future internet

MDPI

Article

EsCorpiusBias: The Contextual Annotation and Transformer-Based Detection of Racism and Sexism in Spanish Dialogue

Ksenia Kharitonova [1,†], David Pérez-Fernández [2,†], Javier Gutiérrez-Hernando [1], Asier Gutiérrez-Fandiño [3], Zoraida Callejas [1,4] and David Griol [1,4,*]

[1] Department Software Engineering, University of Granada, 18071 Granada, Spain; ksenia@ugr.es (K.K.); javier.gutierrez@ugr.es (J.G.-H.); zoraida@ugr.es (Z.C.)
[2] Department of Mathematics, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, 28049 Madrid, Spain; david.perez@inv.uam.es
[3] LHF Labs, 48007 Bilbao, Spain; asier@lhf.ai
[4] Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, 18071 Granada, Spain
* Correspondence: dgriol@ugr.es
† These authors contributed equally to this work.

Abstract

The rise in online communication platforms has significantly increased exposure to harmful discourse, presenting ongoing challenges for digital moderation and user well-being. This paper introduces the *EsCorpiusBias* corpus, designed to enhance the automated detection of sexism and racism within Spanish-language online dialogue, specifically sourced from the Mediavida forum. By means of a systematic, context-sensitive annotation protocol, approximately 1000 three-turn dialogue units per bias category are annotated, ensuring the nuanced recognition of pragmatic and conversational subtleties. Here, annotation guidelines are meticulously developed, covering explicit and implicit manifestations of sexism and racism. Annotations are performed using the Prodigy tool (v1. 16.0) resulting in moderate to substantial inter-annotator agreement (Cohen's Kappa: 0.55 for sexism and 0.79 for racism). Models including logistic regression, SpaCy's baseline n-gram bag-of-words model, and transformer-based BETO are trained and evaluated, demonstrating that contextualized transformer-based approaches significantly outperform baseline and general-purpose models. Notably, the single-turn BETO model achieves an ROC-AUC of 0.94 for racism detection, while the contextual BETO model reaches an ROC-AUC of 0.87 for sexism detection, highlighting BETO's superior effectiveness in capturing nuanced bias in online dialogues. Additionally, lexical overlap analyses indicate a strong reliance on explicit lexical indicators, highlighting limitations in handling implicit biases. This research underscores the importance of contextually grounded, domain-specific fine-tuning for effective automated detection of toxicity, providing robust resources and methodologies to foster socially responsible NLP systems within Spanish-speaking online communities.

Keywords: hate speech detection; bias; natural language processing; corpus annotation; sexism and racism detection; machine learning for toxicity; annotated dialogue corpora; Spanish

Academic Editor: Paulo Quaresma

Received: 17 June 2025
Revised: 20 July 2025
Accepted: 21 July 2025
Published: 28 July 2025

Citation: Kharitonova, K.; Pérez-Fernández, D.; Gutiérrez-Hernando, J.; Gutiérrez-Fandiño, A.; Callejas, Z.; Griol, D. EsCorpiusBias: The Contextual Annotation and Transformer-Based Detection of Racism and Sexism in Spanish Dialogue. *Future Internet* **2025**, *17*, 340. https://doi.org/10.3390/fi17080340

1. Introduction

The rapid expansion of online communication platforms has significantly intensified exposure to various forms of harmful content, including hate speech, posing considerable

challenges to digital well-being and effective content moderation. At the same time, rapid advances in large language models (LLMs) have revolutionized the processing, understanding, and generation of human-like text, leading to its increasing integration into systems that profoundly affect our social interactions [1]. Despite their remarkable success, a critical concern is that these models may learn, perpetuate, and even amplify detrimental social biases present in their training data. Addressing this problem requires rigorous research focused on measuring and mitigating social bias to ensure fairness in natural language processing (NLP) systems. However, this effort fundamentally requires precise definitions of the various types of social biases that can arise from LLMs, going beyond general "safety" concepts that may lack explicit definitions [2–4].

A common flaw in current research on bias and fairness in LLMs is insufficient precision in describing the specific harms caused by model behaviors [5–7]. This includes failing to clearly articulate who is harmed, the underlying reasons for harmful behavior, and how such harm reflects and reinforces existing social principles or power hierarchies. This challenge is especially acute when it comes to toxicity and offensive language, which are intrinsically intertwined with historical and structural power asymmetries. Moreover, implicit forms of toxicity, such as coded language, sarcasm, and micro-aggressions, remain notoriously difficult to detect, and very few existing corpora provide reliable and consistent annotation of such subtleties [5,8]. Specifically naming these harms, understanding their origins in social relations and histories, and recognizing the assumptions made in their conceptualization are crucial steps to effectively defining the role of NLP technologies in reproducing inequity and injustice [9,10].

In recent years, there has been considerable expansion in the development of annotated Spanish-language datasets specifically aimed at identifying toxicity, hate speech, and aggression. Previous studies, such as that of Taulé et al. [11], have systematically categorized these existing resources, detailing their characteristics, annotation schemes, and main limitations. However, while contemporary corpora like contextualized hate speech [12], NewsCom-TOX [11], and DETESTS-Dis [13] incorporate context, their contextual depth is predominantly monological or limited to short discussion threads. Moreover, it is imperative to expand the scope of bias and fairness considerations beyond the typically assumed American and Western contexts and to languages other than English, necessitating the creation of additional linguistic resources that accurately reflect the diverse linguistic features and representations of bias [1].

While the creation of annotated Spanish-language datasets has progressed in recent years, significant gaps remain in capturing both rich dialogical context and cultural diversity in expressions of bias and fairness. To design future linguistic resources that overcome these limitations, several strategies can be adopted, including cross-cultural sampling ensuring representation of local slang, references, and socio-political contexts that shape the manifestation of bias, incorporate dialogues from multiple platforms, and involve annotators from multiple cultural backgrounds providing guidelines that explicitly address region-specific stereotypes, taboo language, and forms of humor or sarcasm, thus improving annotation reliability for implicit and culturally nuanced cases.

Another important aspect is to consider conversational windows that consider multi-turn dialogues, enabling models to leverage pragmatic cues, and subtle forms of context-dependent toxicity that are lost in single turns. In this line, this paper presents the *EsCorpiusBias* corpus, a novel dataset of annotated multi-turn, multi-user dialogues specifically sourced from discussions on the Mediavida forum https://www.mediavida.com/foro/, (accessed on 16 June 2025).

We originally focused on sexism, racism, homophobia, and aporophobia, the last of which relies on a split between "deserved" and "undeserved" poverty to flag cues such

as "the poor are responsible for their poverty" or "social aid only encourages laziness". Pilot annotation on Mediavida data, however, revealed that clear cases of aporophobia and homophobia were scarce (less than two dozen positives in several thousand dialogues), so we reduced the final release to the two well-represented biases: sexism and racism, with approximately 1000 dialogue units each. Each dialogue unit consisted of three turns in order to preserve humor, irony, and quoted speech, signals that are often lost in single-comment corpora.

To concretely illustrate the challenges and nuances of online racism and sexism in Spanish, Section 3.2 presents full examples of annotated multi-turn Mediavida forum dialogues in Spanish, alongside their English translations. These examples highlight how context can reveal both explicit and subtle expressions of prejudice that may be missed in isolated comments.

The main contributions of our paper are as follows:

- A novel context-aware corpus: We release *EsCorpiusBias*, the first large Spanish dataset of annotated multi-turn, multi-user forum dialogues for sexism and racism detection. Unlike many previous datasets in Spanish that focus on isolated Twitter comments or decontextualized comment fragments, our approach focuses on contextual grounding, recognizing that toxicity often arises from discursive interaction.
- Rich contextual annotation protocol: Three-turn dialogues were annotated following meticulously developed guidelines that covered both explicit and implicit manifestations of sexism and racism.
- Reliable annotation quality: Annotations were conducted following a well-defined protocol using the Prodigy tool, resulting in moderate to substantial inter-annotator agreement (Cohen's Kappa: 0.55 for sexism and 0.79 for racism). Discrepancies were resolved using a manual adjudication protocol.
- Comprehensive model evaluation: We trained and evaluated models such as logistic regression, SpaCy's n-gram bag-of-words model (TextCatBOW) and the BETO transformer-based model. Our experiments show that contextualized transformer-based approaches (BETO) significantly outperform baseline models (such as logistic regression and TextCatBOW). Better recall and F1 score performance were observed for the contextualized variants of our models, underscoring the critical role of the preceding dialog context. Comparison with external models (such as piuba-bigdata/beto-contextualized-hate-speech and unitary/multilingual-toxic-xlm-roberta from Hugging Face) without domain-specific fine-tuning confirmed their shortcomings, especially in recall, highlighting the need to tune the models specifically for the linguistic characteristics and nuances of forum dialogues.
- Error and lexical bias analysis: We provide confusion matrices, error examples, and lexical overlap analysis, revealing current model strengths and limitations in detecting implicit bias.

The remainder of the paper is structured as follows. Section 2 reviews the main datasets and models developed to date for toxicity and hate speech detection, with a focus on Spanish-language resources. Section 3 details the data sources and collection procedures, specifying the focus on Mediavida forum discussions due to ethical compliance. This section also thoroughly describes the annotation procedure, including the development of guidelines for four distinct types of bias: sexism, racism, homophobia, and aporophobia, and emphasizing the contextual grounding in multiturn dialogues using tools like Prodigy. Section 4 presents the annotation statistics and distribution of the dataset, along with a comparative analysis of the effectiveness of the model. Finally, we describe the main findings of our work in Section 5 and present the conclusions and future research lines in Section 6. This paper contains explicit examples of offensive or upsetting language used

solely for illustrative purposes. These examples do not reflect the views or opinions of the authors.

2. Related Work

Despite their success, bias is a fundamental concern in LLMs, which often inherit it from the vast and largely unmoderated training datasets with which they are fed. These models learn, perpetuate, and amplify harmful social biases because they are trained with vast multimodal datasets that include text, images, audio, and video, and often inherit social biases embedded in human-generated content. As society evolves, the biases present in these archived data sources persist and inadvertently permeate LLMs.

Bias has broad implications, including compromising efficiency in business decision making, exacerbate health care inequities, lead to misclassification of human-generated text (especially from non-native speakers), perpetuate stereotypes in specific business applications (such as educational recommendations, employment, financial management, and health consultations), affect the quality of interactions in multi-agent systems (especially in creative and emotionally sensitive contexts), and influence the reliability of models in numerical interpretation and reasoning (where a bias in the generation of smaller digits has been identified, contributing to numerical hallucinations) [1,8,14].

Social bias is broadly defined as the differing treatment or outcomes between social groups arising from historical and structural power asymmetries. Two main types of harms have been taxonomized [1]. On the one hand, representational harms imply denigrating and subordinating attitudes towards a social group. They include derogatory language (insults or phrases that denigrate a social group), disparate system performance (demeaning comprehension or generation among social groups or linguistic variations), erasure (omission or invisibility of the language and experiences of a social group), exclusionary norms (reinforcement of the normativity of the dominant social group and implicit exclusion or devaluation of other groups), and toxicity (offensive language that attacks, threatens, or incites hatred or violence against a social group).

On the other hand, allocational harms involve a disparate distribution of resources or opportunities among social groups. They include direct discrimination (disparate treatment explicitly due to membership in a social group) and indirect discrimination (disparate treatment despite superficially neutral consideration toward social groups, due to proxies or other implicit factors). Specifically naming these harms, understanding their origins in social relations and histories, and recognizing the assumptions made in their conceptualization is crucial to effectively defining the role of NLP technologies in the reproduction of inequity and injustice.

Gender bias is one of the most studied types of bias [5,15,16]. LLMs have been observed to exhibit gender stereotypes, for example, associating “she” with nurse and “he” with doctor, or preferring male pronouns in contexts of leadership and professional success. Previous research has also identified gender stereotypes in multilingual training corpora, such as associating women with being “beautiful, empathetic, and neat” and men with being “tough, professional, leaders, and strong”. The potential to perpetuate racial bias in different domains has also been assessed in different studies, with the need for rigorous equity assessments in key application domains such as healthcare [17].

Several studies also highlight the importance of analyzing biases on multiple demographic axes and their interactions, such as gender, race, ethnicity, sexual orientation, religion, marital status, and number of children [15,18,19]. For example, a dataset has been developed to assess intersectional bias in South Asian LLM, considering religion, gender, marital status, and the number of children [20].

