Anatomy of Data Repositories for the Analysis and Detection of Toxicity in Portuguese

Lorena Souza Moreira[1], Paula Teresa M. Gibrim[1], Leonardo Rocha[2], Julio C. S. Reis[1]

[1]Universidade Federal de Viçosa (UFV), Viçosa - MG - Brazil
[2]Universidade Federal de São Jõao Del-Rei (UFSJ), São João del-Rei - MG - Brazil
Correspondence: lorenasouzamoreiraa@gmail.com

Abstract

The proliferation of online hate speech requires a rigorous examination of the datasets used to train detection models. In this work, we analyze six Brazilian Portuguese datasets annotated for hate speech or toxicity to investigate how their lexical "anatomy" and domain characteristics affect cross-domain generalization. We combine HurtLex-based lexical profiling with cross-dataset evaluation in a feature-based transfer-learning setup, using BERTimbau embeddings and an XGBoost classifier. Our analysis shows that, although the datasets share a broadly similar macro-level focus, they diverge substantially in how specific terms are used and labeled across platforms and topics. Results indicate that lexical breadth and annotation practices strongly predict transferability: datasets with broader and more heterogeneous toxic vocabulary yield better cross-domain performance, whereas resources with narrow, profanity-centered labeling lead to severe generalization gaps, even when lexical overlap is high. These findings underscore the impact of collection and labeling strategies on the curation and evaluation of Portuguese hate speech datasets. ***Warning!*** *This work and the referenced datasets contain examples of offensive and hateful language.*

1 Introduction

Social media platforms are used by approximately 5.41 billion people worldwide, representing about 65.7% of the global population (Kepios, 2025; Chaffey, 2025). These platforms foster diverse communities but have also enabled the spread of harmful practices, most notably hate speech (Boyd and Ellison, 2007; Gagliardone et al., 2015). According to the United Nations, hate speech includes communication that attacks or uses discriminatory language based on religion, ethnicity, nationality, race, gender, or other identity markers. In Brazil, although specific legislation is still evolving, existing legal frameworks address crimes resulting from prejudice, which has stimulated a growing body of research on automated detection (Poletto et al., 2021).

To operationalize hate speech detection, researchers typically use supervised machine learning models trained on annotated datasets. However, a critical challenge arises: the literature predominantly treats detection as a binary task, often overlooking the fact that datasets are artifacts of their collection strategies. A dataset collected during a political election, for example, may conflate "hate" with "political insults", while one focused on influencers may emphasize "harassment". These thematic biases can result in brittle models that fail to accurately capture nuances or generalize to new contexts (Fortuna et al., 2020; Cheng et al., 2022).

While recent studies have explored the characteristics of hate and toxicity datasets, few focus specifically on Portuguese or empirically measure how dataset composition affects cross-domain generalization. This study addresses this gap. Inspired by the "anatomy" analysis proposed by Guimarães et al. (2023), we investigate the structural and lexical composition of six prominent and widely explored Portuguese hate speech datasets. We extend prior work by moving beyond descriptive analysis: we combine lexical profiling—using HurtLex (Bassignana et al., 2018), a multilingual lexicon of hate-related words—with a robust cross-domain evaluation based on transfer learning to quantify the transferability of knowledge across domains.

We adopt the term "anatomy" to describe the distribution of hate categories and vocabulary within each dataset. While we acknowledge the theoretical distinction between "hate speech" and "offensive language" (Davidson et al., 2017), the analyzed datasets use varying definitions. Therefore, for this cross-domain analysis, we consider these categories under the broader scope of toxic content, characterized by negative, rude, and/or disrespectful messages (Aroyo et al., 2019), while respecting the

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 456–466

original labels provided by the dataset creators.

Specifically, in this work, we focus on answering the following research questions (RQs):

- **RQ1:** How do Portuguese hate speech and toxicity datasets differ in terms of their lexical "anatomy", considering the distribution of HurtLex categories across toxic and non-toxic messages in each domain?
- **RQ2:** How are differences in lexical composition across datasets related to variations in cross-domain performance?
- **RQ3:** To what extent do models trained on one dataset (and its associated domain, such as political, news comments, or YouTube discussions) generalize to other datasets when evaluated in a cross-domain transfer learning setup?

Our results reveal several interesting findings related to our research questions. Regarding **RQ1**, we show that, although the datasets share a broadly similar lexical anatomy—with toxicity concentrated in categories such as moral defects and negative stereotypes—they differ significantly in how specific terms are used and labeled across domains. For **RQ2**, we find that these lexical and annotational differences are reflected in cross-domain performance: corpora with broader and more heterogeneous toxic vocabulary yield models that transfer better than those with stricter, profanity-centered labeling. Finally, in relation to **RQ3**, our cross-dataset experiments reveal that models trained on certain datasets (e.g., ToLD-BR) generalize comparatively well to other platforms and topics, whereas models trained on others (e.g., OLID-BR) suffer from substantial performance drops, underscoring that successful transfer depends not only on lexical overlap but also on the alignment of domain and labeling practices.

