determinar a classificação binária final de cada tweet.13
    - **Leite et al. (2020) / ToLD-Br:** Um dataset extensivo com 21.000 tweets em português brasileiro, anotados por 129 voluntários (cada tweet avaliado por três pessoas) em sete categorias: não tóxico, LGBTfobia, obsceno, insulto, racismo, misoginia e xenofobia. Para alinhar com a definição operacional de discurso de ódio do TuPyE, textos rotulados exclusivamente como 'ofensivos' e/ou 'obscenos' foram excluídos. A classificação final também foi determinada por votação majoritária.13
    - **Vargas et al. (2022) / HateBR:** Uma coleção de 7.000 comentários do Instagram de políticos brasileiros, rotulados por especialistas. Estes dados já haviam passado por um processo de classificação por maioria simples, não necessitando de protocolos adicionais de classificação para integração.13
  + **Novo Componente de Dados (TuPy):** Além dos datasets consolidados, o TuPyE incorpora 10.000 documentos originais e inéditos, anotados, provenientes do TuPy-Dataset base.20 Estes são tweets coletados em 2023.27 O processo de anotação para este componente envolveu três avaliações por documento, com a classificação final decidida por maioria (dois ou mais votos idênticos resultam em 1, caso contrário 0).27 É relevante notar que a criação deste novo componente proprietário, que decorreu ao longo de sete meses, contou com a colaboração de uma equipa de especialistas, incluindo um linguista, um advogado de direitos humanos, psicólogos comportamentais e investigadores de PLN/ML.13
  + **Diretrizes de Integração:** Foi seguida uma metodologia rigorosa para a integração, incluindo a aplicação consistente de votação por maioria e o mapeamento ou exclusão de categorias dos datasets originais, para garantir a consistência e compatibilidade dentro do corpus unificado TuPyE.13
  + **Heterogeneidade Resultante da Consolidação:** A estratégia central de consolidação do TuPyE, embora eficaz para alcançar uma escala significativa, introduz inevitavelmente heterogeneidade no dataset final. Os corpora constituintes (Fortuna, Leite, Vargas, TuPy) diferem nas suas origens: fontes de dados distintas (Twitter vs. Instagram) 13, metodologias de anotação variadas (especialistas vs. voluntários vs. não-especialistas) 5, e potenciais variações subtis nas definições iniciais das categorias, mesmo após os esforços de harmonização.13 Esta diversidade interna pode ser vista como uma faca de dois gumes: por um lado, pode tornar os modelos treinados no TuPyE mais robustos a diferentes estilos de linguagem e fontes; por outro, pode introduzir sinais potencialmente conflituosos durante o treino, levando a uma representação do discurso de ódio que, embora ampla, pode ser menos precisa do que a aprendida a partir de um dataset anotado de forma homogénea.
* **2.3. Características dos Dados:**
  + **Tamanho:** O dataset final compreende 43.668 documentos meticulosamente anotados.20 É explicitamente apresentado como o maior corpus público anotado em português para a deteção de discurso de ódio.13
  + **Fontes:** Os dados provêm de diversos contextos de redes sociais 20, principalmente do Twitter 15 e do Instagram 13, procurando refletir o discurso online real.
  + **Foco Linguístico:** O foco principal é o português brasileiro 1, embora o componente do dataset de Fortuna et al. também inclua português europeu.1
* **2.4. Esquema de Anotação:**
  + **Estrutura em Dois Níveis:** O processo de anotação aplicado ao corpus consolidado segue uma estrutura de dois níveis 13:
    - **Nível 1: Classificação Binária:** Agressivo vs. Não-Agressivo. A categoria "Agressivo" é definida como englobando tanto o discurso de ódio quanto a linguagem ofensiva.13
    - **Nível 2: Categorias de Discurso de Ódio Multi-Rótulo:** Documentos classificados como "Agressivos" no primeiro nível são subsequentemente atribuídos a uma ou mais categorias específicas de discurso de ódio.13
  + **Categorias de Ódio:** As categorias multi-rótulo específicas definidas são: ageísmo (etarismo), aporofobia (desprezo por pessoas pobres), body shame (vergonha do corpo), capacitismo, LGBTfobia, político (ódio baseado em ideologia política), racismo, intolerância religiosa, misoginia, xenofobia e 'outros'.13
  + **Processo de Anotação:** A anotação final representa o consenso dos anotadores.25 Para o componente TuPy original, utilizaram-se 3 anotadores com votação majoritária.27 Para os componentes fundidos, a votação majoritária foi aplicada para padronizar os rótulos.13 A definição do esquema e a anotação do novo componente contaram com a equipa de especialistas.13
  + **Granularidade vs. Hierarquia:** O TuPyE adota uma estrutura de multi-rótulos plana para as categorias de ódio, o que o diferencia da abordagem hierárquica utilizada no dataset de Fortuna et al..15 Enquanto a descrição do TuPyE lista múltiplas categorias que podem ser aplicadas simultaneamente a um único texto 13, o dataset de Fortuna é explicitamente caracterizado pelo seu "esquema de anotação hierárquico".5 Esta escolha de design representa um compromisso: a estrutura plana do TuPyE pode simplificar o processo de anotação e o treino de modelos de classificação multi-rótulo, mas perde as relações semânticas e as nuances entre categorias que uma hierarquia pode capturar (por exemplo, sub-tipos de discriminação). Por outro lado, a hierarquia de Fortuna oferece maior profundidade semântica, mas pode ser mais complexa para anotar de forma consistente e para modelar eficazmente.
* **2.5. Disponibilidade e Acesso:**
  + O dataset está disponível publicamente no GitHub.25 Recomenda-se a leitura do artigo de investigação associado para obter detalhes completos sobre o dataset e sua criação.20

