# Balancing the Scales: Addressing Gender Bias in Social Media Toxicity Detection

**Beatriz Botella-Gil** CENID University of Alicante
**Juan Pablo Consuegra-Ayala** CENID University of Alicante
**Alba Bonet-Jover** DLSI University of Alicante
**Paloma Moreda-Pozo** DLSI University of Alicante
{beatriz.botella, juan.consuegra, alba.bonet, moreda}@ua.es

## Abstract

The detection of toxic content in social media has become a critical task in Natural Language Processing (NLP), particularly given its intersection with complex issues like subjectivity, implicit language, and cultural context. Among these challenges, bias in training data remains a central concern—especially as language models risk reproducing and amplifying societal inequalities. This paper investigates the interplay between toxicity and gender bias on Twitter/X by introducing a novel dataset of violent and non-violent Spanish tweets, annotated not only for violence but also for gender. We conduct an exploratory analysis of how biased data can distort toxicity classification and present algorithms to mitigate these effects through dataset balancing and debiasing. Our contributions include four new dataset splits—two balanced and two debiased—that aim to support the development of fairer and more inclusive NLP models. By foregrounding the importance of equity in data curation, this work lays the groundwork for more ethical approaches to automated violence detection and gender annotation.

## 1 Introduction

Technological advances have exponentially increased digital communication speed, transforming social interaction. People can now express opinions and emotions in real-time, generating vast amounts of digital information. This overload, known as "infoxication" (Navas-Martin et al., 2012; Leyton, 2016), leads to anxiety and limits constructive exchange (Vania and Ruíz, 2015). Similarly, "infodemiology" or "infodemic" (Eysenbach, 2002; Gallotti et al., 2020; Moyano et al., 2024) describes the rapid spread of unreliable information, especially during periods of high information abundance.

This excessive, often toxic, data hinders users' ability to identify relevant content. As Navas-Martin et al. note, more information can mean less knowledge when time is spent processing irrelevant content. The spread of unverified, biased information—particularly on social media—fosters polarization, as these platforms allow immediate, unfiltered, and often anonymous exchanges. The resulting toxic language reflects and amplifies societal biases such as racism, sexism, and xenophobia. When this biased content is used to train Language Models (LM), it risks replicating and normalizing toxic discourse.

This highlights the need for ethical filters to clean data before it feeds Artificial Intelligence (AI) systems. Natural Language Processing (NLP) plays a key role in detecting and mitigating toxicity and bias, ensuring models represent inclusive and respectful communication.

This paper addresses toxicity and bias in social media, focusing on Twitter/X, by introducing a novel Spanish dataset about violence, annotated with gender, and balanced to support the development of fairer and more inclusive NLP models. The main contributions are:

- A resource consisting of violent and non-violent Spanish tweets, annotated by violence but also by gender, which allows us to determine the level of gender bias with respect to the toxicity of the message.
- An exploratory analysis of the relationship between toxicity in information and biased data.
- The development and application of algorithms to balance and debias the dataset by selectively removing sentences, ensuring fairer representation across categories, and mitigating potential sources of bias in subsequent analyses.



*Proceedings of Recent Advances in Natural Language Processing*,pages 194–203
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-023

The article is structured as follows: Section 2 introduces the concepts of toxicity and bias, and the importance of research into the relationship between toxic datasets and biased data; Section 3 presents the methodology followed, the annotated data employed for this exploratory study and the balancing strategies proposed; Section 4 summarizes the results obtained from our experimental evaluation and the discussion of the research; and finally, Section 5 presents the conclusions and future work.

## 2 Related Work

This section presents some state-of-the art research working on toxicity in NLP either in (1) online discourse, (2) bias, and (3) the interconnection between the two.

### 2.1 Toxicity in NLP

The term toxic is usually defined as an "umbrella term to refer to any form of content, including but not limited to hate speech, cyberbullying, abusive speech, misogyny, sexism, offense, and obscenity" (Garg et al., 2023). Other authors, such as Salehabadi et al., define toxicity as "rude, disrespectful, or unreasonable comment".

