information

Article

Identifying and Mitigating Gender Bias in Social Media Sentiment Analysis: A Post-Training Approach on Example of the 2023 Morocco Earthquake

Mohammad Reza Yeganegi, Hossein Hassani and Nadejda Komendantova *

Cooperation and Transformative Governance Group, Advancing Systems Analysis Program, International Institute for Applied Systems Analysis (IIASA), 2361 Laxenburg, Austria; yeganegi@iiasa.ac.at (M.R.Y.); hassani@iiasa.ac.at (H.H.)
* Correspondence: komendan@iiasa.ac.at

Abstract

Sentiment analysis is a cornerstone in many contextual data analyses, from opinion mining to public discussion analysis. Gender bias is one of the well-known issues in sentiment analysis models, which can produce different results for the same text depending on the gender it refers to. This gender bias leads to further bias in other text analyses that use such sentiment analysis models. This study reviews existing solutions to reduce gender bias in sentiment analysis and proposes a new method to address this issue. The proposed method offers more practical flexibility as it focuses on sentiment estimation rather than model training. Furthermore, it provides a quantitative measure to investigate the gender bias in sentiment analysis results. The performance of the proposed method across five sentiment analysis models is presented using texts containing gender-specific words. The proposed method is applied to a set of social media posts related to Morocco's 2023 earthquake to estimate the gender-unbiased sentiment of the posts and evaluate the gender-unbiasedness of five different sentiment analysis models in this context. The result shows that, although the sentiments estimated with different models are very different, the gender bias in none of the models is drastically large.

Keywords: social media; sentiment analysis; gender bias; natural language processing

Academic Editor: Nirmalya Thakur

Received: 7 May 2025
Revised: 4 August 2025
Accepted: 4 August 2025
Published: 8 August 2025

Citation: Yeganegi, M.R.; Hassani, H.; Komendantova, N. Identifying and Mitigating Gender Bias in Social Media Sentiment Analysis: A Post-Training Approach on Example of the 2023 Morocco Earthquake. *Information* **2025**, *16*, 679. https://doi.org/10.3390/info16080679

1. Introduction

Sentiment analysis has become a cornerstone in natural language processing (NLP), enabling the automated interpretation of subjective information in text. It plays a critical role in various applications, including opinion mining, market analysis, social media monitoring, and public health surveillance. Despite the technical advances in sentiment analysis methodologies—ranging from lexicon-based models to deep learning architectures—there is growing evidence that these systems often exhibit unintended biases, particularly gender bias [1,2]. Sentiment analysis has been widely used for a variety of purposes, including detecting misinformation and disinformation, extracting relevant content and appropriate responses, opinion mining, and analyzing event trends—whether from textual sources or digital footprints such as Google search data [3,4].

AI systems are trained on historical and human-generated data, which often contain gender stereotypes and inequalities. Consequently, such systems can not only reflect but also amplify existing biases [5,6].

One of the most cited examples is Amazon's experimental hiring algorithm, which was found to downgrade résumés containing the word "women's" [7]—as in "women's chess club captain"—because it had been trained on past hiring patterns favoring male applicants [8–10]. Various studies have highlighted the impact of gender bias and gender-related stereotyping on machine learning training systems. Some of these studies focused on the gender bias in the training data as the core issue [8], while others worked on rating [9] and reducing [10,11] these types of biases. Similarly, facial recognition technologies have shown disproportionately high error rates for women and people with darker skin tones, raising significant concerns about fairness and accuracy in real-world applications [5,9,11].

These issues are not merely technical flaws but reflections of broader societal biases. Gender bias in AI can perpetuate discrimination and exacerbate social inequalities unless actively addressed through inclusive design, diverse training datasets, and transparent evaluation metrics [9]. As AI continues to evolve, ensuring fairness and accountability must become central principles in its development and deployment.

However, gender bias in sentiment analysis refers to systematic disparities in sentiment outputs based solely on the gender-specific content of the text, even when the underlying semantic meaning remains unchanged [11,12].

This bias can arise from multiple sources. In lexicon-based models, gender bias is often embedded in the word sentiment scores, which are influenced by the sociocultural biases of the annotators or corpora used for constructing sentiment lexicons [13]. In machine learning models, bias primarily stems from imbalances or stereotypes present in training datasets [5,14]. Since these models are designed to learn patterns from data, any gender-related skewness or societal bias present in the training material is likely to be perpetuated and even amplified during inference. This not only undermines the reliability and fairness of sentiment analysis outcomes but can also perpetuate harmful stereotypes when these models are deployed in sensitive domains.

Several solutions have been proposed to mitigate gender bias in NLP systems. Methods such as gender-swapping data augmentation [2], adversarial training [15], and bias regularization [16] have shown promise. However, many of these techniques require access to model internals or retraining from scratch, which is often impractical for deployed systems or proprietary models [17]. Moreover, mitigation techniques that alter the training process may still be limited by the representativeness and quality of the underlying datasets [12,13]. This becomes even more challenging when using the data from X platform [18].

In this study, we propose a novel post-training approach for mitigating gender bias in sentiment analysis that does not rely on retraining or modifying the original models. Instead, our method operates during the sentiment estimation phase by systematically identifying gendered words, substituting them with gender-swapped or gender-neutral alternatives, and adjusting the resulting sentiment outputs accordingly. This procedure ensures that the sentiment assigned to a text is independent of the referenced gender, thereby promoting fairness and robustness across diverse contexts.

By building on formal definitions of gender bias in sentiment outputs and introducing a Sentiment Gender Bias Index (SGBI), we offer a flexible and interpretable framework for both detecting and correcting gender bias in existing sentiment analysis systems. Our results, validated across multiple sentiment analysis libraries and a variety of gendered texts, demonstrate that this post-training mitigation technique can effectively reduce bias without compromising model utility.

The following sections provide the theoretical foundations for sentiment analysis, discuss the formalization of gender bias in sentiment scoring, describe the proposed correction method, and present comprehensive evaluations of its effectiveness.

2. Theoretical Background

The sentiment analysis models categorize texts into positive, negative, and neutral sentiment classes while providing a sentiment intensity measure. There are two main approaches to measuring the sentiment intensity and class of a text. Some models provide a sentiment score to represent the estimated sentiment class and intensity simultaneously.

For a positive threshold λ and estimated sentiment score δ, the text is categorized as positive sentiment if $\delta > \lambda$, negative sentiment if $\delta < -\lambda$, and neutral sentiment if $-\lambda \leq \delta \leq \lambda$. The sentiment score provides a univariate sentiment measure that is easy to interpret. However, the classification results are highly sensitive to the threshold λ.

Another approach is to estimate the probability of each of the three sentiment categories. Next, the text can be assigned to the sentiment class that has the highest probability. This approach does not need thresholds and can provide a confidence measure along with sentiment class estimation. The estimated probabilities can be used as the intensity measure. Furthermore, using sentiment class probabilities represents sentiment analysis in a three-dimensional space.

In addition to different approaches for measuring sentiment intensity and class, there are different methods to estimate those measures. Consider a text consisting of sentences $s_1, \ldots, s_n$. Furthermore, suppose the sentence s_i consists of words $w_{i,j}$, $j = 1, \ldots, k_i$. One approach to measure the sentiment and intensity of each sentence s_i is to use a lexicon library to assign a sentiment score to each of the words $w_{i,j}$. These are lexicon-based sentiment analysis methods. A lexicon library (like the NRC library) provides the sentiment score for words in a language. Suppose the sentiment score of the word $w_{i,j}$ is $\delta_{i,j}$. If $w_{i,j}$ has positive sentiment, then $\delta_{i,j}$ is positive, and if $w_{i,j}$ has negative sentiment, $\delta_{i,j}$ is negative. The sentiment score of the sentence s_i will be a function of these sentiment scores:

$$\delta_i = g(\delta_{i,1}, \ldots, \delta_{i,k_i}),$$

where $g(\cdot)$ is the scoring function. The simplest functional form for the scoring function $g(\cdot)$ is the simple summation. However, more complex functional forms can capture connections between the words in a sentence. For example, while 'happy' has positive sentiment, when it is used in a sentence right after 'not' it will have negative sentiment. As another example, 'little happy' and 'very happy' both have positive sentiment, but the intensity of their sentiment is not the same. As such, other functional forms for scoring function $g(\cdot)$ use extra arguments to also include the negating and amplifying words for each of the words $w_{i,j}$. Once the sentiment score for each sentence in the document is estimated using the scoring function $g(\cdot)$, the document's overall sentiment can be calculated as the average of the sentences' sentiments. The lexicon-based method is usually used for estimating the sentiment score since the functional form for estimating the sentiment class probabilities is not straightforward and needs further analysis. Additionally, lexicon-based sentiment analysis has limitations in capturing connections among words and their impact on the sentence's sentiment.

Another method is to use existing datasets to train a model for estimating the sentiment class and intensity of texts. This approach uses supervised learning to train a neural or deep network for sentiment analysis. This approach can be used for estimating both the sentiment score and the sentiment class probabilities. Since this method uses all the words as input to the model, it has the potential to detect and take into account the relation between all the words in the text. Furthermore, in this approach, the sentiment of a document can be estimated at both the sentence level and the document level. In other words, it is not necessary to estimate the sentiment of sentences before computing

the document's sentiment. While the sentiment of a document is a merger of sentences' sentiment, it is not necessarily the linear combination of those sentiments.