**3. Levantamento de Datasets Comparáveis de Discurso de Ódio**

* **3.1. Datasets em Língua Portuguesa:**
  + **Fortuna et al. (2019) / HPHS Dataset:**
    - **Descrição:** Considerado um dataset pioneiro para o português.15 Contém 5.668 tweets 5 provenientes do Twitter 15, abrangendo tanto português europeu quanto brasileiro.1
    - **Anotação:** Utilizou dois esquemas distintos: uma anotação binária ('hate' vs. 'no-hate') realizada por não-especialistas, e um esquema hierárquico multi-rótulo detalhado (com 81 categorias) aplicado por especialistas.5 Cada tweet foi anotado por três pessoas.5 O trabalho destacou a subtileza do discurso de ódio e a variação no acordo inter-anotadores (IAA) entre as categorias.15
    - **Disponibilidade:** Disponível no GitHub 31, embora o texto dos tweets seja omitido devido aos termos da API do Twitter.32 Foi utilizado em investigações subsequentes.5
    - **Característica Chave:** Introdução de um esquema de anotação hierárquico para o português.15
  + **Leite et al. (2020) / ToLD-Br:**
    - **Descrição:** Um dataset de larga escala.23 Contém 21.000 tweets 1 do Twitter, em português brasileiro.1 A coleta foi realizada usando palavras-chave e visando tweets direcionados a figuras influentes.5
    - **Anotação:** Realizada por crowdsourcing, com 129 voluntários (42 selecionados visando diversidade demográfica), e cada tweet anotado por três pessoas.5 As categorias incluem: Não tóxico, LGBTfobia, Obsceno, Insulto, Racismo, Misoginia, Xenofobia.1 Está disponível em formato binário (tóxico/não-tóxico) e multi-rótulo (contagem de votos por categoria).21
    - **Disponibilidade:** Acessível via GitHub 22 e Hugging Face.21 Utilizado em várias pesquisas posteriores.14
    - **Característica Chave:** Grande escala e processo de anotação por crowdsourcing com foco na diversidade demográfica dos anotadores.21
  + **Vargas et al. (2022) / HateBR:**
    - **Descrição:** Primeiro dataset de larga escala anotado por especialistas para comentários do Instagram em português brasileiro.28 Contém 7.000 comentários 13 extraídos de posts de políticos brasileiros.13
    - **Anotação:** Anotado manualmente por três especialistas 13, com alto IAA reportado.28 Possui três camadas de anotação: (1) Classificação binária (ofensivo/não-ofensivo), (2) Nível de ofensividade (altamente, moderadamente, levemente ofensivo), (3) Alvos do discurso de ódio (xenofobia, racismo, homofobia, misoginia, intolerância religiosa, partidarismo, etc.).13
    - **Disponibilidade:** Disponível no GitHub.28 Utilizado em trabalhos subsequentes.14 Existe uma versão atualizada (HateBR 2.0) que substituiu comentários muito similares ou de uma única palavra.35
    - **Característica Chave:** Anotação por especialistas, esquema de anotação em camadas e fonte de dados específica (Instagram político).28
  + **Salles et al. (2025) / HateBRXplain:**
    - **Descrição:** Uma extensão ou versão do HateBR focada em explicabilidade.1 Baseado nos 7.000 comentários do Instagram do HateBR, com 3.500 comentários ofensivos anotados com justificativas (rationales).1
    - **Anotação:** Anotação manual por dois anotadores, adicionando justificativas baseadas em humanos (segmentos de texto que suportam a classificação).1
    - **Característica Chave:** Primeiro benchmark em português para a deteção *explicável* de discurso de ódio.1
  + **Carvalho et al. (2022) / FIGHT:**
    - **Descrição:** Dataset grande focado em alvos e contexto específicos.39 Contém 63.450 tweets 39 de utilizadores localizados em Portugal.40 Abrange o período pré e pós-declaração da pandemia de Covid-19 (Agosto 2018 - Outubro 2021).40
    - **Coleta:** Dados geolocalizados, filtrados com base em léxicos, focando nas comunidades afrodescendente, cigana e LGBTQ+.3
    - **Anotação:** Realizada uma anotação manual inicial numa amostra para distinguir entre discurso de ódio, ofensivo, ambíguo e não relevante.40 Utilizado em experiências de anotação semi-supervisionada 5 e para desenvolver modelos específicos de domínio como o HateBERTimbau.3
    - **Característica Chave:** Foco em alvos específicos (minorias em Portugal), contexto temporal (Covid-19) e dados geolocalizados.39
  + **Outros Recursos em Português:**
    - **OFFCOMBR (de Pelle & Moreira, 2017):** Um trabalho anterior.1 Tamanho: 1.250 comentários. Anotação: Binária (ofensivo/não-ofensivo) e 6 categorias de ódio.1
    - **Portuguese Hate Speech Superset (Manueltonneau):** Um esforço recente para unificar datasets disponíveis (ToLD-Br, Fortuna, HateBR, TuPyE) num "superset" de 43.222 posts.14 Os rótulos foram binarizados (ódio/não-ódio) e metadados como alvo, fonte, origem do dataset e número de anotadores foram incluídos.14 Visa facilitar o treino/avaliação de modelos e o estudo do discurso de ódio.14
