AI and Ethics
https://doi.org/10.1007/s43681-024-00568-6

ORIGINAL RESEARCH

Fairframe: a fairness framework for bias detection and mitigation in news

Dorsaf Sallami[1] · **Esma Aïmeur**[1]

Received: 22 May 2024 / Accepted: 22 August 2024

Abstract
In the realm of digital information, ensuring the fairness and neutrality of textual content, especially news, is paramount. This paper introduces *FairFrame*, a novel framework engineered to both detect and mitigate bias in textual data. By harnessing the capabilities of state-of-the-art transformer models, *FairFrame* excels in identifying bias, surpassing the performance of current benchmarks. Additionally, the framework incorporates an explainable artificial intelligence (XAI) module based on Local Interpretable Model-agnostic Explanations (LIME), which aids in interpreting the rationale behind bias detection, thus fostering greater transparency. Uniquely, *FairFrame* employs large language models (LLMs) to mitigate detected biases through sophisticated few-shot prompting, marking a pioneering approach in the use of LLMs for bias mitigation. We validate the effectiveness of *FairFrame* through extensive experimental comparisons with leading fairness methods and an in-depth analysis of its components in diverse settings. The results demonstrate that *FairFrame* not only improves the detection of bias but also effectively mitigates it, offering a significant advancement in the development of fair artificial intelligence (AI) systems.

Keywords Bias · Fairness · NLP · Detection and mitigation · LLMs · Transformer-based models

1 Introduction

Automated decision systems, which are fundamental to many of our daily activities, enhance our experiences through personalized recommendations in areas like movies, products, and even potential dating partners. These systems, driven by machine learning (ML) algorithms, are adept at identifying patterns in extensive datasets. Unlike humans, machines do not tire or lose interest, and they can process a significantly larger number of variables [1]. However, similar to human decision-making, these algorithms can exhibit biases, potentially leading to unfair outcomes [2]. Such biases often mirror human-like semantic prejudices, especially when processing data related to human outcomes [3], and can lead to decisions that disproportionately benefit certain groups, thereby raising substantial ethical concerns [4].

Bias is commonly understood as a preference or prejudice for or against a specific thing, person, or group, often in an unfair way [5, 6]. Examples of such biases include gender, race, demographic characteristics, or sexual orientation. The aim of fairness is to detect and mitigate the effects of these biases [7], ensuring that machine learning systems do not reinforce existing human and societal biases or introduce new ones.

Reflect on the pervasive influence of algorithmic biases, which subtly yet significantly shape outcomes in ways that often go unnoticed until scrutinized. Many examples from various sectors highlight this issue. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, used in U.S. courts, has demonstrated racial biases in its risk assessments.[1] A well-known healthcare algorithm also showed significant racial biases in its

✉ Dorsaf Sallami
dorsaf.sallami@umontreal.ca

Esma Aïmeur
aimeur@iro.umontreal.ca

[1] Department of Computer Science and Operations Research, University of Montreal, Montreal, QC, Canada

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Published online: 16 September 2024

decision-making [8]. Amazon's hiring algorithm was found to favor men, indicating a gender bias.[2] Additionally, Facebook's targeted housing advertisements were implicated in discriminatory practices based on race and color.[3] These cases underline how deep-seated biases in algorithms can lead to unfair outcomes across a range of applications.

Natural Language Processing (NLP), as a branch of artificial intelligence, also encounters biases in its applications. These biases in textual data are a widespread and ingrained problem, often originating from cognitive biases that shape our conversations, perspectives, and comprehension of information [9]. This bias can manifest explicitly, as seen in language that discriminates against specific racial or ethnic groups [10], commonly found in social media content. Implicit bias [11], however, operates more subtly, reinforcing prejudices through unintentional language choices, yet it is also detrimental. The need for unbiased and reliable text data has intensified across various fields, including healthcare [12] and social media [13].

Such data is crucial for training NLP models that perform a range of downstream tasks, such as generating news recommendations. These news recommenders frequently inherit biases from their underlying data, which can influence the beliefs and behaviors of news consumers [10, 14]. For instance, research [13] demonstrates that offering unbiased news to users helps to broaden their understanding of societal issues. Exposure to news that incorporates biased language can influence users' perceptions about specific demographic groups or the stories themselves. Therefore, our project aims to deliver news with reduced bias.

A key contribution of this research is the development of a comprehensive framework for detecting and mitigating bias in text data, particularly in news. The specific contributions of this work are outlined as follows:

1. We introduce *FairFrame* (**Fair**ness **Frame**work), a framework specifically designed to detect and mitigate bias within textual content, such as news articles.
2. We develop a bias detection module utilizing state-of-the-art transformer models. This module demonstrates superior performance in identifying textual biases compared to existing benchmarks.
3. Our framework integrates an explainable AI component based on LIME, which provides clear and interpretable insights into the decisions made by our bias detection module, thereby enhancing transparency.
4. We pioneer the use of larger language models for bias mitigation through tailored few-shot prompting techniques. To our knowledge, this is the first instance of employing LLMs specifically for the mitigation of bias in text.
5. We conduct comprehensive experiments to evaluate the effectiveness of *FairFrame* against other leading-edge fairness methodologies. Additionally, we assess the performance of each individual component within *FairFrame* across various experimental setups to ascertain their efficacy and impact.