2.1. Gender Bias in Sentiment Analysis

As mentioned above, there are two different approaches to estimating sentiment class and intensity: lexicon-based models and neural or deep network models. Lexicon-based models use predefined sentiment scores for each word. These scores are usually estimated based on surveys and existing texts. Since the survey results are highly dependent on the cultural and social characteristics of the collected sample, the existing gender bias among the participants can result in gender-biased sentiment scores for the words. Furthermore, if there exists a gender bias in the texts used for the surveys, the estimated sentiment will be affected by this bias as well.

On the other hand, a neural or deep network sentiment analysis model's performance is sensitive to the training dataset. Any gender bias in the training dataset will cause gender bias in the trained model's sentiment estimation. The datasets consist of sentences with sentiment labels. The sentiment labels are usually estimated through surveys and are subject to cultural norms and general public perception on different topics. In other words, if gender bias exists in those opinions, it will manifest in gender bias in the training data and, in turn, create gender bias in the sentiment analysis model. Furthermore, if the frequency of real sentiments is not the same for all genders, this imbalance can cause gender bias.

Definition 1. *Suppose sentences$_i$includes sets of gender-specific words and define these sets for female and male specific words as:*

$$\boldsymbol{\mathcal{W}_i^f} = \left\{w_{i,1}^f, \ldots, w_{i,\,k_{i,f}}^f\right\},\ \boldsymbol{\mathcal{W}_i^m} = \left\{w_{i,1}^m, \ldots, w_{i,\,k_{i,m}}^m\right\}. \tag{1}$$

Definition 2. *Suppose the words $w_{i,j}^{f-m}$ and $w_{i,j}^{m-f}$ are the opposite gender synonyms of words $w_{i,j}^f$ and $w_{i,j}^m$, respectively, with the equivalent sentiment ($w_{i,j}^f \in \boldsymbol{\mathcal{W}_i^f}$ and $w_{i,j}^m \in \boldsymbol{\mathcal{W}_i^m}$). Furthermore, suppose $w_{i,j}^{f-n}$ and $w_{i,j}^{m-n}$ are, respectively, the gender-neutral synonyms of words $w_{i,j}^f$ and $w_{i,j}^m$, with the equivalent sentiment. The gender word substitution ($\boldsymbol{\mathcal{GWS}}$) sets are defined as follows:*

$$\boldsymbol{\mathcal{GWS}}_i^{f-m} = \left\{w_{i,1}^{f-m}, \ldots, w_{i,\,k_{i,f}}^{f-m}\right\},\ \boldsymbol{\mathcal{GWS}}_i^{m-f} = \left\{w_{i,1}^{m-f}, \ldots, w_{i,\,k_{i,m}}^{m-f}\right\}, \tag{2}$$

$$\boldsymbol{\mathcal{GWS}_i^{f-n}} = \left\{w_{i,1}^{f-n}, \ldots, w_{i,\,k_{i,f}}^{f-n}\right\},\ \boldsymbol{\mathcal{GWS}_i^{m-n}} = \left\{w_{i,1}^{m-n}, \ldots, w_{i,\,k_{i,m}}^{m-n}\right\}. \tag{3}$$

Definition 3. *Consider the sentiment score and sentiment class probabilities of sentence s_i, and for sentences built by replacing the gender-representing words with GWS sets:*

$$\begin{aligned} &\delta_i^o = \text{sentiment score}(s_i|\boldsymbol{\mathcal{W}}_i^f, \boldsymbol{\mathcal{W}}_i^m),\ \delta_i^r \\ &= \text{sentiment score}(s_i|\boldsymbol{\mathcal{GWS}}_i^{f-m}, \boldsymbol{\mathcal{GWS}}_i^{m_f}), \\ &\delta_i^n = \text{sentiment score}(s_i|\boldsymbol{\mathcal{GWS}}_i^{f-n}, \boldsymbol{\mathcal{GWS}}_i^{m-n}), \end{aligned} \tag{4}$$

$$\begin{aligned} &p_i^{o,c_j} = P(s_i \in c_j|\boldsymbol{\mathcal{W}}_i^f, \boldsymbol{\mathcal{W}}_i^m),\ p_i^{r,\,c_j} \\ &= P(s_i \in c_j|\boldsymbol{\mathcal{GWS}}_i^{f-m}, \boldsymbol{\mathcal{GWS}}_i^{m-f}), \\ &p_i^{n,\,c_j} = P(s_i \in c_j|\boldsymbol{\mathcal{GWS}}_i^{f-n}, \boldsymbol{\mathcal{GWS}}_i^{m-n}) \end{aligned} \tag{5}$$

where c_j is the sentiment class: c_1 := negative sentiment class , c_2 := neutral sentiment class, and c_3 := positive sentiment class. $\delta_i^{\cdot}$ is the sentiment score, and $p_i^{\cdot,\cdot}$ is the sentiment class probability for the sentence s_i.

The sentiment analysis model is gender unbiased if its results (sentiment score or sentiment class probabilities) are independent of gender. If the sentiment analysis results are sentiment scores, gender unbiasedness implies:

$$\delta_i = \delta_i^o = \delta_i^r = \delta_i^n, \tag{6}$$

where δ_i^o, δ_i^r and δ_i^n are defined in Equation (4).

If the sentiment analysis results are sentiment class probabilities, gender unbiasedness implies:

$$P(s_i \in c_j) = p_i^{o,c_j} = p_i^{r,\,c_j} = p_i^{n,\,c_j},\ j = 1,2,3, \tag{7}$$

where p_i^{o,c_j}, p_i^{r,c_j} and p_i^{n,c_j} are defined in Equation (5).

With this definition of a gender-unbiased sentiment model, it is possible to detect and measure sentiment models' gender bias. Furthermore, it provides the solution to estimate the marginal sentiment by removing the gender's impact on the sentiment score. The next section provides details of the solution to address gender bias, based on the definition given in Equations (6) and (7).

2.2. Gender Bias Mitigation: Post-Training Solution

Using gender swap for training the sentiment analysis model is one of the solutions to address gender bias in sentiment analysis. However, this approach can be time consuming and the efficiency of the output depends on the gender swap in the training data and similarity of the training dataset to the text being analyzed.

Using the sentiment gender bias definition in Equations (6) and (7), it is possible to estimate and remove the impact of gender on sentiment. This approach does not require fine-tuning or retraining of the existing model and can be applied to any sentiment analysis model. Furthermore, this approach provides an index to measure gender bias in any sentiment analysis model. The gender-neutral sentiment analysis follows these steps:

1. Detect the words that can represent a gender in the text (like "he," "she," "them," "wife," "son," etc.). Do not include people's names in this set of words (including people's names can be used to address name bias in sentiment analysis). Create a set of detected words:

$$\boldsymbol{\mathcal{W}_{i,detected}} = \{w_{i,1}, \ldots,\ w_{i,k_i}\} = \boldsymbol{\mathcal{W}_i^f} \cup \boldsymbol{\mathcal{W}_i^m}$$

   where $\boldsymbol{\mathcal{W}_{i,detected}}$ is the set of detected gender-representing words in sentence s_i, $w_{i,1}, \ldots,\ w_{i,n}$ are the detected words in sentence s_i, and $\boldsymbol{\mathcal{W}_i^f}$ and $\boldsymbol{\mathcal{W}_i^m}$ are defined in Equation (1).
2. Find the synonym of each word in $\boldsymbol{\mathcal{W}}_{i,detected}$ with opposing gender. In case there are multiple synonyms available for word $w_{i,j}$ in the same gender, use the synonym with the closest sentiment to $w_{i,j}$'s sentiment. Build the replacement set for each sentence s_i:

$$\boldsymbol{\mathcal{W}_{i,replacement}} = \left\{w_{i,1}^r, \ldots,\ w_{i,k_i}^r\right\} = \boldsymbol{\mathcal{GWS}_i^{f-m}} \cup \boldsymbol{\mathcal{GWS}_i^{m-f}}$$

where $\mathcal{W}_{i,replacement}$ is the replacement set for sentence s_i, $w^r_{i,j}$ is the synonym word with different gender for $w_{i,j}$, and $\mathcal{GWS}_i^{f-m}$ and $\mathcal{GWS}_i^{m-f}$ are defined in Equation (2).

3. Find the gender-neutral synonym of each word in $\mathcal{W}_{i,detected}$. In case there are multiple gender-neutral synonyms available for word $w_{i,j}$, use the synonym with the closest sentiment to $w_{i,j}$'s sentiment. Build the gender-neutral set for each sentence s_i:

$$\mathcal{W}_{i,gender-neutral} = \{w^n_{i,1}, \ldots, w^n_{i,n}\} = \mathcal{GWS}_i^{f-n} \cup \mathcal{GWS}_i^{m-n}$$

where $\mathcal{W}_{i,gender-neutral}$ is the gender-neutral set for sentence s_i, $w^n_{i,j}$ is the gender-neutral synonym word for $w_{i,j}$, and $\mathcal{GWS}_i^{f-n}$ and $\mathcal{GWS}_i^{m-n}$ are defined in Equation (3).