The detection of toxicity through NLP can be approached from various perspectives, each with specific advantages and challenges. One commonly employed strategy involves the use of heuristic techniques, where rule-based systems classify messages according to predefined patterns established by human experts. These rules are often supported by lexicons that act as indicators of violent content, enabling explicit classification of text. However, while this approach is effective in terms of interpretability, it presents challenges due to the significant time investment required for designing the rules (Huang et al., 2023). Despite these challenges, heuristic techniques offer predictions that are more interpretable compared to advanced methods such as Machine Learning (ML) or Deep Learning (DL), which, although precise, often lack transparency in their internal workings (Clarke et al., 2023).

ML, on the other hand, has emerged as a fundamental tool in the detection of toxicity, with approaches classified into supervised (Plaza-del Arco et al., 2021; Haddad et al., 2019), semi-supervised (Wiegand and Siegel, 2018; Rosenthal et al., 2020), and unsupervised learning (De la Peña Sarracén and Rosso, 2022; Balakrishnan et al., 2023).

### 2.2 Bias in Toxicity Detection

In recent years, research on biases in NLP, particularly in the context of toxic or violent speech, has gained significant relevance (Costa-jussà et al., 2023; Sahoo et al., 2022). Various studies have analyzed how NLP models, when trained on large amounts of data from online platforms, can perpetuate and amplify pre-existing social biases, particularly those related to race, gender, and other demographic attributes. These biases can affect the effectiveness and fairness of automatic toxicity detection systems in online discourse.

Several studies have worked on mitigating bias in toxicity detection and have used toxic speech datasets to evaluate and mitigate the presence of bias, such those presented in the survey of Garg et al., who presented a key study in this area by conducting a comprehensive review of how to address biases in toxic speech detection. This work provides an overview of the existing strategies and methods for mitigating biases in NLP models, highlighting the need for more systematic approaches to address biases related to the demographic and social characteristics of users. The authors argue that biases in training data, stemming from how data is collected and labeled, are one of the main sources of bias in toxicity detection.

In another approach, Mozafari et al. focus on hate speech detection and racial bias mitigation using BERT-based models. This work demonstrates how NLP models, especially those trained with neural networks such as BERT, can exhibit racial biases when analyzing discourse on social media platforms. The authors propose a racial bias mitigation approach by fine-tuning BERT models to reduce biases against racial minorities and improve the accuracy of hate speech detection without amplifying racial discrimination. This work highlights how the use of pre-trained models can be both a challenge and an opportunity to improve fairness in online content moderation.

Furthermore, an innovative approach is presented by Cheng et al., who propose a sequential decision-making approach for bias mitigation in toxicity detection. According to this study, NLP models can enhance their ability to mitigate biases through incremental adaptation, where decisions about the toxicity of comments are made in successive stages, allowing biases to be adjusted as more data is received. This dynamic approach contrasts with traditional models, which assume all

information is available upfront and make a single decision.

Finally, the work of Sap et al. focuses on racial bias in hate speech detection. This study emphasizes how models trained on large online text corpora can acquire and propagate racial biases due to statistical correlations in the data. The authors show that tweets in African American English (AAE) and tweets written by self-identified African Americans are significantly more likely to be labeled as offensive, reflecting a racial bias in automated labeling. The authors propose methods to reduce these biases, such as dialect priming and racial priming, which help annotators become more aware of different dialects and avoid racial bias in data labeling.

Together, these studies highlight the complexity and importance of addressing biases in toxic and violent speech detection. While the methodologies vary, from general strategy reviews to innovative approaches such as sequential decision-making or racial bias analysis, all agree on the need to develop fairer and more impartial NLP models.

## 3 Methodology

In this section, we will explain the methodology we used to conduct our research on biases in a corpus that annotates violence, as well as the annotation process we followed to identify biases.