Overall, these insights contribute to the field by emphasizing that model robustness is intrinsically tied to the lexical and structural composition of training data. By mapping the lexical boundaries and biases of each dataset, this work provides essential guidance for selecting training resources that go beyond surface-level patterns, ultimately fostering more reliable toxicity detection systems.

The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 describes the datasets, HurtLex-based profiling, data cleaning and processing, vocabulary analysis, and cross-domain setup; Section 4 presents our empirical findings and related discussion; and Section 5 summarizes our contributions and outlines future work.

2 Related Work

A growing number of studies have leveraged data from digital platforms to understand and mitigate online hate speech (Lima et al., 2020; Poletto et al., 2021; Castaño-Pulgarín et al., 2021; Buturoiu and Corbu, 2026). These studies range from sociocultural analyses of how hate manifests in specific communities and historical moments (Matamoros-Fernández, 2017) to the development of computational frameworks and resources for automatic detection (Fortuna and Nunes, 2018). Together, they emphasize that hate speech is not only a linguistic phenomenon but is also shaped by sociopolitical and platform-specific contexts.

At the empirical level, several labeled datasets have been created to monitor specific scenarios, such as the rise of hate against Asians during the COVID-19 pandemic (He et al., 2021), enabling studies of the dynamics of racial prejudice in crisis situations (Vishwamitra et al., 2020; He et al., 2021). While these initiatives advance our understanding of how hate emerges and evolves, their strong dependence on particular events and communities (e.g., health crises, electoral disputes, fandoms) means that models trained on them may learn context-dependent patterns that do not necessarily transfer to other domains.

In parallel, the literature has explored specialized lexicons and cross-lingual transfer strategies to support data filtering and analysis. From a lexical–semantic perspective, the interpretation of potentially hateful expressions is complicated by polysemy, i.e., the coexistence of multiple related senses for the same lexical item (Ravin and Leacock, 2000), which poses challenges for both annotation and automatic processing. To expand coverage and semantic granularity, researchers have translated and adapted consolidated resources, such as Hatebase[1], and proposed multilingual lexicons, such as Multilingual Offensive Lexicon (MOL) (Salles et al., 2025). Lexicons like HurtLex (Bassignana et al., 2018), in turn, organize entries into target-oriented macro-areas (e.g., racism, xenophobia, misogyny, homophobia), supporting cross-dataset and cross-lingual comparisons—an especially relevant feature

[1]Available at: https://hatebase.org/

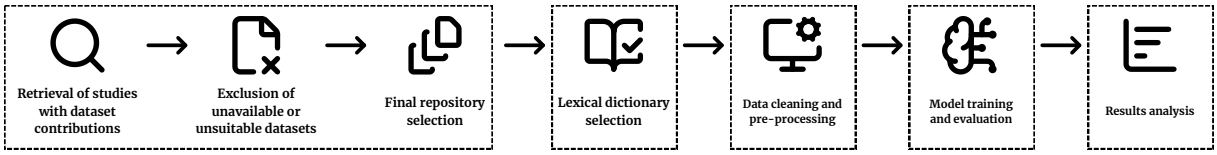


Figure 1: Overview of the proposed methodology.

for languages like Portuguese, where resources are more limited and often derived from cross-lingual adaptations.

A complementary line of research examines how hate speech interacts with platform architectures and moderation regimes. Studies on Gab and Reddit, for example, show how low moderation and echo chambers can amplify hateful content, shape "hate ecologies", and challenge the robustness of detection models across environments (Chandrasekharan et al., 2017; Lima et al., 2018; Zannettou et al., 2018; Lima et al., 2024). In the Brazilian context, legal and regulatory debates highlight tensions between freedom of expression, *fake news*, and hate speech (Da Silva et al., 2021), underscoring the need for tools that are sensitive to jurisdictional and sociopolitical specificities, particularly for Portuguese-speaking countries.

Closer to our work, Guimarães et al. (2023) analyze the composition of English-language repositories, exploring overlaps and gaps across datasets. We extend this research in two main ways: (i) we focus on a set of Portuguese-language repositories, examining their "anatomy" in terms of structural and lexical composition, with particular attention to the domains from which messages were collected (e.g., political discussions or content about public figures and influencers); and (ii) we go beyond descriptive analysis by combining lexical profiling based on HurtLex with a cross-domain transfer learning evaluation, quantifying how domain-specific these datasets are and how their composition affects the generalization ability of automated hate speech detection systems.

3 Methodology

The methodology proposed herein employs a computational approach to characterize the lexical structure and evaluate the cross-domain generalization of prominent Portuguese-language hate speech and toxicity datasets. As shown in Figure 1, the research workflow is organized into five main stages: (i) curation of six public datasets (Section 3.1); (ii) selection of the structured lexicon HurtLex (Bassignana et al., 2018) for term categorization (Section 3.2); (iii) data pre-processing and normalization (Section 3.3); (iv) cross-dataset evaluation experiments using transfer learning (Section 3.5); and (v) a multi-faceted quantitative analysis combining lexical and performance-based perspectives (Section 3.4). A detailed description of each stage is provided below.