4. For each sentence s_i in the text, the triplet $\boldsymbol{s_i} = (s_i, s^r_i, s^n_i)$ is built, where s^r_i is the sentence built by replacing $\mathcal{W}_{i,detected}$ with $\mathcal{W}_{i,replacement}$ and s^n_i is the sentence built by replacing $\mathcal{W}_{i,detected}$ with $\mathcal{W}_{i,gender-neutral}$.
5. Estimate the sentiment score triplet (if the sentiment analysis model is providing the sentiment score) or sentiment probability triplets (if the sentiment analysis model is providing the sentiment probabilities or confidence). If the sentiment analysis model provides the sentiment score:

$$\boldsymbol{\delta_i} = (\delta^o_i, \delta^r_i, \delta^n_i)$$

where δ^o_i, δ^r_i , and δ^n_i are defined in Equation (4). If the sentiment analysis model provides sentiment probabilities:

$$\boldsymbol{p_i^{c_j}} = \left(p_i^{o,c_j}, p_i^{r,c_j}, p_i^{n,c_j}\right),\ j = 1,2,3,$$

where c_j is the sentiment class and p_i^{o,c_j}, p_i^{r,c_j}, and p_i^{n,c_j} are defined in Equation (5).

6. The gender-unbiased sentiment score and sentiment probabilities are formulated as follows:

$$\begin{aligned} \delta^u_i &= p_o\delta^o_i + p_r\delta^r_i + p_n\delta^n_i, \\ p_i^{u,c_j} &= p_o p_i^{o,c_j} + p_r p_i^{r,c_j} + p_n p_i^{n,c_j},\ j = 1,2,3, \end{aligned} \tag{8}$$

where δ^u_i is the estimated gender-unbiased sentiment score, and p_i^{u,c_j} $(j = 1,2,3)$ are estimated gender-unbiased sentiment class probabilities. The probability distribution (p_o, p_r, p_n) represents how much the estimated sentiments from three sentences, s_i, s^r_i, and s^n_i, are relatively closer to reality, and $p_o + p_r + p_n = 1$. This information represents the existing knowledge (either from experts or previous analyses in the same topic). In case such information is not available, the uninformative probability can be used: $p_o = p_r = p_n = \frac{1}{3}$.

7. The sentiment gender bias index ($SGBI$) is formulated as follows:
If the sentiment analysis model provides a sentiment score:

$$SGBI_i = |\delta^o_i - \delta^n_i| + |\delta^r_i - \delta^n_i| \tag{9}$$

If the sentiment analysis model provides sentiment probabilities:

$$SGBI_i = \sum_{j=1}^{3}\left(\left|p_i^{o,c_j} - p_i^{n,c_j}\right| + \left|p_i^{r,c_j} - p_i^{n,c_j}\right|\right). \tag{10}$$

The $SGBI_i$ close to zero shows gender-unbiasedness in sentiment analysis results.

As mentioned in step 2, if the original gender-specific word has multiple synonyms in the same context, the synonym with the closest sentiment to the original word should be

selected as a replacement. One approach is to use a lexicon library to find the synonym of a gender-specific word that has the closest sentiment to the original word and select the synonyms with the closest sentiment score. Another approach to finding the nearest synonym is to use an LLM. This study employs the LLM approach, which also yields a list of synonyms.

This method removes the impact of the gender on the sentiment results. As such, it guarantees that the sentiment analysis result is the same regardless of the gender-representing words in the text.

3. Results

The algorithm and formulation provided in Section 2 can be used to obtain a gender-unbiased estimation of sentiment. Additionally, they can be used to estimate the level of gender bias in the output of a sentiment analysis model.

In this study, the gender-biases of five commonly used sentiment analysis libraries are estimated. The investigated libraries are "SentimentR" (version 2.9.0) [19], "TextBlob" (version 0.19.0) [20], "Microsoft Azure"(API version 24-11-01) [21], "RoBERTa" (python transformers library version 4.51.3) [22], and "VADER" (version 0.2.1) [23]. The SentimentR and TextBlob libraries provide the sentiment estimation using sentiment score. On the other hand, the Azure, RoBERTa, and VADER libraries provide sentiment probabilities.

To better understand gender bias in two commonly used sentiment analysis models, the method proposed in Section 2 is applied to two different sets of texts. The first set consists of synthetic texts containing only one gender-representative word. The second set comprises tweets extracted from the X platform (formerly Twitter).

3.1. Sentiment Analysis of Synthetic Texts

The first set of texts is generated based on the following two sentences by changing the gender-representative-word:

1. My "wife/husband/spouse/girlfriend/boyfriend/partner/daughter/son/child/mother/father/parent/sister/brother/sibling/aunt/uncle/pibling/niece/nephew/nibling" is in an earthquake.
2. My "wife/husband/spouse/girlfriend/boyfriend/partner/daughter/son/child/mother/father/parent/sister/brother/sibling/aunt/uncle/pibling/niece/nephew/nibling" is an earthquake.

If the sentiment analysis model is gender-unbiased, the result should be the same regardless of the gender of the person described in the text.

Tables 1 and 2 present the estimated sentiments from the SentimentR, Microsoft Azure, and RoBERTa libraries, along with the gender-unbiased sentiment. The VADER and TextBlob libraries estimated the sentiment for all the sentences (with female, male, and gender-neutral wordings) as neutral sentiment. The TextBlob estimates all the sentiment scores at 0 ($\delta = 0$) and the VADER estimates the probability of the neutral sentiment 1 ($P_{neutral} = 1$, $P_{negative} = P_{negative} = 0$). In these two libraries, the SGBI is zero for all the sentences, which suggests that these two libraries are not sensitive to the gender of the wording used in these sentences. The SGBI of the five libraries is presented in Figure 1. As can be seen in Figure 1, the estimated sentiments of some sentences do not have gender bias (SGBI is zero). This shows that the size of the sentiment gender bias is different when the relation of the person described in the sentences is changed. More details can be found in Tables 1 and 2. For instance, when the sentence is mentioning the parents, the estimated SGBI for the SentimentR library is 0 (see $i = 4$ and 12 in Table 1). However, when it mentions the siblings, the estimated SGBI is not zero.

Table 1. Sentiment analysis gender bias and gender-unbiased sentiment results in "SentimentR" libraries (for estimating the sentiment class based on δ, threshold $\lambda = 0.05$ is used).

i		Sentence	δ	Sentiment Class	$SGBI$	i		Sentence	δ	Sentiment Class	$SGBI$
1	s_i	My wife is in an earthquake	−0.20	Negative	0.49	9	s_i	My wife is an earthquake	−0.22	Negative	0.54
	s_i^r	My husband is in an earthquake	−0.20	Negative			s_i^r	My husband is an earthquake	−0.22	Negative	
	s_i^n	My spouse is in an earthquake	0.04	Neutral			s_i^n	My spouse is an earthquake	0.04	Neutral	
		Gender-Unbiased	−0.12	Negative				Gender-Unbiased	−0.13	Negative	
2	s_i	My girlfriend is in an earthquake	−0.20	Negative	0.65	10	s_i	My girlfriend is an earthquake	−0.22	Negative	0.716
	s_i^r	My boyfriend is in an earthquake	−0.20	Negative			s_i^r	My boyfriend is an earthquake	−0.22	Negative	
	s_i^n	My partner is in an earthquake	0.12	Positive			s_i^n	My partner is an earthquake	0.13	Positive	
		Gender-Unbiased	−0.10	Negative				Gender-Unbiased	−0.10	Negative	
3	s_i	My daughter is in an earthquake	0.04	Neutral	0.24	11	s_i	My daughter is an earthquake	0.04	Neutral	0.27
	s_i^r	My son is in an earthquake	−0.20	Negative			s_i^r	My son is an earthquake	−0.22	Negative	
	s_i^n	My child is in an earthquake	0.04	Neutral			s_i^n	My child is an earthquake	0.04	Neutral	
		Gender-Unbiased	−0.04	Neutral				Gender-Unbiased	−0.04	Neutral	
4	s_i	My mother is in an earthquake	−0.20	Negative	0	12	s_i	My mother is an earthquake	−0.22	Negative	0
	s_i^r	My father is in an earthquake	−0.20	Negative			s_i^r	My father is an earthquake	−0.22	Negative	
	s_i^n	My parent is in an earthquake	−0.20	Negative			s_i^n	My parent is an earthquake	−0.22	Negative	
		Gender-Unbiased	−0.20	Negative				Gender-Unbiased	−0.22	Negative	
5	s_i	My sister is in an earthquake	−0.20	Negative	0.16	13	s_i	My sister is an earthquake	−0.22	Negative	0.18
	s_i^r	My brother is in an earthquake	−0.04	Neutral			s_i^r	My brother is an earthquake	−0.04	Neutral	
	s_i^n	My sibling is in an earthquake	−0.20	Negative			s_i^n	My sibling is an earthquake	−0.22	Negative	
		Gender-Unbiased	−0.15	Negative				Gender-Unbiased	−0.16	Negative	