### 3.1 Corpus

Our research is based on the VILLANOS corpus (Botella-Gil et al., 2024), a dataset comprising over 6,000 tweets and containing both violent and non-violent messages. To ensure the quality of the dataset and evaluate the accuracy and effectiveness of the annotation guidelines applied during its construction, an inter-annotator agreement analysis was conducted. This analysis employed Cohen's Kappa formula (Cohen, 1960), a widely recognized metric commonly used for two annotators, as it accounts for agreement due to chance.

The main classification level (violent/non-violent) yielded a Kappa index of 0.95, which is considered almost perfect agreement. This high level of annotator consistency was a key factor in selecting this corpus for our bias research. Furthermore, the dataset not only provides a high degree of reliability but also offers a valuable opportunity to work with violent messages, which are essential for addressing the objectives of our study.

### 3.2 Annotation

To build our corpus ToBi[1], focused on bias and toxicity, we annotated a total of 1,000 tweets randomly selected from the VILLANOS corpus. This dataset was carefully balanced, comprising 500 tweets categorized as violent and 500 as non-violent, ensuring representativeness across both categories.

The annotation process was carried out by two annotators, experts in linguistics and criminology. They followed a structured set of guidelines designed to systematically capture relevant information. These guidelines consisted of two main steps:

1. **Identification of the mention**: At this stage, annotators analyzed whether the message explicitly referred to a man, a woman, or both. This step aimed to identify potential gender biases in the content of the tweets, enabling a more detailed analysis of how mentions were distributed in relation to the message's violent or neutral nature. For example, the following sentence mentions a female: *"Isabel es una gran actriz"* —Isabel is a great actress.

2. **Annotation of the target**: For non-violent messages, annotators evaluated whether the message was directly addressed to a specific individual or represented a general statement without a clear addressee. This distinction helped differentiate personal communications from broader abstract ideas or contexts. For example, the following sentence targets just a female: *"Irene, la mujer de Pepe, no sabe cocinar"* —Irene, Pepe´s wife, doesn't know how to cook.

   For violent messages, the annotators recorded who was the target of the violence. This could include an explicitly mentioned individual or a broader group, depending on the tweet's context. This step was crucial to understand the intentionality behind the violent content and its potential impact. For example, the following sentence targets a female: *"Ana, eres idiota."* —Ana, you're an idiot.

To ensure consistency in the annotation process, annotators participated in training sessions before starting the annotation task. Additionally, periodic reviews were conducted to resolve potential discrepancies in the interpretation of the guidelines. This

[1]Provisional name for peer review. The final version of the corpus will be available to the scientific community in case the article is accepted.

methodological approach allowed us to construct a highly precise annotated corpus, suitable for investigating the relationship between bias, toxicity, and violence in language.

Finally, two tweets were removed as they contained the same textual message, and the goal was to include different texts. As a result, we have 499 violent messages and 499 non-violent messages.

### 3.3 Measurements

A systematic approach is employed to quantify fairness and balance in the annotated corpus. The analysis considers two key aspects: mentions of specific gender groups and their role as targets of violence. These aspects enable the observation of how fairness and balance manifest in contexts where individuals of a specific gender are mentioned in violent or non-violent texts and, in particular, in situations where they are targets of expressed violence, each presenting distinct characteristics.

Two primary metrics are used to assess fairness and balance: *statistical parity* and *representation parity*.

- **Statistical parity** measures whether a gender group (male or female) is equally likely to be assigned to the non-violent category compared to the other gender group. Maximum balance is achieved when the probability of being classified as violent is the same for both genders. Conversely, maximum disparity occurs when all examples from one gender are classified as violent, while none from the other gender are classified as violent. See Equation 1.

$$P(\textit{Y=non-violent} \mid \textit{P=female}) = P(\textit{Y=non-violent} \mid \textit{P=male}) \quad (1)$$

- **Representation parity** captures the imbalance in how often males and females are mentioned or targeted, across the *violent* and *non-violent* categories. Maximum balance is achieved when both classes (violent and non-violent) contain an equal number of examples for each gender. In contrast, maximum disparity occurs when one gender is entirely absent from one of the classes. See Equations 2 and 3.

$$P(\textit{Y=violent}, \textit{P=female}) = P(\textit{Y=violent}, \textit{P=male}) \quad (2)$$

$$P(\textit{Y=non-violent}, \textit{P=female}) = P(\textit{Y=non-violent}, \textit{P=male}) \quad (3)$$

While statistical parity focuses on fairness by examining the distribution of violence across gender groups, representation parity addresses the balance within the corpus itself by studying the representation of gender across the violent and non-violent classes.