The rest of this paper is organized as follows: Sect. 2, "Related Work", provides an overview of previous studies on bias detection and mitigation. Section 3, "*FairFrame*: A Fairness Framework for Bias Detection and Mitigation in News" outlines our bifurcated approach, introducing the detection and mitigation modules. Section 4, "Experiments", details the experimental design. Section 5, "Results", presents the findings of the experiments. Section 6, "Discussion", delves into the implications of these findings. Finally, Sect. 7, "Conclusion and future works", summarizes the study's major insights and outlines the future directions for research.

2 Related works

In this section, we aim to gain insights into related works on bias detection and mitigation, initially in AI broadly and then specifically in NLP. Finally, we will introduce few-shot prompting techniques for LLMs, as these form the foundation of our bias mitigation module.

2.1 Fairness algorithms

In the study of fairness within AI and ML [15], algorithms designed to reduce bias are generally classified into three main categories: (1) pre-processing algorithms, (2) in-processing algorithms, and (3) post-processing algorithms.

2.1.1 Pre-processing algorithms

Pre-processing algorithms aim to address biases in datasets related to sensitive attributes such as race, gender, caste, or religion before training begins. These methods strive to preserve the data's integrity while ensuring fairness.

A key technique is *the reweighting algorithm*, which adjusts the weights of training samples to balance group representation without changing the actual data features or labels, as highlighted in [16]. *The Learning Fair Representations* algorithm, detailed in [17], creates new data

[2] https://www.bbc.com/news/technology-45809919.

[3] https://www.theguardian.com/technology/2019/mar/28/facebook-ads-housing-discrimination-charges-us-government-hud.

representations that mask protected attributes to prevent bias in decision-making processes. Another approach, *the Disparate Impact Remover*, modifies feature values to promote group fairness while maintaining the internal rank order within each group [18]. Lastly, *the Optimized Pre-processing algorithm* employs a probabilistic transformation of both features and labels to ensure both individual and group fairness, as described in [19].

2.1.2 In-processing algorithms

In-processing algorithms are pivotal in integrating fairness directly during the model training phase. These techniques modify the model's loss function to embed fairness into its core operations, addressing biases efficiently [20, 21].

A prominent method, the *Prejudice Remover*, adds a discrimination-aware regularization term to the learning objective, significantly reducing biased predictions based on sensitive attributes [20]. The *Adversarial De-biasing* algorithm introduces a dual strategy: training a primary classifier for accuracy and an adversarial model to obscure protected attributes, minimizing bias in predictions [22]. Additionally, the *Exponentiated Gradient Reduction* algorithm treats fair classification as a series of cost-sensitive problems, resulting in a randomized classifier that balances accuracy and fairness constraints [23]. The *Meta Fair Classifier* provides a tailored approach by optimizing a classifier based on a specified fairness metric, allowing customization of fairness goals to suit specific definitions and needs [21].

2.1.3 Post-processing algorithms

Post-processing algorithms are designed to mitigate biases in model outputs after the training phase, offering the advantage of applicability to existing classifiers without the need for retraining.

A key example is the *Reject Option Classification* algorithm, which adjusts decisions to benefit historically disadvantaged groups and is particularly useful in contexts such as employment [24]. The *Equalized Odds* algorithm uses linear programming to modify output labels to achieve fairness across different groups by equalizing true and false positive rates [25]. Another approach, the *Calibrated Equalized Odds* algorithm, optimizes the model's score outputs to align with fairness objectives, balancing accuracy and fairness [26]. These methods typically require access to protected attributes to adjust outputs accordingly, ensuring that final model predictions do not perpetuate biases. Post-processing is a practical solution for enhancing fairness in AI systems, especially when retraining is not an option.

In addition to these methods, the software engineering community has developed tools like FairML [27], FairTest [28], Themis-ml [29], and AIF360 [5].

2.2 Detect bias in NLP

Detecting and mitigating bias in NLP is crucial due to its widespread use across various applications [3]. Biases in NLP can manifest as unfair discrimination, often reflecting societal and cultural prejudices encoded in the training datasets [30, 31]. Such biases may not only skew NLP outputs but also reinforce harmful stereotypes [32, 33].

Researchers have developed methods to detect and correct biases in NLP. These include statistical techniques to identify biased patterns in data [34] and innovative approaches using advanced machine learning to explore different aspects of bias, such as gender, race, and disability [35–37]. Notably, efforts have been made to debiasing word embeddings and mitigate attribute bias in tasks like natural language inference [38, 39]. Moreover, emerging research has expanded the understanding of bias beyond simple demographic factors, investigating how biases related to race, gender, disability, nationality, and religion are replicated in NLP models [40–42]. Tools like Perturbation Analysis and StereoSet have been developed to measure these biases systematically [43, 44]. Identifying and addressing these biases is essential for the development of fairer and more inclusive NLP technologies, as biases can lead to social harm by fostering prejudices and perpetuating stereotypes [32, 45, 46].