  + **Tendência de Consolidação e Especialização:** A existência tanto do TuPyE 20 quanto do Portuguese Hate Speech Superset 14 aponta para uma tendência de agregar datasets menores em recursos maiores, potencialmente para melhorar o treino de modelos. Ao mesmo tempo, datasets como o FIGHT 39 e o HateBRXplain 1 demonstram uma especialização crescente, focando em contextos específicos (geotemporal, grupos-alvo) ou tarefas particulares (explicabilidade). Esta evolução dupla sugere que o campo está a mover-se simultaneamente para a criação de recursos amplos e de larga escala através da agregação, e para o desenvolvimento de datasets especializados que permitem análises mais profundas ou abordam tarefas específicas que os datasets agregados maiores podem não suportar tão bem. A escolha entre escala e especificidade torna-se assim uma consideração importante para os investigadores.
* **3.2. Datasets Relevantes em Outras Línguas (Significado Metodológico):**
  + **Davidson et al. (2017) - Inglês:**
    - **Descrição:** Um dataset inicial amplamente citado.7 Tamanho: aprox. 24.800 tweets.7 Fonte: Twitter.7 Coleta: Baseada em léxico (HateBase).7
    - **Anotação:** Crowdsourced (Figure-Eight/Crowdflower).7 Três classes: Discurso de Ódio, Linguagem Ofensiva, Nenhum.7
    - **Relevância Metodológica:** Frequentemente criticado por misturar linguagem "de ódio" e "ofensiva" 9, evidenciando a ambiguidade e a dificuldade na definição e anotação destas categorias.30 O seu esquema de anotação influenciou trabalhos posteriores (ex: HateXplain 44). Demonstra os desafios da amostragem baseada em léxico e do crowdsourcing.7 Foi extensivamente utilizado em estudos de generalização entre datasets.8
  + **ETHOS (Mollas et al., 2020) - Inglês:**
    - **Descrição:** Dataset focado em anotação multi-rótulo e na abordagem do desequilíbrio de classes.10 Tamanho: Relativamente pequeno; Binário: 998 comentários (56.6% não-ódio, 43.4% ódio); Multi-rótulo: 433 comentários de ódio.13 Fonte: Comentários do YouTube e Reddit.10
    - **Anotação:** Validada via crowdsourcing (Figure-Eight).10 Categorias multi-rótulo: violência, direcionamento (pessoa/grupo), género, raça, origem nacional, deficiência, religião, orientação sexual.49 Menciona um procedimento de amostragem ativa.10
    - **Relevância Metodológica:** Explora a anotação multi-rótulo para além da simples classificação binária.10 Aborda explicitamente o desequilíbrio de classes.10 Utiliza fontes diversas para além do Twitter.48 Evidencia os compromissos entre o tamanho do dataset e a granularidade/qualidade da anotação.10
  + **Learning From the Worst (Vidgen et al., 2021) - Inglês:**
    - **Descrição:** Dataset criado usando uma abordagem dinâmica e adversarial.46 Tamanho: aprox. 40.000-41.000 entradas.46 Fonte: Gerado sinteticamente por anotadores treinados que tentam enganar modelos.46
    - **Anotação:** Processo "human-and-model-in-the-loop" ao longo de 4 rondas.46 Rótulos detalhados para entradas de ódio (tipo, alvo, pivô).54 Dataset equilibrado (54% ódio).53 Inclui exemplos desafiadores de perturbação/contrastivos.54
    - **Relevância Metodológica:** Representa uma abordagem fundamentalmente diferente para a criação de datasets, focando na robustez do modelo e em exemplos desafiadores, em vez de apenas coletar dados existentes.46 Demonstra como a coleta dinâmica de dados pode melhorar o desempenho do modelo e reduzir as taxas de erro em entradas adversariais.46 Oferece uma forma de gerar datasets mais equilibrados.54
  + **Expansão por Consolidação vs. Geração Adversarial:** A comparação entre a abordagem de "expansão" do TuPyE (consolidação) 13 e a metodologia do "Learning From the Worst" 46 revela filosofias distintas na criação de datasets. O TuPyE expande agregando dados reais existentes, visando escala e representatividade do discurso online atual.13 Em contraste, "Learning From the Worst" gera dados novos e sintéticos, especificamente concebidos para desafiar modelos e melhorar a sua robustez.46 O TuPyE é, portanto, mais adequado para modelar o discurso de ódio típico encontrado atualmente nas redes sociais em português. A metodologia de "Learning From the Worst" fornece uma abordagem (e dados) para construir modelos potencialmente mais resilientes a formulações de discurso de ódio novas ou enganosas, embora os dados em si possam ser menos representativos do conteúdo online médio.