Entropy bias is defined as the variation in the information content of the generated text under different values of sensitive attributes in a prompt, e.g., showing a gender bias based on the amount of information provided in the responses [21]. LLMs, like humans, tend to favor additive over subtractive changes when solving problems or enhancing content, which can lead to unnecessarily complex or inefficient solutions [22]. This addition bias aligns with the tendency of models to favor longer outputs. The digit distribution in the LLM pretraining corpus can also induce a bias in digit generation, leading to "numerical hallucinations" and overgeneration of smaller digits [23].

These biases can manifest themselves in a variety of ways [1,8]. LLMs embedded in recommender systems can prioritize recent and popular content, reinforcing existing user preferences and limiting exposure to diverse options, which can negatively affect sales and user satisfaction [8]. Bias can appear locally or globally in text generation systems [1]. Local bias is a property of word-context associations (e.g., the probability of the following token for "The man was known by…" vs. "The woman was known by…") while global bias is a property of an entire text fragment, such as the sentiment of several generated sentences. As for machine translation systems, gender bias persists, especially when translating between languages with natural gender (such as English) and without gender (such as Persian, Indonesian, or Finnish), with models favoring male pronouns in professional contexts [24].

Detecting and quantifying bias is a critical and challenging research area in information management [1,8]. Several methodologies have been established, many of which build on and evolve from each other, reflecting the complexity and need for a multifaceted approach. These methodologies are mainly organized around metrics and datasets.

Evaluation metrics can be classified according to the level at which they operate: embeddings, probabilities or generated text. Embeddings-based metrics use dense vector representations to measure bias by calculating conceptual distances between target words (e.g., nationalities) and attributes (e.g., races). They include the Word Embedding Association Test (WEAT) [25] and the Sentence Encoder Association Test (SEAT) [26]. As a main drawback, these metrics have a weak or inconsistent relationship with biases in subsequent tasks [1].

Probability-based metrics measure systematic deviations from unbiased behavior using the token probabilities assigned by the model (e.g., masked token methods such as DisCo, LPBS, CBS; pseudo-log-likelihood methods such as CrowS-Pairs Score, CAT, AUL, LMB) [8]. Like embeddings-based metrics, they can also have weak correlations with biases in downstream tasks and are often based on simplified notions of bias, such as binary groups [1].

Generated text-based metrics analyze model-generated text from a prompt (e.g., distribution metrics such as social group substitutions, co-occurrence bias score, demographic representation, or stereotypical associations) [1]. Classifier metrics (such as perspective API, score parity, counterfactual sentiment bias, and regard score) rely on an auxiliary model to score toxicity, sentiment or any other bias dimension of the generated text. Bias is detected if text generated from similar prompts, but with different social groups, is classified differently (e.g., using the Perspective API to detect toxicity) [1]. Lexicon metrics such as HONEST [27] compare each word in the output to a pre-compiled list of words (e.g., derogatory language) to generate a bias score.

There are different datasets that have been generated with the main objective of bias assessment, mainly in English, although some are multilingual [1]. They can be mainly classified according to the application tasks. For masked token tasks (where the LLM predicts the most likely word), the most relevant ones include Winogender [28], WinoBias [29], WinoBias+ [30], GAP [31], GAP-Subjective [32], BUG [33], StereoSet [34], or BEC-Pro [35]. For unmasked sentences (where the LLM predicts the most probable

sentence), the main corpora include CrowS-Pairs [36], WinoQueer [37], RedditBias [38], Bias-STS-B [39], PANDA [40], Equity Evaluation Corpus [41], or Bias NLI [42]. For sentence completions (where the LLM provides a continuation to a sentence), the main examples include RealToxicityPrompts [43], BOLD [44], HolisticBias [45], and TrustGPT [46]. A relevant example of a multilingual corpus is HONEST [27], which measures the completion of hurtful sentences in English, Italian, French, Portuguese, Spanish, and Romanian. Despite their usefulness, many of these datasets, especially counterfactual input datasets, face limitations in reliability and validity, with ambiguities about the stereotypes they capture, inconsistencies in the treatment of social groups, and limited generalizability to contexts beyond the US [1,8,47].

HESEIA is a large-scale dataset co-designed in Latin American school settings, capturing intersectional biases across multiple demographic axes and school subjects, reflecting local contexts [18]. TWC (Translate-with-Care) has been created to assess gender bias and logical consistency in machine translation, especially between naturally gendered and non-gendered languages [24]. EuroGEST is an expansion of the GEST dataset to 30 European languages for a more holistic assessment of how gender biases are embedded in multilingual models [15].

In recent years, the development of annotated Spanish-language datasets aiming at the identification of toxicity, hate speech, and aggression has expanded considerably. Ref. [11] provides a systematic categorization of existing resources, highlighting their main features, annotation schemes, and limitations. Table 1 synthesizes this overview, positioning the newly developed *EsCorpiusBias* corpus within this evolving landscape.

Table 1. Spanish corpora annotated for toxicity or hate speech (adapted from [11]) and the new *EsCorpiusBias* forum corpus.

Dataset	Source	Size	% Toxic	Phenomenon/ Task	Main Target	Annotation Scheme	Context *	Annotators	References
AMI-2018	Twitter	4138	49.8	Misogyny	Women	multi-level	—	crowd + 3 exp.	[48]
MEX-A3T (18/20)	Twitter	11,856/ 10,475	29.6/ 28.7	Aggressiveness	Generic/ Women	binary	—	2 exp.	[49,50]
HateEval-2019	Twitter	6600	41.5	Hate speech	Women, migration	multi-level	—	crowd + 2 exp.	[51]
HaterNet-2019	Twitter	6000	26	Hate speech	—	binary	—	4 exp.	[52]
EXIST 2021/2022	Twitter, Gab	5701/ 6226	≈50	Sexism	Women	multi-class	—	crowd + ≥5 exp.	[53,54]
OffendES	Tw., YT, IG	30416	12.8	Offensiveness	Generic	5-class	—	3–10 exp.	[55]
OffendMEX	Twitter	7319	27.6	Offensiveness	Generic	multi-class	—	3 exp.	[55]
Context. Hate Speech	Twitter	56,869	15.3	Hate speech	Women, migration, LGBTI+, disabled...	multi-class	✓	6 exp.	[12]
NewsCom-TOX	News comments	4359	31.9	Toxicity	Immigration	multi-level	—	4 exp.	[11]
DETESTS-Dis	News, Digital media	10,978	≈40	Stereotype detection (explicit/implicit)	Immigration	binary + implicitness	✓	3 exp.	[13,56]
EsCorpiusBias	Mediavida forum	1990 †	≈26	Sexism/Racism	Women, migration	binary	✓	2 exp.	This work

† 1001 dialogues annotated for sexism and 989 for racism. *✓indicates that multi-turn conversational context is provided for each example.

Initial datasets such as AMI-2018 [48] and MEX-A3T [49,50] primarily leveraged Twitter data, focusing explicitly on misogyny and aggressiveness via binary or simplified annotation schemes. These datasets typically lacked substantial discourse context despite having significant proportions of toxic content.

Subsequent efforts such as HateEval-2019 [51] and HaterNet-2019 [52] further targeted hate speech on Twitter, extending their scope to xenophobia and immigration-related hate. More recent initiatives, including EXIST 2021/2022 and OffendES/OffendMEX [53–55], adopted richer classification frameworks to identify various forms of offensiveness across platforms such as YouTube and Instagram.

Notably, recent corpora like contextualized hate speech, NewsCom-TOX, and DETESTS-Dis [11–13,56] explicitly incorporate context by annotating news comments and social media posts, evaluating stereotypes, stance, sarcasm, implicitness, and insults.

The availability of these annotated datasets has directly facilitated advances in machine learning approaches to detect toxicity in the Spanish language. Recent research extensively applies transformer-based models fine-tuned on these datasets, showing notable improvements over traditional approaches. Table 2 summarizes representative deep learning models, highlighting their base architectures, target phenomena, and distinguishing methodological aspects. BETO-based [57] models (e.g., [12,55]) tend to outperform multilingual transformers (e.g., [58,59]), highlighting the importance of context-aware embeddings and domain-specific fine-tuning to effectively capture linguistic subtleties in Spanish.

Table 2. Deep learning models fine-tuned for toxicity and hate speech detection in Spanish.

Model	Base Transformer	Target Phenomena (Labels)	Additional Features
BETO Offensiveness [55]	BETO (Spanish BERT)	Offensiveness (5 classes: Non-offensive, Offensive, etc.)	Evaluated on multiple platforms: Twitter, YouTube, Instagram
Contextualized Hate Speech [12]	BETO (Spanish BERT)	Hate speech (multiclass: sexism, racism, LGBTI+ hate, disability hate, etc.)	Context-aware embeddings; trained on news-site comments (multi-turn context)
HaterNet [52]	CNN + linguistic features	Hate speech (binary: hate, non-hate)	Uses user-level metadata and linguistic features; specifically tailored for Twitter
Multilingual Transformers [58]	mBERT, XLM-RoBERTa	Hate speech, aggressiveness, offensive language (binary and multiclass)	Compares multilingual transformers to BETO; highlights performance benefits of BETO
Multilingual Toxic-XLM-RoBERTa [59]	XLM-RoBERTa	General toxicity (multiple: toxicity, severe_toxicity, obscene, threat, etc.)	Trained on multilingual Wikipedia talk page comments; effective cross-lingual generalization capabilities

Despite this progress, research on toxicity detection in Spanish still suffers from several key deficiencies. First, there is a lack of consistency in the definitions and taxonomies used for toxicity-related phenomena, which presents challenges in comparing models and evaluation benchmarks. Second, the predominance of Twitter-based datasets continues to limit the generalizability of models to other discourse genres such as forums, long-form discussions, and multimodal contexts. Third, multilingual models, while promising, tend to underperform in domain-specific tasks compared to monolingual alternatives like BETO, due to the reduced linguistic and cultural alignment. Finally, implicit forms of toxicity, such as coded language, sarcasm, and microaggressions, remain difficult to detect, and few corpora provide reliable annotation of such subtleties.

To proactively address these limitations and significantly advance the automated detection of toxicity and bias in Spanish online dialogues, this paper presents the EsCorpiusBias corpus. This innovative dataset stands out by meticulously annotating multi-turn, multi-user dialogues specifically extracted from Mediavida forum discussions. This approach allows for more nuanced and interactional analyses of toxicity, sexism, and racism, capturing the dynamic development of prejudice in conversation.

Overall, the continuous refinement of annotated resources such as *EsCorpiusBias*, combined with advanced transformer-based methodologies, provides an essential foundation for developing robust, contextually sensitive, and socially responsible NLP systems to detect and mitigate toxic discourse in Spanish online communities.

In addition, to improve the reliability and cultural inclusivity of bias detection in LLMs, especially for under-represented languages and non-Western communities, future research should prioritize the following: engaging local experts and communities to capture culturally relevant language and bias; tailoring annotation guidelines to local hierarchies, discourse, and intersectional identities beyond binary taxonomies; adopting interdisciplinary methods to account for nuanced phenomena like code-switching, humor, and microaggressions; leveraging cross-lingual and active learning to support low-resource settings; ensuring transparency through open data, code, and ongoing community involvement. These steps are vital to addressing global diversity in social bias and building equitable LLM-based systems.

3. Materials and Methods

This section details the data sources and collection procedures, specifying the focus on Mediavida forum discussions due to ethical compliance. We also thoroughly describe the annotation procedure, including the development of guidelines for four distinct types of bias: sexism, racism, homophobia, and aporophobia, emphasizing the contextual grounding in multiturn dialogues using tools like Prodigy.

3.1. Data Sources and Collection Procedures

We focused on capturing authentic user-generated dialogue from Spanish-speaking online communities. Although multiple public platforms were initially considered, such as Meneame, Reddit, and Usenet newsgroups, we ultimately restricted our dataset to content from Mediavida, a Spanish internet forum, due to its terms of service regarding data reuse and sharing for research purposes. This decision ensured full compliance with ethical and legal standards for data collection and publication.

Mediavida is a long-standing, high-traffic online forum in which users engage in threaded discussions on topics ranging from technology and gaming to social issues. The platform is characterized by its multi-user, multi-threaded structure, where users, often pseudonymous, contribute comments that branch off from an original post and often reply to one another in nested hierarchies. This creates a tree-like structure of conversation, where the root node represents the original post, intermediate nodes represent comments with replies, and leaves represent terminal comments.