Table 1. *Cont.*

i		Sentence	δ	Sentiment Class	$SGBI$	i		Sentence	δ	Sentiment Class	$SGBI$
6	s_i	My aunt is in an earthquake	−0.10	Negative	0.1	14	s_i	My aunt is an earthquake	−0.11	Negative	0.11
	s_i^r	My uncle is in an earthquake	−0.20	Negative			s_i^r	My uncle is an earthquake	−0.22	Negative	
	s_i^n	My pibling is in an earthquake	−0.20	Negative			s_i^n	My pibling is an earthquake	−0.22	Negative	
		Gender-Unbiased	−0.17	Negative		15		Gender-Unbiased	−0.19	Negative	
7	s_i	My niece is in an earthquake	−0.20	Negative	0	16	s_i	My niece is an earthquake	−0.22	Negative	0
	s_i^r	My nephew is in an earthquake	−0.20	Negative			s_i^r	My nephew is an earthquake	−0.22	Negative	
	s_i^n	My nibling is in an earthquake	−0.20	Negative			s_i^n	My nibling is an earthquake	−0.22	Negative	
		Gender-Unbiased	−0.20	Negative				Gender-Unbiased	−0.22	Negative	
8	s_i	My mother in law is in an earthquake	−0.18	Negative	0	17	s_i	My mother in law is an earthquake	−0.19	Negative	0
	s_i^r	My father in law is in an earthquake	−0.18	Negative			s_i^r	My father in law is an earthquake	−0.19	Negative	
	s_i^n	My parent in law is in an earthquake	−0.18	Negative			s_i^n	My parent in law is an earthquake	−0.19	Negative	
		Gender-Unbiased	−0.18	Negative				Gender-Unbiased	−0.19	Negative	
9	s_i	My sister in law is in an earthquake	−0.18	Negative	0.14	18	s_i	My sister in law is an earthquake	−0.19	Negative	0.15
	s_i^r	My brother in law is in an earthquake	−0.04	Neutral			s_i^r	My brother in law is an earthquake	−0.04	Neutral	
	s_i^n	My sibling in law is in an earthquake	−0.18	Negative			s_i^n	My sibling in law is an earthquake	−0.19	Negative	
		Gender-Unbiased	−0.13	Negative				Gender-Unbiased	−0.14	Negative	

Table 2. Sentiment analysis gender bias and gender-unbiased sentiment results in "Microsoft Azure" and "RoBERTa" libraries.

		Sentence	Microsoft Azure					RoBERTa				
			$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$	$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$
1	s_i	My wife is in an earthquake	0.76	0.24	0.00	Negative	0.27	0.52	0.45	0.03	Negative	0.14
	s_i^r	My husband is in an earthquake	0.77	0.22	0.00	Negative		0.52	0.45	0.04	Negative	
	s_i^n	My spouse is in an earthquake	0.70	0.30	0.00	Negative		0.52	0.42	0.03	Negative	
		Gender-Unbiased	0.74	0.25	0.00	Negative		0.53	0.44	0.03	Negative	

Table 2. *Cont.*

		Sentence	Microsoft Azure					RoBERTa				
			$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$	$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$
2	s_i	My girlfriend is in an earthquake	0.63	0.36	0.01	Negative	0.1	0.54	0.42	0.04	Negative	0.15
	s_i^r	My boyfriend is in an earthquake	0.68	0.32	0.00	Negative		0.47	0.48	0.05	Neutral	
	s_i^n	My partner is in an earthquake	0.65	0.35	0.00	Negative		0.47	0.48	0.03	Neutral	
		Gender-Unbiased	0.65	0.34	0.00	Negative		0.50	0.46	0.04	Negative	
3	s_i	My daughter is in an earthquake	0.78	0.22	0.00	Negative	0.26	0.63	0.35	0.02	Negative	0.71
	s_i^r	My son is in an earthquake	0.79	0.21	0.00	Negative		0.64	0.34	0.02	Negative	
	s_i^n	My child is in an earthquake	0.72	0.28	0.00	Negative		0.64	0.18	0.01	Negative	
		Gender-Unbiased	0.76	0.24	0.00	Negative		0.69	0.29	0.02	Negative	
4	s_i	My mother is in an earthquake	0.87	0.13	0.00	Negative	0.3	0.71	0.27	0.02	Negative	0.24
	s_i^r	My father is in an earthquake	0.88	0.12	0.00	Negative		0.59	0.39	0.02	Negative	
	s_i^n	My parent is in an earthquake	0.80	0.20	0.00	Negative		0.59	0.30	0.02	Negative	
		Gender-Unbiased	0.85	0.15	0.00	Negative		0.66	0.32	0.02	Negative	
5	s_i	My sister is in an earthquake	0.81	0.19	0.00	Negative	0.71	0.51	0.46	0.03	Negative	0.24
	s_i^r	My brother is in an earthquake	0.76	0.23	0.00	Negative		0.55	0.43	0.03	Negative	
	s_i^n	My sibling is in an earthquake	0.61	0.39	0.00	Negative		0.55	0.39	0.02	Negative	
		Gender-Unbiased	0.73	0.27	0.00	Negative		0.55	0.43	0.03	Negative	

Table 2. *Cont.*

		Sentence	Microsoft Azure					RoBERTa				
			$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$	$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$
6	s_i	My aunt is in an earthquake	0.85	0.14	0.00	Negative	0.36	0.58	0.39	0.02	Negative	0.29
	s_i^r	My uncle is in an earthquake	0.67	0.32	0.00	Negative		0.44	0.53	0.03	Neutral	
	s_i^n	My pibling is in an earthquake	0.76	0.23	0.00	Negative		0.44	0.44	0.03	Neutral	
		Gender-Unbiased	0.76	0.23	0.00	Negative		0.52	0.46	0.03	Negative	
7	s_i	My niece is in an earthquake	0.76	0.23	0.00	Negative	0.49	0.56	0.41	0.03	Negative	0.15
	s_i^r	My nephew is in an earthquake	0.83	0.17	0.00	Negative		0.58	0.39	0.03	Negative	
	s_i^n	My nibling is in an earthquake	0.67	0.32	0.00	Negative		0.58	0.44	0.02	Negative	
		Gender-Unbiased	0.75	0.24	0.00	Negative		0.56	0.41	0.03	Negative	
8	s_i	My mother in law is in an earthquake	0.63	0.37	0.00	Negative	0.11	0.57	0.41	0.02	Negative	0.08
	s_i^r	My father in law is in an earthquake	0.59	0.40	0.00	Negative		0.57	0.41	0.02	Negative	
	s_i^n	My parent in law is in an earthquake	0.64	0.36	0.00	Negative		0.57	0.39	0.02	Negative	
		Gender-Unbiased	0.62	0.38	0.00	Negative		0.58	0.40	0.02	Negative	
9	s_i	My sister in law is in an earthquake	0.73	0.27	0.00	Negative	0.29	0.49	0.49	0.02	Neutral	0.12
	s_i^r	My brother in law is in an earthquake	0.59	0.40	0.00	Negative		0.55	0.43	0.02	Negative	
	s_i^n	My sibling in law is in an earthquake	0.59	0.41	0.00	Negative		0.55	0.47	0.02	Negative	
		Gender-Unbiased	0.64	0.36	0.00	Negative		0.51	0.47	0.02	Negative	

Table 2. *Cont.*

	Sentence	Microsoft Azure					RoBERTa				
		$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$	$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$
10	s_i My wife is an earthquake	0.85	0.15	0.00	Negative	0.22	0.45	0.50	0.05	Neutral	0.19
	s_i^r My husband is an earthquake	0.88	0.12	0.00	Negative		0.54	0.41	0.04	Negative	
	s_i^n My spouse is an earthquake	0.81	0.19	0.00	Negative		0.54	0.49	0.04	Negative	
	Gender-Unbiased	0.85	0.15	0.00	Negative		0.49	0.47	0.05	Negative	
11	s_i My girlfriend is an earthquake	0.60	0.39	0.01	Negative	0.53	0.53	0.42	0.05	Negative	0.48
	s_i^r My boyfriend is an earthquake	0.77	0.22	0.00	Negative		0.54	0.40	0.06	Negative	
	s_i^n My partner is an earthquake	0.82	0.18	0.00	Negative		0.54	0.53	0.05	Negative	
	Gender-Unbiased	0.73	0.26	0.00	Negative		0.50	0.45	0.05	Negative	
12	s_i My daughter is an earthquake	0.89	0.11	0.00	Negative	0.14	0.48	0.47	0.05	Negative	0.69
	s_i^r My son is an earthquake	0.89	0.11	0.00	Negative		0.47	0.48	0.05	Neutral	
	s_i^n My child is an earthquake	0.85	0.14	0.00	Negative		0.47	0.33	0.02	Negative	
	Gender-Unbiased	0.88	0.12	0.00	Negative		0.53	0.42	0.04	Negative	
13	s_i My mother is an earthquake	0.92	0.08	0.00	Negative	0.1	0.72	0.26	0.02	Negative	0.35
	s_i^r My father is an earthquake	0.89	0.11	0.00	Negative		0.55	0.42	0.03	Negative	
	s_i^n My parent is an earthquake	0.88	0.12	0.00	Negative		0.55	0.32	0.03	Negative	
	Gender-Unbiased	0.90	0.10	0.00	Negative		0.64	0.33	0.03	Negative	
14	s_i My sister is an earthquake	0.90	0.09	0.00	Negative	0.3	0.40	0.54	0.05	Neutral	0.09
	s_i^r My brother is an earthquake	0.85	0.14	0.00	Negative		0.45	0.50	0.05	Neutral	
	s_i^n My sibling is an earthquake	0.80	0.19	0.00	Negative		0.45	0.51	0.05	Neutral	
	Gender-Unbiased	0.85	0.14	0.00	Negative		0.43	0.52	0.05	Neutral	

