---

**Proceedings Paper:**
Kansaon, D., Benevenuto, F. and Maynard, D. (2026) Us vs. them: a dataset of political othering in Brazilian WhatsApp discourse. In: Brooker, S., Benatti, F., Pisarski, M. and Adamou, A., (eds.) HT '26: Proceedings of the 37th ACM Conference on Hypertext. HT '26: 37th ACM Conference on Hypertext, 14-18 Sep 2026, London, England. ACM, New York, pp. 219-231. ISBN: 9798400725647.

https://doi.org/10.1145/3800935.3830842

---

# Us vs. Them: A Dataset of Political Othering in Brazilian WhatsApp Discourse

Daniel Kansaon
Federal University of Minas Gerais
Belo Horizonte, Brazil
daniel.kansaon@dcc.ufmg.br

Fabricio Benevenuto
Federal University of Minas Gerais
Belo Horizonte, Brazil
fabricio@dcc.ufmg.br

Diana Maynard
School of Computer Science
The University of Sheffield
Sheffield, United Kingdom
d.maynard@sheffield.ac.uk

## Abstract

Political polarization is increasingly expressed through discursive strategies that construct moral and symbolic boundaries between "us" and "them". Among these strategies, othering is particularly important because it portrays sociopolitical groups as fundamentally different, inferior, or threatening. However, despite extensive research on hate speech and misinformation, othering remains less studied, partly due to the lack of annotated resources that capture this phenomenon. In this paper, we introduce a large-scale labeled dataset for political othering from Brazilian public WhatsApp groups. Starting from a manually annotated gold set, we develop an LLM-assisted annotation pipeline with systematic human validation, resulting in 8.5k labeled messages. Our analysis shows that othering is not limited to explicit insults or traditional hate-speech targets. Instead, it frequently targets ideological, partisan, and institutional opponents through negative affect and the discursive construction of group boundaries. These findings highlight political othering as a distinct dimension of harmful discourse in online environments.

## CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Information systems** → **Social networks**; • **Applied computing** → *Sociology*.

## Keywords

hate speech, fear speech, othering, natural language processing, social networks, whatsapp, dataset

**ACM Reference Format:**
Daniel Kansaon, Fabricio Benevenuto, and Diana Maynard. 2026. Us vs. Them: A Dataset of Political Othering in Brazilian WhatsApp Discourse. In *37th ACM Conference on Hypertext (HT '26), September 14–18, 2026, London, United Kingdom.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3800935.3830842

## 1 Introduction

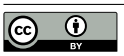
Recent elections around the world have unfolded amid growing concerns about the health of democratic debate in digital environments. These concerns are not limited to misinformation [19]. They also include intensified political polarization [10, 35], coordination [23], and other harmful online behaviors that reshape how political actors and their opponents are perceived [37].

In highly polarized contexts, such antagonism may extend beyond political disagreement and take the form of discursive exclusion, in which out-groups are portrayed as illegitimate, dangerous, immoral, or fundamentally outside the boundaries of belonging. This process is especially concerning because it can normalize hostility toward political out-groups and be a precursor to more extreme forms of aggression, including physical violence [17, 31, 32].

We conceptualize this form of discursive exclusion as othering: the process through which social or political out-groups are constructed as fundamentally different, morally inferior, threatening, or outside the boundaries of legitimate belonging [5, 22, 28]. In sociological terms, othering sustains a symbolic distinction between "us" and "them", often through positive portrayals of the in-group and negative portrayals of the out-group [17].

Othering is also a common rhetorical mechanism in political communication [33], often organized around moral boundaries, making out-group construction a central mechanism of political conflict, reinforcing exclusionary views, and contributing to a hostile social climate with extreme consequences. For example, in Germany, anti-immigrant narratives were mobilized in protests that framed immigrants and asylum seekers as a threatening out-group [3]. In Myanmar, the Rohingya minority became the target of hate and incitement to violence, being portrayed as foreigners and denied citizenship and belonging to the national community [41]. A similar dynamic was observed in Brazil, where growing hostility toward Venezuelan migrants culminated in attacks on encampments and the forced flight of many migrants across the border [2]. Together, these cases show that othering is not only a narrative process, but also one that can legitimize real-world practices of violence.

While harmful online phenomena such as hate speech, political polarization, and misinformation have been extensively studied, the underlying process of out-group construction that sustains many exclusionary messages remains comparatively underexplored. One key reason for this gap is the limited availability of public datasets specifically designed to capture this phenomenon. As a result, little is known about the prevalence of othering across online platforms and how it is structured within political discourse. Moreover, othering often appears in subtle forms that do not rely on explicit slurs, insults, or direct incitement, making it harder to detect using conventional approaches focused on toxicity or hate [17]. Messaging applications such as WhatsApp play a central role in political communication in countries such as Brazil [4], making them particularly suitable environments for studying these dynamics. In the Brazilian context, users often organize themselves into public political




*HT '26, London, United Kingdom*

ACM ISBN 979-8-4007-2564-7/26/09
https://doi.org/10.1145/3800935.3830842

groups aligned with opposing ideological positions, creating highly antagonistic environments in which discursive boundaries between "us" and "them" become especially salient. This structure provides a unique opportunity to observe and systematically capture othering in naturally occurring political discourse.

In this work, we address this gap by introducing the first dataset of political othering in the Brazilian context, focused on highly engaged political activist groups on WhatsApp. We propose an LLM-assisted pipeline specifically designed to capture subtle forms of exclusion in political discourse. Starting from an expert-annotated gold set, we combine confidence-based expansion with manual auditing to construct a final labeled resource of 8.5k messages. Our goal is to provide an initial and high-reliability resource for studying political othering in an underexplored discursive context. This represents an important first step toward studying political othering, thereby enabling detection and comparative analyses across platforms. Our initial analyses reveal that political othering is primarily organized around ideological and partisan divisions, while also targeting state institutions and marginalized social groups. We also find important lexical and psycholinguistic differences between othering and non-othering messages, particularly in negative affect and in how out-groups are referenced and positioned.

This paper is organized as follows. Section 2 introduces the theoretical background on othering and related work on harmful online discourse. Section 3 presents our methodological pipeline, covering data collection, expert annotation, and preprocessing. Section 4 reports the experimental results and motivates the model adopted for large-scale labeling. Section 5 discusses a qualitative error analysis, and Section 6 explores the main targets and linguistic patterns associated with othering. Finally, Section 7 outlines the limitations, and Section 8 presents the conclusion.

## 2 Theoretical and Empirical Foundations

To contextualize our research, we provide a theoretical overview of how othering language facilitates inter-group hostility. We first examine the concept of othering language and then review recent efforts in detecting these narratives.

### 2.1 Defining Political Othering

Othering is a social and discursive process through which an out-group is constructed as fundamentally different from and outside the boundaries of an in-group [5, 17, 22, 28]. For instance, a sentence such as "Our country is no place for people like them" frames the out-group as excluded from the community.

Othering is not simply the recognition of difference, but the construction of exclusionary distinctions between "us" and "them". This is a process through which groups construct themselves as good, virtuous, and legitimate, while portraying others as inferior, threatening, dangerous, or morally wrong [5, 28]. In this sense, othering is closely tied to the perception of threat, the morality of the out-group, and the reinforcement of in-group belonging [5, 17].

Othering is often discussed in relation to race, ethnicity, religion, or migration. However, its manifestations are context-dependent and may become more complex in highly polarized political contexts. In such contexts, processes of othering may also be strongly structured by ideological and partisan affiliations, allowing political opponents to be framed not merely as adversaries, but as morally corrupt, dangerous, or incompatible with the values of the in-group. In this sense, the out-group is not restricted to traditional social categories. It may also include groups associated with particular political positions or ideological leanings. For example, the expression "Thieving president and those who still support that rat" targets not only a political leader but also his supporters, constructing an out-group defined by their support for that leader. Prior work has similarly shown that online othering can create moral boundaries between users, even when explicit group labels are absent or less clearly defined [28], as illustrated by statements such as "Those on the left or right have lost the right to an opinion, they only defend garbage", in which political identities themselves can become the basis for exclusion and delegitimization [28]. In highly polarized environments, people who support a different political camp may be represented as enemies of the nation, threats to social order, or actors whose values are seen as incompatible with those of the in-group [1, 28].