Both the difference and ratio versions of these metrics are included, providing distinct perspectives on potential disparities. The difference version highlights disparities by showing the absolute differences in the probabilities between gender groups, offering a clear view of the magnitude of imbalance in concrete terms. In contrast, the ratio version emphasizes relative magnitudes, illustrating the proportional relationship between the probabilities of assignment for the two gender groups, which can be particularly informative for understanding the scale of disparity. The difference is computed as $max_{g \in G}(prob_g) - min_{g \in G}(prob_g)$, where $G = \{\textit{male}, \textit{female}\}$ The ratio is computed as $\frac{min_{g \in G}(prob_g)}{max_{g \in G}(prob_g)}$.

The results of these analyses, including statistical parity and representation parity, are presented in detail in Section 4.2. These metrics provide a foundation for identifying imbalances in the corpus and for evaluating the effectiveness of the proposed balancing strategies.

### 3.4 Balancing Strategies

The objective of this step is to generate fair and balanced versions of the manually annotated dataset. By addressing imbalances in representation and fairness across gender groups and categories, the aim is to create dataset splits that can be selected based on the requirements of specific use cases. This process ensures that the models trained on these datasets are less prone to propagate existing biases and can adapt to different fairness objectives.

Four balanced dataset splits are created as part of this process:

**Balanced by Mentions:** Ensures that mentions of males and females are equally represented, irrespective of whether they are targets.

**Balanced by Targets:** Focuses on equal representation of males and females as targets, independent of mentions.

**Debiased by Mentions:** Balances fairness by ensuring statistical parity in labels (violent/non-

violent) for males and females when they are mentioned.

**Debiased by Targets:** Similar to the previous split, but ensures statistical parity in labels when males and females are targets.

These splits allow flexibility in choosing a dataset that best aligns with the downstream application's fairness and representation goals.

#### 3.4.1 Balancing Algorithm

The balancing process involves automatically adjusting the dataset to meet the objectives of representation or fairness. Algorithms[2] outline the balancing procedures implemented for this purpose. The first algorithm, **balance**, creates datasets with equal representation, while the second algorithm, **debias**, emphasizes fairness by ensuring balanced category-specific distributions.

The key distinction between the two lies in how examples are included for each combination of gender: exclusively female-related, exclusively male-related, both female and male-related, or unrelated to gender. The balancing algorithm ensures an equal number of male- and female-related examples. In contrast, the debiasing algorithm focuses solely on ensuring that each gender is equally represented within violent and non-violent examples, without requiring an equal total number of male and female examples. Specifically, the debiasing algorithm sets the probability ratio to 0.5, ensuring that male and female examples have an equal likelihood of being classified as violent.

The splits resulting from applying these algorithms to the manually annotated datasets are detailed in Section 4.2, offering a comprehensive overview of their statistics.

## 4 Experiments

This section summarizes the findings from our experimental evaluation. Section 4.1 outlines the experimental setup, including the software used (Section 4.1.1). Section 4.2 then presents the experimental results, offering quantitative insights into the performance and efficacy of our approach. Finally, Section 4.3 explores the implications of these findings.

[2]The algorithms used in this study are available in the supplementary resources https://github.com/bfair-ml/bias-in-toxicity-supplementary-materials

### 4.1 Experimental Setup

Our experimental setup aims to assess the fairness and balance of gender representation in a corpus under various experimental conditions. The final version of the corpus is publicly available online[3].