**4. Análise Comparativa**

* **4.1. Tabela Comparativa:**

| **Dataset Name** | **Língua(s)** | **Fonte(s) Primária(s)** | **Tamanho (Exemplos)** | **Método Anotação** | **Esquema Anotação** | **Característica/Objetivo Chave** | **Técnica Expansão/Geração** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TuPyE** | PT-BR (maioria), PT-EU | Twitter, Instagram | 43.668 | Misto (Consol.) | Binário (Agressivo); Multi-rótulo (10+ categorias) | Maior corpus PT consolidado; Multi-rótulo específico | Consolidação |
| Fortuna et al. (HPHS) | PT-BR, PT-EU | Twitter | 5.668 | Misto (Exp/NãoExp) | Binário (Ódio); Hierárquico Multi-rótulo (81 cats) | Pioneiro PT; Anotação hierárquica | N/A |
| ToLD-Br | PT-BR | Twitter | 21.000 | Crowdsourcing | Binário (Tóxico); Multi-rótulo (6 cats, votos) | Larga escala PT; Anotadores diversos | N/A |
| HateBR | PT-BR | Instagram | 7.000 | Especialistas | Binário (Ofensivo); Nível Ofens.; Alvos Ódio (Camadas) | Anotação especialista; Camadas; Fonte específica (política) | N/A |
| FIGHT | PT-EU | Twitter | 63.450 | Investigadores (trial) | Ódio/Ofensivo/Ambíguo/Não-relevante (trial) | Foco minorias PT; Contexto COVID; Geolocalizado | N/A |
| Davidson et al. (2017) | EN | Twitter | ~24.800 | Crowdsourcing | Ternário (Ódio, Ofensivo, Nenhum) | Benchmark EN; Influente; Conflação Ódio/Ofensivo | N/A |
| ETHOS (2020) | EN | YouTube, Reddit | 998 (Bin), 433 (Multi) | Crowdsourcing | Binário (Ódio); Multi-rótulo (8 cats) | Multi-rótulo; Fontes diversas; Foco desequilíbrio | N/A |
| Learning From the Worst | EN | Sintético | ~40.000 | Treinados (Dinâmico) | Multi-rótulo (Tipo, Alvo); Equilibrado | Geração dinâmica/adversarial; Robustez; Equilibrado | Geração Dinâmica |

\*Nota: PT-BR = Português Brasileiro, PT-EU = Português Europeu, EN = Inglês. "Consol." refere-se à consolidação de múltiplos datasets. "N/A" indica Não Aplicável.\*

* **4.2. Tamanho e Diversidade do Dataset:**
  + **Escala:** Em termos de escala absoluta para o português, TuPyE 25 e FIGHT 40 destacam-se como os maiores datasets discutidos, ultrapassando significativamente Fortuna et al. 25, HateBR 13 e OFFCOMBR.1 ToLD-Br 25 também apresenta um tamanho considerável. O Portuguese Hate Speech Superset 14 atinge uma escala semelhante à do TuPyE através da fusão. Estes tamanhos são comparáveis a alguns datasets influentes em inglês, como o de Davidson et al. 7, mas podem ser menores que outros esforços de larga escala em inglês. A obtenção de grande escala para dados de discurso de ódio em português parece depender fortemente da combinação de esforços existentes ou de campanhas de coleta extensivas, talvez refletindo um número menor de projetos individuais de anotação em larga escala em comparação com o que ocorreu inicialmente para o inglês.14
  + **Diversidade de Fontes:** O TuPyE combina dados do Twitter e Instagram 25, oferecendo uma mistura de plataformas. Outros datasets focam-se predominantemente numa única fonte: Twitter (Fortuna 15, ToLD-Br 29, FIGHT 40, Davidson 7), Instagram (HateBR 28), ou YouTube/Reddit (ETHOS 48). "Learning From the Worst" é sintético.54 A diversidade de fontes no TuPyE pode expor os modelos a uma gama mais ampla de estilos linguísticos, enquanto datasets de fonte única podem ser mais adequados para estudar fenómenos específicos de uma plataforma.
* **4.3. Complexidade e Granularidade da Anotação:**
  + **Evolução:** Observa-se uma evolução nos esquemas de anotação em português, partindo de classificações binárias simples (presentes como uma camada em Fortuna 15 e outros) para esquemas mais complexos: multi-categoria (ToLD-Br 21, OFFCOMBR 1), multi-rótulo (TuPyE 25, ETHOS 48), hierárquico (segunda camada de Fortuna 15), em camadas com múltiplos níveis de análise (HateBR 28), e incorporando justificativas (HateBRXplain 1).
  + **Profundidade vs. Amplitude:** Existe um espectro na granularidade. Alguns datasets oferecem grande profundidade, como as 81 categorias hierárquicas de Fortuna 15, os níveis de ofensividade e alvos de HateBR 28, ou as justificativas de HateBRXplain.1 Outros, como TuPyE 25 e ToLD-Br 21, oferecem uma amplitude de categorias de ódio, mas numa estrutura mais plana. A escolha entre profundidade e amplitude depende dos objetivos da investigação.
  + **Ódio vs. Ofensivo:** A distinção (ou a sua ausência) entre discurso de ódio e linguagem ofensiva geral é um ponto crítico e variável entre os datasets. Davidson et al. 9 é frequentemente apontado por misturá-los. TuPyE tenta separá-los no Nível 1 (Agressivo vs. Não-Agressivo) 13, enquanto HateBR os trata em camadas distintas.28 Fortuna foca-se mais estritamente em 'ódio'.15 As dificuldades inerentes à definição e distinção destes conceitos são um desafio recorrente na área.3
* **4.4. Abordagens Metodológicas:**
  + **Coleta:** As estratégias de coleta variam: métodos baseados em léxicos (Fortuna 5, Davidson 7, FIGHT 40) podem introduzir viés de léxico 45; o direcionamento a plataformas/utilizadores específicos (ToLD-Br 29, HateBR 28) pode capturar contextos particulares; e a geração dinâmica (Learning From the Worst 54) foca na robustez. Cada método tem implicações nos tipos de dados coletados e nos potenciais vieses introduzidos.
  + **Anotação:** A escolha entre anotação por especialistas (HateBR 28, especialistas de Fortuna 15, novo componente TuPy 13) e crowdsourcing por não-especialistas ou voluntários (ToLD-Br 21, não-especialistas de Fortuna 15, Davidson 7, ETHOS 48) envolve compromissos entre custo, escala, consistência e qualidade. A anotação por especialistas pode oferecer maior precisão, enquanto o crowdsourcing permite maior escala, mas pode requerer validação mais rigorosa e estar sujeito a maior variabilidade ou viés dos anotadores.3 Considerações sobre o acordo inter-anotadores (IAA) são importantes em ambos os casos.7
  + **Expansão/Geração:** A abordagem de consolidação do TuPyE 20 e do Portuguese Superset 14 contrasta com a geração adversarial de "Learning From the Worst" 54, representando diferentes estratégias para aumentar a quantidade ou a robustez dos dados de treino.