3.1 Datasets Selection

The selection of datasets entailed a comprehensive literature search (e.g., via Google Scholar and scientific databases) for studies presenting Portuguese-language collections annotated for the presence or absence of any level of toxicity. Following an initial screening, only publicly accessible resources available through official webpages or GitHub were retained. Collections that provided only message identifiers were discarded; priority was given to those containing full textual content, even if anonymized. Consequently, six reference corpora for Brazilian Portuguese were consolidated for this study. A descriptive summary detailing their sources, collection methods, and class distributions is presented in Table 1.

These corpora also differ significantly in terms of domain and collection strategy. They cover distinct platforms (e.g., Twitter, Instagram, YouTube, news portals) and thematic focuses (e.g., political figures, minority groups, controversial topics), which we treat as contextual domains in our analysis. Notably, there is significant conceptual divergence among these datasets: terms such as "offensive", "toxic", and "hate speech" are not semantically interchangeable across resources. While some collections (e.g., OLID-BR, ToLD-BR) employ "toxic" or "offensive" as broad umbrella terms, others (e.g., HateBRXplain) focus more narrowly on offensive language. Despite these nuances, the primary binary labels available in each dataset—as outlined in Table 1—were utilized to establish a common ground, standardizing all instances under a unified definition of toxic content for the purposes of cross-domain comparison.

Dataset	General Description	Raw Size	Cleaned Size (% Toxic)
Fortuna [16]	Tweets retrieved from **Twitter** based on a specific vocabulary of potential hate speech terms, annotated as hate speech or non-hate speech (column `hatespeech_comb`).	5,670	5,666 (31.52%)
HateBRXplain [34]	Comments collected from the **Instagram** accounts of six Brazilian politicians, annotated as offensive or non-offensive (column `offensive_label`).	7,000	7,000 (50%)
OffComBR3 [15]	Comments collected from the **G1** news portal covering various topics, annotated as offensive or non-offensive (column `@@class`).	1,033	1,029 (19.53%)
OlidBR [36]	Comments collected mostly from **YouTube** videos covering controversial topics (e.g., politics, LGBTQ+ rights), annotated as offensive or non-offensive (column `is_offensive`).	6,952	6,952 (85.39%)
ToLDBR [26]	Tweets retrieved from **Twitter** using keywords related to minority groups, annotated as toxic or non-toxic (column `hate`).	21,000	16,806 (30.74%)
TuPy [14]	Tweets collected from **Twitter** covering diverse domains (e.g., politics, sports), annotated as hate speech or non-hate speech (column `hate`).	10,000	9,917 (10.41%)

Table 1: Quantitative summary of data repositories after pre-processing.

3.2 Lexical Dictionary: HurtLex

For the identification and categorization of toxic language, we selected HurtLex (Bassignana et al., 2018), a lexicon widely used in the literature (Pamungkas and Patti, 2019; Koufakou et al., 2020; Chiril et al., 2022). This resource was chosen primarily due to its comprehensive taxonomy, which offers a wide range of semantic classes, and its versatility for fine-grained lexical analysis across domains. Available in over 50 languages, including Portuguese, and derived from the Italian lexicon *Parole per Ferire*, HurtLex organizes entries into 17 semantic categories (Table 2) grouped into three macro-themes:

- **(i) Negative stereotypes and ethnic slurs:** Derogatory terms targeting nationality, ethnicity, religion, sexual orientation, or disabilities (categories: *ps*, *rci*, *pa*, *ddf*, *ddp*, *dmc*, *is*);
- **(ii) Non-stereotyped hateful words and insults:** Offensive terms related to social taboos or dehumanization, including animal references or sexual slurs (categories: *or*, *an*, *asm*, *asf*, *pr*, *om*);
- **(iii) Other insults and negative connotations:** General pejorative terms related to crime, morality, and negative behaviors (categories: *qas*, *cds*, *re*, *svp*).

Additionally, for each category, we also provide in Table 2 examples of associated terms. Last, or this study, it is important to mention that the "conservative" version of HurtLex was adopted. This variant prioritizes precision by restricting entries to terms validated through back-translation, a strategy designed to minimize false positives during the analytical process and to provide a more reliable basis for comparing lexical "anatomies" across domains.

3.3 Data Cleaning and Processing

Before analysis, a pre-processing pipeline was implemented to normalize content and reduce sparsity while preserving semantically relevant information. Initially, all repository files (originally in formats such as `.parquet` and `.arff`) were first standardized to `.csv`. Class labels were then binarized to a uniform schema: `1` (toxic) and `0` (non-toxic), harmonizing discrepancies in original labeling conventions (e.g., *off*/*not*, *yes*/*no*).