We conducted the experiments on the manually annotated dataset, as well as on four additional splits to explore fairness-related scenarios (see Section 3.4). Metrics to quantify fairness and balance in the annotated corpus (see Section 3.3) are reported alongside disaggregated results for females and males, enabling a detailed analysis of gender disparities. By systematically modifying the dataset composition, we aim to evaluate the impact of these interventions on gender disparities and overall fairness metrics.

#### 4.1.1 Libraries

All resources analyzed in this study are accessible through our publicly available Python library, `BFair`[4].

### 4.2 Results

Table 1 summarizes the distribution of data across five different experimental splits: *manually annotated*, *balanced-by-mentions*, *balanced-by-targets*, *debiased-by-mentions*, and *debiased-by-targets*. For each split, the table presents the total number of instances, along with a breakdown of violent and non-violent cases under the *Violence* category. Additionally, the *Mentions* category shows the number of instances where no gender was mentioned, or where female, male, or both genders were referenced. Similarly, the *Targets* category provides counts of instances targeting no one, targeting females, targeting males, or targeting both genders. This comprehensive overview highlights variations in data composition across the splits, illustrating the impact of balancing and debiasing methods on gender representation and violence classification.

Table 2 summarizes the fairness and balance measurements derived from the manually annotated corpus. The table presents both **Mentions** and **Targets** categories, with measurements computed for statistical parity and representation parity across violent ($\mathcal{V}$) and non-violent ($\mathcal{NV}$) contexts. Specifically, the metrics include differences (Diff), ratios (Ratio), and disaggregated results for females and males within each category. These metrics offer

[3]https://doi.org/10.5281/zenodo.14264921

[4]https://github.com/bfair-ml/bfair

| Split | Total | Violence | | Mentions | | | | Targets | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Yes | No | N | F | M | M&F | N | F | M | M&F |
| Manually Annotated | 998 | 499 | 499 | 281 | 525 | 51 | 141 | 582 | 328 | 59 | 29 |
| Balanced-by-Mentions | 228 | 114 | 114 | 76 | 38 | 38 | 76 | 145 | 37 | 36 | 10 |
| Balanced-by-Targets | 120 | 60 | 60 | 20 | 55 | 20 | 25 | 40 | 38 | 38 | 4 |
| Debiased-by-Mentions | 868 | 434 | 434 | 250 | 504 | 38 | 76 | 522 | 297 | 38 | 11 |
| Debiased-by-Targets | 842 | 421 | 421 | 242 | 466 | 42 | 92 | 504 | 296 | 38 | 4 |

Table 1: Summary of the manually annotated dataset and its balanced and debiased splits. Columns **N**, **F**, **M**, and **M&F** stand for no one, females, males, both genders, respectively.

insights into the relative disparities and proportions observed in the dataset, providing a quantitative basis for evaluating gender fairness in the corpus.

Tables 3 and 4 summarize the fairness and balance measurements computed from two different splits of the source dataset: *balanced-by-mentions* and *balanced-by-targets*. These tables provide a detailed breakdown of statistical parity and representation parity for both mentions and targets, with measurements further categorized by gender (female and male) and contextual attributes (violent, $\mathcal{V}$, and non-violent, $\not\mathcal{V}$). Table 3 focuses on the corpus balanced by mentions, where the balance is maintained with respect to the gender distribution of mentions. Table 4, on the other hand, presents analogous metrics for the corpus balanced by targets, highlighting the differences in gender and contextual representation when the balance is shifted to the targets.

Tables 5 and 6 present the fairness and balance measurements obtained from two corpora processed with debiasing strategies: *debiased-by-mentions* and *debiased-by-targets*. These tables report metrics such as statistical parity and representation parity, split across mentions and targets, and further categorized by gender (female and male) and contextual attributes (violent, $\mathcal{V}$, and non-violent, $\not\mathcal{V}$). Table 5 focuses on debiasing applied at the level of mentions, highlighting how this approach affects the gender and contextual distribution. In contrast, Table 6 provides analogous measurements for the corpus debiased by targets, showcasing the implications of this alternative debiasing strategy.