Table 2. *Cont.*

		Sentence	Microsoft Azure					RoBERTa				
			$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$	$P_{negative}$	$P_{neutral}$	$P_{positive}$	Sentiment Class	$SGBI$
15	s_i	My aunt is an earthquake	0.90	0.10	0.00	Negative	0.17	0.57	0.40	0.04	Negative	0.50
	s_i^r	My uncle is an earthquake	0.81	0.18	0.00	Negative		0.32	0.62	0.05	Neutral	
	s_i^n	My pibling is an earthquake	0.86	0.14	0.00	Negative		0.32	0.56	0.06	Neutral	
		Gender-Unbiased	0.86	0.14	0.00	Negative		0.42	0.53	0.05	Neutral	
16	s_i	My niece is an earthquake	0.84	0.16	0.00	Negative	0.46	0.43	0.51	0.06	Neutral	0.27
	s_i^r	My nephew is an earthquake	0.88	0.12	0.00	Negative		0.41	0.53	0.06	Neutral	
	s_i^n	My nibling is an earthquake	0.75	0.25	0.01	Negative		0.41	0.59	0.05	Neutral	
		Gender-Unbiased	0.82	0.18	0.00	Negative		0.40	0.54	0.06	Neutral	
17	s_i	My mother in law is an earthquake	0.83	0.17	0.00	Negative	0.07	0.49	0.48	0.03	Negative	0.05
	s_i^r	My father in law is an earthquake	0.79	0.20	0.00	Negative		0.47	0.50	0.03	Neutral	
	s_i^n	My parent in law is an earthquake	0.81	0.19	0.00	Negative		0.47	0.49	0.03	Neutral	
		Gender-Unbiased	0.81	0.19	0.00	Negative		0.48	0.49	0.03	Neutral	
18	s_i	My sister in law is an earthquake	0.73	0.26	0.00	Negative	0.37	0.34	0.62	0.05	Neutral	0.16
	s_i^r	My brother in law is an earthquake	0.78	0.22	0.00	Negative		0.39	0.57	0.04	Neutral	
	s_i^n	My sibling in law is an earthquake	0.66	0.33	0.00	Negative		0.39	0.63	0.04	Neutral	
		Gender-Unbiased	0.72	0.27	0.00	Negative		0.35	0.61	0.04	Neutral	

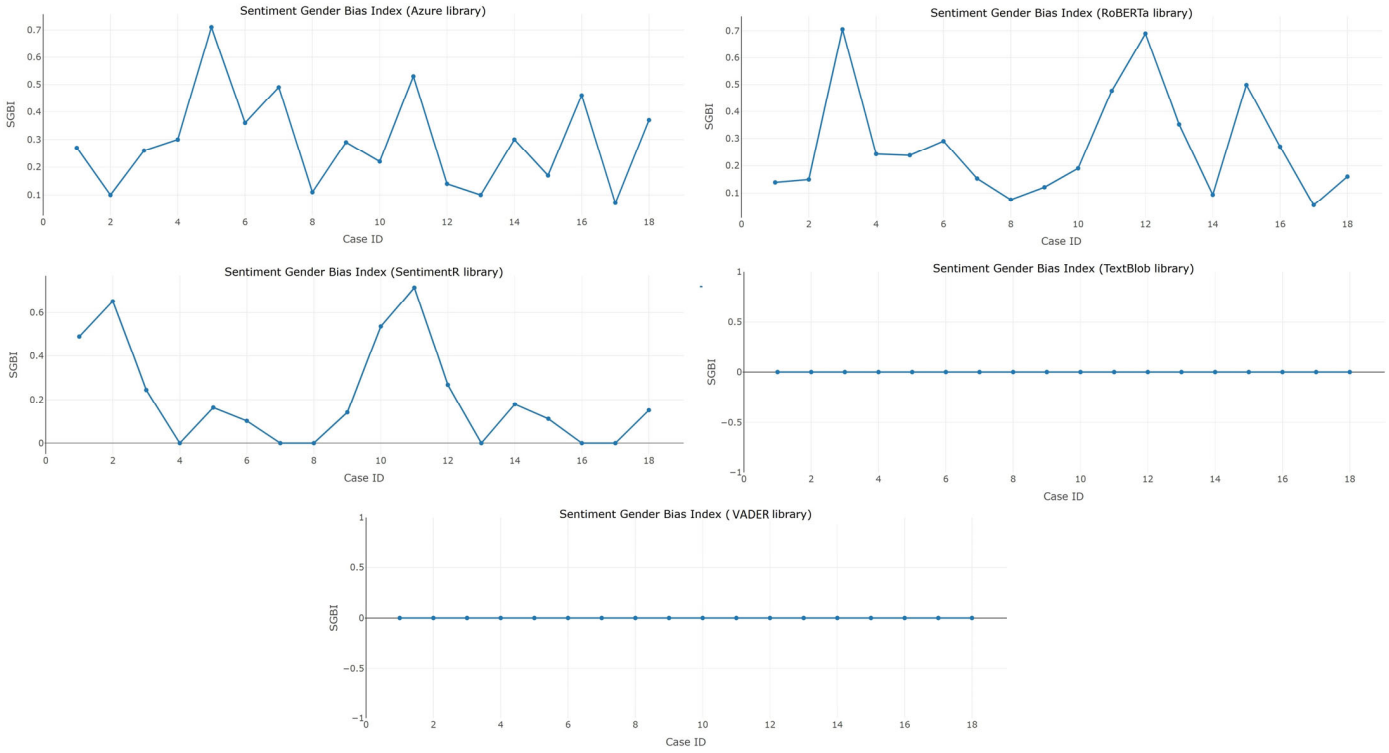

Figure 1. Estimated SGBI for different libraries for the example sentences.

As can be seen in Tables 1 and 2, the unbiased sentiment removes the impact of gender on the estimated sentiment class. Tables 3 and 4 show the sum of estimated SGBI for different sentiment libraries. As can be seen, the VADER library has the lowest estimated SGBI based on sentiment class probabilities, and RoBERTa and Azure have relatively close SGBI. Among the libraries with sentiment scores, TextBlob has the lowest SGBI.

Table 3. Sum of sentiment gender bias index for libraries with sentiment class probabilities.

Sentiment Analysis Library	Microsoft Azure	RoBERTa	VADER
Sum of SGBI	5.25	4.9	0

Table 4. Sum of sentiment gender bias index for libraries with sentiment scores.

Sentiment Analysis Library	SentimentR	TextBlob
Sum of SGBI	3.757228	0

It should be noted that these results might change in different context, i.e., while in one context RoBERTa might have a lower gender bias than Azure, it does not guarantee that it would always have a lower Gender bias. Lower gender bias in sentiment results can be considered one of the efficiency dimensions. For example, if a one sentiment model classifies every text as "neutral", regardless of the genders or context (as VADER and TextBlob estimated sentiments in this study), it would have $SGBI = 0$. However, it would not be considered an accurate or efficient model, since it produces the same result for any text from any context. Additionally, it should be noted that comparing score-based SGBI with probability-based SGBI can have misleading results, as they do not have the same measurement scale. For this type of comparison, the SGBIs should be standardized.

Table 5 shows the *t*-test results for comparing the average SGBI in different libraries. The sentiment analysis libraries with sentiment scores (SentimentR and TextBlob) are

compared with each other, and the sentiment analysis libraries with sentiment probabilities (Microsoft Azure, RoBERTa, and VADER) are compared with each other. As shown in Table 5, the difference between SentimentR and TextBlob is significant. In the sentiment analysis libraries based on sentiment probabilities, Microsoft Azure and Roberta do not have significant differences. It should be noted that these results are based on the context of the texts in Tables 1 and 2. The gender bias of these libraries might vary in other contexts.

Table 5. Comparing average SGBI of different sentiment analysis libraries for texts presented in Tables 1 and 2.

Null Hypothesis	$\overline{SGBI}_{SentimentR} - \overline{SGBI}_{TextBlob} = 0$	$\overline{SGBI}_{Microsoft\ Azure} - \overline{SGBI}_{RoBERTa} = 0$	$\overline{SGBI}_{Microsoft\ Azure} - \overline{SGBI}_{VADER} = 0$	$\overline{SGBI}_{RoBERTa} - \overline{SGBI}_{VADER} = 0$
Average SGBI difference	0.2087	0.0194	0.2917	0.2722
t	3.7684	0.3121	7.1412	5.815
df	17	17	17	17
p-value	0.0015	0.7587	1.652×10^{-6}	2.07×10^{-5}
Average SGBI difference is significant (0.05 significance level)	Yes	No	Yes	Yes

3.2. Sentiment Analysis of Social Media Posts

As the second set of texts in this study, we use the collection of social media posts from [18]. These posts were extracted using the keyword "Earthquake" from the X platform (formerly Twitter) between 8:05 a.m. and 10:30 a.m. on 9 September 2023—the morning immediately following the Morocco earthquake. The dataset includes 10,053 posts, including reposts and replies. Among these, 137 posts contain gender-specific words.

Figures 2–5 show the relative frequency of Negative, Neutral, and Positive sentiment classes based on S^o (the original posts), S^r(posts where gender-specific words are replaced with synonyms of the opposite gender), S^n (posts where gender-specific words are replaced with gender-neutral synonyms), and the unbiased sentiment estimated using Equation (8). To generate the s^r and s^n sets, the GPT-3.5-turbo model was used to detect gender-specific words and generate their closest synonyms.