Textual pre-processing focused on normalizing URLs and user mentions to ensure consistent anonymization across all datasets. Hashtags were preserved because of their potential semantic and contextual relevance, especially in political and fandom-related domains. To maintain data integrity, duplicate messages within each dataset were identified and removed to prevent redundancy bias. The final dataset statistics are presented in Table 1.

Furthermore, for the lexical frequency analysis (e.g., word clouds), a rigorous stopword filtering process was applied using the `spaCy` library (model `pt_core_news_lg`). This process excluded standard high-frequency function words and included an additional manual filtering step to remove informal terms (e.g., *"pra", "vc", "ta"*) that remained after standard processing, ensuring the prominence of content-bearing vocabulary that better reflects domain-specific toxic expressions.

Label	Description	Examples of Terms (in Portuguese)
ps	negative stereotypes ethnic slurs	*preto, crioulo, bárbaros*
rci	locations and demonyms	*selvagem, camponês, barbárie*
pa	professions and occupations	*cafajeste, pedante, churl*
ddf	physical disabilities and diversity	*anão, nanus, feia*
ddp	cognitive disabilities and diversity	*imbecil, ignorância, trouxa*
dmc	moral and behavioral defects	*falso, mal, farsante*
is	words related to social and economic disadvantage	*pedinte, miserável, maltrapilho*
or	plants	*viado, imbecil, burro*
an	animals	*sanguessuga, medricas, bosta*
asm	male genitalia	*caralho, pinto, punheteiro*
asf	female genitalia	*bosseta, mulher, xana*
pr	words related to prostitution	*gigolô, puta, piranha*
om	words related to homosexuality	*veado, homo, lésbica*
qas	with potential negative connotations	*azarado, miserável, rústico*
cds	derogatory words	*safado, gente, tagarela*
re	felonies and words related to crime and immoral behavior	*calúnia, chantagem, vilanismo*
svp	words related to the seven deadly sins of the Christian tradition	*avarenta, preguiçoso, lascivo*

Table 2: Description of semantic categories in the HurtLex lexicon, with illustrative examples.

3.4 Vocabulary Analysis

After pre-processing, a lexical-matching procedure was applied to identify occurrences of HurtLex terms in the messages of each repository. For every match, the term was associated with the class label of the corresponding message (*toxic* or *non-toxic*). This procedure allows us not only to quantify the frequency of hate-related terminology and estimate how often such terms appear in *toxic* versus *non-toxic* content within each corpus.

Matched terms were then mapped to their semantic categories in the lexicon, enabling category-level frequency aggregation per dataset and per class. In this way, we obtain a lexical "anatomy" for each repository, characterized by: (i) which HurtLex categories are most prevalent, and (ii) how these categories are distributed across toxic and non-toxic messages. Comparing these profiles across datasets provides a corpus-level view of how different domains (e.g., political debate, news comments, YouTube discussions) emphasize distinct types of toxic vocabulary.

Finally, these lexical insights are contrasted with the machine learning experiments described below, which assess cross-domain generalization. Together, they enable us to investigate how differences in lexical composition relate to the transferability of detection models across datasets.

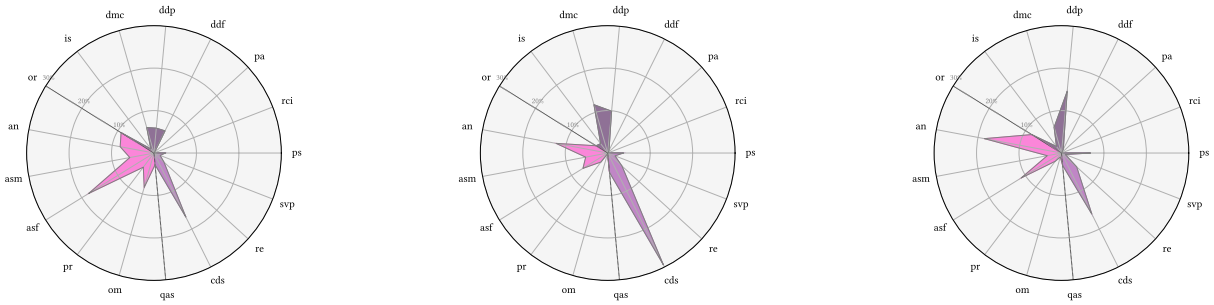
3.5 Cross-Dataset Evaluation Setup

To empirically measure the impact of dataset characteristics on generalization, we implemented a cross-dataset evaluation strategy in which a detection model is trained on each source repository and tested on all remaining target datasets. This yields a matrix of train–test combinations that allows comparison of *within-domain* performance (using a 70/30 train–test split on the same corpus, preserving the original imbalance between toxic and non-toxic messages) and *cross-domain* performance (training on one corpus and testing on another).

The feature extraction process employed BERTimbau (Souza et al., 2020), a pre-trained Portuguese language model, to generate dense vector representations of the messages. Specifically, embeddings were extracted from the `[CLS]` token of the last hidden layer (768 dimensions), leveraging transfer learning to capture semantic context uniformly across all datasets.