### 4.3 Discussion

As shown in Table 2, the manually annotated corpus exhibits some imbalance and bias (all differences and ratios are larger than $10\,\%$ and $20\,\%$, respectively). Although the violent and non-violent classes are equally represented (see Table 1), the gender distribution within these classes is uneven (see representation parity, columns *Female* and *Male* in Table 2). In terms of the total number of messages, there is a greater representation of messages that mention and/or target females. However, when examining relative distributions, messages that mention or target males are more likely to be violent compared to those involving females (see statistical parity, columns *Female* and *Male* in Table 2). This imbalance poses challenges for automating violence classification, as models may inadvertently form incorrect associations between male-related terms and violence or non-violence. To address these challenges, balanced and debiased splits of the dataset have been created. These curated, though smaller, versions of the original dataset are better suited for specific tasks, effectively mitigating bias and imbalance.

As shown in Tables 3 and 4, applying Balance Algorithm generates two splits of the manually annotated dataset, where gender mentions or targets are balanced across the violent and non-violent classes, resulting also in a debiased version of the dataset (see columns *Diff* and *Ratio* for mentions or targets depending on the split). In contrast, applying Debias Algorithm creates two splits of the dataset that ensure equal probabilities of violence and non-violence for both genders, without enforcing representation balance (see statistical parity, columns *Diff* and *Ratio* for mentions or targets depending on the split). This method results in a larger dataset compared to the balanced version, making it better suited for fairness-sensitive tasks, especially when the ML model and training process are resilient to data imbalance.

When comparing the distribution of gender across mentions and targets in the manually annotated dataset, it is evident that targets exhibit greater representation and fairness disparities than mentions (see columns *Diff* and *Ratio* across *Mentions* and *Targets* in Table 2). This indicates that, while

| **Measurements** *(manually annotated)* | **Mentions** | | | | **Targets** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Diff** | **Ratio** | **Female** | **Male** | **Diff** | **Ratio** | **Female** | **Male** |
| Statistical Parity | 0.102 | 0.219 | 0.467 | 0.365 | 0.182 | 0.432 | 0.420 | 0.239 |
| Representation P. $[\mathcal{V}]$ | 0.467 | 0.656 | 0.711 | 0.244 | 0.281 | 0.676 | 0.415 | 0.134 |
| Representation P. $[\not\mathcal{V}]$ | 0.483 | 0.775 | 0.623 | 0.140 | 0.259 | 0.860 | 0.301 | 0.042 |

Table 2: Fairness and balance measurements from the *manually annotated* corpus. Symbols $\mathcal{V}$ and $\not\mathcal{V}$ stand for violent and non-violent, respectively.

| **Measurements** *(balanced-by-mentions)* | **Mentions** | | | | **Targets** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Diff** | **Ratio** | **Female** | **Male** | **Diff** | **Ratio** | **Female** | **Male** |
| Statistical Parity | 0.000 | 0.000 | 0.500 | 0.500 | 0.099 | 0.234 | 0.426 | 0.326 |
| Representation P. $[\mathcal{V}]$ | 0.000 | 0.000 | 0.500 | 0.500 | 0.035 | 0.129 | 0.237 | 0.272 |
| Representation P. $[\not\mathcal{V}]$ | 0.000 | 0.000 | 0.500 | 0.500 | 0.044 | 0.250 | 0.175 | 0.132 |

Table 3: Fairness and balance measurements from the *balanced-by-mentions* corpus. Symbols $\mathcal{V}$ and $\not\mathcal{V}$ stand for violent and non-violent, respectively.

| **Measurements** *(balanced-by-targets)* | **Mentions** | | | | **Targets** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Diff** | **Ratio** | **Female** | **Male** | **Diff** | **Ratio** | **Female** | **Male** |
| Statistical Parity | 0.001 | 0.003 | 0.487 | 0.489 | 0.000 | 0.000 | 0.500 | 0.500 |
| Representation P. $[\mathcal{V}]$ | 0.300 | 0.439 | 0.683 | 0.383 | 0.000 | 0.000 | 0.350 | 0.350 |
| Representation P. $[\not\mathcal{V}]$ | 0.283 | 0.436 | 0.650 | 0.367 | 0.000 | 0.000 | 0.350 | 0.350 |