**5. Características Distintivas do TuPyE**

Com base na análise comparativa, as seguintes características distinguem o TuPyE no panorama dos datasets de discurso de ódio:

* **Liderança em Escala Consolidada para Português:** O TuPyE destaca-se como o maior recurso *público* resultante da *consolidação* de datasets existentes para a deteção de discurso de ódio em português.13
* **Estratégia de Consolidação Explícita:** A sua metodologia primária de "expansão" é a fusão deliberada de vários datasets portugueses conhecidos e estabelecidos.13
* **Esquema Multi-Rótulo Específico:** Apresenta um conjunto particular de categorias de ódio multi-rótulo numa estrutura plana (incluindo ageísmo, aporofobia, etc.), diferenciando-se dos esquemas puramente binários, hierárquicos ou baseados em níveis de outros datasets portugueses.13
* **Fontes de Dados Mistas:** A sua composição inclui dados do Twitter e do Instagram, oferecendo potencialmente uma cobertura mais ampla dos estilos linguísticos das redes sociais em comparação com datasets focados numa única plataforma.13
* **Anotação em Dois Níveis:** A estrutura de classificação inicial Agressivo/Não-Agressivo, seguida pela categorização específica do tipo de ódio, representa uma escolha arquitetónica distinta para a anotação.13

**6. Avaliação: Pontos Fortes e Limitações**

A análise dos diferentes datasets revela que não existe um recurso "perfeito"; cada um envolve compromissos inerentes entre fatores como tamanho, profundidade e qualidade da anotação, representatividade da fonte, custo de criação e objetivos específicos da investigação. O TuPyE, por exemplo, prioriza a escala e a amplitude através da consolidação para o português, enquanto outros datasets podem priorizar a anotação por especialistas (HateBR), a profundidade hierárquica (Fortuna), contextos específicos (FIGHT), explicabilidade (HateBRXplain) ou robustez através da geração (Learning From the Worst). A seleção do dataset mais adequado é, portanto, dependente do contexto da investigação, exigindo uma avaliação cuidadosa das características de cada recurso em relação à questão de investigação específica, aos recursos computacionais disponíveis e à tolerância a diferentes tipos de vieses ou limitações potenciais.8