In this study, political othering is defined as the discursive construction of a collective out-group as morally inferior, threatening, or not belonging to the in-group [28]. Unlike personal attacks or general criticism, othering requires a process of generalization that encompasses an entire social or political group. Political othering can overlap with hate speech but is not limited to it. Hate speech commonly refers to abusive or derogatory attacks [16, 30], whereas othering captures the broader discursive construction of an out-group as inferior and outside the boundaries of belonging, often without relying on explicit slurs or hateful expressions [28]. This distinction is particularly important in political discourse, where hostility may target partisan or geopolitical out-groups without constituting identity-based hate [15].

### 2.2 Related Work

The study of online speech, particularly on social networks, has become a major research area in recent years. The anonymity provided by these platforms can contribute to the prevalence of toxic [6], abusive [29], discriminatory [39], extremist [29], and hateful [16] comments. The growing research interest is driven by the significant impact of online speech, particularly its role in attacks against minority groups, raising serious social and ethical concerns.

This harmful language tends to intensify during periods of social instability, such as economic crises, political polarization, or global health crises, in which externalized fears lead to the blame of specific groups [14]. In these contexts, the perception of an out-group threat becomes a primary psychological driver, justifying hostility as a defensive necessity. These dynamics are often reflected in both hate speech and fear speech. While hate speech is typically characterized by explicit dehumanization and the vilification of protected groups [26, 30], fear speech portrays an out-group as an existential threat to the in-group's survival or values [36]. Othering, however, is a broader discursive process through which groups are constructed as fundamentally different, illegitimate, inferior, or threatening. In this sense, hate and fear speech may be understood as possible manifestations of broader othering dynamics, although othering may also occur through subtler forms of political exclusion that do not contain explicit hate or threat language. These

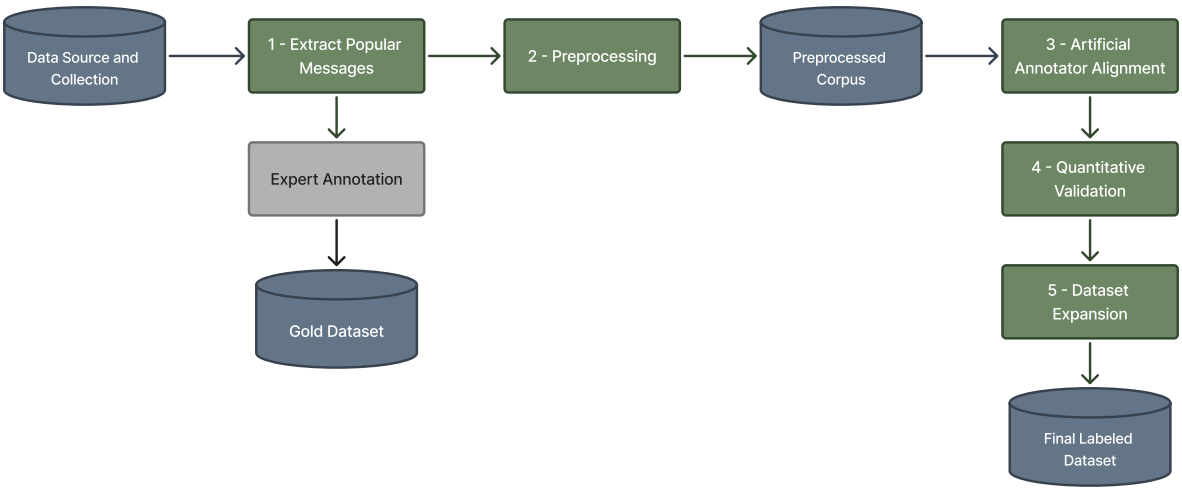



**Figure 1: Pipeline for constructing a labeled dataset of othering messages.**

rhetorical strategies facilitate exclusion by reducing individuals to a single perceived group membership [21].

Beyond its general definition, othering language encompasses a broad sociological process of out-group construction that manifests through diverse social identities and contextual triggers [17]. Research indicates that the thematic focus of othering often shifts depending on the geopolitical landscape and the targeted group. For instance, studies analyzing conflict in Ukraine identified specific dimensions of othering, such as threats to cultural identity, perceived risks to physical survival, and explicit dehumanization through vilification [17]. Similarly, [36] observed that religious othering often centers on existential threats to faith communities. Furthermore, [1] proposed a framework for the detection of "enemy", measuring how othering targets race, religion, sexism, and disability, highlighting the nature of these biases in extremist rhetoric.

However, the targets of these exclusionary narratives also vary by social domain. Recent work by [28] highlights that political disaffiliation can also be a primary driver of othering. This is particularly salient in Latin America, where othering often manifests through intense ideological divisions. In these regions, the main enemy, the out-group, consists of neither foreigners nor immigrants, but the left, homosexuals, and feminists, adversaries that tend to merge into a single melting-pot [28] in which ideological opposition is framed as an existential threat to the nation's values.

**Research Gap.** While hate speech, and to some extent fear speech, have been extensively studied [6, 11, 36], othering remains comparatively understudied as a computational phenomenon. A central reason for this gap is the limited availability of datasets specifically designed to capture how exclusionary boundaries are constructed in discourse, particularly in political contexts. This gap is especially relevant to WhatsApp group communication. In Brazil, WhatsApp plays a central role in political debate, where public political groups often serve as highly polarized spaces for attacks on opponents [12, 24]. The platform offers a particularly suitable setting for studying how political opponents are discursively constructed as out-groups. We address this gap by introducing, to the best of our knowledge, the first Portuguese-language resource for studying political othering in Brazilian WhatsApp discourse. We aim to bridge the gap between sociological theory and computational linguistics, providing an important resource for analyzing the structure of ideological polarization and othering attacks.

## 3 Methodology

In this study, our goal is not only to build a large annotated corpus, but also to develop a pipeline for building a political othering dataset within Brazilian WhatsApp discourse. We employ a two-stage strategy: first, we build an expert-annotated gold set used to operationalize the concept of political othering, refine the annotation criteria, and calibrate candidate models. Next, we use this gold set to support a conservative large-scale expansion step, in which LLM-generated labels are retained only under a high-confidence threshold and later checked through manual auditing, as shown in Figure 1. This design prioritizes label reliability over full corpus coverage, aiming to produce a higher-quality labeled dataset for subsequent exploratory analyses.

### 3.1 Dataset Source and Collection

This study builds on an existing source corpus of public political WhatsApp groups [23, 24]. The corpus was collected from public Brazilian political WhatsApp groups monitored between January 2022 and January 2023, a period covering the months before and

after the 2022 presidential election in Brazil. This interval is particularly relevant due to intense political mobilization, misinformation surrounding the electoral process, and post-election protests. Political keywords were used to search for public WhatsApp invitation links shared on websites and social media platforms [23], and candidate groups were manually verified to ensure they were indeed political. In total, approximately 1,400 public political WhatsApp groups were monitored during this period.

### 3.2 Extract Popular Messages

In this work, we build on that corpus by identifying, annotating, and analyzing messages related to political othering. We focus specifically on text messages labeled by WhatsApp as "forwarded many times." A key feature of WhatsApp is message forwarding, which allows users to share content across multiple chats and groups. The "forwarded many times" label is automatically displayed by the platform when a message exceeds a predefined forwarding threshold, indicating that it has circulated widely across group communities. We focus on these because, unlike organic messages, such content is often strategically crafted to maximize virality, disseminate political narratives at scale, mobilize activists, and coordinate partisan narratives across group communities [27]. By filtering the source corpus for messages marked as "forwarded many times", we obtained 21,667 unique messages. This subset constitutes the dataset used in this study.

### 3.3 Expert Annotation

From this dataset, we first constructed a gold annotation set of 150 messages through manual annotation. This gold set was designed primarily as a calibration resource for the evaluated models. To establish a gold standard political othering dataset, we combined random sampling with keyword-based upsampling. We first generated an initial list of candidate keywords, including dehumanizing, demonizing, segregation, and threat-framing expressions, as well as election-related entities and themes identified through prior literature and domain knowledge (see Appendix A for the complete keyword list). These keywords were only used as an aid for sampling and to retrieve a pool of candidate messages from the corpus for manual labeling. Because othering is relatively sparse under random sampling, keyword-assisted sampling was used to ensure sufficient positive instances for reliable evaluation.