To collect the data, we used a custom web scraping pipeline built with `requests-html` and `BeautifulSoup`. For each thread, we extracted the original post along with its full comment tree. Each entry included (i) a unique comment identifier, (ii) the identifier of the parent comment or root post, (iii) the raw text content, and (iv) an anonymized user ID.

We stored the data in JSON format, preserving the hierarchical structure of each thread. To extract linear dialogue sequences from these trees, we implemented a breadth-first search (BFS) algorithm. The BFS traversal allowed us to collect all root-to-leaf paths, each corresponding to a coherent conversational trajectory. To avoid redundancy, we

filtered out any dialogues that were strict subsets of longer paths, a step limited to posts with fewer than 100,000 dialogue paths to maintain computational efficiency.

During preprocessing, we applied several refinement steps to clean and standardize the dataset:

1. Removal of URLs, email addresses, and phone numbers.
2. Discarding of comments shorter than 10 characters.
3. Anonymization of usernames and user mentions.
4. Application of a basic profanity filter adapted from [60] to exclude overtly toxic language not relevant for initial model training.

This procedure resulted in a high-quality, ethically sourced dataset of Spanish online dialogues suitable for annotation and machine learning tasks related to hate speech and bias detection.

3.2. Annotation Framework and Theoretical Foundations

Toxicity in online communication refers broadly to language, behavior, or attitudes expressed through digital interactions that cause harm, reinforce negative stereotypes, or perpetuate discrimination against specific groups or individuals. According to recent studies, toxicity encompasses diverse phenomena including hate speech, offensive language, aggressiveness, and implicit forms of prejudice such as stereotypes and microaggressions [11,12]. Given the complexity and context-dependence of toxic behavior, precise and operational definitions are crucial for accurate annotation and effective automated detection.

For the purpose of our dataset annotation, we specifically focused on four main categories of toxicity: sexism, racism/xenophobia, homophobia, and aporophobia. Each category was meticulously defined based on prior literature, ensuring clarity and consistency across annotations.

To ensure both transparency and replicability of our annotation process, Table 3 provides a structured overview of all primary annotation labels, their operational definitions, concise annotation guidelines, and illustrative examples drawn directly from our corpus.

Table 3. Label definitions, annotation guidelines, and sample dialogues.

Label	Definition	Annotation Guidelines	Sample Dialogue (with Translation)
Sexism	Discrimination or prejudiced statements based on gender, reinforcing stereotypes or inequalities.	Annotators evaluate dialogues for manifestations such as hostile, benevolent, objectifying, ideological, or stereotypical sexism. Annotation is context-dependent and requires assessing subtle cues within the conversation.	Sexist Example: "Es que es obvio, los videojuegos de siempre han sido cosa de hombres…" ("It's obvious, videogames have always been a guy thing…")
Non-Sexism	Absence of gender-based discriminatory or prejudiced statements.	Annotators confirm no sexist elements exist within dialogue context, ensuring neutral or inclusive expressions.	Non-Sexist Example: "Que yo sepa la mayoría de competiciones permite competir a ambos sexos…" ("As far as I know, most competitions allow both sexes…")
Racism	Expressions involving prejudice or discrimination based on race, ethnicity, or national origin, whether overt or implicit.	Annotators identify dialogues containing affective, evaluative, judgmental, overt or covert racism, and stereotyping. Contextual understanding is crucial to detect subtle or ambiguous manifestations.	Racist Example: "El camarero no puso mesa de infraseres…ha puesto mesa acorde a lo que son, gitanos…" ("The waiter didn't write 'table of subhumans'…he wrote a table according to what they are, gypsies…")

Table 3. *Cont.*

Label	Definition	Annotation Guidelines	Sample Dialogue (with Translation)
Non-Racism	Absence of racially prejudiced or discriminatory statements.	Annotators ensure no racist or xenophobic elements are present, confirming the dialogue context is neutral or inclusive.	Non-Racist Example: "Sí que es verdad que aquí en Francia cuando eres autónomo hay diferentes categorias y a lo mejor tu categoría sería diferente a la mía." ("It is true that here in France when you are self-employed there are different categories and maybe your category would be different from mine.")

Given that toxic language and bias frequently manifest in nuanced or implicit forms only detectable within multi-turn conversations, the importance of dialogical context is emphasized in both our guidelines and example selection. We present two examples in Tables 4 (racism) and 5 (sexism) that highlight how context shapes annotation decisions, illustrating why single utterances may be insufficient to reliably capture subtle or pragmatic forms of bias.

Table 4. Sample annotated dialogue for racism (Spanish and English Translation).

Spanish Dialogue
<Context> Claro claro, cuéntame más. A mi si un gitano me viene de buenas, le voy a contestar de buenas, pero nunca vienen de buenas.
<Context> Cuando habláis de gitanos en el estudio os referís a la escoria entiendo yo (. . .) el 99% son escoria, eso es así (. . .), ya dije que a un gitano que me trate normal lo voy a tratar normal, pero todavía no he conocido a ningún gitano que lo haga.
<Turn being annotated> Además de estar de acuerdo con vosotros, un poco cutre los primeros histogramas del estudio donde ni siquiera ponen nada en los ejes. Debería darle vergüenza al grupo de investigadores que está elaborando los datos.
English Translation
<Context> Of course, tell me more. If a gypsy is nice to me, I'm going to answer him nicely, but they never come nicely.
<Context> When you talk about gypsies in the study you mean the scum I understand (. . .) 99% are scum, that's how it is (. . .), I've already said that I will treat a gypsy who treats me normally, but I haven't met any gypsies who do that yet.
<Turn being annotated> Besides agreeing with you, the first histograms of the study where they don't even put anything on the axes are a bit crappy. Shame on the group of researchers who are producing the data.

This example shows how the racist undertones emerge only through the context of the ongoing conversation about ethnic labels in a public setting. The annotation as "Racism" was only possible by considering the interaction across turns, not from an isolated utterance.

Table 5. Sample annotated dialogue for sexism (Spanish and English translation).

Spanish Dialogue
<Context> Quería decir italianas pero me pudo la emoción. Mujeres italianas y hombres españoles. Eslovenia tiene las mujeres más guapas por mucho que diga esa web.
<Context> las eslovenas y las checas. . . y encima la cerveza buena y barata.
<Turn being annotated> pruebas de las eslovenas para afirmar eso. Checoslovaquia no se tuvo que separar nunca.

Table 5. *Cont.*

English Translation
<Context> I wanted to say Italian women but I was overcome with emotion. Italian women and Spanish men. Slovenia has the most beautiful women, no matter what that website says.
<Context> Slovenian and Czech ones... and good, cheap beer on top of that.
<Turn being annotated> Slovenian evidence to support this. Czechoslovakia never had to split.

Note: The sexist assumptions about gender in the last turn become apparent only when considering the argumentative progression across the previous turns. Annotation relies on dialogical context, not isolated statements.

3.2.1. Annotation of Sexism

Sexism has been defined in our work as a form of discrimination characterized by establishing stereotypes, roles, and expectations based on a person's sex. In social media contexts, sexism manifests through messages, comments, and attitudes that sustain gender inequalities and unjust treatment [53,61].

Three main forms of sexism have been considered and explained to annotators:

- Hostile sexism: Openly negative or demeaning attitudes towards women. Example: "Women don't belong in the workplace; they should stay at home".
- Benevolent sexism: Seemingly positive but patronizing beliefs that reinforce traditional roles. Example: "Women are delicate, the angel in the house".
- Objectification: Reducing women to sexual objects or physical appearance, ignoring their dignity. Example: "Women exist solely for our enjoyment".

Furthermore, inspired by the EXIST campaigns [53], we described nuanced manifestations of sexism as follows:

- Ideology and inequality: Comments discrediting feminism, denying gender inequality, or portraying men as victims of gender-based oppression.
- Stereotyping and domination: False beliefs suggesting women are naturally suited for certain roles or unfit for others, reinforcing male superiority.
- Sexual violence: Comments suggesting, soliciting, or implying sexual aggression or harassment.
- Misogyny and non-sexual violence: Explicit expressions of hatred or non-sexual violence towards women.

During annotation, annotators followed the systematic steps below:

1. Read the entire dialogue for overall context and dynamics.
2. Evaluate the targeted comment based on sexism definitions and examples.
3. Annotate the target comment within the context of the dialogue using a binary classification framework: sexist/non-sexist.

This structured approach, informed by previous empirical studies and multilingual annotation efforts [62,63], ensured comprehensive and consistent identification of sexism within online interactions.

3.2.2. Annotation of Racism/Xenophobia

Racism covers racial offenses, tribalism, regionalism, xenophobia, and nativism—manifesting as hostility toward immigrants, refugees, and minority groups based on ethnicity, region, skin color, or physical traits [64]. Xenophobia is defined as fear or animosity towards those seen as outsiders, often driven by beliefs in cultural or national purity and threats to group identity [65].

Based on these broad definitions, annotators were guided to categorize racist and xenophobic content according to specific manifestations:

- Affect: Negative emotions and reactions, particularly those expressing hate or anger based on racial, ethnic, or religious differences. Example: "Take your damn piece of pizza and go back to Africa".
- Evaluation: Negative judgments regarding inherent characteristics attributed to specific groups, used dogmatically as reasons for discrimination. Example: "These people aren't even citizens of this country".
- Judgment: Negative assessments about behaviors and actions perceived as typical or representative of specific racial, ethnic, or religious groups. Example: "Crime, welfare, immigration—these issues always involve the Black, Hispanic, or Asian communities".

Additionally, we incorporated distinctions based on the explicitness of racist aggression, adapted from the aggression-annotated corpus methodologies [66]:

- Overt racism/aggression: Direct, explicit expressions of racial or ethnic prejudice, including derogatory terms, negative stereotypes, and calls for discrimination or violence. Example: "Immigrants don't adopt our values".
- Covert racism/aggression: Subtle or superficially neutral comments containing implicit racial prejudices or assumptions, often questioning someone's belonging or origins. Example: "You don't look Spanish," or "Where are your parents really from?"
- Stereotypes: Comments or humor reinforcing stereotypical views. Example: "All Asians are good at math," or compliments based on stereotypes: "You speak Spanish really well for someone from…"
- Environmental racism: Statements implicitly accepting racial inequalities in housing, employment, or service accessibility, often justified through meritocratic rhetoric. Example: "People are in their situations through their own efforts; we don't need policies to balance racial inequalities".

Annotators followed a structured procedure to ensure accurate classification:

1. Read the entire dialogue carefully to comprehend the overall context and interactions.
2. Evaluate whether the targeted comment aligns with the predefined categories of racial bias and xenophobia.
3. Annotate the target comment within a dialogue context using a binary classification framework: xenophobic/racist or non-xenophobic/non-racist.

This detailed guideline allowed for systematic annotation, supporting consistent and reliable detection of racist and xenophobic content in Spanish online interactions.

3.2.3. Annotation of Homophobia

Homophobia is defined as negative attitudes and reactions directed toward individuals who identify as homosexual, including gay, lesbian, bisexual, queer, or gender non-conforming individuals. It is generally characterized by hostility and negative stereotypes toward these groups [67–69]. Specifically, homophobic expressions frequently include pejorative labels as well as derogatory phrases like "that's so gay" or "don't be a homo" [70]. According to [69], homophobia fundamentally reflects an attitude of hostility toward homosexual individuals.

Homophobia includes specific forms such as lesbophobia, gayphobia, and biphobia, each targeting groups within the LGBTQ+ community [71]. For analysis, these are generally grouped under homophobia, which can appear overtly or implicitly, ranging from direct insults to subtle reinforcement of negative stereotypes about LGBTQ+ individuals' behaviors or characteristics [71].

Homophobic threatening language explicitly incites or supports violence against LGBTQ+ individuals, including direct threats of physical or sexual harm [71]. This form of homophobia goes beyond denigration to advocate harmful actions.

The annotation for homophobia followed a hierarchical taxonomy: content is first labeled as homophobic or non-homophobic, and, if homophobic, further subclassified by specific targets (e.g., gayphobic, lesbophobic, biphobic, and transphobic) [72]. This structured approach covers both implicit and explicit forms, enabling robust and accurate annotation. However, we found a not significant number of cases in our corpus, which is why homophobia is not considered in EsCorpiusBias.