* **6.1. Dataset TuPyE:**
  + **Pontos Fortes:**
    - **Tamanho:** Sendo o maior recurso público consolidado para o português, oferece um volume de dados substancial que pode permitir o treino de modelos mais generalizáveis e robustos.13
    - **Consolidação:** Reutiliza esforços de anotação anteriores, economizando recursos e combinando dados de fontes diversas e estabelecidas na comunidade de investigação portuguesa.20
    - **Granularidade Multi-Rótulo:** Fornece uma classificação mais fina do que rótulos binários, capturando diferentes facetas do ódio.25 Inclui categorias relevantes e específicas como aporofobia, body shame e capacitismo.25
    - **Diversidade de Fontes:** A combinação de dados do Twitter e Instagram pode oferecer uma representação mais ampla da linguagem utilizada nas redes sociais.25
    - **Acessibilidade:** A sua disponibilidade aberta no GitHub facilita o acesso e a utilização pela comunidade de investigação.25
  + **Limitações:**
    - **Heterogeneidade:** A fusão de datasets com diferentes protocolos de anotação originais, tipos de anotadores e origens de plataforma pode introduzir inconsistências. A qualidade geral depende da qualidade dos datasets subjacentes que foram fundidos.
    - **Profundidade da Anotação:** O esquema multi-rótulo plano, embora informativo, carece da informação relacional de uma hierarquia (como a de Fortuna 15) ou do poder explicativo das justificativas (como em HateBRXplain 1).
    - **Artefatos da Consolidação:** A aplicação de votação majoritária e mapeamento de categorias pode obscurecer nuances presentes nos datasets originais ou introduzir vieses. A exclusão de categorias como 'ofensivo'/'obsceno' do ToLD-Br pode remover casos limítrofes relevantes.13
    - **Vieses Potenciais:** O dataset herda potenciais vieses das plataformas de origem e dos métodos originais de coleta e anotação.45 Há informação limitada sobre a diversidade demográfica dos anotadores do *novo* componente TuPy em comparação com o objetivo explícito de diversidade do ToLD-Br.21
* **6.2. Outros Datasets (Visão Geral Breve):**
  + **Datasets Portugueses:**
    - **Fortuna:** Ponto forte (pioneirismo, profundidade hierárquica 15); Limitação (tamanho menor 15, problemas de acesso ao texto 32).
    - **ToLD-Br:** Ponto forte (grande tamanho 29, diversidade de anotadores 21); Limitação (variabilidade na qualidade da anotação por voluntários, categorias mais simples que Fortuna/HateBR 21).
    - **HateBR:** Ponto forte (anotação por especialistas 28, esquema em camadas 28, contexto específico da fonte 28); Limitação (menor que TuPyE/ToLD-Br/FIGHT 28, específico do Instagram 28).
    - **HateBRXplain:** Ponto forte (explicabilidade 1); Limitação (foco no subconjunto ofensivo 1, herda limitações do HateBR).
    - **FIGHT:** Ponto forte (muito grande 40, alvos/contexto específicos 40, geolocalizado 40); Limitação (coleta potencialmente ruidosa 40, foco específico pode limitar a generalização).
    - **Superset:** Ponto forte (grande tamanho através da fusão 14); Limitação (rótulos binarizados perdem granularidade 14, herda problemas de heterogeneidade 14).
  + **Datasets de Outras Línguas:**
    - **Davidson:** Ponto forte (benchmark amplamente utilizado 8); Limitação (conflação ódio/ofensivo 9, problemas de crowdsourcing 7).
    - **ETHOS:** Ponto forte (multi-rótulo 48, fontes diversas 48, foco no desequilíbrio 13); Limitação (tamanho muito pequeno 48).
    - **Learning From the Worst:** Ponto forte (foco na robustez 54, equilibrado 54, metodologia inovadora 54); Limitação (representatividade dos dados sintéticos 53, custo/complexidade da geração 54).

**7. Conclusão e Direções Futuras**

* **Sumário das Descobertas:** O TuPyE representa um avanço significativo para a investigação em deteção de discurso de ódio em português, principalmente devido à sua escala, alcançada através de uma estratégia de consolidação de recursos existentes. As suas características incluem um esquema de anotação multi-rótulo plano com categorias específicas e a incorporação de dados de múltiplas redes sociais. Os seus principais pontos fortes residem na sua dimensão e acessibilidade pública, enquanto as suas limitações estão associadas à heterogeneidade inerente ao processo de consolidação e a um esquema de anotação que privilegia a amplitude em detrimento da profundidade hierárquica ou explicativa.
* **Contribuição do TuPyE:** Ao agregar vários datasets conhecidos e adicionar novo material anotado, o TuPyE fornece um recurso valioso e de larga escala que pode reduzir a barreira de entrada para investigadores que trabalham com o português, facilitando o treino e a avaliação de modelos de ML mais poderosos, apesar da sua heterogeneidade intrínseca.
* **Tendências Gerais:** A análise do TuPyE e de datasets comparáveis revela tendências importantes na área, como a busca por maior escala através da consolidação de recursos, a crescente especialização de datasets para abordar contextos ou tarefas específicas (como explicabilidade ou análise de grupos-alvo), e a diversificação das metodologias de criação de datasets, incluindo abordagens dinâmicas e adversariais.
* **Direções Futuras:**
  + **Diversificação Contínua:** São necessários esforços contínuos para criar datasets diversificados em português, explorando potencialmente mais plataformas (além de Twitter e Instagram), diferentes dialetos ou socioletos, e cobrindo uma gama ainda mais ampla de alvos de ódio.13
  + **Generalização e Transferência:** Dada a prevalência de abordagens de consolidação e a existência de múltiplos datasets, a investigação sobre a generalização de modelos entre diferentes datasets e a transferência de conhecimento entre línguas (cross-lingual transfer) torna-se crucial.8
  + **Profundidade Semântica e Contexto:** Há valor em incorporar uma compreensão semântica mais profunda, informações contextuais 28 e mecanismos de explicabilidade 1 em futuros datasets e modelos, indo além da simples classificação de textos isolados.
  + **Definição e Anotação:** O desafio contínuo de definir e anotar consistentemente formas subtis e complexas de discurso de ódio permanece uma área ativa de investigação.3
  + **Análise de Heterogeneidade:** Investigar especificamente o impacto da heterogeneidade do TuPyE no desempenho, robustez e vieses dos modelos treinados neste dataset seria uma via de investigação valiosa.