Using this sampled set, two trained annotators independently labeled 150 messages using the GATE Teamware 2 platform [45]. Before annotation, both annotators completed a training stage with curated examples of othering and non-othering, and were required to pass a brief qualification check to ensure their understanding of the guidelines. Annotation was conducted in five batches of 30 messages. For each batch, the annotators first labeled the messages independently. After each batch, disagreements were discussed to refine the guidelines, and final labels were subsequently assigned through consensus. Inter-annotator agreement, measured with Cohen's Kappa [9], was 0.62. This is consistent with the variability typically observed in subjective content annotation tasks, where agreement may range from moderate to substantial depending on label definition and task complexity [25, 43, 44]. We selected 150 messages to keep expert annotation feasible, allowing the annotators to carefully examine each message and iteratively refine the annotation guidelines. This decision also reflects the length and contextual complexity of the messages, which contain an average of 240 words and 14.55 sentences. The resulting gold annotation set contains 101 othering and 49 non-othering messages. Its class imbalance results from the keyword-assisted sampling strategy, which increased the number of positive instances. We therefore treat this gold set as a carefully calibrated reference that provides a reliable basis for evaluating and comparing model performance.

### 3.4 Preprocessing

Even after filtering the corpus for unique messages via MD5 hash matching, substantial redundancy remains. Due to the nature of viral content, which is frequently forwarded across groups, the same core message often reappears with minor modifications, added emojis, spelling changes, or the inclusion/removal of short contextual phrases. As a result, exact deduplication alone is insufficient to eliminate repeated content, motivating an additional near-duplicate removal step based on semantic similarity.

To deal with this, we first computed sentence embeddings for all text messages and indexed them with FAISS [13], a library for efficient similarity search over dense vectors, to perform approximate nearest-neighbor retrieval at scale. For each message, we retrieved its $k = 25$ nearest neighbors and their cosine similarity scores. We then built a similarity graph by connecting pairs of messages with a cosine similarity $\geq 0.90$. Within this graph, each connected component represents a set of messages linked by high-similarity relationships. Finally, we performed deduplication by retaining only a single representative message from each cluster.

Furthermore, to reduce noise unrelated to substantive political discourse, we applied additional text-quality filters to remove flooding and spam-like content, which are often characterized as malicious attacks intended to disrupt WhatsApp's conversational environment [24]. First, we discarded extremely long entries (> 5,000 characters), which are typically dominated by copy-pasted blocks or repeated sequences and are less suitable for manual and model-based annotation. Second, we removed very short entries (< 5 words), as these rarely contain enough linguistic context to reliably identify discursive framing. Finally, we filtered messages with a highly repetitive structure by removing messages in which more than 90% of the words were duplicates. After applying these filtering and deduplication criteria, we obtained a preprocessed corpus of 12,972 messages. This dataset provides the basis for our subsequent labeling phase and systematic analysis of othering.

## 4 Artificial Annotator Alignment and Expansion

In this section, we evaluate candidate models as artificial annotators for political othering and use this evaluation to support the expansion of the dataset beyond the manually annotated subset. We first present the candidate models (Section 4.1), then assess their performance against the expert-annotated gold set (Section 4.2), and finally apply the selected model to the full corpus to construct the final silver-labeled resource (Section 4.3).

| Model | Non-Othering | | | Othering | | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 | Macro-F1 | $\kappa$ |
| Gemma-2-9B | 0.844±0.000 | 0.551±0.000 | 0.667±0.000 | 0.813±0.002 | 0.950±0.001 | 0.876±0.001 | 0.771±0.001 | 0.550±0.001 |
| Llama-3.1-8B | 0.848±0.023 | 0.680±0.011 | 0.755±0.007 | 0.858±0.003 | 0.941±0.011 | 0.898±0.005 | 0.826±0.006 | 0.654±0.012 |
| Qwen-2.5-7B | 0.602±0.019 | 0.776±0.000 | 0.678±0.012 | 0.873±0.003 | 0.751±0.020 | 0.807±0.013 | 0.743±0.013 | 0.490±0.023 |
| Ministral-8B | 0.580±0.019 | **0.796±0.014** | 0.671±0.010 | 0.879±0.005 | 0.719±0.025 | 0.791±0.015 | 0.731±0.012 | 0.470±0.021 |
| Gemma-2-9B$^{\dagger}$ | 0.854±0.016 | **0.794±0.017** | **0.823±0.010** | **0.886±0.005** | 0.921±0.012 | 0.903±0.006 | **0.863±0.006** | **0.726±0.013** |
| Llama-3.1-8B$^{\dagger}$ | **0.936±0.020** | 0.665±0.028 | 0.777±0.019 | 0.879±0.011 | **0.982±0.006** | **0.927±0.006** | 0.852±0.012 | 0.707±0.023 |
| Qwen-2.5-7B$^{\dagger}$ | 0.602±0.019 | 0.776±0.000 | 0.678±0.012 | 0.873±0.003 | 0.751±0.020 | 0.807±0.013 | 0.743±0.013 | 0.490±0.023 |
| Ministral-8B$^{\dagger}$ | 0.580±0.019 | **0.796±0.014** | 0.671±0.010 | 0.879±0.005 | 0.719±0.025 | 0.791±0.015 | 0.731±0.012 | 0.470±0.021 |
| Random Forest | 0.524±0.046 | 0.465±0.052 | 0.492±0.043 | 0.754±0.019 | 0.794±0.033 | 0.773±0.021 | 0.633±0.029 | 0.267±0.059 |
| Log. Regression | 0.660±0.106 | 0.278±0.046 | 0.389±0.052 | 0.726±0.014 | 0.929±0.031 | 0.815±0.017 | 0.602±0.032 | 0.242±0.060 |
| SVM | 0.699±0.090 | 0.249±0.021 | 0.366±0.027 | 0.722±0.007 | 0.947±0.022 | 0.819±0.012 | 0.593±0.018 | 0.234±0.036 |
| XGBoost | 0.449±0.047 | 0.441±0.046 | 0.444±0.035 | 0.730±0.018 | 0.735±0.059 | 0.732±0.035 | 0.588±0.031 | 0.177±0.061 |
| BERTimbau | 0.476±0.349 | 0.220±0.184 | 0.287±0.210 | 0.706±0.014 | 0.900±0.108 | 0.790±0.041 | 0.539±0.098 | 0.136±0.143 |

**Table 1: Model performance for othering classification. Precision (Prec.), Recall (Rec.), F1-score (F1), and Cohen's $\kappa$ are reported as mean±95% confidence interval. LLMs are evaluated across six independent runs, while statistical baselines use five repeated stratified 5-fold cross-validation evaluations. $^{\dagger}$ denotes High-Confidence filtering (conf $\geq 0.90$). Best values are shown in bold.**

## 4.1 Artificial Annotator Alignment

To expand the gold-standard dataset and obtain labels at scale for the larger corpus, we adopted an LLM-assisted labeling strategy inspired by prior work on artificial annotators [18, 40]. In this approach, an automated classifier is calibrated against a small set of human-annotated examples and then used to propagate labels to a much larger collection of unlabeled messages. The goal is to approximate the decisions of trained human annotators while enabling scalable annotation. For this task, we use classifier models to propagate labels beyond the manually annotated subset, using the gold set as the reference for evaluation. To select a suitable model for this role, we benchmarked multiple candidate LLMs under a consistent experimental protocol, comparing them against traditional statistical classifiers, including Logistic Regression, Support Vector Machines (SVM), Random Forests, and XGBoost.

All candidate models were evaluated on the gold annotation set before being applied to the preprocessed corpus. The evaluation procedure differed according to the modeling strategy. The supervised baselines were evaluated using repeated stratified 5-fold cross-validation, whereas the LLMs were evaluated through repeated inference over the gold set using a fixed prompt configuration. To reduce stochastic variation, we fixed the random seed (42) across experiments. Class imbalance was handled by using class-weighted training for the supervised baselines, assigning a higher weight to the minority class during optimization.

**Strategy for Classification Models.** For the classical baselines, texts were represented using TF–IDF features based on unigrams and bigrams. Following the same text preprocessing pipeline, we evaluated Logistic Regression (*liblinear*, *max_iter* = 2000), Linear SVM ($C$=1), Random Forest ($n$=1000, *min_samples_split*=5), and XGBoost ($n$=800, $\eta$=0.05). We also evaluated a transformer baseline for Brazilian Portuguese (BERTimbau) [38], using the base pretrained model. We fine-tuned BERTimbau on the training partition of each fold in a stratified 5-fold cross-validation setup. Fine-tuning was performed for five epochs with mixed precision (FP16), per-device batch size of 8, and gradient accumulation over two steps. We report its performance alongside the classical baselines and the LLM-based classifiers.