3.2.4. Annotation of Aporophobia

Aporophobia denotes the rejection or aversion toward the poor, especially those perceived as unable to offer material benefit [73]. Naming this entrenched form of exclusion has enabled more precise analysis, particularly within an intersectional framework [74], which highlights how aporophobia can intersect with other biases, such as xenophobia, especially in attitudes toward impoverished migrants.

The annotation guide emphasized that, although aporophobia is a multidimensional social phenomenon [75], the focus was on interpersonal attitudes and insulting expressions (e.g., "The poor are lazy").

Following work like [76], our annotation framework distinguished stereotypes of the "deserving" versus "undeserving" poor, with stigmatizing claims such as "the poor are responsible for their poverty" or "social aid only encourages laziness" understood as systemic dehumanization of the poor rather than individual critique [77,78]. Example aporophobic expressions include the following: "The poor don't want to work", "They are parasites of the state", or "They chose that life". Such utterances reflect a denial of collective responsibility and a moral blindness, often rooted in societal values that demand productivity and self-sufficiency [79].

In annotating instances of aporophobia, annotators were instructed to look for the following:

- generalizations associating poverty with personal failure or criminality;
- expressions of disgust or inferiority toward poor individuals;
- denial of structural causes of poverty, often replaced by narratives of meritocracy;
- language that dehumanizes or blames the poor for systemic issues.

This detailed theoretical and practical framing enabled the consistent identification and annotation of aporophobic content across Spanish-language online dialogues, although the number of samples detected was very low, so it was finally not considered for the corpus.

3.3. Annotation Procedure and Contextual Grounding in Dialogue

To construct a high-quality corpus for toxic language detection in Spanish-language online discourse, we implemented a systematic and context-aware annotation protocol. The dataset comprises approximately 1000 three-turn dialogue samples for each of two primary bias categories: sexism and racism. These samples were drawn from multi-user conversational threads on the Mediavida forum. Each unit consisted of the target comment and its two preceding turns, preserving the flow of interaction and enabling nuanced interpretation of pragmatics, tone, and intention. Additionally, we initially launched annotation efforts for two further bias dimensions, homophobia and aporophobia, using the same three-turn contextual structure. However, as detailed in Section 4.1, the frequency of positively labeled examples in these categories proved extremely low, which ultimately led us to exclude them from model training and evaluation due to insufficient representation.

To construct the dataset, we initially sampled dialogues at random; however, this approach yielded very few positive instances of bias. To address this, we developed targeted lists of slurs and keywords for racism, sexism, homophobia, and aporophobia, and used these terms to filter and pre-select dialogues more likely to contain the phenomena of interest. This strategy substantially increased the number of positive examples for racism and sexism, while instances of homophobia and aporophobia remained extremely rare despite targeted filtering. Ultimately, the final datasets for sexism and racism comprised a mix of naturally occurring, randomly sampled dialogues and examples identified through lexical filtering, ensuring both diversity and adequate representation of toxic content in these two categories.

Unlike traditional hate speech datasets that present isolated comments, our approach centers on contextual grounding. This methodology acknowledges that toxicity often emerges from discursive interplay: a phrase may appear innocuous when decontextualized but reveal discriminatory implications when situated within a broader interaction. For example, humor, irony, or quoted speech can mask latent sexism or racism, which are patterns only reliably detected by examining the turn-taking structure and preceding provocations or justifications.

We selected a three-turn context window (the annotated comment plus two preceding turns) after a preliminary analysis of Mediavida dialogues. Shorter windows often lacked sufficient discourse information for interpreting nuanced or implicit bias, while longer contexts increased annotator load and topical drift. Thus, three turns offered an optimal balance, capturing key interactional context without sacrificing annotation quality. Tables 4 (racism) and 5 (sexism) provide representative examples demonstrating the necessity and sufficiency of this window.

To ensure consistency and analytical rigor, annotation was conducted using Prodigy, a modern annotation tool built on the spaCy NLP library. Prodigy was configured in manual text classification mode (`textcat.manual`) and customized to present full dialogue units, allowing annotators to assess semantic and pragmatic dependencies between turns. Its support for custom JSONL schemas enabled us to format and render three-turn dialogues clearly and coherently, which was essential for preserving interactional structure during the annotation task.

Two human annotators with prior experience in bias detection and online moderation, completed the labeling process following a detailed annotation guide grounded in contemporary definitions of the biases considered. Each comment was assigned a binary label "toxic" or "non-toxic" for the relevant phenomenon. Prodigy's lightweight interface facilitated focused, uninterrupted work while simultaneously tracking metadata and versioning for reproducibility.

To clarify our quality assurance and annotation methodology, Figure 1 summarizes the complete annotation workflow used in this study. The process is divided into three main stages: before, during, and after annotation. Before annotation, all annotators were provided with comprehensive guidelines, including clear definitions and practical examples for each bias category (racism and sexism), ensuring a shared understanding of the phenomena. During annotation, annotators read each entire dialogue, assessed whether the target comment fitted the label definition, and annotated it within its conversational context, following a structured, step-by-step protocol.

After annotation, we computed inter-annotator agreement using Cohen's Kappa, obtaining $\kappa = 0.55$ for sexism (moderate agreement) and $\kappa = 0.79$ for racism (substantial agreement), in line with accepted standards for subjectivity-prone linguistic annotation. Divergent cases were resolved via a manual adjudication protocol involving both annotators

and a supervising linguist, thereby producing a fully curated and consensus-driven dataset suitable for training and evaluation.

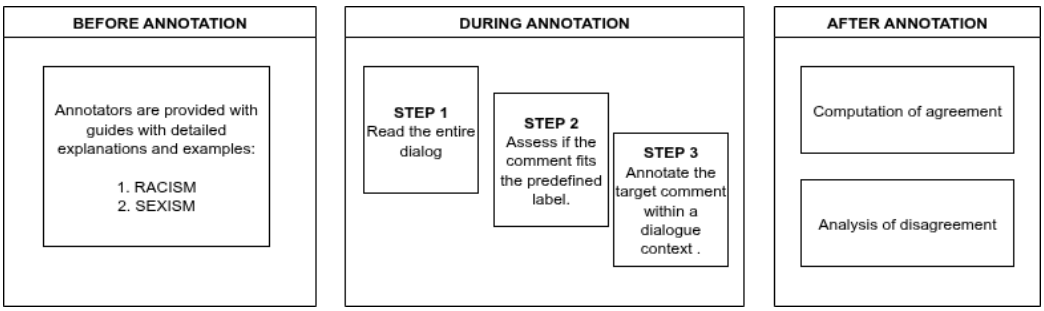

Figure 1. Overview of the annotation workflow.

The inclusion of dialogical context and the use of interactive, task-specific annotation tooling distinguish this work from many prior Spanish-language toxicity corpora, which have primarily focused on single-turn Twitter posts or decontextualized comment fragments. In contrast, our annotation captures the dynamic unfolding of prejudice in conversation, enabling future models to leverage discourse-level signals that are critical for the robust detection of subtle, pragmatic, and evolving forms of online bias.

3.4. Models and Experimental Set-Up

In order to rigorously evaluate the performance and robustness of models detecting sexism and racism in Spanish-language online dialogues, we developed and trained multiple classifiers using EsCorpiusBias.

Two variants were trained for each model:

- Single-turn: Trained on individual comments without dialogue context, thus evaluating the model's ability to detect hate speech solely based on isolated utterances.
- Contextualized: Incorporating preceding dialogue turns to provide context, this model addressed the conversational nature of online interactions, capturing discursive nuances critical for accurate classification.

Our experiments included three distinct modeling approaches:

- Logistic regression baseline: A traditional logistic regression model trained on TF-IDF vectorized text features served as an interpretable baseline, helping to ascertain whether simpler linear methods could effectively capture the linguistic features indicative of sexist and racist content.
- SpaCy TextCatBOW pipeline: As a lightweight neural baseline, we fine-tuned SpaCy's TextCatBOW architecture, which employs bag-of-words hash embeddings pooled via mean aggregation and fed into a single-layer feedforward classifier. Optimized with Adam and early stopping, this transformer-free model offers high inference speed and serves as a useful benchmark for isolating the impact of contextual embeddings.
- Transformer-based Pipeline (SpaCy + BETO): To leverage advanced contextual embeddings, we implemented and fine-tuned a transformer-based model utilizing SpaCy's transformer integration. The configuration incorporated the widely recognized BETO model (`dccuchile/bert-base-spanish-wwm-cased`, https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased, (accessed on 16 June 2025)), specifically configured as follows (see Appendix A for more details):
  - Used the SpaCy pipeline component `transformer` combined with a `textcat_multilabel` classifier.

- Implemented subtoken pooling strategies (mean pooling) and strided windows of 128 tokens with a stride of 224 tokens, ensuring comprehensive coverage of long conversational texts.
- Fine-tuned with the Adam optimizer (learning rate of 1×10^{-5}, dropout of 0.1), trained for up to 20 epochs or until convergence.

Table 6 summarizes the number of examples in each split. For the racism detection task in the contextual (CTX) setting, the training set consists of 792 dialogues, while oversampling the positive class increases it to 1096 examples. The corresponding development set contains 197 examples. Similarly, for sexism detection, the training set has 800 dialogues, increasing to 1258 when oversampled, with a development set of 199 examples. Oversampling was applied to mitigate class imbalance and enhance model sensitivity to the underrepresented positive class.

Table 6. Dataset splits: number of trainings, oversampled training, and development examples for racism and sexism detection tasks.

Dataset	Training	Oversampled Training	Development
Racism	792	1096	197
Sexism	800	1258	199

Models were trained using SpaCy's built-in training commands, specifying configurations (`--textcat-multilabel`, `--lang es`) to accurately tailor the training environment to Spanish-language multi-label classification tasks. Detailed hyperparameters, including batch sizes, gradient accumulation strategies, and GPU allocation, were explicitly documented to ensure reproducibility and transparency.

For comparative benchmarking and cross-domain generalization analysis, we also evaluated two publicly available state-of-the-art hate speech detection models:

- `piuba-bigdata/beto-contextualized-hate-speech`, https://huggingface.co/piuba-bigdata/beto-contextualized-hate-speech, (accessed on 16 June 2025) [12]: This BETO-based transformer model, originally fine-tuned on Spanish news comment sections, provided insights into model performance when applied to conversational data from different domains.
- `unitary/multilingual-toxic-xlm-roberta`, https://huggingface.co/unitary/multilingual-toxic-xlm-roberta, (accessed on 16 June 2025) [59]: Based on XLM-RoBERTa, this model was trained on multilingual Wikipedia talk page comments for toxic vs. non-toxic classification. Its broad multilingual scope enables evaluation of cross-lingual generalization to Spanish without fine-tuning, providing a benchmark for assessing how well generic models handle context-sensitive toxicity in domain-specific forums like Mediavida.

Model performance was systematically evaluated using four established classification metrics: precision, recall, F1 score, and ROC-AUC. Precision and recall measure model accuracy and sensitivity, particularly relevant given the dataset's imbalanced classes. The F1 score offered a balanced assessment combining precision and recall. ROC-AUC assessed overall discriminative power independently of classification thresholds, providing robust comparative insights across models.

Our comprehensive experimental framework enabled a rigorous comparative analysis of model effectiveness, interpretability, and domain transferability, delivering critical insights into the capabilities and limitations of contemporary Spanish-language hate speech detection systems, particularly within complex and context-dependent online dialogues.

3.5. Keyword Overlap Analysis

To systematically evaluate the lexical grounding of model predictions in explicit bias, we constructed a comprehensive Spanish slur dictionary by combining multiple high-coverage resources. The primary source was the Spanish lexicon from HurtLex [80], a multilingual lexicon of offensive and hateful words annotated by semantic category. To further increase coverage of colloquial and internet-specific slurs, we integrated terms from the BigScience ROOTS Corpus's [60] offensive language list.

After merging these resources, we manually curated the resulting list to remove duplicates and harmonize token forms (e.g., stripping diacritics, converting to lowercase, and standardizing variants). In order to reduce false positives arising from homonyms or benign polysemic words, we developed a custom whitelist of Spanish terms that are misidentified as slurs but are not actually offensive in context (for example, animal names or technical terms appearing in HurtLex). All whitelist entries were explicitly excluded from the final dictionary used for model evaluation.