Em suma, o TuPyE é um recurso importante que reflete tanto os progressos quanto os desafios atuais na criação de datasets para a deteção de discurso de ódio em português. A sua análise e comparação com outros corpora fornecem um roteiro útil para investigadores que navegam neste campo complexo e em rápida evolução.

#### Referências citadas

1. HateBRXplain: A Benchmark Dataset with Human-Annotated Rationales for Explainable Hate Speech Detection in Brazilian Portuguese - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2025.coling-main.446.pdf>
2. Exploring Portuguese Hate Speech Detection with Transformers - LatinX in AI (LXAI) Research, acessado em abril 29, 2025, <https://research.latinxinai.org/papers/naacl/2024/pdf/Gabriel_Assis.pdf>
3. Automatic Hate Speech Detection in Portuguese Social Media Text, acessado em abril 29, 2025, <https://fenix.tecnico.ulisboa.pt/downloadFile/1126295043839755/resumo_tese.pdf>
4. Unveiling Patterns of Hate Speech in the Portuguese Sphere: A Social Network Analysis Approach⋆ - IPMU 2024, acessado em abril 29, 2025, <https://ipmu2024.inesc-id.pt/files/paper_1083.pdf>
5. Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains - DROPS, acessado em abril 29, 2025, <https://drops.dagstuhl.de/storage/01oasics/oasics-vol104-slate2022/OASIcs.SLATE.2022.11/OASIcs.SLATE.2022.11.pdf>
6. Automatic Hate Speech Detection in Portuguese Social Media Text - Fenix, acessado em abril 29, 2025, <https://fenix.tecnico.ulisboa.pt/downloadFile/1126295043839753/Tese.pdf>
7. In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2020.alw-1.18.pdf>
8. Towards generalisable hate speech detection: a review on obstacles and solutions - PMC, acessado em abril 29, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8237316/>
9. Hate Speech Detection Using Cross-Platform Social Media Data in English and German Language - SciTePress, acessado em abril 29, 2025, <https://www.scitepress.org/Papers/2024/130700/130700.pdf>
10. ETHOS: a multi-label hate speech detection dataset - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/357620424_ETHOS_a_multi-label_hate_speech_detection_dataset>
11. [2006.08328] ETHOS: an Online Hate Speech Detection Dataset - arXiv, acessado em abril 29, 2025, <https://arxiv.org/abs/2006.08328>
12. Hate speech detection: A comprehensive review of recent works - SenticNet, acessado em abril 29, 2025, <http://ww.sentic.net/hate-speech-detection.pdf>
13. arxiv.org, acessado em abril 29, 2025, <https://arxiv.org/pdf/2312.17704>
14. manueltonneau/portuguese-hate-speech-superset · Datasets at Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/datasets/manueltonneau/portuguese-hate-speech-superset>
15. A Hierarchically-Labeled Portuguese Hate Speech Dataset - ACL ..., acessado em abril 29, 2025, <https://aclanthology.org/W19-3510/>
16. View of The Impact of Data Augmentation on the Hate Speech Detection in Portuguese Language - Florida Online Journals, acessado em abril 29, 2025, <https://journals.flvc.org/FLAIRS/article/view/135307/139554>
17. (PDF) A Hierarchically-Labeled Portuguese Hate Speech Dataset - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/335783854_A_Hierarchically-Labeled_Portuguese_Hate_Speech_Dataset>
18. Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains, acessado em abril 29, 2025, <https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2022.11>
19. A Hierarchically-Labeled Portuguese Hate Speech Dataset - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/W19-3510.pdf>
20. TuPyE-Dataset Dataset - Papers With Code, acessado em abril 29, 2025, <https://paperswithcode.com/dataset/tupye-dataset>
21. JAugusto97/told-br · Datasets at Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/datasets/JAugusto97/told-br>
22. Toxic Language Dataset for Brazilian Portuguese (ToLD-Br) - GitHub, acessado em abril 29, 2025, <https://github.com/JAugusto97/ToLD-Br>
23. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2020.aacl-main.91/>
24. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis | Request PDF - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/348351425_Toxic_language_detection_in_social_media_for_Brazilian_Portuguese_new_dataset_and_multilingual_analysis>
25. Silly-Machine/TuPyE-Expanded-Brazilian-Hate-Speech-Dataset - GitHub, acessado em abril 29, 2025, <https://github.com/Silly-Machine/TuPyE-Expanded-Brazilian-Hate-Speech-Dataset>
26. Machine Learning Datasets - Papers With Code, acessado em abril 29, 2025, [https://paperswithcode.com/datasets?q=&v=lst&o=newest&lang=portuguese&mod=texts&page=1](https://paperswithcode.com/datasets?q&v=lst&o=newest&lang=portuguese&mod=texts&page=1)
27. Silly-Machine/TuPy-Brazilian-Hate-Speech-Dataset - GitHub, acessado em abril 29, 2025, <https://github.com/Silly-Machine/TuPy-Dataset>
28. franciellevargas/HateBR: HateBR is the first large-scale ... - GitHub, acessado em abril 29, 2025, <https://github.com/franciellevargas/HateBR>