**Strategy for LLMs.** These were evaluated under an in-context learning (ICL) setup, without specific fine-tuning or gradient updates. We employed a single, fixed prompt template and varied only the number of in-context examples, with $k \in \{2, 4, 8\}$ to perform 2-shot, 4-shot, and 8-shot prompting. For each setup, the $k$ examples were balanced, being evenly divided between the two target classes. Preliminary comparisons across these settings indicated modest and unstable differences in Macro-F1 on the gold dataset. We therefore adopt the $k = 2$ configuration as a fixed reference condition for the comparative evaluation. Prompts were executed with low-temperature sampling ($T = 0.2$) to balance output consistency with some flexibility in generation. Each message was evaluated independently using the same prompt template, with no conversational history. In the few-shot setting, labeled examples preceded the target message and served as task demonstrations. All experiments were conducted using a Portuguese prompt (see Appendix B for the full template).

Unlike setups that restrict outputs to a single binary label, our prompt required a structured output containing: (i) a binary decision for othering (`is_othering` $\in$ {`yes`, `no`}); (ii) a confidence score; and (iii) auxiliary fields including a brief justification and supporting evidence span, for auditing and error analysis. For quantitative evaluation, we used only the label (`is_othering`), with the remaining fields treated as metadata. Model outputs were parsed into our structured schema, and we additionally employed a high-confidence filtering strategy for dataset expansion.

| Model | $N$ | Non-Othering | | | Othering | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 | Accuracy | Macro-F1 | $\kappa$ |
| Gemma-2-9B$^{\dagger}$ | 8,592 | 1.000 | 0.929 | 0.963 | 0.719 | 1.000 | 0.836 | **0.940** | **0.900** | **0.801** |
| Llama-3.1-8B$^{\dagger}$ | 9,663 | 0.989 | 0.825 | 0.900 | 0.636 | 0.972 | 0.769 | 0.860 | 0.834 | 0.675 |

**Table 2: Manual audit of high-confidence labels produced after applying each model to the full dataset. The audit uses a manually labeled sample of 150 messages. $\kappa$ denotes Cohen's kappa between model predictions and audit labels. $^{\dagger}$ denotes high-confidence filtering (conf $\geq 0.90$) and $N$ denotes the number of messages retained after applying the filter.**

## 4.2 Quantitative Validation and Benchmarking

Table 1 reports performance of the tested models using Macro-F1 and Cohen's Kappa ($\kappa$) as our primary metrics, complemented by class-specific precision, recall, and F1-score. Given the unequal distribution in the gold annotation set, we emphasize Macro-F1, which captures balanced performance across the two classes, whereas $\kappa$ measures agreement between the model predictions and the human gold labels. Intuitively, $\kappa$ can be interpreted as the extent to which the model behaves like an additional annotator. While results vary across all models, the main differences arise between LLM-based approaches and classical baselines. For LLM-based classifiers, we repeated inference six times for each model and report the mean and 95% confidence interval across runs. For the statistical baselines, we used repeated stratified 5-fold cross-validation with five random seeds. For each seed, predictions from the five folds were concatenated before computing the evaluation metrics. We then report the mean and 95% confidence interval across the five repetitions.

As shown in Table 1, among the large language models, Llama-3.1-8B achieves the strongest overall performance ($\kappa = 0.654 \pm 0.012$, Macro-F1 $= 0.826 \pm 0.006$), outperforming Gemma-2-9B, Qwen-2.5-7B, and Ministral-8B. By contrast, classical statistical models show substantially lower agreement ($\kappa \approx 0.267$ for Random Forest) and lower Macro-F1 ($\approx 0.633$), suggesting that TF–IDF $n$-gram features have limited capacity to capture the contextual dependencies involved in othering. The classical baselines tend to over-predict the othering class, resulting in a substantial number of false positives (see Appendix C for details). Although the LLMs capture relevant patterns of othering, their unfiltered predictions still exhibit non-negligible disagreement with the gold labels. Applying these predictions directly to a substantially larger corpus could propagate classification errors during dataset expansion.

To mitigate this limitation, we investigate whether the confidence value generated by each model can be used as an abstention criterion. For each message, the models were instructed to return a binary label together with a confidence score between 0 and 1. This score was generated directly by the model as part of the structured output and was not derived from token probabilities. We interpret this value as the model's self-reported certainty rather than as a calibrated probability. As models may differ in how they report confidence, we interpret the resulting scores within each model rather than comparing them directly across models.

The limited resolution of continuous self-reported confidence is particularly evident for Qwen-2.5-7B and Ministral-8B, whose confidence values for predictions labeled as othering are concentrated at or above 0.90. Consequently, the threshold provides little discrimination between more and less reliable positive predictions for these models. By contrast, the scores generated by Gemma-2-9B and Llama-3.1-8B show a more informative relationship with prediction quality, with higher-confidence subsets exhibiting greater agreement with the human annotations.

We therefore use self-reported confidence as a filtering criterion evaluated separately for each model. Given a confidence threshold, predictions below the threshold are excluded rather than reassigned to the non-othering class. We evaluated thresholds of 0.80, 0.85, and 0.90 on the manually annotated data, examining the trade-off between agreement and coverage for each model. We adopted 0.90 because it produced the highest agreement among the evaluated thresholds while retaining sufficient coverage for Gemma and Llama. In Table 1, rows marked with $^{\dagger}$ report performance only on predictions retained at the selected threshold.

As shown in Table 1, when evaluated on the predictions retained at this threshold, Gemma-2-9B achieves $\kappa = 0.726$ and a Macro-F1 of 0.863, compared with $\kappa = 0.550$ and a Macro-F1 of 0.771 when all predictions are considered. Similarly, Llama-3.1-8B achieves $\kappa = 0.707$ and a Macro-F1 of 0.852, compared with $\kappa = 0.654$ and a Macro-F1 of 0.826 on the full set of predictions. These results indicate that confidence-based filtering provides a useful trade-off between label quality and coverage for both Gemma and Llama.

## 4.3 Dataset Expansion

Based on the confidence-filtered benchmark, we retained Gemma-2-9B$^{\dagger}$ and Llama-3.1-8B$^{\dagger}$ as the two strongest candidate artificial annotators. We applied both models to the full corpus using the same confidence threshold ($\geq 0.90$) and manually audited a sample of their predictions, as shown in Table 2. Applying the confidence criterion to the preprocessed corpus of 12,972 messages, Gemma-2-9B$^{\dagger}$ retained 8,592 predictions (66.2% coverage), whereas Llama-3.1-8B$^{\dagger}$ retained 9,663 predictions (74.5% coverage).

To evaluate the quality of the retained labels beyond the original gold set, we manually audited an additional random sample of 150 predictions from each model. This audit provided an independent assessment of the reliability of the expanded dataset and enabled a careful inspection of the models' errors and overall behavior. As shown in Table 2, Gemma-2-9B$^{\dagger}$ achieved a Macro-F1 of 0.900, and $\kappa = 0.801$. Llama-3.1-8B$^{\dagger}$ achieved a Macro-F1 of 0.834, and $\kappa = 0.675$. Together with the gold-set results, the audit indicates that confidence filtering produces a reliable subset of silver labels, with Gemma showing stronger agreement with human judgments. The audit also reveals model-specific differences: Gemma achieves

| Error category | Original text (PT-BR) | English translation |
|---|---|---|
| *Personal vilification* **(FP)** | "Esse psicopata safado, ordinário e bandido" | "That despicable, vile, and criminal psychopath." |
| *Partisan attack and negative campaigning* **(FP)** | "Votar em um candidato que defende esse crimes é o maior pecado que o brasileiro pode cometer. Um dia isso poderá acontecer com você ou com alguém da sua família. Não VOTE em candidato que defende LADRAO" | "Voting for a candidate who defends this crime is the greatest sin a Brazilian can commit. One day this could happen to you or someone in your family. Do NOT VOTE for a candidate who defends THIEVES." |
| *Threat and conspiracy rhetoric* **(FP)** | "O dever patriótico do povo foi cumprido! Agora é hora de muita oração em casa! Porque as coisas vão ficar feias... Pela nossa liberdade, pela ordem e pelo progresso!" | "The patriotic people's duty was fulfilled! Now it is time for lots of prayer at home, because things are going to get ugly... For our freedom, for order, and for progress!" |
| *Reported hostility and irony* **(FP)** | "Para o presidente argentino, a inflação de 60% e a grave crise econômica são provocadas pelos idosos, que insistem em continuar vivos." | "According to the Argentine president, 60% inflation and the severe economic crisis are caused by the elderly, who insist on staying alive." |
| *Implicit group boundary* **(FN)** | "Eles brincaram com nossas vidas, nossa liberdade e nada aconteceu." | "They played with our lives, our freedom, and nothing happened." |
| *Subtle moral boundary* **(FN)** | "A política é a capacidade de dizer a mesma merda de uma maneira diferente para enganar os apoiadores fanáticos de um político que não enxergam além de sua própria devoção." | "Politics is the ability to say the same shit in different ways to deceive a politician's fanatical supporters, who cannot see beyond their own devotion." |