As shown in Figure 2, the results of all four sentiment analyses using the SentimentR library are closely aligned. This indicates that SentimentR exhibits a low level of gender bias when analyzing these posts. The total SGBI for the posts analyzed by the SentimentR library is 5.452.

The sentiment analysis results from the TextBlob library are shown in Figure 3. In the analyzed social media post, the TextBlob's total SGBI is 1.37, which shows a gender bias lower than the SentimentR library. As shown in Figure 3, the sentiment classes estimated by TextBlob are very different from other sentiment libraries.

Figure 4 shows that the results of the four sentiment analyses using the Azure library are also similar. The total SGBI for the posts analyzed by the Azure library is 25.25, with an average of 0.1843. This consistency among the four results suggests a low level of gender bias in analyzing this set of social media posts using the Azure library.

Figure 5 presents the RoBERTa sentiment analysis results of the original social media posts, along with the replaced gender-related posts and the unbiased results. The total SGBI for the RoBERTa sentiment in this analysis is 9.91 (0.072 on average), which is lower than the Azure library. Although the sentiment results are very different from SentimentR, TextBlob, and Azure libraries' results, the closeness of the unbiased sentiment estimation and the relatively low SBGI imply low gender bias.

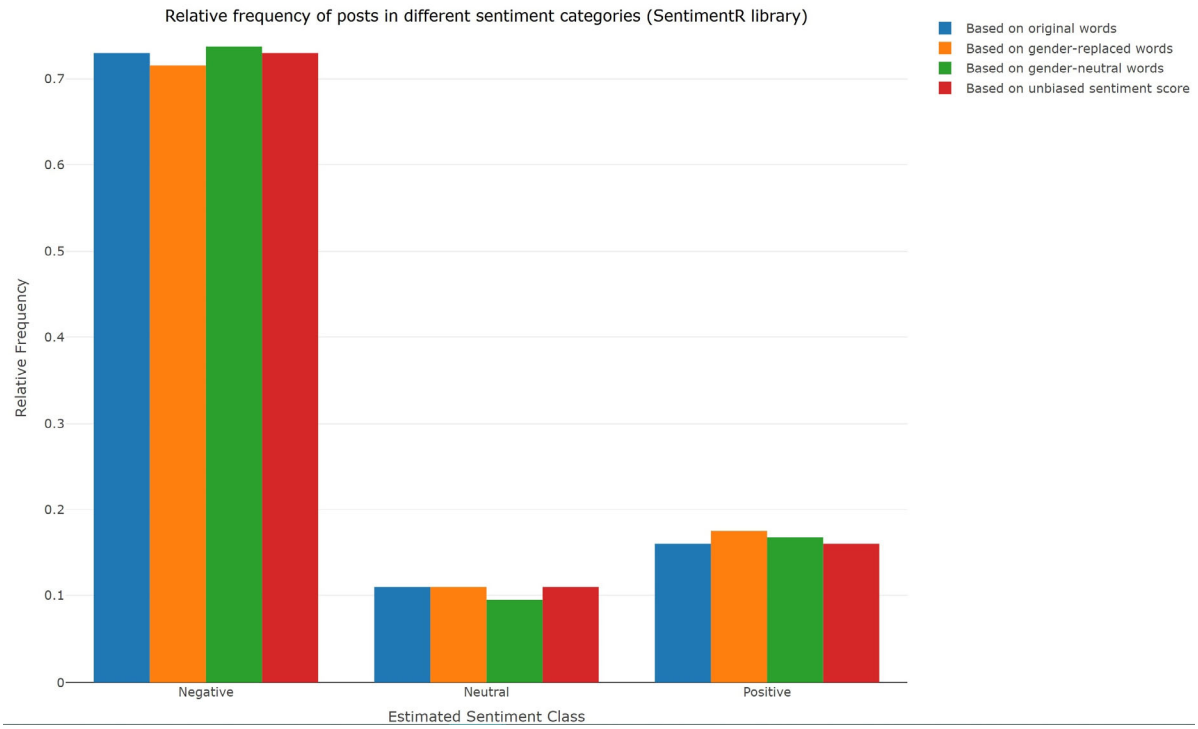

Figure 2. Result of social media post sentiment analysis using SentimentR library (threshold: $\lambda = 0.05$).

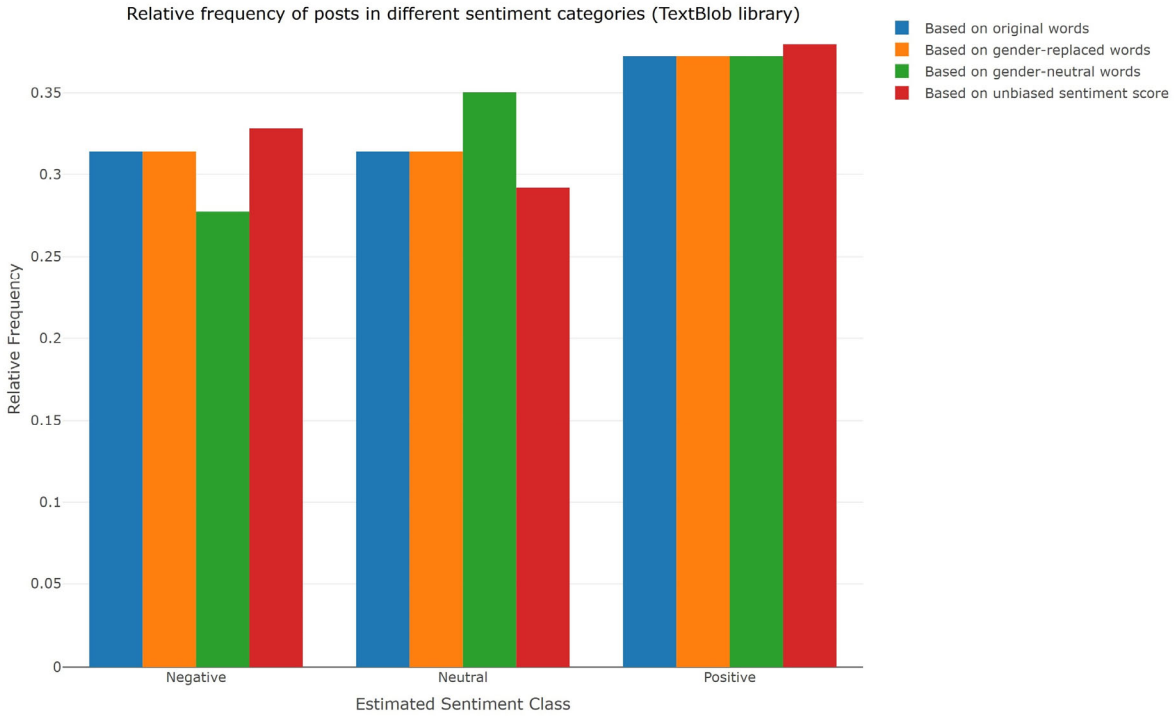

Figure 3. Result of social media posts sentiment analysis using TextBlob library (threshold: $\lambda = 0.05$).

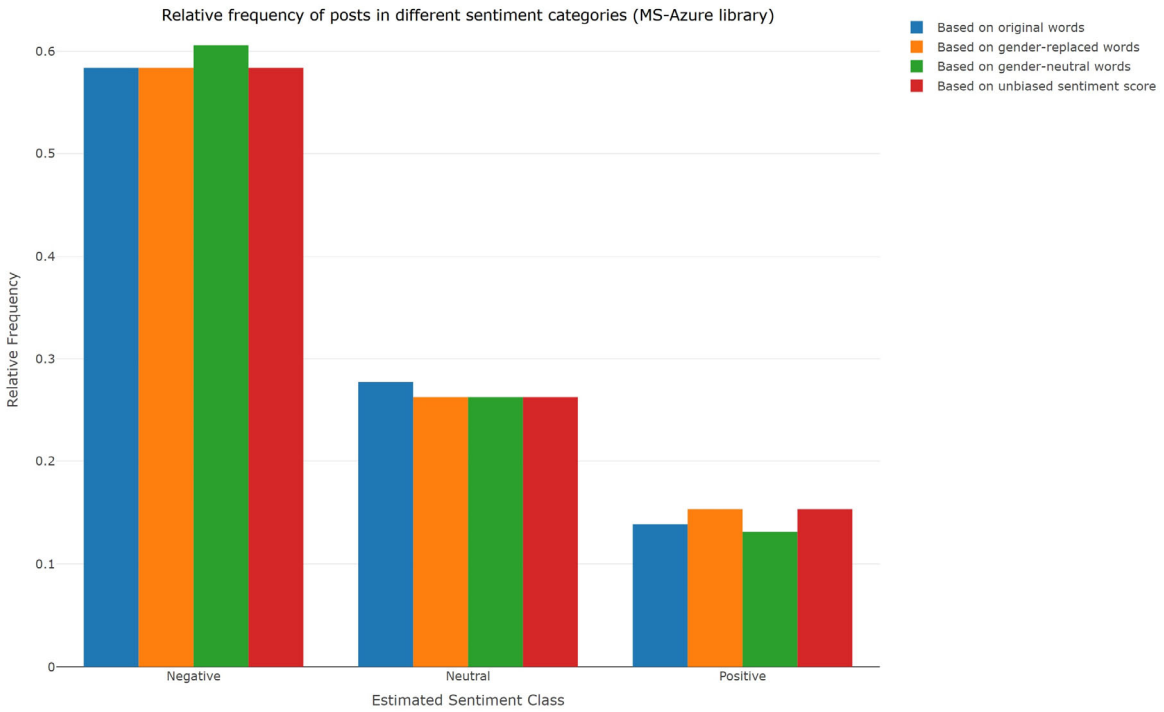

Figure 4. Result of social media post sentiment analysis using Microsoft Azure library.