Table 4: Fairness and balance measurements from the *balanced-by-targets* corpus. Symbols $\mathcal{V}$ and $\not\mathcal{V}$ stand for violent and non-violent, respectively.

| **Measurements** *(debiased-by-mentions)* | **Mentions** | | | | **Targets** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Diff** | **Ratio** | **Female** | **Male** | **Diff** | **Ratio** | **Female** | **Male** |
| Statistical Parity | 0.000 | 0.000 | 0.500 | 0.500 | 0.125 | 0.276 | 0.451 | 0.327 |
| Representation P. $[\mathcal{V}]$ | 0.537 | 0.803 | 0.668 | 0.131 | 0.313 | 0.805 | 0.389 | 0.076 |
| Representation P. $[\not\mathcal{V}]$ | 0.537 | 0.803 | 0.668 | 0.131 | 0.283 | 0.885 | 0.320 | 0.037 |

Table 5: Fairness and balance measurements from the *debiased-by-mentions* corpus. Symbols $\mathcal{V}$ and $\not\mathcal{V}$ stand for violent and non-violent, respectively.

| **Measurements** *(debiased-by-targets)* | **Mentions** | | | | **Targets** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Diff** | **Ratio** | **Female** | **Male** | **Diff** | **Ratio** | **Female** | **Male** |
| Statistical Parity | 0.021 | 0.043 | 0.491 | 0.470 | 0.000 | 0.000 | 0.500 | 0.500 |
| Representation P. $[\mathcal{V}]$ | 0.506 | 0.750 | 0.675 | 0.169 | 0.306 | 0.860 | 0.356 | 0.050 |
| Representation P. $[\not\mathcal{V}]$ | 0.501 | 0.770 | 0.651 | 0.150 | 0.306 | 0.860 | 0.356 | 0.050 |

Table 6: Fairness and balance measurements from the *debiased-by-targets* corpus. Symbols $\mathcal{V}$ and $\not\mathcal{V}$ stand for violent and non-violent, respectively.

differences exist between male and female mentions, these disparities are even more pronounced when examining individuals targeted by violence. In the sampled messages from the source dataset (VILLANOS), violence is more frequently associated with female-related text (see representation parity, columns *Female* and *Male* in Table 2), whereas male-related texts are more likely to be violent in nature (see statistical parity, columns *Female* and *Male* in Table 2). However, these findings cannot be generalized to the broader universe of messages on the internet. The source dataset was specifically curated to study a particular time period marked by spikes in violence, during which discussions happened to center on women. However, it is not necessarily representative of typical behavior on social networks.

To gain a more comprehensive understanding of this topic in society, a more uniform and representative sample is needed. This would involve studying a random sample of messages from broader media sources.

The created splits serve distinct purposes, with the balanced splits being particularly suitable for tasks such as automatic gender annotation, where an even representation of classes is critical to ensure fairness and reduce bias in model training. In contrast, the debiased splits are more appropriate for applications like violent versus non-violent content classification, where the removal of confounding factors minimizes unintended correlations and enhances the model's focus on the primary classification task. This approach ensures that the splits are optimized for their respective applications, promoting both fairness in gender-related tasks and robustness in detecting violence-related content.

## 5 Conclusion and Future Work

Mitigating biases in NLP models is a crucial task, particularly given the continued growth of violence in the digital sphere. It is essential that toxicity analysis focuses not only on identifying violent discourse but also on biases toward certain social groups. Due to the need to analyze toxic behavior across social groups, this research focuses on carrying out a preliminary study of bias in online toxicity information.

The main contribution of this work is the creation of a balanced and unbiased dataset from a base corpus in Spanish, using bias mitigation strategies. It is important to ensure the creation of resources that do not amplify human biases and to train fair and representative predictors for all the society, that do not marginalize anyone. For that reason, the development of this resource may enable the training and evaluation of models for future violence predictors or automatic gender annotators.