**Table 3: Representative examples of model misclassifications identified in the qualitative error analysis. FP and FN indicate false-positive and false-negative predictions, respectively. Examples are shortened for readability.**

a precision of 1.00 for non-othering and 0.719 for othering, whereas Llama achieves a lower othering precision of 0.636.

Differences between the gold-set evaluation and the manual audit may partly reflect their distinct sampling strategies. The gold dataset was constructed via keyword-based upsampling to ensure sufficient positive instances, whereas the audit sample was randomly drawn from the high-confidence predictions retained from the full corpus. These different procedures may produce samples with different class distributions and levels of ambiguity. Our error inspection further suggests that Llama tends to over-predict othering in negative messages that target individuals or institutions but do not generalize to an out-group. An example of this pattern is shown in Table 3, where the model misclassifies a case of personal vilification as othering even though the message does not construct a group-based exclusionary boundary. In other words, the model appears to rely more on aggressive tone and toxic language and has greater difficulty identifying the generalization step that distinguishes political othering from individual criticism, which leads to false positives in the large-scale labeling.

Based on the manual audit, we selected the high-confidence predictions from Gemma-2-9B$^{\dagger}$ as the final labels used in the subsequent analyses. Although Llama retains a larger proportion of the corpus, Gemma achieves substantially stronger agreement with human judgments and fewer false positives. The confidence threshold retains approximately two-thirds of the corpus, providing a favorable balance between label quality and coverage.

The final labeled dataset contains 8,592 messages, including 7,203 non-othering and 1,389 othering messages. A limitation of this strategy is that high-confidence filtering may exclude more ambiguous or difficult examples, reducing the representation of borderline cases. Nevertheless, we consider this an acceptable trade-off, since our primary goal is to obtain labels with higher reliability. The anonymized dataset is available on Zenodo.[1]

## 5 Qualitative Error Analysis

To complement the quantitative evaluation, we conducted a qualitative error analysis of misclassified messages, examining both false positives and false negatives across the evaluated models. This analysis helps to clarify the conceptual limits of the task, showing that the main difficulty lies in distinguishing group-based exclusion from other forms of hostility, since exclusion is not always conveyed through explicit hostility or dehumanizing language [20]. It is often expressed indirectly, subtly, or in context-dependent ways that are difficult to capture [42]. Table 3 provides representative examples of these recurrent false-positive and false-negative patterns. We sampled messages from the human-annotated validation dataset, which were flagged as false positives (FP) and false negatives (FN).

***Personal vilification (FP).*** False positives frequently arise when the model interprets strong attacks against individuals as othering. In these cases, the message targets a particular person rather than

[1] https://doi.org/10.5281/zenodo.21500268

| # | Target Category | #Messages | Example Target Groups (translated) |
|---|---|---|---|
| #0 | Left-wing Supporters | 596 | communists, socialists, leftists, workers' party voters, lula voters |
| #1 | Generic Political Opponents | 268 | opposition, political opponents, opposing political group, enemies |
| #2 | State Institutions and Authorities | 144 | judges, supreme court justices, federal court, electoral court, senators |
| #3 | Right-wing Supporters | 78 | bolsonaro supporters, conservatives, bolsonaro voters, right-wing, far right |
| #4 | Journalists and Media | 55 | journalists, press, mainstream media, globo tv, band tv, international media |
| #5 | Social Movements and Organized Groups | 51 | landless movement, black bloc, antifa, criminal factions, labor unions |
| #6 | Racial and Ethnic Groups | 44 | immigrants, indigenous peoples, venezuelans, argentinians, cubans, black people |
| #7 | Military and Security Forces | 29 | generals, military, military who support the election, armed forces, federal police |
| #8 | Socioeconomic Groups | 27 | businesspeople, rural producers, farmers, elites, companies, bankers |
| #9 | Gender and Sexuality Groups | 26 | LGBTQIA+ people, transgender people, travesti people, feminists, women |
| #10 | Religious Groups | 10 | protestants, catholics, christians, afro-brazilian, non-religious, non-protestants |
| #11 | Others | 135 | NGOs, globalists, professors, people who get vaccinated, truck drivers, young people |

**Table 4: Distribution of consolidated target categories in othering messages. Frequencies indicate the number of othering messages in each category. Messages may include targets from multiple categories. The final column presents target expressions.**

a broader social or political group, and thus falls outside the scope of othering. Although such messages may include insults or moral degradation, they do not necessarily target a collective out-group.

***Partisan attack and negative campaigning (FP).*** Political messages framed as ideological criticism, denunciations, or electoral attack are adversarial and represent standard political contestation. In some cases, the model may overgeneralize, mistaking partisan hostility as othering, even when no clear exclusion exists. Othering is distinct from legitimate political disagreement, in that it requires identifying a shift toward group-based delegitimization [34]. However, the boundary between routine partisan attacks and exclusionary discourse is often intentionally blurred through "calculated ambivalence" [46], creating a level of nuance that remains challenging even for expert human annotators (as seen in the second example in Table 3).

***Threat and conspiracy rhetoric (FP).*** Political slogans and narratives centered on censorship, persecution, fraud, or war may trigger the classifier without clearly constructing an out-group. Terms such as "war" or appeals to "freedom" and "order" can intensify political antagonism without necessarily producing othering.

***Reported hostility and irony (FP).*** Some misclassifications arise in messages whose interpretation depends on irony, context, or reported speech, including news-like content that reproduces hostile or controversial claims without necessarily endorsing them. In such cases, the model may fail to capture the real intention.

***Implicit group boundary (FN).*** In some false negatives, othering is expressed through an implicitly constructed out-group, without explicit target naming. These cases are particularly challenging for the model because they require contextual interpretation. In other cases, the external group is evoked through references to institutions or collective actors, such as the media or other public entities, rather than through a clearly defined social group.

***Subtle moral boundary (FN).*** Some messages construct the out-group through subtle and softened forms of criticism rather than explicit hostility or dehumanization. These cases often depend on shared context, evaluative insinuation, or weakly signaled judgments, which makes them harder to capture.

In summary, these models had difficulty detecting implicit and context-dependent forms of othering. This convergence suggests that the main challenges arise less from model architectural limitations and more from the conceptual and contextual complexity of political othering. Most errors involve aggressive or highly negative language that does not construct an out-group boundary. Such attacks may target individuals, events, or institutions without generalizing blame to a group. Because models often rely heavily on toxic or strongly negative lexical cues, they can misclassify individual criticism as othering even when the message lacks the generalization. In practice, an attack on a single person may be misclassified as othering simply because of its aggressive tone, even when the message does not engage in the generalization that characterizes othering. In this sense, the qualitative errors reinforce that political othering should be treated as a boundary-making phenomenon rather than a synonym for toxicity or hate speech.