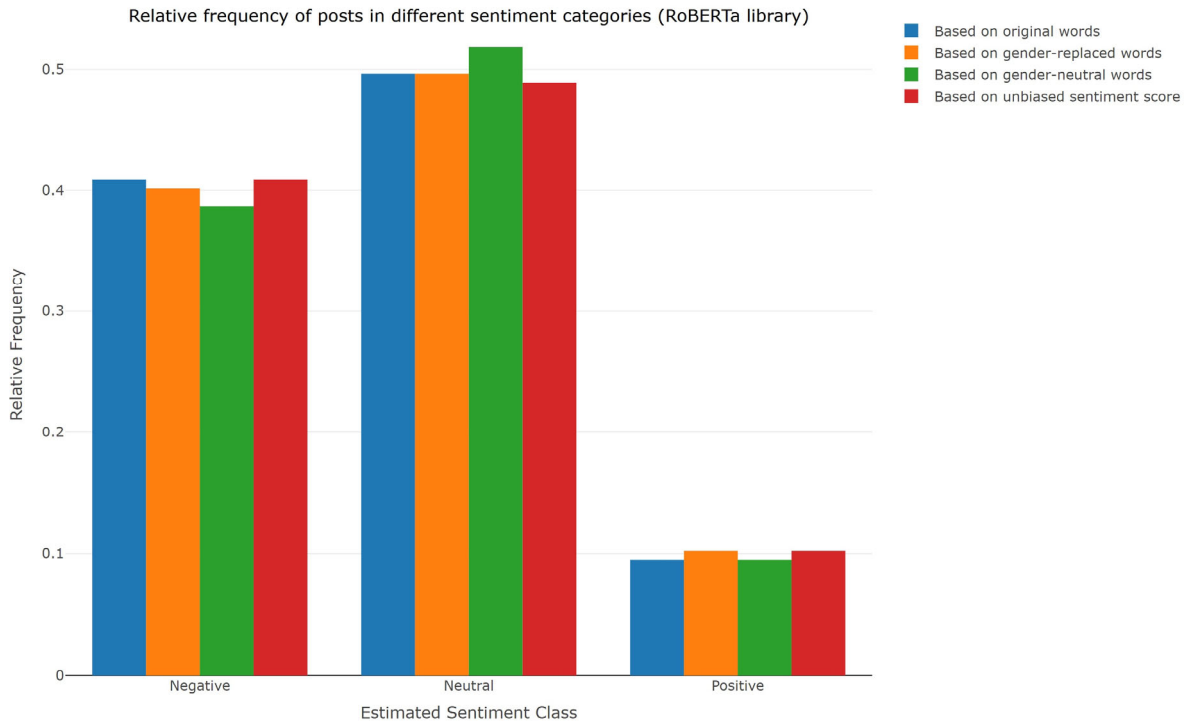

Figure 5. Result of social media posts sentiment analysis using RoBERTa library.

Figure 6 shows the estimated sentiment class for Morocco earthquake social media posts, estimated with the VADER library. As can be seen, the VADER library categorized all the posts as 'Neutral.' However, there have been small differences among estimated sentiment probabilities. As such, the total SGBI for the VADER library in the analyzed posts was 0.01.

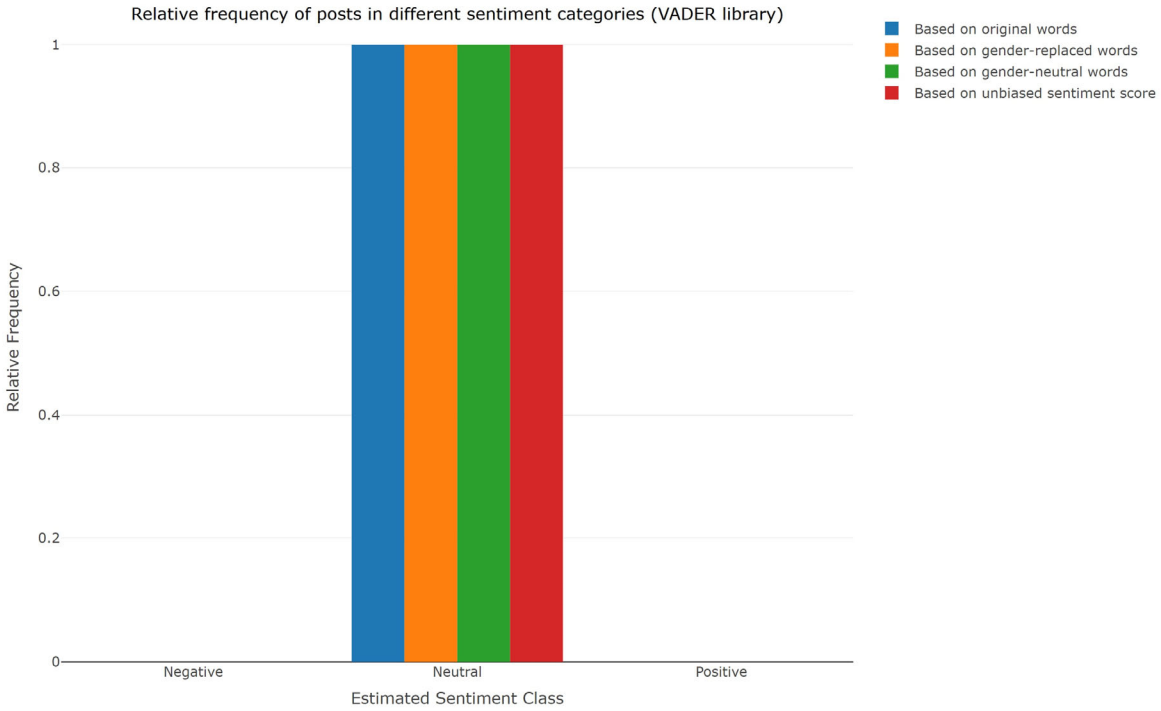

Figure 6. Result of social media post sentiment analysis using VADER library.

According to Figures 2–6, the five libraries analyzed show low gender bias in their sentiment analysis of social media posts related to the Morocco earthquake. However, neither library is completely free of gender bias. It should be noted that these findings do not imply that these five libraries perform equally well in analyzing these social media posts overall. Table 6 shows the comparison of the SGBI in social media posts related to the Morocco earthquake. These results show that in these posts, neither of the two libraries has the same SGBI. Among the sentiment analysis libraries with sentiment scores, TextBlob has significantly lower SGBI, on average. Among the libraries based on sentiment probabilities, Microsoft Azure has the highest average SGBI. These results show how the SGBI can be employed for analyzing the gender bias in sentiment analysis. However, the comparison should be performed in each analysis so that the gender bias is in the scope of that study. In other words, the generalization of these results needs the calculation of the SGBI on a comprehensive corpus to cover a large variety of contexts and sentiments.

Table 6. Comparing average SGBI of different sentiment analysis libraries for text presented in Morocco earthquake tweets.

Null Hypothesis	$\overline{SGBI}_{SentimentR} - \overline{SGBI}_{TextBlob} = 0$	$\overline{SGBI}_{Microsoft\ Azure} - \overline{SGBI}_{RoBERTa} = 0$	$\overline{SGBI}_{Microsoft\ Azure} - \overline{SGBI}_{VADER} = 0$	$\overline{SGBI}_{RoBERTa} - \overline{SGBI}_{VADER} = 0$
Average SGBI difference	0.0298	0.1120	0.1842	0.0722
t	3.4775	5.3909	8.9917	11.508
df	136	136	136	136
p-value	0.0007	3.01×10^{-7}	1.842×10^{-15}	2.2×10^{-16}
Average SGBI difference is significant (0.05 significance level)	Yes	Yes	Yes	Yes

As shown in Figures 2–6, the sentiment results of the five libraries are very different. This difference can be related to the difference in performance among these libraries. Evaluating other aspects of their performance is beyond the scope of this study. It should be noted that the results of SGBI can be different in different contexts and batches of text, since the severity of gender bias in the training datasets can be different in different topics. As such, it is necessary to run the gender-bias analysis and compute the unbiased sentiment results and SGBI every time that gender unbiasedness is being addressed in a sentiment analysis.

4. Discussion

As illustrated in Section 3, the gender-unbiased estimation of sentiment (Equation (8)) can be used regardless of the presence or magnitude of gender bias in the sentiment analysis model. As demonstrated using synthetic texts and social media posts, when gender bias is minimal or nonexistent, the gender-unbiased sentiment closely matches the sentiment of the original text.