The resulting slur lexicon was stored in a plain text file, with one term per line, and served as the reference for keyword overlap analysis.

To detect the presence of slurs in model-predicted positive and negative examples, we implemented a substring-matching approach in Python 3.14. For each annotated dialogue, the script lowercases the input text and checks whether any slur in the lexicon appears as a substring. Each example is thus flagged as containing or not containing a known slur, enabling both aggregate and case-by-case analyses of lexical overlap. This method, while simple, provides high recall for explicit toxic language and offers a transparent, reproducible basis for quantifying the reliance of models on explicit lexical cues.

This procedure ensures that our keyword overlap analysis meaningfully distinguishes between lexical and context-dependent toxicity, providing insight into both the strengths and weaknesses of the evaluated classifiers when exposed to diverse forms of bias in Spanish-language online dialogues.

4. Results

This section details the annotation statistics and distribution of the dataset, along with a comparative analysis of the effectiveness of the model.

4.1. Annotation Statistics and Dataset Distribution

The finalized annotated datasets from the Mediavida forum contained a total of 1001 dialogues annotated for sexism and 989 dialogues annotated for racism (see Table 7). The sexism dataset exhibited a class imbalance, comprising 221 positive examples (22.1%) and 780 negative examples (77.9%). The racism dataset had a slightly higher proportion of positive instances, with 299 examples labeled as racist (30.2%) and 690 labeled non-racist (69.8%).

Table 7. Annotation statistics for sexism and racism dialogues.

Metric	Sexism Dataset	Racism Dataset
Total annotated dialogues	1001	989
Positive examples (%)	22.1%	30.2%
Negative examples (%)	77.9%	69.8%
Mean tokens per example	133.3	123.7
Median tokens	103	94
Maximum tokens	957	784
Cohen's Kappa (κ)	0.55	0.79

In terms of textual characteristics, dialogues annotated for sexism had an average length of 133.3 tokens per example, with a median of 103 tokens and a maximum length of 957 tokens. Dialogues annotated for racism were shorter on average, containing 123.7 tokens per example, a median of 94 tokens, and a maximum of 784 tokens. The observed distribution highlights that sexism-related discussions tended to be more extensive, reflecting possibly deeper conversational contexts.

Inter-annotator agreement measured through Cohen's Kappa yielded moderate agreement for sexism ($\kappa = 0.55$) and substantial agreement for racism ($\kappa = 0.79$), confirming annotation reliability and guideline effectiveness.

In addition to sexism and racism, initial annotation efforts were undertaken for two other bias categories: homophobia and aporophobia. The homophobia dataset consisted of 2153 examples but contained only 20 positive instances (0.9%), while the aporophobia dataset had 1352 examples with just 9 positive cases (0.7%). Due to the extremely limited presence of these labels in the data, indicative of the relatively infrequent explicit appearance of these biases within the analyzed dialogues, further dataset creation and subsequent model training for these two categories were not pursued.

4.2. Model Performance: Sexism and Racism Detection

Table 8 summarizes the classification results obtained from our models and compares their performance against the latest models from Hugging Face. Domain-specific models, fine-tuned using the SpaCy (v3.8.7) pipeline on the Mediavida corpus, exhibited varying effectiveness based on the use of contextual information.

Table 8. Evaluation results * for sexism and racism detection across multiple models. LogReg baseline, TextCatBOW, and transformer (BETO) models were fine-tuned on the Mediavida corpus using SpaCy, with single-turn (ST) and contextual (CTX) variants.

Model	Train/FT	Evaluate	Precision	Recall	F1 Score	ROC–AUC
Sexism Detection						
LogReg baseline	ST	ST	0.52	0.24	0.33	0.77
LogReg baseline	CTX	CTX	0.72	0.43	0.54	0.82
TextCatBOW (SpaCy)	ST	ST	0.69	0.20	0.31	0.76
TextCatBOW (SpaCy)	CTX	CTX	0.64	0.51	0.57	0.81
HF BETO (SpaCy)	ST	ST	0.59	0.76	0.67	0.85
HF BETO (SpaCy)	CTX	CTX	0.64	0.71	0.67	0.87
HF PIUBA-CONTEXTUALIZED	–	ST	0.86	0.13	0.23	0.84
HF PIUBA-CONTEXTUALIZED	–	CTX	1.00	0.08	0.15	0.82
HF MULTILINGUAL-TOXIC-XLM-ROBERTA	–	ST	0.30	0.07	0.11	0.72
HF MULTILINGUAL-TOXIC-XLM-ROBERTA	–	CTX	0.50	0.12	0.20	0.62
Racism Detection						
LogReg baseline	ST	ST	0.86	0.53	0.66	0.87
LogReg baseline	CTX	CTX	0.93	0.46	0.61	0.89
TextCatBOW (SpaCy)	ST	ST	0.91	0.50	0.64	0.88
TextCatBOW (SpaCy)	CTX	CTX	0.89	0.60	0.72	0.88
HF BETO (SpaCy)	ST	ST	0.84	0.79	0.81	0.94
HF BETO (SpaCy)	CTX	CTX	0.75	0.75	0.75	0.90
HF PIUBA-CONTEXTUALIZED	–	ST	1.00	0.07	0.13	0.82
HF PIUBA-CONTEXTUALIZED	–	CTX	0.50	0.02	0.04	0.79
HF MULTILINGUAL-TOXIC-XLM-ROBERTA	–	ST	0.40	0.10	0.16	0.70
HF MULTILINGUAL-TOXIC-XLM-ROBERTA	–	CTX	0.79	0.20	0.32	0.65

* Results obtained using oversampled datasets for positive classes.

For sexism detection, the single-turn TextCatBOW (SpaCy) model yielded a precision of 69%, a recall of 20%, an F1 score of 31%, and a ROC-AUC of 0.76. The contextualized variant of TextCatBOW improved significantly in recall (51%) and F1 score (57%), despite slightly reduced precision (64%), and exhibited a higher ROC-AUC of 0.81. The Transformer-based BETO model trained within the SpaCy pipeline delivered strong overall performance, particularly notable in its single-turn variant, achieving a precision of 59%, a high recall of 76%, an F1 score of 67%, and a robust ROC-AUC of 0.85. Its contextual variant similarly performed well, with balanced precision (64%), recall (71%), F1 score (67%), and the highest ROC-AUC among our models at 0.87.

For racism detection, the single-turn TextCatBOW model demonstrated high precision (91%) but moderate recall (50%), resulting in an F1 score of 64% and ROC-AUC of 0.88. The contextualized TextCatBOW variant showed slightly lower precision (89%) but higher recall (60%), thus improving the F1 score to 72%, maintaining a ROC-AUC of 0.88. The Transformer-based BETO models significantly outperformed the other variants, particularly in the single-turn setting, obtaining a precision of 84%, a recall of 79%, an F1 score of 81%, and the highest ROC-AUC at 0.94. The contextualized BETO variant had a balanced precision and recall of 75%, resulting in an F1 score of 75% and ROC-AUC of 0.90.

The Logistic Regression baseline, trained with TF-IDF features, presented competitive but generally lower performance. For sexism detection, the single-turn variant showed limited effectiveness with precision at 52%, recall at 24%, and F1 at 33% (ROC-AUC of 0.77). The contextual variant showed improvement across metrics (precision 72%, recall 43%, F1 54%, ROC-AUC 0.82). For racism detection, Logistic Regression delivered notably better results, achieving a precision of 86%, recall of 53%, F1 score of 66%, and ROC-AUC of 0.87 in the single-turn setup. The contextual variant improved precision to 93% but reduced recall to 46%, resulting in a slightly lower F1 score of 61% and an ROC-AUC of 0.89.

4.3. Statistical Significance of Model Comparisons

We performed formal statistical significance testing to assess whether BETO's observed performance improvements over the baseline models are statistically reliable. Specifically, we applied McNemar's exact test (two-tailed) to paired model predictions on the development sets for both racism and sexism tasks, under both single-turn (ST) and contextual (CTX) evaluation settings. Table 9 reports F1 scores and p-values for each model pair and configuration.

Overall, BETO achieved the highest F1 scores in every setting, most notably in single-turn racism (F1 = 0.814) and single-turn sexism (F1 = 0.667), outperforming both the TextCatBOW (BOW) and logistic regression (LogReg) baselines. However, McNemar's test indicated that none of the pairwise differences between models reached statistical significance at the conventional $\alpha = 0.05$ level. The smallest p-values were found for the single-turn racism setting, where BETO's advantage over BOW ($p = 0.0522$) and over LogReg ($p = 0.0614$) was nearly significant, suggesting a strong but not statistically confirmed benefit.

In contextual settings, and for sexism detection overall, p-values were higher (all > 0.20), indicating no significant differences in error patterns between BETO and the baselines, despite higher F1 for BETO. This is likely attributable to moderate dataset size, as well as the difficulty of the tasks.

In summary, BETO consistently outperforms baseline models in absolute F1 score across all configurations; however, the improvements did not reach statistical significance in pairwise McNemar's tests. These results highlight BETO's practical utility for toxic language detection, while also suggesting that larger datasets may be needed to robustly establish significance for future model comparisons.

Table 9. McNemar's exact test *p*-values (two-tailed) comparing paired model predictions for single-turn (ST) and contextual (CTX) settings on sexism and racism dev sets. Bold marks $p < 0.05$ (statistically significant).

Setting	Model Pair	F1 Difference	*p*-Value
	Racism Detection		
ST	BETO vs. BOW	0.814 vs. 0.644	0.0522
ST	BETO vs. LogReg	0.814 vs. 0.660	0.0614
ST	BOW vs. LogReg	0.644 vs. 0.660	1.0000
CTX	BETO vs. BOW	0.745 vs. 0.717	0.8506
CTX	BETO vs. LogReg	0.745 vs. 0.610	0.5966
CTX	BOW vs. LogReg	0.717 vs. 0.610	0.2101
	Sexism Detection		
ST	BETO vs. BOW	0.667 vs. 0.305	0.4799
ST	BETO vs. LogReg	0.667 vs. 0.328	0.2203
ST	BOW vs. LogReg	0.305 vs. 0.328	0.4807
CTX	BETO vs. BOW	0.673 vs. 0.568	0.5966
CTX	BETO vs. LogReg	0.673 vs. 0.538	0.8506
CTX	BOW vs. LogReg	0.568 vs. 0.538	0.8318

4.4. Qualitative and Quantitative Error Analysis

To comprehensively assess model performance, we conducted qualitative and quantitative analyses of model errors using confusion matrices. This evaluation provided insights into frequent sources of false positives (FP) and false negatives (FN).

Confusion matrices (see Table 10) showed BETO models generally had better recall and balanced precision, particularly evident in the single-turn variants for both racism and sexism. For instance, BETO single-turn racism achieved a precision of 83.6% and a recall of 79.3%, highlighting robustness in capturing racist dialogues. Conversely, BOW models exhibited higher precision but notably lower recall, as demonstrated in BOW single-turn sexism with a precision of 69.2% but a low recall of 19.6%.

Table 10. Confusion matrices for selected BETO, BOW, and logistic regression models.

Model	Variant	TP	FP	FN
	Racism Detection			
LogReg baseline	ST	31	5	27
LogReg baseline	CTX	25	2	30
HF BETO (SpaCy)	ST	46	9	12
HF BETO (SpaCy)	CTX	41	14	14
TextCatBOW (SpaCy)	ST	29	3	29
TextCatBOW (SpaCy)	CTX	33	4	22
	Sexism Detection			
LogReg baseline	ST	11	10	35
LogReg baseline	CTX	21	8	28
HF BETO (SpaCy)	ST	35	24	11
HF BETO (SpaCy)	CTX	35	20	14
TextCatBOW (SpaCy)	ST	9	4	37
TextCatBOW (SpaCy)	CTX	25	14	24

Next, we illustrate the typical errors with selected qualitative examples:

- False negative (racism, BETO CTX): "Que ahora tengamos 1123 hijos por pareja no quiere decir que en 2090 vayamos a tener 0. Seguimos siendo superiores intelectual-

mente (Que es lo que hace predominar una raza sobre otra)…". English translation: "The fact that we have 1123 kids per couple does not mean that in 2090 we will have 0. We continue to be superior intellectually (Which is what makes a race predominate over another)". This dialogue explicitly mentions intellectual racial superiority, yet BETO CTX failed to detect racism.