These embeddings were then used as input features for an XGBoost classifier (Chen and Guestrin, 2016), selected for its efficiency and robustness with high-dimensional data. To keep the comparison focused on dataset and domain discrepancies rather than on algorithmic tuning, all models were trained with default hyperparameters across all train–test configurations.

For performance evaluation, the Macro-F1 score was used as the primary metric. Given the severe class imbalance in some datasets, as shown in Table 1, this metric provides a stricter assessment by treating both classes equally, preventing the majority non-toxic class from dominating the results. By jointly analyzing the Macro-F1 scores across domains and the lexical profiles obtained in Section 3.4, we can examine how the "anatomy" of each dataset influences the cross-domain generalization capacity of hate speech detection models.



(a) Fortuna (b) HateBRXplain (c) OffComBR

(d) OlidBR (e) ToLD-BR (f) TuPy

Figure 2: Percentage distribution of hate categories per dataset based on HurtLex.

	HateBR	OffComBR	OLID-BR	Fortuna	ToLD-BR
OffComBR	0.9879	-	-	-	-
OLID-BR	0.9238	0.9399	-	-	-
Fortuna	0.8780	0.8611	0.9136	-	-
ToLD-BR	0.8202	0.8529	0.9202	0.8276	-
TuPy	0.9658	0.9713	0.9454	0.8732	0.9085

Table 3: Cosine Similarity Matrix based on HurtLex Category Distributions.

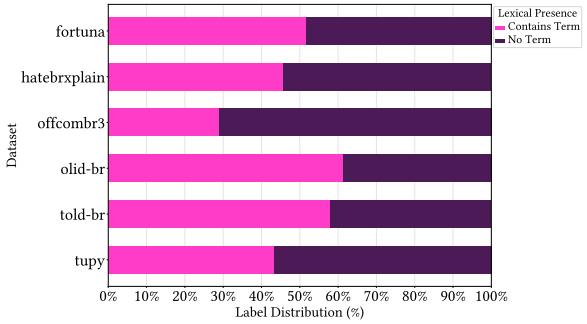


Figure 3: Presence of HurtLex terms (Contains Term vs. No Term) within the subset of toxic messages across datasets.

4 Results and Discussion

In this section, we combine quantitative metrics, lexical profiling, and model performance to characterize the landscape of Portuguese hate speech and toxicity datasets. The analysis is organized into three dimensions that directly address our research questions: (i) lexical and thematic alignment across corpora (RQ1), (ii) contextual divergence in the use and labeling of toxic vocabulary (RQ1, RQ2), and (iii) implications of these patterns for cross-dataset generalization (RQ2, RQ3).

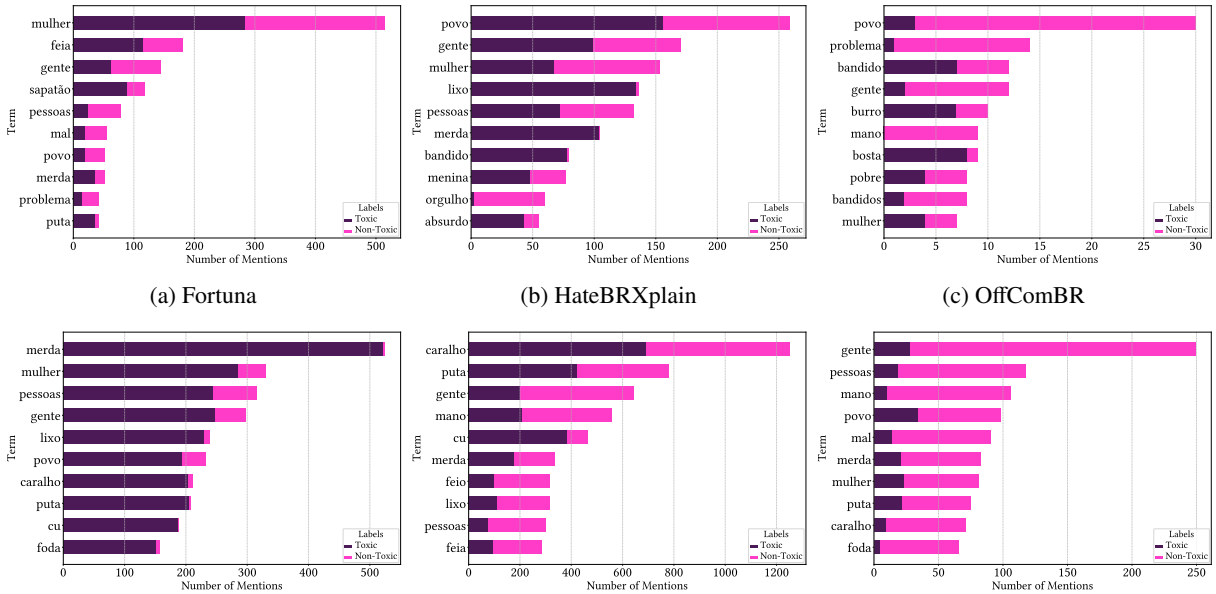
4.1 Lexical and Thematic Alignment

To quantify vocabulary overlap between datasets, we computed the *Jaccard Similarity Index* (Karabiber, 2025) for all repository pairs. This metric measures the intersection of unique tokens, regardless of their frequency, and serves as a proxy for lexical proximity between domains.