## 6 Discursive Patterns of Political Othering

In this section, we examine how political othering is expressed in Brazilian WhatsApp discourse through two complementary dimensions: its main targets (Section 6.1) and the linguistic patterns distinguishing othering from non-othering messages (Section 6.2). Together, these analyses show who is constructed as an out-group and how this construction is linguistically expressed.

| LIWC (othering) | $\|r\|$ | LIWC (non-othering) | $\|r\|$ |
|---|---|---|---|
| negemo | 0.188 | relativ | 0.073 |
| anger | 0.182 | hear | 0.069 |
| conj | 0.128 | insight | 0.068 |
| they | 0.117 | time | 0.063 |
| affect | 0.100 | ingest | 0.054 |
| pronoun | 0.100 | percept | 0.042 |
| ppron | 0.095 | past | 0.042 |
| you | 0.091 | work | 0.030 |
| function | 0.085 | i | 0.029 |
| excl | 0.078 | motion | 0.029 |

**Table 5: Top 10 LIWC categories for othering and non-othering messages, ranked by Pearson correlation ($|r|$) between normalized frequencies and binary labels. All correlations are significant at $p < 0.01$.**

## 6.1 Othering Targets

In this section, we analyze the most frequent targets of political othering, focusing on the othering subset of the final labeled dataset, which comprises 1,389 messages. We use the target labels generated by the model as part of its structured output. We employ these labels as an interpretive layer to understand which types of groups the model associates with political othering when assigning positive labels. This allows us to examine both the frequency of different targets and the main groups that become the focus of othering attacks. Because these labels often contained multiple targets and different formulations referring to the same group, we manually reviewed all othering messages and assigned one or more broader target categories. During this process, semantically equivalent expressions, such as references to leftists and left-wing voters, were consolidated into the same category. We identified 12 distinct target categories, as detailed in Table 4. For each target category, the frequency corresponds to the number of othering messages associated with that category. However, a single message may include multiple target categories and, therefore, contribute to more than one frequency count. Consequently, the sum of the category frequencies may exceed the total of 1,389 othering messages.

Most instances of political othering are directed at explicitly political targets, especially Left-wing Supporters, Right-wing Supporters, and Generic Political Opponents. This suggests that othering is used primarily as a strategy to delegitimize political adversaries by portraying them as morally inferior, threatening, or fundamentally opposed to the in-group. For example, one message states that "leftists are bums, the kind of bums who depend on government assistance", explicitly framing political opponents as lazy, dependent, and inferior. In this sense, political othering appears closely tied to partisan conflict and ideological polarization, functioning as a discursive mechanism that transforms political opponents into a hostile and stigmatized group.

At the same time, the presence of categories such as Gender and Sexuality Groups, Racial/Ethnic Groups, and Social Movements shows that othering is not restricted to electoral competition alone. Attacks against minorities, immigrants, Black and Indigenous people reflect a more classical form of exclusion, in which vulnerable or historically marginalized groups are framed as threats or as incompatible with the in-group. In this sense, political othering in our data also intersects with broader patterns of social exclusion.

This broader distribution of targets also reveals that they are not always framed as explicit social groups. Often, the target is formulated through the name of an institution, such as a federal court, legislative body, or media organization, and serves as a proxy for the broader social or collective political enemy associated with it. In these cases, institutional criticism may function as a proxy for stigmatizing an external group. While this highlights the analytical importance of identifying political othering, it also introduces subjectivity, since the boundary between legitimate criticism and out-group construction is not always straightforward.

This pattern is particularly evident in attacks against journalists and the media. References to outlets such as Globo TV and Band TV illustrate how institutional actors are repositioned as political antagonists. For example, one message states that "The media is trying to create leaks on a ship that we are all on board", portraying the press as an internal enemy acting deliberately against the collective good. By framing the press as a corrupt or biased entity aligned with "the enemy", these messages transform media criticism into a tool for group-based othering. Such discourse also creates a favorable environment for conspiracy narratives and disinformation.

A similar dynamic appears in the category State Institutions and Authorities, highlighting a particularly relevant feature of the Brazilian context. This is especially significant in Brazil, where the Supreme Court and electoral authorities have become prominent targets in narratives concerning electoral fraud and political persecution. For example, one message states that "The Supreme Court justices are actually communists and leftists who want the destruction of Brazil", portraying state institutions and their representatives as political enemies through their perceived association with leftist ideologies. In these messages, othering is used to discredit judicial actors by framing them not as neutral institutions, but as enemies aligned with a political side. In this polarized narrative, not only were the justices vilified, but their defenders were also labeled as complicit enemies of the people, a framing also reflected in repeated protests against the Supreme Court in Brazil [8].

These findings highlight the broad scope of political othering. It is mobilized against a wide range of actors, including political opponents, marginalized groups, social movements, journalists, state institutions, and security forces. This diversity suggests that political othering is not merely an expression of partisan hostility, but a flexible discursive strategy that adapts to the dynamics of political conflict. In this sense, the variety of targets reflects specific features of the political environment where different actors can be constructed as threatening and incompatible with the in-group.

## 6.2 Linguistic Analysis

In this analysis, we characterize each message using the psycholinguistic categories of the Portuguese LIWC [7], comparing the othering subset with the non-othering messages of the final labeled dataset. For each message, we calculate the proportion of words assigned to each LIWC category and measure its Pearson correlation

with the binary othering label. Table 5 presents the ten categories most strongly associated with othering and non-othering messages.

The most distinctive pattern among othering messages is the prominence of negative affect and interpersonal reference. The categories *negemo* and *anger* indicate that these messages more frequently employ vocabulary associated with negative emotional evaluation and hostility. Although negative language is expected in antagonistic communication, its association with pronoun categories such as *they* and *you* suggests that negative evaluations are often directed toward specific targets. Othering messages therefore appear to be organized around a relational structure in which a speaker confronts or attributes negative characteristics to the out-group. Othering messages are also associated with the categories *conj*, *function*, and *excl*. This indicates that differences between the classes extend beyond explicitly emotional terms and also involve the sentence's functional organization. These exclusion terms may contribute to constructions that distinguish one group from another. The prominence of exclusion-related terms, such as *but* and *except*, may point to the role of boundary-making in othering discourse. Such constructions may linguistically reinforce distinctions by defining who belongs to the in-group and who is positioned outside it. From this perspective, the in-group is constructed not only through shared characteristics, but also through its opposition to and exclusion of an out-group.

In contrast, non-othering messages are more strongly associated with categories such as *relativ*, *time*, *past*, and *motion*. These categories point toward a greater emphasis on temporal relations, events, and actions. The presence of *hear* and *insight* further suggests that non-othering messages more frequently report information and describe experiences. This does not imply that such messages are neutral. This means that their linguistic focus seems to be less centered on the negative characterization of social actors.

The contribution of this analysis is to show that othering is characterized not simply by negative emotion, but by the combination of negative evaluation and explicit references to the out-group. These findings reveal consistent aggregate differences between othering and non-othering messages. Although individual LIWC categories should not be interpreted as isolated deterministic markers of othering, their combined psycholinguistic pattern highlights that this discourse relies on a systematic intersection of negative evaluation and the structural marking of boundaries.

## 7 Limitations and Ethical Issues

The LLM-assisted expansion may also introduce propagation bias, since systematic classification errors or interpretive tendencies of the selected model can be reproduced across the final labeled dataset. We mitigate this risk by comparing multiple candidate models against the adjudicated gold set, applying confidence filtering, conducting an independent manual audit of retained predictions, and examining recurrent error patterns. Therefore, the expanded labels should be interpreted as a high-confidence silver resource rather than as human-verified ground truth. The gold annotation set is also limited to 150 expert-annotated messages and was constructed using keyword-assisted sampling. While this strategy ensured a sufficient number of positive instances for task operationalization and model comparison, the limited sample size and lexical sampling procedure may reduce representativeness and introduce selection bias. Despite these limitations, the adjudicated gold set provides an essential human reference for operationalizing the task and guiding the construction of the expanded dataset. Finally, the dataset is restricted to public political WhatsApp groups, and some findings may not generalize to non-political contexts.

We acknowledge that the analyzed messages may contain personal opinions and sensitive information. Therefore, we implemented rigorous measures to ensure the privacy and anonymity of all participants. All sensitive identifiers, including usernames and phone numbers, were anonymized for the released dataset using the Microsoft Presidio framework and custom strategies to address edge cases specific to the Brazilian personal data format.