- False positive (sexism, BOW CTX): "…En la empresa privada, sí hay que demostrar más, porque si no, te crujen, que siempre hay alguien jugándose la pasta. Sobre todo en puestos de cierta responsabilidad. Anda que no hay becarios haciendo el trabajo a "personas hechas a si mismas" que son hijos del jefe". English translation: "…In the private sector, you do have to prove more, because if you don't, they'll crack you, because there's always someone who's putting their money on the line. Especially in positions of some responsibility. There are no interns doing the work of 'self-made people' who are the boss's kids". While the content might be controversial, it does not inherently reflect sexism. However, the BOW CTX model mistakenly flagged it as sexist due to possible lexical overlap.

These examples both show the difficulty of the task and indicate a clear area for model improvement, especially emphasizing better contextual understanding and differentiation of implicit versus explicit bias expressions.

4.5. Comparative Analysis with External Models

Comparison with external Hugging-Face models underscores the significant advantage of fine-tuning domain-specific models for the Mediavida dataset. External models without fine-tuning, specifically `piuba-bigdata/beto-contextualized-hate-speech` and `unitary/multilingual-toxic-xlm-roberta`, exhibited consistent shortcomings when applied to our dataset.

For sexism detection, the `piuba-bigdata/beto-contextualized-hate-speech` model achieved high precision—86% in the single-turn variant and perfect precision (100%) in the contextual variant—but struggled considerably with recall, achieving only 13% and 8%, respectively. Consequently, the F1 scores were notably low at 23% (ST) and 15% (CTX), despite relatively high ROC-AUC scores (0.84 and 0.82, respectively). The multilingual `unitary/multilingual-toxic-xlm-roberta` model performed even less effectively, with a precision of 30% (ST) and 50% (CTX), and similarly low recall scores of 7% and 12%, resulting in F1 scores of merely 11% (ST) and 20% (CTX), accompanied by modest ROC-AUC scores of 0.72 and 0.62.

In racism detection, similar patterns emerged. The `piuba-bigdata/beto-contextualized-hate-speech` model maintained excellent precision (100%) in the single-turn scenario but failed with recall (7%), yielding an F1 score of 13%. Its contextual variant dropped significantly in performance, attaining only 50% precision and 2% recall, resulting in a minimal F1 score of 4% and a ROC-AUC of 0.79. Meanwhile, the `unitary/multilingual-toxic-xlm-roberta` model displayed limited precision (40% in ST and 79% in CTX) coupled with very low recall rates of 10% (ST) and 20% (CTX), producing F1 scores of 16% and 32%, and ROC-AUC values of 0.70 and 0.65, respectively.

These results confirm a substantial performance gap when relying on general-purpose models without domain-specific fine-tuning, highlighting the necessity of training models specifically for the characteristics, linguistic nuances, and interactional complexity inherent to forum-based dialogues like those found on Mediavida.

4.6. Keyword Overlap Analysis

We assessed the lexical sensitivity of our models by comparing their predictions against an extensive dictionary of Spanish slurs and biased expressions. The results (see Table 11)

revealed a very high prevalence of known slurs in predicted positive samples across all models, particularly evident in contextual variants, which achieved perfect (100%) slur coverage. This confirms that model predictions strongly rely on explicit lexical indicators of bias. Conversely, a substantial proportion of predicted negatives also included known slurs (between 66.20% and 98.12%), especially for contextual models. This indicates potential model limitations in capturing nuanced or contextually dependent biases, highlighting the importance of continued emphasis on context-sensitive annotation and training strategies to better handle implicit forms of online toxicity.

Table 11. Lexical overlap analysis: proportion of predicted positive and negative examples containing at least one slur from the HurtLex- [80] and the BigScience Roots-based [60] dictionary.

Model	Variant	Positives w/ Slur (%)	Negatives w/ Slur (%)
		Sexism	
BETO (SpaCy)	ST	89.83% (53/59)	69.29% (97/140)
BETO (SpaCy)	CTX	100.00% (55/55)	97.92% (141/144)
TextCatBOW	ST	92.31% (12/13)	74.19% (138/186)
TextCatBOW	CTX	100.00% (39/39)	98.12% (157/160)
		Racism	
BETO (SpaCy)	ST	96.36% (53/55)	66.20% (94/142)
BETO (SpaCy)	CTX	100.00% (55/55)	96.48% (137/142)
TextCatBOW	ST	96.88% (31/32)	70.30% (116/165)
TextCatBOW	CTX	100.00% (37/37)	96.88% (155/160)

5. Discussion

The results presented demonstrate critical insights into the automated detection of sexism and racism in Spanish-language forum dialogues. The *EsCorpiusBias* corpus offers important methodological advances through its careful consideration of contextual dynamics within dialogues, significantly contributing to the wider field of toxicity detection in NLP.

5.1. Annotation Challenges and Guideline Effectiveness

One of the key challenges encountered during annotation was capturing implicit or context-dependent forms of toxicity. Annotators frequently faced difficulties distinguishing between subtle humor, irony, or indirect references and genuine bias. For instance, detecting benevolent sexism and covert racism required extensive contextual understanding. The detailed annotation guidelines, supported by clear theoretical foundations and exemplified instances, considerably facilitated consistent labeling, although achieving high inter-annotator agreement for sexism ($\kappa = 0.55$) proved challenging due to inherent subtleties. A higher agreement for racism ($\kappa = 0.79$) was achieved. The work presented in this paper could be further improved by replicating the process with a higher number of annotators, reflecting the qualified opinions of a council of experts in sexism and racism.

5.2. Error Analysis and Interpretation of Results

The quantitative evaluation highlights that Transformer-based models (HF BETO) substantially outperformed traditional baseline methods (LogReg, TextCatBOW) in identifying racist and sexist comments. Particularly noteworthy is the contextualized BETO variant's balanced performance, underscoring the critical role that preceding dialogue context plays in accurately detecting nuanced expressions of bias. Conversely, externally-trained Hugging Face models like `piuba-bigdata/beto-contextualized-hate-speech` and `unitary/multilingual-toxic-xlm-roberta` displayed significant deficiencies in recall, revealing a strong reliance on lexical cues rather than contextually nuanced features.

Qualitative error analysis indicated that most classification errors arose from implicit forms of sexism or racism, especially in cases involving sarcasm or multi-turn ironic exchanges. Models struggled notably in differentiating genuine expressions of subtle bias from sarcastic or satirical comments, emphasizing the complexity of pragmatic phenomena and pointing to the necessity for advanced pragmatic modeling in future research.

5.3. Limitations of Current Models

Despite the observed successes, several limitations persist. Firstly, although contextual models demonstrated better performance, they remain susceptible to conversational nuances such as irony, indirect references, or implicit bias, which are common in longer online dialogues. Secondly, the class imbalance in the annotated dataset necessitated oversampling, potentially introducing biases or reducing the generalizability of the models to naturally balanced datasets. Additionally, current models demonstrated a strong dependency on explicit lexical features, highlighting a critical gap in handling implicit bias effectively.

Moreover, the application of general-purpose models to the Mediavida corpus revealed the limitations of transfer learning in hate speech detection across different domains and conversational contexts. Models trained predominantly on news comments or short social media texts exhibited insufficient adaptability to forum-based dialogues, underscoring the importance of domain-specific fine-tuning.

5.4. Effects and Limitations of Incorporating Multi-Turn Dialogical Context

Incorporating multi-turn dialogue context into annotation and model training is motivated by the need to capture forms of toxicity, such as irony, sarcasm, veiled hostility, or cumulative microaggressions, which are often invisible or ambiguous in isolated utterances. Unlike single-turn approaches, which risk missing the discursive and pragmatic cues surrounding a toxic comment, multi-turn context enables both annotators and models to evaluate intent, provocation, and evolving interpersonal dynamics.

For example, a reply like “claro, porque los de siempre nunca fallan” (“sure, because the usual suspects never fail”) may seem innocuous out of context, but when situated after a sequence of xenophobic exchanges, it reveals clear alignment with biased discourse. Similarly, indirect allusions, defensive humor, or quoted speech often only become legible as toxic when read within a chain of preceding comments, as shown by our dataset and examples of dialogues as shown in Tables 4 and 5.

Empirically, our contextualized models demonstrated improved recall and F1 scores over their single-turn counterparts (see Table 8), confirming that multi-turn inputs do help capture some subtle and pragmatic bias phenomena. Annotators also reported that the three-turn window frequently enabled more confident and nuanced labeling decisions, especially for cases involving sarcasm or allusion.

However, the observed improvements were modest. Several factors likely limit the full potential of context-aware modeling in our setting:

1. Insufficient contextual richness: The Mediavida forum often features short, rapidly-shifting dialogues, where adjacent turns may not always provide enough semantic or pragmatic information to reveal hidden toxicity.
2. Fixed context window size: Our two-turn window (preceding the target) may be too narrow for some conversations and not necessary for overly sexist or racist turns, so dynamic window sizes may be worth exploring.
3. Model reliance on lexical features: Our error analysis confirmed that, despite contextual input, models tend to prioritize explicit lexical markers (e.g., slurs), with subtle cues from surrounding turns often underweighted.

Thus, while multi-turn dialogical context offers clear qualitative advantages over traditional single-turn annotation, enabling both annotators and models to better detect pragmatic and implicit toxicity, the realized quantitative gains are currently bounded by the dataset's conversational depth and modeling limitations. Future work should explore variable-sized and adaptive context windows, as well as advanced attention-based architectures that can more effectively exploit nuanced dialogical structure.

5.5. Implications for Automatic Hate Speech Moderation

The implications of these findings are significant for automated moderation practices. The demonstrated advantages of contextually-aware, domain-specific fine-tuning underline the necessity of tailoring moderation systems specifically to the dialogue and interactional dynamics of targeted platforms. Incorporating contextually nuanced annotation guidelines and leveraging transformer-based models could substantially improve moderation accuracy, especially for subtle and implicit forms of bias. Such improvements hold considerable promise for reducing false negatives and false positives in automated content moderation, thereby enhancing digital community health and user experiences.

5.6. Future Work

Building on these findings, several promising directions for future research and system development could be considered:

- Cross-lingual and transfer learning: Future studies should investigate cross-lingual transfer learning approaches that leverage annotated data from multiple languages or domains to improve model generalizability and robustness, especially in languages or platforms with limited labeled data. Multilingual transformers and transfer learning can help bridge resource gaps and facilitate rapid adaptation to new domains.
- Semi-supervised and active learning: To address annotation scarcity and improve coverage of rare or subtle phenomena, employing semi-supervised learning (leveraging large amounts of unlabeled data) and active learning (prioritizing the most informative or uncertain samples for human annotation) could significantly improve model performance and annotation efficiency.
- Adaptive context modeling: Exploring architectures that dynamically select or weight relevant turns, rather than relying on fixed context windows, may yield better contextual understanding, especially for implicit bias and sarcasm. Techniques such as hierarchical attention or memory networks could be considered.
- Rich pragmatic and multimodal signals: Incorporating pragmatic cues (e.g., speaker intent, conversation roles, or thread structure) and multimodal information (e.g., accompanying images and metadata) could improve detection of implicit and nuanced forms of bias.
- Bias mitigation and fairness evaluation: Systematic analysis of model and annotation biases, including the cultural perceptions and subjectivities of annotators, should be incorporated, with transparent reporting and fairness audits.
- Multiple expert annotators: Engaging a higher number of annotators with varying expertise coming not only from computer science, but also from social science (to understand systemic sexism/racism and their social dynamics), language (to analyze nuanced language use), and community representatives with life experience from affected groups, among others.

These directions will not only address the limitations of the present study but also propel future research in automated, fair, and contextually grounded detection of toxicity and bias in online discourse across languages and platforms.

6. Conclusions

This paper introduced the *EsCorpiusBias* corpus, a contextually grounded and rigorously annotated resource specifically designed for detecting and mitigating online toxicity and bias in Spanish-language forums. Our methodological approach, focusing on multi-turn dialogue interactions sourced from the Mediavida forum, allowed us to capture nuanced manifestations of sexism and racism that emerge distinctly within conversational contexts.

Through extensive annotation, we demonstrated moderate to substantial inter-annotator agreement (Cohen's Kappa of 0.55 for sexism and 0.79 for racism), underscoring both the effectiveness of our detailed guidelines and the complexity inherent in identifying subtler forms of toxicity. The final dataset, comprising approximately 1000 dialogues each for sexism and racism, reflects realistic class distributions and provides a robust basis for training sophisticated NLP models.