The results reveal a cohesive subset of social media-based corpora. HateBRXplain, OLID-BR, Fortuna, ToLD-BR, and TuPy exhibit relatively high lexical similarity among themselves, with scores ranging from 54% to 60%. This pattern is consistent with the informal register shared by platforms such as Twitter, Instagram, and YouTube. In sharp contrast, OffComBR3 shows substantially lower overlap with the other datasets (≈26–32%), reflecting its distinct source: as a corpus of news comments from the G1 portal[2], its vocabulary diverges from the slang-heavy and conversational style observed in social networks.

Beyond raw vocabulary overlap, the radar charts

[2]Available at: https://g1.globo.com/



(a) Fortuna (b) HateBRXplain (c) OffComBR

(d) OLID-BR (e) ToLD-BR (f) TuPy

Figure 4: Top-10 HurtLex terms (in Portuguese) per repository: number of mentions in toxic and non-toxic messages.

in Figure 2 indicate that the overall "anatomy of hate" is remarkably stable across corpora. This observation is quantitatively supported by the *Cosine Similarity matrix* (Lahitani et al., 2016) in Table 3, which shows alignment scores consistently above 0.85 when comparing distributions of HurtLex categories. Structurally, all datasets gravitate towards similar lexical macro-areas, particularly *dmc* (moral and behavioral defects) and *ps* (negative stereotypes). These results suggest that, at a high level, toxicity in the analyzed Brazilian corpora is consistently centered on attacks against character and reputation, indicating strong thematic compatibility despite differences in specific lexical items and platforms.

4.2 Contextual Divergence and Vocabulary

While the datasets exhibit macro-level alignment, a more granular analysis reveals important divergences in how toxic vocabulary is linguistically realized and labeled across domains.

Figure 3 illustrates the presence of HurtLex terms within the subset of toxic messages for each dataset. The variation in lexical presence—ranging from approximately 30% in OffComBR3 to over 60% in OLID-BR—reveals distinct characteristics in how toxicity is manifested across datasets. This heterogeneity challenges the notion of a single, universal definition of toxicity and is consistent with the arguments of Fortuna et al. (2020).

In OLID-BR, toxicity is often explicit and strongly relies on the lexicon. Conversely, in OffComBR3 and TuPy, most toxic content does not contain any explicit HurtLex terms ($\approx$ 70% and $\approx$55% of toxic messages lack these terms, respectively). This absence of explicit vocabulary in toxic instances aligns with Waseem and Hovy (2016), illustrating that toxicity is not solely defined by specific keywords. Instead, in these environments, abuse is likely conveyed through implicit means, irony, or context-dependent hostility that a static lexicon cannot capture.

Figure 4 presents the top-10 most frequent HurtLex terms for each repository, distinguishing their occurrences in toxic and non-toxic messages. Here, the contrasts in lexical usage across domains become even more evident.

A key source of variation is lexical polysemy (Ravin and Leacock, 2000), especially in terms that can function both as slurs and as intensifiers. A salient example is the term "*puta*" (literally "whore"). Although it is a strong gendered insult, in Brazilian Portuguese it is also widely used as a colloquial intensifier (e.g., "*puta vontade*" / "huge urge") or interjection (e.g., "*puta que pariu*" / "holy shit").

This ambiguity is clearly visible in social media-oriented datasets such as ToLD-BR and TuPy, where "*puta*" appears with a substantial proportion of non-toxic labels (represented by the lighter bars in Figure 4). In these cases, simple lexicon lookup cannot discriminate between misogynistic attacks

(a) Fortuna (b) HateBRXplain (c) OffComBR

(d) OLID-BR (e) ToLD-BR (f) TuPy

Figure 5: Word clouds of the most frequent terms (in Portuguese) in each repository in messages categorized as hate speech. Words in pink are present in the HurtLex lexicon.

and harmless slang. In contrast, in OLID-BR the term is almost exclusively labeled as toxic, suggesting an annotation bias towards marking explicit profanity as toxic regardless of its syntactic or pragmatic function. Together, these findings reinforce that models must leverage contextual and syntactic information—e.g., distinguishing a noun used as a direct slur from an adjective or interjection used as an intensifier—rather than relying solely on static lists of prohibited words.