## 8 Conclusion

In this paper, we introduce an initial resource for studying political othering in Brazilian WhatsApp discourse. Building on messages collected from public political activist groups, we propose a two-stage dataset construction pipeline combining expert annotation with conservative LLM-assisted expansion. This approach enabled us to operationalize political othering as a computational task and to construct, to the best of our knowledge, the first Portuguese dataset. Our findings show that political othering in this context extends beyond traditional targets such as race, ethnicity, or religion, and can also emerge through ideological and partisan boundary-making. Our findings also indicate that othering and non-othering messages exhibit distinct lexical and psycholinguistic patterns. The dataset is particularly valuable for capturing exclusionary political discourse that may not be identified by existing resources centered on explicit hate speech or misinformation, while the proposed construction approach provides a framework for developing similar resources in other communities. Its extension to new contexts can incorporate local political dynamics, allowing the framework to support comparative analyses across different contexts.

More broadly, this work opens new directions for future research by enabling the analysis of harmful political discourse through the perspective of othering. This perspective is complementary to, and partly orthogonal to, existing approaches focused on misinformation, hate speech, and polarization, and may help clarify the exclusionary processes that cut across these phenomena. In this sense, the main contribution of this resource is not only to support the study of political othering itself but also to provide a foundational dataset for investigating a range of relevant online harms through the shared lens of othering narratives.

## Acknowledgments

This work was supported by the EU ATRIUM project (Grant Agreement No. 101132163), whose Transnational Access (TNA) programme funded the lead author's visit to the University of Sheffield and enabled this collaboration. This work was partially supported by the Kunumi Institute, by FAPEMIG (Grant No. APQ-04803-25) and by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILD-IAR) under Grant No. 408490/2024-1, and by individual grants from CNPq.

## References


[1] Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L Williams. 2019. "The enemy among us" detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)* 13, 3 (2019), 1–26.

[2] BBC News. 2018. Brazil mob attacks camps of Venezuelan migrants. Accessed July 23, 2026. https://www.bbc.com/news/world-latin-america-45242786

[3] BBC News. 2018. Chemnitz protests: Germany probes banned Nazi salutes. https://www.bbc.com/news/world-europe-45330168. Accessed July 23, 2026.

[4] Fabrício Benevenuto and Philipe Melo. 2024. Misinformation Campaigns through WhatsApp and Telegram in Presidential Elections in Brazil. *Commun. ACM* 67, 8 (2024), 72–77. doi:10.1145/3663359

[5] Lajos L Brons. 2015. Othering, an analysis. *Transcience: A Journal of Global Studies* 6, 1 (2015), 69–90.

[6] Pete Burnap and Matthew L. Williams. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet* 7, 2 (2015), 223–242. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/poi3.85 doi:10.1002/poi3.85

[7] Flavio Carvalho, Fabio Paschoal Junior, Eduardo Ogasawara, Lilian Ferrari, and Gustavo Guedes. 2023. Evaluation of the Brazilian Portuguese version of linguistic inquiry and word count 2015 (BP-LIWC2015). *Language Resources and Evaluation* (2023), 1–20.

[8] CNN Brasil. 2022. *Manifestantes hostilizam ministros do STF na porta de hotel em Nova York*. CNN Brasil. https://www.cnnbrasil.com.br/politica/manifestantes-hostilizam-ministros-do-stf-na-porta-de-hotel-em-nova-york/

[9] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[10] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2021. Political Polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (Aug. 2021), 89–96. doi:10.1609/icwsm.v5i1.14126

[11] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media* 11 (03 2017). doi:10.1609/icwsm.v11i1.14955

[12] Philipe de Freitas Melo, Daniel Kansaon, Joao MM Couto, Julio CS Reis, and Fabricio Benevenuto. 2025. A Sticker is Worth a Thousand Words: Characterizing the Use and Abuse of Stickers on WhatsApp Political Groups in Brazil. *Proceedings of the International AAAI Conference on Web and Social Media* 19, 1 (June 2025), 1210–1223. doi:10.1609/icwsm.v19i1.35869

[13] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data* (2025).

[14] Victoria M Esses and Leah K Hamilton. 2021. Xenophobia and anti-immigrant attitudes in the time of COVID-19. *Group Processes & Intergroup Relations* 24, 2 (2021), 253–259.

[15] Emilio Ferrara. 2026. Manufactured Divisiveness: Decomposing the Hostile Content of Seven Social Media Influence Operations. arXiv:2607.14491 [cs.SI] https://arxiv.org/abs/2607.14491

[16] Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.* 51, 4, Article 85 (July 2018), 30 pages. doi:10.1145/3232676

[17] Patrick Gerard, Tim Weninger, and Kristina Lerman. 2025. Fear and loathing on the frontline: Decoding the language of othering by russia-ukraine war bloggers. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 615–635.

[18] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision* 129, 6 (2021), 1789–1819.

[19] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.

[20] Emily Harmer and Rosalynd Southern. 2021. Digital microaggressions and everyday othering: an analysis of tweets sent to women members of Parliament in the UK. *Information, Communication & Society* 24, 14 (2021), 1998–2015.

[21] Jolanda Jetten, Russell Spears, and Antony SR Manstead. 1997. Strength of identification and intergroup differentiation: The influence of group norms. *European journal of social psychology* 27, 5 (1997), 603–609.

[22] Hélène Joffe. 1999. *Risk and'the Other'*. Cambridge University Press.

[23] Daniel Kansaon, Philipe de Freitas Melo, Savvas Zannettou, and Fabricio Benevenuto. 2025. From Fake News to Real Protests: WhatsApp's Role in Brazilian Political Coordination. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 1007–1020.

[24] Daniel Kansaon, Philipe F. Melo, Savvas Zannettou, Anja Feldmann, and Fabrício Benevenuto. 2024. Strategies and Attacks of Digital Militias in WhatsApp Political Groups. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 813–825.

[25] 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-Vol. 1*. 472–482.

[26] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14867–14875.

[27] Philipe F. Melo, Mohamad Hoseini, Savvas Zannettou, and Fabricio Benevenuto. 2024. Don't Break the Chain: Measuring Message Forwarding on WhatsApp. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'24, Vol. 18)*. 1054–1067.

[28] Esteban Morales, Jaigris Hodson, Victoria O'Meara, Anatoliy Gruzd, and Philip Mai. 2025. Online toxic speech as positioning acts: Hate as discursive mechanisms for othering and belonging. *new media & society* (2025), 14614448251338493.

[29] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[30] María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open* 10, 4 (2020), 2158244020973022.

[31] James Piazza and Natalia Van Doren. 2023. It's about hate: Approval of Donald Trump, racism, xenophobia and support for political violence. *American Politics Research* 51, 3 (2023), 299–314.

[32] James A Piazza and Lauren O'Rourke. 2025. Hostile sexism, social dominance orientation, political illiberalism, and support for political violence in the United States. *Politics & Gender* 21, 2 (2025), 144–168.

[33] Ines Rehbein. 2025. Americans are dreamers–Generic statements and stereotyping in political tweets. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*. 128–137.

[34] Martin Reisigl. 2008. Analyzing Political Rhetoric. In *Qualitative Discourse Analysis in the Social Sciences*, Ruth Wodak and Michał Krzyżanowski (Eds.). Palgrave Macmillan, 96–120. doi:10.1007/978-1-137-04798-4_5

[35] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. 2020. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 131–141.

[36] Punyajoy Saha, Binny Mathew, Kiran Garimella, and Animesh Mukherjee. 2021. "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In *Proceedings of the Web conference 2021*. 1110–1121.

[37] Daniel Thilo Schroeder, Meeyoung Cha, Andrea Baronchelli, Nick Bostrom, Nicholas A Christakis, David Garcia, Amit Goldenberg, Yara Kyrychenko, Kevin Leyton-Brown, Nina Lutz, et al. 2026. How malicious AI swarms can threaten democracy. *Science* 391, 6783 (2026), 354–357.

[38] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*.

[39] Neil Thompson. 2020. *Anti-discriminatory practice: Equality, diversity and social justice*. Bloomsbury Publishing.

[40] Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588* (2023).

[41] United Nations. 2025. Rohingya plight in Myanmar, a 'test for humanity'. https://news.un.org/en/story/2025/09/1166004. Accessed July 23, 2026.

[42] Teun A Van Dijk. 1993. *Elite discourse and racism*. Vol. 6. Sage.

[43] Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 7174–7183.

[44] Francielle Vargas, Samuel Guimarães, Shamsuddeen Hassan Muhammad, Diego Alves, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Diallo Mohamed, Thiago Pardo, and Fabrício Benevenuto. 2024. HausaHate: An expert annotated corpus for Hausa hate speech detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*. Mexico City, Mexico, 52–58.