Experimental results highlighted the critical role of domain-specific fine-tuning and contextual embedding strategies. Our transformer-based models, notably the BETO architecture fine-tuned within the SpaCy pipeline, consistently outperformed simpler logistic regression and TextCatBOW baselines across metrics (precision, recall, F1, and ROC-AUC). The findings further emphasized the inadequacy of externally trained general-purpose models in accurately identifying nuanced toxicity, reinforcing the importance of targeted training datasets and methods.

Keyword overlap analysis confirmed models' heavy reliance on explicit lexical cues, suggesting ongoing challenges in accurately capturing implicit and context-dependent toxicity. This points toward critical avenues for future research, including the development of advanced pragmatic modeling and more effective handling of implicit bias.

However, our approach is not without limitations. One notable challenge is the inherent subjectivity in contextual labeling, particularly in distinguishing subtle or implicit forms of bias, such as sarcasm, irony, and microaggressions. Annotators, despite rigorous training, might still be influenced by individual biases, cultural perceptions, and subjective interpretations, potentially affecting label consistency. Additionally, the three-turn context window, while effective for capturing immediate conversational dynamics, might occasionally fail to encompass sufficient contextual depth necessary for interpreting complex or evolving interactions.

Furthermore, our dataset reflects a specific online community (Mediavida forum), potentially limiting the generalizability of our findings to other contexts, platforms, or demographic groups. The scarcity of annotated examples for homophobia and aporophobia also highlights challenges related to the availability and representation of certain bias categories within our corpus.

Overall, our contributions provide significant methodological and resource advancements for socially responsible NLP, facilitating improved automated moderation systems tailored explicitly to Spanish-language online interactions. Future work should continue to refine contextual modeling capabilities, integrate richer pragmatic reasoning frameworks, and explore multilingual and multimodal approaches to enhance the adaptability and effectiveness of hate speech detection systems.

Author Contributions: Conceptualization: Z.C., D.G. and D.P.-F.; methodology: Z.C., D.G., A.G.-F., K.K. and D.P.-F.; software: K.K. and D.P.-F.; validation: Z.C., D.G. and J.G.-H.; formal analysis: Z.C., D.G., A.G.-F., J.G.-H., K.K. and D.P.-F.; investigation: Z.C., D.G., A.G.-F., J.G.-H., K.K. and D.P.-F.; resources, Z.C. and D.G.; data curation, K.K. and J.G.-H.; writing—original draft preparation, K.K. and D.G.; writing—review and editing, Z.C., D.G. and D.P.-F.; visualization, K.K.; supervision, Z.C. and D.G.; project administration, Z.C. and D.G.; funding acquisition: Z.C. and D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This dataset and publication is part of the "CONVERSA: Effective and efficient resources and models for transformative conversational AI in Spanish and co-official languages" project with reference (Agencia Estatal de Investigación) TED2021-132470B-I00, funded by MCIN/AEI/10.13039/501100011033 and by the European Union 'NextGenerationEU/PRTR'.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The EsCorpiusBias corpus, including all annotated Spanish dialogue data and relevant metadata generated and analyzed during this study, is publicly available via Zenodo at https://doi.org/10.5281/zenodo.15637906. This dataset is provided under an open license CC-BY-NC-ND 4.0 to support transparency, reproducibility, and further research in the community. Researchers are encouraged to access, use, and cite the resource in accordance with the terms of use provided at the repository.

Acknowledgments: We gratefully acknowledge our annotators and Prodigy's tooling for efficient data curation. We would also like to thank Juan Albacete-Maza (from IQS-Universitat Ramon Llull) for his expert advice on aporophobia. During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-4o, 2025) for the purposes of generating LaTeX code, reformulating author-written passages for clarity and adjusting bibliographic references to meet the formatting requirements of the journal. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
BETO	A BERT model trained on a big Spanish corpus
BOW	Bag-of-words
CTX	Dialogue in context with two preceding turns
FN	False negative
FP	False positive
FT	Fine-tuning
LLM	Large Language Model
NLP	Natural Language Processing
ROC-AUC	The area under the ROC curve
ST	Single-turn comment

Appendix A

BETO uses exactly the same architecture as BERT-Base, as Table A1 shows. BERT (bidirectional encoder representations from transformers) consists of stacked transformer encoder layers that produce contextualized embeddings for each input token.

BETO is a bidirectional model, which means that it takes into account both the left and right context of a word when representing it.

For this work, the BETO model, structurally identical to BERT-base but pretrained on large-scale Spanish corpora, is employed and further fine-tuned for toxicity detection.

The transformer-based BETO model was fine-tuned within the spaCy pipeline using the spaCy transformers library. Specifically, we utilized the pre-trained BETO `dccuchile/bert-base-spanish-wwm-cased` model from Hugging Face, renowned for its efficacy on a broad range of Spanish NLP tasks.

Table A1. Architecture and main parameters of BETO.

Parameter	Value
Model type	Transformer (encoder only)
Layers	12
Attention heads	12
Hidden dimension	768
Total parameters	110 million
Tokens per input	Up to 512 tokens

Appendix A.1. Data Preparation

The annotated dialogues were first pre-processed into spaCy's `DocBin` format, optimized for efficient storage and processing. Each document contained the relevant single or three-turn dialogue unit, preserving contextual structure and labels. Positive and negative samples were balanced through oversampling of the minority class to mitigate class imbalance during training.

Appendix A.2. Model Configuration and Hyperparameters

For fine-tuning, we employed the standard spaCy pipeline configuration with transformers, as detailed in Table A2.

Table A2. Main hyperparameters for fine-tuning the BETO (`dccuchile/bert-base-spanish-wwm-cased`) model in SpaCy's contextualized TextCat pipeline.

Component	Hyperparameter	Value
Transformer (BETO)	Pretrained model name	`dccuchile/bert-base-spanish-wwm-cased`
	Maximum word-piece tokens per window	512 (model default), split into windows of 128 WP tokens
	Stride between windows	224 WP tokens
Tokenizer/Batching	SpaCy pipeline batch size	16
	Batcher schedule (words per batch)	Compounding from 100 to 1000 (factor 1.001)
	Discard oversize examples	`true`
	Batcher tolerance	0.1
Training Schedule	Dropout	0.1
	Patience (early stopping)	0
	Maximum epochs	20
	Evaluation frequency	Every 200 updates
Optimizer (Adam)	Learning rate (α)	1×10^{-5}
	L_2 regularization	0.01
	Gradient clipping	1.0
	β_1	0.9
	β_2	0.999
	ϵ	1×10^{-8}
TextCat_Multilabel	Classification threshold	0.5
	Tok2vec pooling strategy	mean pooling over transformer outputs
	Linear BoW branch	Enabled (ngram size = 1, vocabulary length = 262,144)

Appendix A.3. Training Procedure

The training was executed using spaCy's recommended `spacy train` CLI command with a custom configuration file (`config.cfg`), specifying the pipeline components and hyperparameters mentioned above. The best-performing model checkpoint was automatically selected based on the lowest validation loss.

Appendix A.4. Evaluation Metrics and Validation

Performance was systematically evaluated using precision, recall, F1 score, and ROC-AUC on an independent validation dataset (20% of total samples), consistent with the standard practices for binary text classification. Additionally, confusion matrices and error analyses were conducted to understand false positives and false negatives, informing further model refinements.

Appendix A.5. Reproducibility

To ensure reproducibility, all training configurations, seeds for random number generators, and dataset splits have been documented and stored. The trained models and configuration files have been archived and made available upon request for replicating the results presented in this study.

References

1. Gallegos, I.O.; Rossi, R.A.; Barrow, J.; Tanjim, M.M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; Ahmed, N.K. Bias and Fairness in Large Language Models: A Survey. *Comput. Linguist.* **2024**, *50*, 1097–1179. [CrossRef]
2. Blodgett, S.L.; Lopez, G.; Olteanu, A.; Sim, R.; Wallach, H. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing (IJCNLP), Bangkok, Thailand, 1–6 August 2021; pp. 1004–1015. [CrossRef]
3. Meade, N.; Poole-Dayan, E.; Reddy, S. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022. [CrossRef]
4. Kim, H.; Yu, Y.; Jiang, L.; Lu, X.; Khashabi, D.; Kim, G.; Choi, Y.; Sap, M. ProsocialDialog: A Prosocial Backbone for Conversational Agents. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 4005–4029. [CrossRef]
5. Radaideh, M.I.; Kwon, O.H.; Radaideh, M.I. Fairness and social bias quantification in Large Language Models for sentiment analysis. *Knowl.-Based Syst.* **2025**, *319*, 113569. [CrossRef]
6. Khalatbari, L.; Bang, Y.; Su, D.; Chung, W.; Ghadimi, S.; Sameti, H.; Fung, P. Learn What NOT to Learn: Towards Generative Safety in Chatbots. *arXiv* **2023**, arXiv:2304.11220. [CrossRef]
7. Blodgett, S.L.; Barocas, S.; Daumé, H., III; Wallach, H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5454–5476. [CrossRef]
8. Wei, X.; Kumar, N.; Zhang, H. Addressing bias in generative AI: Challenges and research opportunities in information management. *Inf. Manag.* **2025**, *62*, 104103. [CrossRef]
9. Campbell, H.; Goldman, S.; Markey, P.M. Artificial intelligence and human decision making: Exploring similarities in cognitive bias. *Comput. Hum. Behav. Artif. Humans* **2025**, *4*, 100138. [CrossRef]
10. Savoldi, B.; Bastings, J.; Bentivogli, L.; Vanmassenhove, E. A decade of gender bias in machine translation. *Patterns* **2025**, *6*, 101257. [CrossRef]
11. Taulé, M.; Nofre, M.; Bargiela, V.; Bonet, X. NewsCom-TOX: A corpus of comments on news articles annotated for toxicity in Spanish. *Lang. Resour. Eval.* **2024**, *58*, 1115–1155. [CrossRef]
12. Pérez, J.M.; Luque, F.M.; Zayat, D.; Kondratzky, M.; Moro, A.; Serrati, P.S.; Zajac, J.; Miguel, P.; Debandi, N.; Gravano, A.; et al. Assessing the Impact of Contextual Information in Hate Speech Detection. *IEEE Access* **2023**, *11*, 30575–30590. [CrossRef]
13. Ariza-Casabona, A.; Schmeisser-Nieto, W.S.; Nofre, M.; Taulé, M.; Amigó, E.; Chulvi, B.; Rosso, P. Overview of DETESTS at IberLEF 2022: DETEction and Classification of Racial Stereotypes in Spanish. *Proces. Leng. Nat.* **2022**, *69*, 217–228.
14. Kostikova, A.; Wang, Z.; Bajri, D.; Pütz, O.; Paaßen, B.; Eger, S. LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models. *arXiv* **2025**, arXiv:2505.19240. [CrossRef]
15. Rowe, J.; Klimaszewski, M.; Guillou, L.; Vallor, S.; Birch, A. EuroGEST: Investigating gender stereotypes in multilingual language models. *arXiv* **2025**, arXiv:2506.03867. [CrossRef]
16. Chandna, B.; Bashir, Z.; Sen, P. Dissecting Bias in LLMs: A Mechanistic Interpretability Perspective. *arXiv* **2025**, arXiv:2506.05166. [CrossRef]

17. Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J.A.; Celi, L.A.; Gichoya, J.; Jurafsky, D.; Szolovits, P.; Bates, D.W.; Abdulnour, R.E.E.; et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *Lancet Digit. Health* **2024**, *6*, e12–e22. [CrossRef]
18. Ivetta, G.; Gomez, M.J.; Martinelli, S.; Palombini, P.; Echeveste, M.E.; Mazzeo, N.C.; Busaniche, B.; Benotti, L. HESEIA: A community-based dataset for evaluating social biases in large language models, co-designed in real school settings in Latin America. *arXiv* **2025**, arXiv:2505.24712. [CrossRef]
19. Borenstein, N.; Stanczak, K.; Rolskov, T.; da Silva Perez, N.; Klein Kafer, N.; Augenstein, I. Measuring Intersectional Biases in Historical Documents. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 2711–2730. [CrossRef]
20. Rinki, M.; Raj, C.; Mukherjee, A.; Zhu, Z. Measuring South Asian Biases in Large Language Models. *arXiv* **2025**, arXiv:2505.18466. [CrossRef]
21. Prabhune, S.; Padmanabhan, B.; Dutta, K. Do LLMs have a Gender (Entropy) Bias? *arXiv* **2025**, arXiv:2505.20343. [CrossRef]
22. Santagata, L.; De Nobili, C. More is more: Addition bias in large language models. *Comput. Hum. Behav. Artif. Humans* **2025**, *3*, 100129. [CrossRef]
23. Shao, J.; Lu, Y.; Yang, J. Benford's Curse: Tracing Digit Bias to Numerical Hallucination in LLMs. *arXiv* **2025**, arXiv:2506.01734. [CrossRef]
24. Zahraei, P.S.; Emami, A. Translate with Care: Addressing Gender Bias, Neutrality, and Reasoning in Large Language Model Translations. *arXiv* **2025**, arXiv:2506.00748. [CrossRef]
25. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186. [CrossRef]
26. May, C.; Wang, A.; Bordia, S.; Bowman, S.R.; Rudinger, R. On Measuring Social Biases in Sentence Encoders. *arXiv* **2019**, arXiv:1903.10561. [CrossRef]
27. Nozza, D.; Bianchi, F.; Hovy, D. HONEST: Measuring Hurtful Sentence Completion in Language Models. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Online, 6–11 June 2021; pp. 2398–2406. [CrossRef]
28. Rudinger, R.; Naradowsky, J.; Leonard, B.; Van Durme, B. Gender Bias in Coreference Resolution. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), New Orleans, LA, USA, 1–6 June 2018. [CrossRef]
29. Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; Chang, K.W. Gender Bias in Contextualized Word Embeddings. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 629–634. [CrossRef]
30. Vanmassenhove, E.; Emmery, C.; Shterionov, D. NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender Neutral Alternatives. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 7–11 November 2021; pp. 8940–8948. [CrossRef]
31. Webster, K.; Recasens, M.; Axelrod, V.; Baldridge, J. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 605–617. [CrossRef]
32. Pant, K.; Dadu, T. Incorporating Subjectivity into Gendered Ambiguous Pronoun (GAP) Resolution using Style Transfer. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Seattle, WA, USA, 15 July 2022; pp. 273–281. [CrossRef]
33. Levy, S.; Lazar, K.; Stanovsky, G. Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November, 2021; pp. 2470–2480. [CrossRef]
34. Nadeem, M.; Bethke, A.; Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing (IJCNLP), Bangkok, Thailand, 1–6 August 2021*; Zong, C., Xia, F., Li, W., Navigli, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 5356–5371. [CrossRef]
35. Bartl, M.; Nissim, M.; Gatt, A. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, Barcelona, Spain, 13 December 2020; pp. 1–16. [CrossRef]
36. Nangia, N.; Vania, C.; Bhalerao, R.; Bowman, S.R. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 1953–1967. [CrossRef]

37. Felkner, V.; Chang, H.C.H.; Jang, E.; May, J. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Toronto, ON, Canada, 9–14 July 2023; pp. 9126–9140. [CrossRef]
38. Barikeri, S.; Lauscher, A.; Vulić, I.; Glavaš, G. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing (IJCNLP), Online, 1–6 August 2021; pp. 1941–1955. [CrossRef]
39. Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; Petrov, S. Measuring and Reducing Gendered Correlations in Pre-trained Models. *arXiv* **2021**, arXiv:2010.06032. [CrossRef]
40. Qian, R.; Ross, C.; Fernandes, J.; Smith, E.M.; Kiela, D.; Williams, A. Perturbation Augmentation for Fairer NLP. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 11 December 2022; pp. 9496–9521. [CrossRef]
41. Kiritchenko, S.; Mohammad, S.M. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *arXiv* **2018**, arXiv:1805.04508. [CrossRef]
42. Dev, S.; Li, T.; Phillips, J.; Srikumar, V. On Measuring and Mitigating Biased Inferences of Word Embeddings. *arXiv* **2019**, arXiv:1908.09369. [CrossRef]
43. Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; Smith, N.A. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 3356–3369. [CrossRef]
44. Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.W.; Gupta, R. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, FAccT'21, Online, 3–10 March 2021; pp. 862–872. [CrossRef]
45. Smith, E.M.; Hall, M.; Kambadur, M.; Presani, E.; Williams, A. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 9180–9211. [CrossRef]
46. Huang, Y.; Zhang, Q.; Y, P.S.; Sun, L. TrustGPT: A Benchmark for Trustworthy and Responsible Large Language Models. *arXiv* **2023**, arXiv:2306.11507. [CrossRef]
47. Lin, X.; Li, L. Implicit Bias in LLMs: A Survey. *arXiv* **2025**, arXiv:2503.02776. [CrossRef]
48. Fersini, E.; Rosso, P.; Anzovino, M. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), Seville, Spain, 18 September 2018; CEUR Workshop Proceedings; Volume 2150, pp. 214–228.
49. Álvarez-Carmona, M.; Guzmán-Falcón, E.; Montes-Gómez, M.; Escalante, H.J.; Villaseñor-Pineda, L.; Reyes-Meza, V.; Rico-Sulayes, A. Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In Proceedings of the 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval), Seville, Spain, 18 September 2018; Volume 6.
50. Aragón, M.E.; Jarquín-Vásquez, H.J.; Montes-Gómez, M.; Escalante, H.J.; Pineda, L.V.; Gómez-Adorno, H.; Posadas-Durán, J.P.; Bel-Enguix, G. Overview of MEX-A3T at IberLEF 2020: Fake News and Aggressiveness Analysis in Mexican Spanish. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) at SEPLN, Malaga, Spain, 23–25 September 2020; pp. 222–235.
51. Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F.M.; Rosso, P.; Sanguinetti, M. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 54–63. [CrossRef]
52. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and Monitoring Hate Speech in Twitter. *Sensors* **2019**, *19*, 4654. [CrossRef]
53. Rodríguez-Sánchez, F.; de Albornoz, J.C.; Plaza, L.; Gonzalo, J.; Rosso, P.; Comet, M.; Donoso, T. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Proces. Leng. Nat.* **2021**, *67*, 195–207.
54. Rodríguez-Sánchez, F.; de Albornoz, J.C.; Plaza, L.; Mendieta-Aragón, A.; Marco-Remón, G.; Makeienko, M.; Plaza, M.; Gonzalo, J.; Spina, D.; Rosso, P. Overview of EXIST 2022: sEXism Identification in Social neTworks. *Proces. Leng. Nat.* **2022**, *69*, 229–240.
55. del Arco, F.M.P.; Casavantes, M.; Escalante, H.J.; Martín-Valdivia, M.T.; Montejo-Ráez, A.; y Gómez, M.M.; Jarquín-Vásquez, H.; Villaseñor-Pineda, L. Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants. *Proces. Leng. Nat.* **2021**, *67*, 183–194.

56. Bourgeade, T.; Cignarella, A.T.; Frenda, S.; Laurent, M.; Schmeisser-Nieto, W.S.; Benamara, F.; Bosco, C.; Moriceau, V.; Patti, V.; Taulé, M. A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 674–684. [CrossRef]
57. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish Pre-Trained BERT Model and Evaluation Data. In Proceedings of the PML4DC at ICLR 2020, Online, 26 April 2020. [CrossRef]
58. Paula, A.F.M.D.; Silva, R.F.D.; Schlicht, I.B. Sexism Prediction in Spanish and English Tweets Using Monolingual and Multilingual BERT and Ensemble Models. In Proceedings of the CEUR Workshop Proceedings, Kharkiv, Ukraine, 20–21 September 2021; pp. 356–373. [CrossRef]
59. Hanu, L.; Unitary Team. Detoxify. Github. 2020. Available online: https://github.com/unitaryai/detoxify (accessed on 16 June 2025).
60. Laurençon, H.; Saulnier, L.; Wang, T.; Akiki, C.; Villanova del Moral, A.; Le Scao, T.; Von Werra, L.; Mou, C.; González Ponferrada, E.; Nguyen, H.; et al. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 31809–31826. [CrossRef]
61. Buie, H.; Croft, A. The Social Media Sexist Content (SMSC) Database: A Database of Content and Comments for Research Use. *Collabra Psychol.* **2023**, *9*, 71341. [CrossRef]
62. Rodríguez-Sánchez, F.J.; Carrillo-de Albornoz, J.; Plaza, L. Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. *IEEE Access* **2020**, *8*, 219563–219576. [CrossRef]
63. Bhattacharya, S.; Singh, S.; Kumar, R.; Bansal, A.; Bhagat, A.; Dawer, Y.; Lahiri, B.; Ojha, A.K. Developing a Multilingual Annotated Corpus of Misogyny and Aggression. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, Marseille, France, 16 May 2020; pp. 158–168. [CrossRef]
64. Jahan, M.S.; Oussalah, M. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing* **2023**, *546*, 126232. [CrossRef]
65. Mouka, E.; Saridakis, I. *Racism Goes to the Movies: A Corpus-Driven Study of Cross-Linguistic Racist Discourse Annotation and Translation Analysis*; Language Science Press: Berlin, Germany, 2015; pp. 35–69. [CrossRef]
66. Kumar, R.; Reganti, A.N.; Bhatia, A.; Maheshwari, T. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018*; Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., et al., Eds.; European Language Resources Association: Paris, France, 2018. [CrossRef]
67. Meyer, E.J. Gendered harassment in secondary schools: Understanding teachers' (non) interventions. *Gend. Educ.* **2008**, *20*, 555–570. [CrossRef]
68. Poteat, V.P.; Rivers, I. The use of homophobic language across bullying roles during adolescence. *J. Appl. Dev. Psychol.* **2010**, *31*, 166–172. [CrossRef]
69. Fraïssé, C.; Barrientos, J. The concept of homophobia: A psychosocial perspective. *Sexologies* **2016**, *25*, e65–e69. [CrossRef]
70. Thurlow, C. Naming the "outsider within": Homophobic pejoratives and the verbal abuse of lesbian, gay and bisexual high-school pupils. *J. Adolesc.* **2001**, *24*, 25–38. [CrossRef] [PubMed]
71. Chakravarthi, B.R.; Priyadharshini, R.; Ponnusamy, R.; Kumaresan, P.K.; Sampath, K.; Thenmozhi, D.; Thangasamy, S.; Nallathambi, R.; McCrae, J.P. Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments. *arXiv* **2021**, arXiv:2109.00227. [CrossRef]
72. Vásquez, J.; Andersen, S.; Bel-enguix, G.; Gómez-adorno, H.; Ojeda-trueba, S.l. HOMO-MEX: A Mexican Spanish Annotated Corpus for LGBT+phobia Detection on Twitter. In *Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH), Toronto, ON, Canada, 13 July 2023*; Chung, Y.l., Röttger, P., Nozza, D., Talat, Z., Mostafazadeh Davani, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 202–214. [CrossRef]
73. Orts, A.C. *Aporofobia, el Rechazo al Pobre. Un Desafío Para la Democracia*; Paidós: Barcelona, Spain, 2017; p. 200.
74. Crenshaw, K. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanf. Law Rev.* **1991**, *43*, 1241–1299. [CrossRef]
75. Comim, F.; Borsi, M.T.; Valerio Mendoza, O. *The Multi-Dimensions of Aporophobia*; MPRA Paper 103124; University Library of Munich: Munich, Germany, 2020.
76. Bell, W. Reforming the Poor. *Soc. Work* **1972**, *17*, 119. [CrossRef]
77. Niño Argüelles, Y.L.; Álvarez Santana, C.L.; Giovanni Locatelli, F. Migración Venezolana, Aporofobia en Ecuador y Resiliencia de los Inmigrantes Venezolanos en Manta, Periodo 2020. *Rev. San Gregor.* **2020**, *43*, 92–108.
78. Martínez-Navarro, E. Aporofobia. In *Glosario Para una Sociedad Intercultural*; Conill, J., Ed.; Bancaja: Valencia, Spain, 2002; pp. 17–23.

79. Picado, E.M.V.; Yurrebaso Macho, A.; Guzmán Ordaz, R. Respuesta social ante la aporofobia: Retos en la intervención social. *IDP. Rev. Internet Derecho PolíTica* **2022**, *37*, 1–19. [CrossRef]
80. Bassignana, E.; Basile, V.; Patti, V. Hurtlex: A Multilingual Lexicon of Words to Hurt. In Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, 10–12 December 2018. [CrossRef]