Complementing the term-level analysis, the word clouds in Figure 5 illustrate the most frequent tokens in messages labeled as toxic in each repository, with HurtLex entries highlighted in pink. In ToLD-BR and OLID-BR, the dominant terms are explicit profanities (e.g., "*porra*", "*merda*"), indicating a form of toxicity strongly driven by coarse language. In contrast, in HateBR the most frequent words are largely neutral political entities and topics (e.g., "*Brasil*", "*Lula*", "*presidente*"), which are not necessarily listed in HurtLex. This suggests an entity-oriented toxicity pattern, where ostensibly neutral terms become hateful through their association with polarized political discourse. In other words, datasets such as HateBR rely more heavily on world knowledge and contextual cues, whereas ToLD-BR leans more on easily identifiable vulgarity. This distinction has direct implications for the transferability of models trained on these corpora.

4.3 Implications for Cross-Dataset Generalization

These lexical and contextual characteristics help interpret the cross-dataset performance results reported in Table 4. The model trained on ToLD-BR achieved the highest overall cross-domain performance, including a Macro-F1 score of 0.78 when tested on HateBR. This behavior can be linked to ToLD-BR's lexical breadth: as indicated by the radar and word cloud analyses, it covers a broad spectrum of toxic vocabulary, including profanity patterns similar to those in TuPy and, to a lesser extent, identity-based slurs present in Fortuna. In practice, ToLD-BR behaves as a "lexical superset" that exposes the model to diverse manifestations of toxicity, thereby supporting better generalization to other domains.

In contrast, despite exhibiting high lexical similarity with ToLD-BR (Jaccard ≈55%) and strong thematic alignment in terms of HurtLex categories (Cosine similarity ≈0.92), the model trained on OLID-BR performs poorly when transferred to other datasets (e.g., Macro-F1 of 0.26 on TuPy). The analysis in Section 4.2 sheds light on this phenomenon: OLID-BR encodes a comparatively rigid notion of toxicity, in which the presence of explicit profanity is almost always annotated as toxic. When such a model is applied to domains like TuPy, where the same words often appear in non-toxic, colloquial contexts, it is likely to yield many false positives, severely harming F1-scores.

Train \ Test	HateBRXplain	OffComBR	OLID-BR	Fortuna	ToLD-BR	TuPy
HateBRXplain	0.8738	0.5439	0.6167	0.6374	0.6801	0.5482
OffComBR	0.5273	0.7699	0.3579	0.5696	0.6264	0.5731
OLID-BR	0.5397	0.2426	0.5379	0.3345	0.4170	0.2654
Fortuna	0.6595	0.6474	0.4572	0.6679	0.6554	0.6115
ToLD-BR	0.7792	0.6855	0.5530	0.6439	0.7140	0.6224
TuPy	0.5313	0.5252	0.2402	0.5146	0.4300	0.6356

Table 4: Macro-F1 in the cross-dataset evaluation.

Taken together, these findings suggest that annotation incompatibilities—particularly disagreements over whether slang and profanity should be systematically labeled as toxic—are a major driver of cross-dataset generalization gaps (RQ2, RQ3). More broadly, they reinforce that the success of transfer learning in hate speech detection depends not only on lexical overlap but also on the alignment of contextual norms and labeling practices across domains, echoing (Assis et al., 2024).

5 Conclusion and Future Work

In this paper, we analyzed six Brazilian Portuguese datasets for hate speech and toxicity detection (i.e., HateBRXplain, OffComBR, OLID-BR, Fortuna, ToLD-BR, and TuPy), combining a lexical "anatomy" based on HurtLex with cross-dataset transfer learning experiments. We showed that, although the corpora share a similar macro-level profile—with toxicity concentrated in categories such as moral defects and negative stereotypes—they differ substantially in how toxic expressions are used and labeled across domains and platforms.

Our results indicate that the same term may be annotated as toxic in one dataset but frequently appears in non-toxic contexts in another, reflecting different annotation criteria and interaction norms. These differences help explain the cross-domain results: models trained on corpora with broader and more heterogeneous definitions of toxicity (such as ToLD-BR) generalize better, while models trained on datasets with stricter profanity-based labeling (such as OLID-BR) tend to over-flag slang in other domains. Thus, generalization depends not only on lexical overlap but also on the alignment of labeling practices and contextual language use.

Lexicon-based profiling is useful for describing datasets, but it is limited in handling polysemy, irony, and sarcasm, which often lead to false positives when words are interpreted out of context. Our findings reinforce the need for hate speech detection approaches that combine lexical resources with contextual and discourse-aware modeling, highlighting that high-quality dataset labeling is a crucial prerequisite for developing more robust models in the future.

As future work, we plan to refine the comparison of labeling schemes (e.g., "hate speech" vs. "offensive language") across corpora, explore the expansion and adaptation of lexicons using resources such as MOL (Vargas et al., 2025), and design new collections and models that more explicitly incorporate linguistic diversity, cultural context, and domain information to improve the robustness of hate speech detection in Portuguese.

Acknowledgements

This work was partially supported by CNPq, CAPES (PIBIC/UFV), FAPEMIG, and National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 1100–1105, Stroudsburg, PA, USA. Association for Computing Machinery.