[45] David Wilby, Twin Karmakharm, Ian Roberts, Xingyi Song, and Kalina Bontcheva. 2023. GATE Teamware 2: An open-source tool for collaborative document classification annotation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Danilo Croce and Luca Soldaini (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 145–151. doi:10.18653/v1/2023.eacl-demo.17

[46] Ruth Wodak. 2015. *The politics of fear: What right-wing populist discourses mean*. Sage.

**Table 6: Keyword lists used in the keyword-assisted sampling. We report canonical forms only. Minor orthographic variants (e.g., accents) and plural forms were collapsed into a single entry.**

| Category | Keywords (canonical forms) |
|---|---|
| **Dehumanizing** | verme, escória, lixo humano, carniça, praga, parasita, ratazana, câncer, gentalha, laia, corja, lixo, gado, rato |
| **Demonizing** | do mal, das trevas, satanista, satânico, demônio, demoníaco, capeta, capiroto, anticristo, satanás, destruidores, diabólico, traidor, derrotados, bandidos, maléficos, canalhas, mentirosos |
| **Segregation** | expulsar, expulsão, banir, banimento, limpeza social, limpeza política, limpeza ideológica, higienização, separar o país, dividir o país, varrer do(s/as), infiltrado, cidadão de bem, limpar |
| **Threat frames** | tomar o Brasil, tomada do Brasil, tomar o poder, ditadura comunista, ditadura do partido, ditadura do STF, escravizar, escravidão, dominação, invasão, ameaça real, perigo iminente, inimigos, destruir, destruição |
| **Us vs. Them** | essa gente, esse povo, essa turma, essa laia, esse bando, eles querem, eles vão, eles querem nos, eles vão nos, contra nós, guerra contra eles, nossos, deles, delas, estamos, grupos, queremos, seremos, sabemos, povo brasileiro, eles são, elas são, contra eles, território, esse tipo de, esse grupo de, hoje são, somos mais, todos juntos, outros grupos |

```
Your task is to analyze WhatsApp messages and identify the presence of "Othering".

TECHNICAL DEFINITION:
Othering is the process of labeling and defining a group (Out-group) as fundamentally different and
inferior to the reference group (In-group). In a political context, this manifests when the text:

1. Moralizes difference: Treats the opponents as potentially evil or corrupt.
2. Constructs a Threat: Presents the group as an existential danger to the audience's values.
3. Dehumanization: Uses metaphors of dirt, disease, animals, or objects.
4. Generalization: Attributes negative behaviors of individuals to an entire social or political
category.

2 EXAMPLES (positive and negative)

OUTPUT FORMAT (JSONL):
{"msg_hash":"...","is_othering":"yes/no","reasoning":"...","target":"...","evidence":"...","
confidence":0.0}

MESSAGES TO LABEL:
$INPUTS
```

**Figure 2: Prompt template used in the LLM-assisted annotation stage.**

## A Keyword-Assisted Sampling

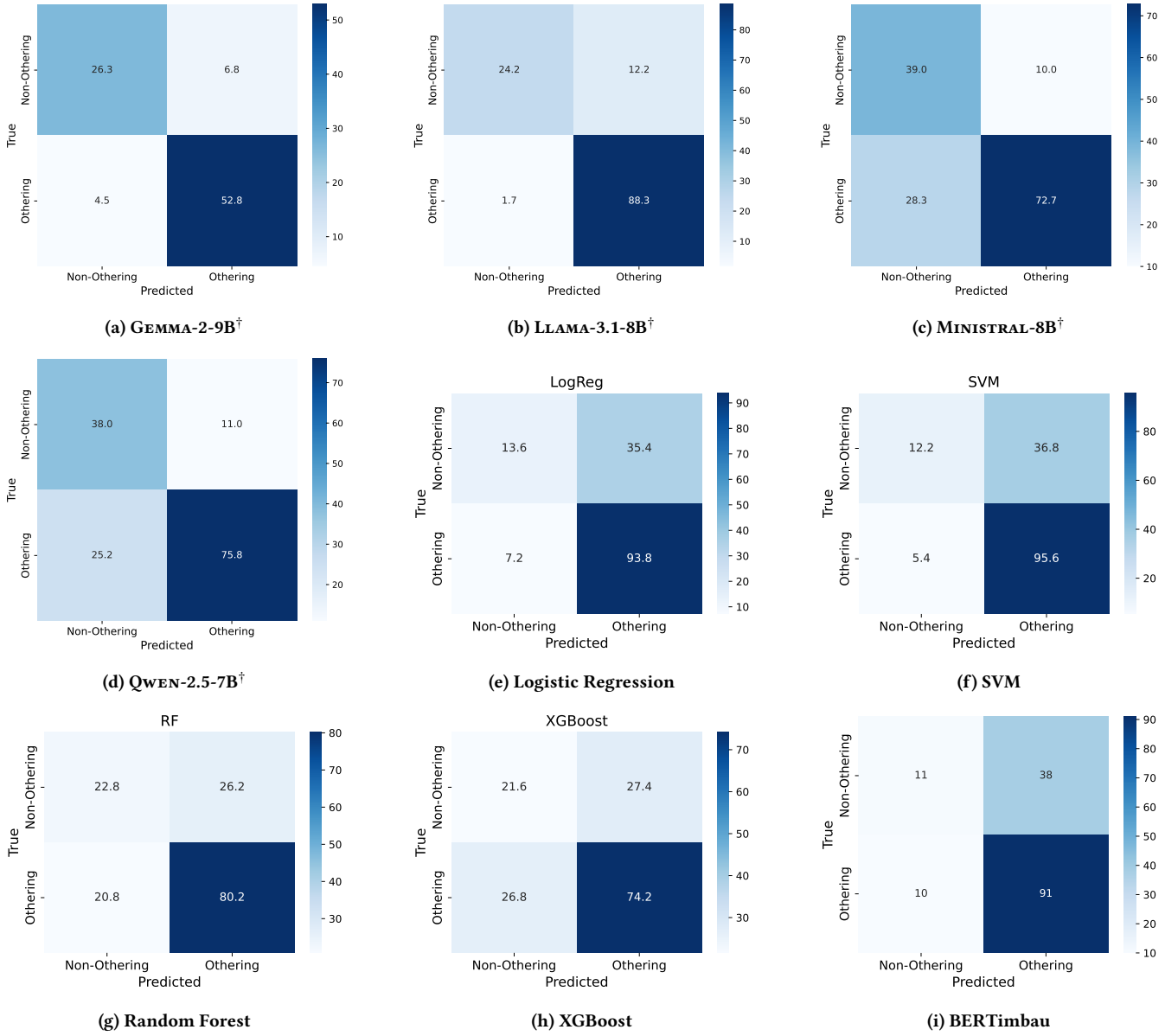
To increase the likelihood of selecting positive instances for expert annotation, we employed a keyword-assisted sampling strategy. Table 6 summarizes the main keyword categories used in this process. These categories were designed to capture expressions associated with dehumanization, threat framing, demonization, and exclusion. The keyword list should be understood as a sampling aid rather than an exhaustive representation of political othering.

## B LLM Prompt Template

To support reproducibility, Figure 2 reports the prompt template used in the LLM-assisted annotation stage. The prompt was designed to classify each message as othering or non-othering while also producing structured metadata, including confidence, target, justification, and evidence span. These auxiliary fields were used for auditing and qualitative inspection, although only the binary label was used for quantitative evaluation.

## C Confusion Matrices

Figure 3 presents the confusion matrices for the evaluated models on the gold dataset. These matrices complement the aggregate metrics reported in Section 4.2 by providing a more detailed view of the distribution of false positives and false negatives in the binary classification of othering and non-othering messages.




(a) Gemma-2-9B† (b) Llama-3.1-8B† (c) Ministral-8B†

(d) Qwen-2.5-7B† (e) Logistic Regression (f) SVM

(g) Random Forest (h) XGBoost (i) BERTimbau

**Figure 3: Confusion matrices for all models evaluated on the gold dataset. For the LLMs, the matrices are averaged across six runs and include only predictions retained after applying the confidence threshold ($\geq 0.90$). For the statistical and supervised-learning baselines, the matrices are averaged across five repeated stratified 5-fold cross-validation evaluations.**