29. Toxic Speech Detection in Portuguese: A Comparative Study of Large Language Models - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2024.propor-1.11.pdf>
30. Paula Fortuna - Semantic Scholar, acessado em abril 29, 2025, <https://www.semanticscholar.org/author/Paula-Fortuna/33621471>
31. A Hierarchically-Labeled Portuguese Hate Speech Dataset - GitHub, acessado em abril 29, 2025, <https://github.com/paulafortuna/Portuguese-Hate-Speech-Dataset>
32. Hate speech dataset annotated for Portuguese - B2SHARE - EUDAT, acessado em abril 29, 2025, <https://b2share.eudat.eu/records/9005efe2d6be4293b63c3cffd4cf193e>
33. ajdavidl/Portuguese-NLP: List of resources and tools developed with focus on Portuguese. - GitHub, acessado em abril 29, 2025, <https://github.com/ajdavidl/Portuguese-NLP>
34. How Good Is ChatGPT For Detecting Hate Speech In Portuguese?, acessado em abril 29, 2025, <https://sol.sbc.org.br/index.php/stil/article/download/25441/25262>
35. OLID annotation schema. Source: Prepared by the author, adapted from (Zampieri et al., 2019a) - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/figure/OLID-annotation-schema-Source-Prepared-by-the-author-adapted-from-Zampieri-et-al_fig2_370494219>
36. ToLD-Br Dataset - Papers With Code, acessado em abril 29, 2025, <https://paperswithcode.com/dataset/told-br>
37. The Impact of Data Augmentation on the Hate Speech Detection in Portuguese Language - Florida Online Journals, acessado em abril 29, 2025, <https://journals.flvc.org/FLAIRS/article/download/135307/139554/258956>
38. Francielle Vargas - GitHub Pages, acessado em abril 29, 2025, <https://franciellevargas.github.io/>
39. Hate Speech Dynamics Against African descent, Roma and LGBTQI Communities in Portugal - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2022.lrec-1.253/>
40. www.lrec-conf.org, acessado em abril 29, 2025, <http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.253.pdf>
41. View of MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection, acessado em abril 29, 2025, <https://ojs.aaai.org/index.php/ICWSM/article/view/31445/33605>
42. MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection - arXiv, acessado em abril 29, 2025, <https://arxiv.org/html/2401.06526v1>
43. Factoring Hate Speech: A New Annotation Framework to Study Hate Speech in Social Media - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2023.woah-1.21.pdf>
44. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection - AAAI Publications, acessado em abril 29, 2025, <https://ojs.aaai.org/index.php/AAAI/article/view/17745/17552>
45. Racial Bias in Hate Speech and Abusive Language Detection Datasets - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/335779647_Racial_Bias_in_Hate_Speech_and_Abusive_Language_Detection_Datasets>
46. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection - ACL Anthology, acessado em abril 29, 2025, <https://aclanthology.org/2021.acl-long.132.pdf>
47. MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection, acessado em abril 29, 2025, <https://irlab.org/metahate.html>
48. ETHOS Dataset - Papers With Code, acessado em abril 29, 2025, <https://paperswithcode.com/dataset/ethos>
49. iamollas/ethos · Datasets at Hugging Face, acessado em abril 29, 2025, <https://huggingface.co/datasets/iamollas/ethos>
50. arXiv:2106.15537v1 [cs.CL] 29 Jun 2021, acessado em abril 29, 2025, <https://arxiv.org/pdf/2106.15537>
51. ETHOS: an Online Hate Speech Detection Dataset - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/342197454_ETHOS_an_Online_Hate_Speech_Detection_Dataset>
52. intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset - GitHub, acessado em abril 29, 2025, <https://github.com/intelligence-csd-auth-gr/Ethos-Hate-Speech-Dataset>
53. Hate Speech Dataset Catalogue | hatespeechdata, acessado em abril 29, 2025, <https://hatespeechdata.com/>
54. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection, acessado em abril 29, 2025, <https://www.researchgate.net/publication/348079550_Learning_from_the_Worst_Dynamically_Generated_Datasets_to_Improve_Online_Hate_Detection>
55. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection, acessado em abril 29, 2025, <https://aclanthology.org/2021.acl-long.132/>
56. Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection | Request PDF - ResearchGate, acessado em abril 29, 2025, <https://www.researchgate.net/publication/353492564_Learning_from_the_Worst_Dynamically_Generated_Datasets_to_Improve_Online_Hate_Detection>
57. Hate Speech Classifiers Learn Normative Social Stereotypes - MIT Press Direct, acessado em abril 29, 2025, <https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00550/115347/Hate-Speech-Classifiers-Learn-Normative-Social>
58. Paula Fortuna, acessado em abril 29, 2025, <https://paulafortuna.github.io/>
59. arXiv:2305.14081v2 [cs.CL] 6 May 2024, acessado em abril 29, 2025, <https://arxiv.org/pdf/2305.14081>