Gabriel Assis, Annie Amorim, Jonnathan Carvalho, Daniel de Oliveira, Daniela Vianna, and Aline Paes. 2024. Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 301–311, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. HurtLex: A multilingual lexicon of words to hurt. In *CEUR Workshop Proceedings*, volume 2253, pages 1–6, Turin, Italy.

Danah M. Boyd and Nicole B. Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230.

Raluca Buturoiu and Nicoleta Corbu. 2026. Online hate speech. In Zizi Papacharissi, editor, *The Routledge Companion to Digital Media and Democracy*. Routledge, London, UK.

Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, and Harvey Mauricio Herrera López. 2021. Internet, social media and online hate speech: Systematic review. *Aggression and Violent Behavior*, 58:101608.

Dave Chaffey. 2025. Global social media statistics research summary 2025 [Feb 2025 update]. Accessed: July 28, 2025.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco, CA, USA. Association for Computing Machinery.

Lu Cheng, Ahmadreza Mosallanezhad, Yasin N. Silva, Deborah L. Hall, and Huan Liu. 2022. Bias mitigation for toxicity detection via sequential decisions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1750–1760, Madrid, Spain. Association for Computing Machinery.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: A multi-target perspective. *Cognitive Computation*, 14(1):322–352.

Gabriela Nunes Pinto Da Silva, Thiago Henrique Costa Silva, and João Da Cruz Gonçalves Neto. 2021. Liberdade de expressão e seus limites: Uma análise dos discursos de ódio na era das fake news. *Revista Argumenta*, 34:415–437.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515, Montreal, Canada. AAAI Press.

Felipe Ramos de Oliveira, Victoria Dias Reis, and Nelson Francisco Favilla Ebecken. 2024. Detecting hate speech on Brazilian social media: New dataset and analysis. In *Proceedings of the XLV Ibero-Latin American Congress on Computational Methods in Engineering*, Maceió, Brazil. ABMEC.

Rogers P. de Pelle and Viviane P. Moreira. 2017. Offensive comments in the Brazilian web: A dataset and baseline results. In *Proceedings of the 6th Brazilian Workshop on Social Network Analysis and Mining*, São Paulo, Brazil. SBC.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. UNESCO Publishing, Paris.

Samuel Guimarães, Gabriel Kakizaki, Philipe Melo, Márcio Silva, Fabricio Murai, Julio C. S. Reis, and Fabrício Benevenuto. 2023. Anatomy of hate speech datasets: Composition analysis and cross-dataset classification. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, pages 33:1–33:11, Rome, Italy. Association for Computing Machinery.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94, Athens, Greece. IEEE Computer Society.

Fatih Karabiber. 2025. Jaccard similarity. Accessed: August 11, 2025.

Kepios. 2025. Global social media statistics. Accessed: July 28, 2025.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *Proceedings of the 4th International Conference on Cyber and IT Service Management*, pages 1–6, Banda Aceh, Indonesia. IEEE.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Lucas Lima, Julio C. S. Reis, Philipe Melo, Fabricio Murai, Leandro Araujo, Pantelis Vikatos, and Fabricio Benevenuto. 2018. Inside the right-leaning echo chambers: Characterizing Gab, an unmoderated social system. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 515–522, Barcelona, Spain. IEEE Computer Society.

Lucas Lima, Julio C. S. Reis, Philipe Melo, Fabrício Murai, and Fabrício Benevenuto. 2020. Characterizing (un)moderated textual data in social systems. In *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 430–434, Virtual Event. IEEE Computer Society.

Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. Toxic content detection in online social networks: A new dataset from brazilian reddit communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 472–482, Santiago de Compostela, Spain. Association for Computational Linguistics.

Ariadna Matamoros-Fernández. 2017. Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, 20(6):930–946.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(3):477–523.

Yael Ravin and Claudia Leacock. 2000. *Polysemy: Theoretical and computational approaches*. Oxford University Press, Oxford, UK.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669, Abu Dhabi, UAE. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems*, pages 403–417, Rio Grande, Brazil. Springer.

Douglas Trajano, Rafael H. Bordini, and Renata Vieira. 2024. OLID-BR: Offensive language identification dataset for Brazilian Portuguese. *Language Resources and Evaluation*, 58(4):1263–1289.

United Nations. What is hate speech? Accessed: July 28, 2025.

Francielle Vargas, Isabelle Carvalho, Thiago A. S. Pardo, and Fabrício Benevenuto. 2025. Context-aware and expert data resources for Brazilian Portuguese hate speech detection. *Natural Language Processing*, 31:435–456.

Nishant Vishwamitra, Ruijia Roger Hu, Feng Luo, Long Cheng, Matthew Costello, and Yin Yang. 2020. On analyzing COVID-19-related hate speech using BERT attention. In *Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications*, pages 669–676, Virtual Event. IEEE.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, CA, USA. Association for Computational Linguistics.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. What is Gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014, Lyon, France. Association for Computing Machinery.