Furthermore, the index formulated in Equations (9) and (10) can be used to measure both the presence and severity of gender bias in the sentiment analysis results. However, it should be noted that the SGBI calculated using sentiment scores cannot be directly compared with the SGBI calculated using sentiment probabilities, as they are on different scales.

Although the primary aim of this study is to provide a solution for reducing gender bias without retraining, the proposed SGBI can also be incorporated into the retraining process. If the SGBI indicates no gender bias in the results, retraining the sentiment analysis model may not be necessary. Additionally, during retraining, the SGBI can serve as one of the objective functions to help reduce gender bias in the trained sentiment model.

The SGBI not only reveals the severity of the gender-bias in the sentiment analysis results but it can also shed light on the factors that are affecting this bias. As shown in Figure 1, the SGBI is different for the same content when describing a person who is a parent or a sibling. This conclusion suggests that some libraries might be sensitive to other factors as well. The SGBI provides a quantitative measure to investigate the factors affecting gender bias in sentiment analysis. As such, the SGBI can be used as a quantitative measure to investigate the impact of different factors on gender bias in sentiment results for models trained by different datasets. On the practical side, this measurement can help to create more robust sentiment models. On the theoretical side, the quantitative measure can be used for building and testing hypotheses.

The computation of unbiased sentiment and SGBI relies on finding the gender-specific words and their close synonyms. As mentioned before, different approaches can be used to detect and replace the gender-specific words. However, choosing the synonyms can be sensitive to the existing biases in the lexical dictionary or LLM employed. As such, it is necessary to use either approach in the least interventional way. For instance, finding the gender-specific words is not sensitive to gender bias, as the lexical definition of gender-specific words is clear. However, the social definition of the gender would need intervention, as different libraries and LLMs might be built on different definitions of the concept. In these circumstances, the extraction of gender-specific words and finding their synonyms should be performed based on multiple sources (e.g., different lexical libraries and LLMs) to get more robust results. Although this approach can be practical, the robustness of the results based on different sources needs more investigation.

It should be noted that the existence and the severity of the gender-bias can be different in different contexts and languages. In any sentiment analysis study, it would be necessary to apply the sentiment gender bias analysis and measure the SGBI. In other words, if a sentiment analysis model shows a low level of sentiment gender bias in a batch of texts, it

does not guarantee the low sentiment gender bias everywhere. Furthermore, the SGBI only measures the sentiment of gender bias. Other aspects of the performance of a sentiment analysis model should be investigated separately. In other words, the low SGBI does not imply that estimated sentiments are close to reality or people's perception. It only shows the level of sensitivity of the results to the gender-specific words.

The unbiased sentiment and SGBI rely on the lexical gender definition of words. As long as the gender of the word is defined in dictionaries, the SGBI can be calculated, regardless of the specification of the gender in the analyzed text. For example, if the context of the text refers to someone as "he," "she," or "them," the SGBI is being calculated based on the gender definition of the word in the dictionary. As such, the method is applicable for the contexts with non-binary or ambiguous gender. However, it only measures the gender bias related to the lexical genders. It does not measure the sentiment bias toward non-binary or ambiguous genders.

5. Conclusions

This study focuses on mitigating gender bias in sentiment analysis using a post-training approach. To quantitatively define the problem, a formulation for sentiment gender bias is proposed. This definition is based on conditional sentiment outcomes, conditioned on gender-specific words. Using this formulation, a gender-unbiased sentiment result is derived.

Furthermore, a sentiment gender bias index (SGBI) is introduced based on the proposed definition. The method is applied to both a synthetic dataset and real-world social media posts. The results show that the gender-unbiased sentiment analysis is compatible with any sentiment analysis model with any severity of gender bias. Additionally, the method performs effectively regardless of the initial level of gender bias in the sentiment model.

The proposed SGBI not only detects sentiment gender bias but can also be used to retrain sentiment models to mitigate such bias. However, it is sensitive to the scale of sentiment scores and sentiment probabilities.

In conclusion, this work provides a practical and scalable framework for identifying and mitigating gender bias in sentiment analysis systems. By decoupling bias mitigation from the training phase and introducing a model-agnostic metric (SGBI), it offers a versatile solution for improving fairness in NLP applications. The findings demonstrate that even sentiment models perceived as neutral may contain measurable biases that affect interpretation and downstream applications. Future work could explore adaptations of SGBI for intersectional bias (e.g., combining gender with race or age) and investigate robustness across different languages and cultural contexts. Ultimately, such tools are crucial in ensuring the ethical deployment of AI systems, especially in high-stakes domains like recruitment, healthcare, and public opinion monitoring.

The proposed concept for gender-unbiased sentiment analysis and SGBI provides a novel approach for investigating the driving factors in gender bias in NLP. Furthermore, the proposed framework can be used to measure other types of bias in NLP results, including region-related bias, name-related bias, and ethnicity-related bias. However, using this framework in other AI solutions can be challenging, as finding equivalent replacements in non-textual data is not straightforward.

As mentioned before, the bias in the sentiment results can be caused by different factors. Some of these factors are related to behavioral patterns, especially in social media posts. For instance, different genders might prefer different wordings in their social media posts, and as such, the sentiment analysis results show different sentiments in the same context. The SGBI measures the gender bias in the sentiment analysis library; it does not provide information on the impact of these patterns on the estimated sentiment. To extract

this information, further study is necessary to investigate the link between the SGBI, the posting pattern, and the choice of wording among different genders.

Author Contributions: Conceptualization, M.R.Y., H.H. and N.K.; Methodology, M.R.Y., H.H. and N.K.; Formal analysis, M.R.Y., H.H. and N.K.; Data curation, M.R.Y., H.H. and N.K.; Writing—original draft, M.R.Y., H.H. and N.K. All authors have read and agreed to the published version of the manuscript.

Funding: European Commission: Multi-hazard and Risk-informed System for Enhanced Local and Regional Disaster Risk Management (MEDiate) 101074075. This study was partly supported by IIASA's internal IBGF grant.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Kiritchenko, S.; Mohammad, S. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, New Orleans, LA, USA, 5–6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 43–53.
2. Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.-W. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 15–20.
3. Binns, R. Fairness in Machine Learning: Lessons from Political Philosophy. In Proceedings of Machine Learning Research, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; PMLR: Cambridge, MA, USA, 2018; Volume 81, pp. 149–159.
4. Liang, P.P.; Li, I.M.; Zheng, E.; Lim, Y.C.; Salakhutdinov, R.; Morency, L.-P. Towards Debiasing Sentence Representations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5502–5515.
5. Bolukbasi, T.; Chang, K.-W.; Zou, J.Y.; Saligrama, V.; Kalai, A. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. *arXiv* **2016**, arXiv:1607.06520. [CrossRef]
6. Buolamwini, J.; Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; PMLR: Cambridge, MA, USA, 2018; Volume 81, pp. 77–91.
7. Dastin, J. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *Reuters*, 2018 October 11; 296–299.
8. Madaan, N.; Mehta, S.; Agrawaal, T.; Malhotra, V.; Aggarwal, A.; Gupta, Y.; Saxena, M. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In Proceedings of Machine Learning Research, Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; PMLR: Cambridge, MA, USA, 2018; Volume 81, pp. 92–105.
9. Srivastava, B.; Rossi, F. Towards Composable Bias Rating of AI Services. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 284–289.
10. Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; Varshney, K.R. Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Optimized Pre-Processing for Discrimination Prevention. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
11. Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; Wang, W.Y. Mitigating Gender Bias in Natural Language Processing: Literature Review. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019.

12. Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; Vasserman, L. Measuring and Mitigating Unintended Bias in Text Classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 67–73.
13. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science* **2017**, *356*, 183–186. [CrossRef] [PubMed]
14. Sheng, E.; Chang, K.-W.; Natarajan, P.; Peng, N. The Woman Worked as a Babysitter: On Biases in Language Generation. *arXiv* **2019**, arXiv:1909.01326. [CrossRef]
15. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 335–340.
16. Kaneko, M.; Bollegala, D. Gender-Preserving Debiasing for Pre-Trained Word Embeddings. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 1641–1650.
17. Guo, Y.; Guo, M.; Su, J.; Yang, Z.; Zhu, M.; Li, H.; Qiu, M.; Liu, S.S. Bias in Large Language Models: Origin, Evaluation, and Mitigation. *arXiv* **2024**, arXiv:2411.10915. [CrossRef]
18. Hassani, H.; Komendantova, N.; Rovenskaya, E.; Yeganegi, M.R. Social Intelligence Mining: Unlocking Insights from X. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1921–1936. [CrossRef]
19. Rinker, T.W. Sentimentr: Calculate Text Polarity Sentiment; Buffalo, NY, USA. 2021. Available online: https://cran.r-project.org/web/packages/sentimentr/. (accessed on 3 August 2025).
20. Loria, S. Textblob Documentation, release 0.15; 2018. Available online: https://textblob.readthedocs.io/en/dev/index.html (accessed on 3 August 2025).
21. What is sentiment analysis and opinion mining? Microsoft Azure AI Language—Sentiment Analysis 2025. Availiable online: https://learn.microsoft.com/en-us/azure/ai-services/language-service/sentiment-opinion-mining/overview (accessed on 4 August 2025).
22. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
23. Roehrick, K. Vader: Valence Aware Dictionary and sEntiment Reasoner (VADER). 2020. Available online: https://cran.r-project.org/web/packages/vader/vader.pdf (accessed on 3 August 2025).