In the future, we plan to expand this work by incorporating other datasets and exploring new domains and data sources to assess the applicability of our approach in diverse contexts. Generating larger corpora is a key step in improving the robustness of bias detection models. Furthermore, we plan to apply our methodology to additional tasks, such as disinformation detection by mitigating bias in fake news or contradictions datasets, where biases also play a significant role.

While we have followed a reduction strategy, in the future, we would like to improve our data using an expansion strategy approach. A larger corpus could be used to further enhance the models' ability to identify and mitigate biases effectively. Ensuring that the data is balanced and that the models are representative of the entire society is essential to providing equitable assessments and services, without marginalizing any group.

## Acknowledgments

This research work is part of the R&D&I projects: NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021) funded by the Generalitat Valenciana; CLEAR.TEXT: Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033. Also, the VIVES: "Pla de Tecnologies de la Llengua per al valencià" project (2022/TL22/00215334) from the Projecte Estratègic per a la Recuperació i Transformació Econòmica (PERTE); SOCIALFAIRNESS.SOCIALTOX/SOCIALTRUST (PDC2022-133146-C21/PDC2022-133146-C21C22), funded by MCIN/AEI/10.13039/501100011033/ and by the European Union NextGenerationEU/PRTR. Also, this work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

## References


Flor Miriam Plaza-del Arco, Arturo Montejo-Ráez, L Alfonso Urena Lopez, and María-Teresa Martín-Valdivia. 2021. OffendES: A new corpus in Spanish for offensive language research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108.

Vimala Balakrishnan, Vithyatheri Govindan, and Kumanan N Govaichelvan. 2023. Tamil offensive language detection: Supervised versus unsupervised learning approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–14.

Beatriz Botella-Gil, Robiert Sepúlveda-Torres, Alba Bonet-Jover, Patricio Martínez-Barco, and Estela Saquete. 2024. Semi-automatic dataset annotation applied to automatic violent message detection. *IEEE Access*, 12:19651–19664.

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

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule by example: Harnessing logical rules for explainable hate speech detection. *arXiv preprint arXiv:2307.12935*.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Gunther Eysenbach. 2002. Infodemiology: The epidemiology of (mis) information. *The American journal of medicine*, 113(9):763–765.

Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of 'infodemics' in response to covid-19 epidemics. *Nature human behaviour*, 4(12):1285–1293.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A Tunisian hate speech and abusive dataset. In *International Conference on Arabic Language Processing*, pages 251–263. Springer.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 90–93.

Wbeimar Antonio Castro Leyton. 2016. El problema de la infoxicación en el manejo de correos electrónicos corporativos. *Tecnología Investigación y Academia*, 4(1):136–141.

Daniela Luz Moyano, María Silveria Agulló-Tomás, and Vanessa Zorrilla-Muñoz. 2024. Género, infodemia y desinformación en salud. revisión de alcance global, vacíos de conocimiento y recomendaciones. *Global Health Promotion*, 31(2):70–79.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Miguel Ángel Navas-Martin, Laura Albornos-Muñoz, and Cintia Escandell-García. 2012. Acceso a fuentes de información sobre salud en españa: cómo combatir la infoxicación. *Enfermería clínica*, 22(3):154–158.

Gretel Liz De la Peña Sarracén and Paolo Rosso. 2022. Unsupervised embeddings with graph auto-encoders for multi-domain and multilingual hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2196–2204, Marseille, France. European Language Resources Association.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. A large-scale semi-supervised dataset for offensive language identification. *arXiv preprint arXiv:2004.14454*.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nazanin Salehabadi, Anne Groggel, Mohit Singhal, Sayak Saha Roy, and Shirin Nilizadeh. 2022. User engagement and the toxicity of tweets. *arXiv preprint arXiv:2211.03856*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Cecilia Tovilla Quesada Vania and Patricia Trujano Ruíz. 2015. Infoxicación, angustia, ansiedad y web semántica. *Razón y palabra*, (92):1–27.

Michael Wiegand and Melanie Siegel. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of KONVENS 2018*.