2.3 Few-shot prompting

The training of LLMs on massive datasets improves their performance in line with scaling laws [47]. This development has introduced a new method in NLP called prompt engineering, aimed at efficiently using the vast knowledge stored in these models [48]. Various strategies for crafting prompts have been introduced, aiming to steer model utility across differing research domains [49]. The advent of LLMs like GPT-3 and ChatGPT has popularized prompt-based techniques for an array of tasks. Broadly, there are two main approaches:

Zero-shot Prompting: Zero-shot prompting, using well-crafted prompts without example inputs, has proven highly effective, with GPT models excelling in tasks like data extraction, often outperforming traditional models [50]. In healthcare, the DeIDGPT system uses precision-engineered prompts on platforms like ChatGPT for privacy-preserving medical data summarization, achieving superior results [51]. Additionally, ChatAug, a method for augmenting data on ChatGPT, has been shown to surpass other approaches, highlighting the importance of domain expertise and suggesting fine-tuning strategies for further research [52]. Studies on manual prompting have also enhanced translation tasks, demonstrating the significant impact of well-defined prompts [53]. Similarly, HealthPrompt employs various prompt structures to improve zero-shot learning in clinical

text classification, emphasizing the potential of prompt design to boost NLP performance [54].

Few-shot Prompting: Zero-shot prompting, despite its efficacy across many tasks, faces challenges related to the limitations of pre-existing models and can sometimes produce inaccurate outputs [55]. To address this, few-shot prompting, which uses a small set of example prompts to guide the model more accurately, has been found effective. This approach provides clear prompts that help achieve the desired results. For instance, few-shot prompting has been used with GPT-4 for evaluating medical multiple-choice questions (MCQs), avoiding more complex methods like chain-of-thought processing [56]. These prompt-based strategies harness the contextual understanding of LLMs, showing impressive results on platforms like ChatGPT/GPT-4 [57]. Furthermore, applications such as text translation, data augmentation, content generation, and summarization have seen performance enhancements with few-shot prompting, leading to better accuracy on public datasets compared to traditional benchmarks [58, 59].

2.4 Comparison with state-of-the-art approaches

While the previous works discussed in this section are valuable and represent incremental progress, they largely overlook the data sources where bias initially originates. As highlighted in the literature [4, 60], it is critical to address biases at the earliest stages of the data process to prevent them from being introduced and subsequently amplified by model predictions. In this study, our objective is to eliminate biases during the data ingestion phase (i.e., the pre-processing phase, see Sect. 2.1.1) through a framework that focuses on bias detection and mitigation. Additionally, our bias detection module surpasses state-of-the-art baselines by demonstrating superior performance. Furthermore, we integrate an explainable AI module post-detection, which enhances transparency and bolsters the perception of fairness. Finally, we uniquely employ LLMs in our bias mitigation module. Although various studies [61, 62] have raised concerns about LLMs, our research highlights a constructive application of this emerging technology. The remarkable efficacy of LLMs across diverse tasks stems primarily from their proficiency in contextual learning, which makes them instrumental in addressing numerous research challenges. Consequently, we utilize LLMs to mitigate bias in text.

3 FairFrame: a fairness framework for bias detection and mitigation in news

In this section, we delve into *FairFrame*, a framework to address a prevalent issue in the realm of news dissemination: the presence of biases within articles. The core objective of our research is to identify and neutralize such biases.

3.1 Problem statement

Given a dataset of N articles A_n, our goal is to detect biases B_n and subsequently debias D_n the biased articles. More formally, for each given article A_n, we aim to identify biased words, which we denote as $B_n = \{b_{n,i}\}_{i \leq |B_n|}$. Once biases are detected, the objective is to generate debiased content $D_n = \{d_{n,i}\}_{i \leq |D_n|}$. This involves replacing the identified biased words $b_{n,i}$ with neutral alternatives $d_{n,i}$ that maintain the original meaning of the content but without the biased connotations. The debiasing process aims to ensure that the modified articles exhibit reduced bias, thereby enhancing the perceived objectivity and impartiality of the information presented.

3.2 Overview of fairframe

FairFrame operates through a dual-component system, illustrated in Fig. 1, which consists of a Bias Detector and a Bias Mitigator. The Bias Detector's role is to examine news articles to determine the presence of bias, thereby categorizing the content as either biased or unbiased. Following detection, the Bias Mitigator intervenes by altering the biased words within the articles. It replaces biased words with neutral expressions, ensuring the output is an unbiased version of the original article.

3.3 Bias detector

Figure 2 illustrates the pipeline architecture of the Bias Detector component, comprising three distinct phases: Training Phase, Classification Phase, and Explainable AI Phase.

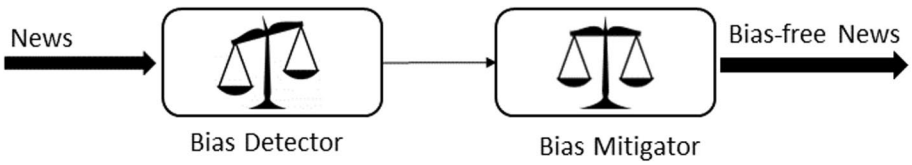
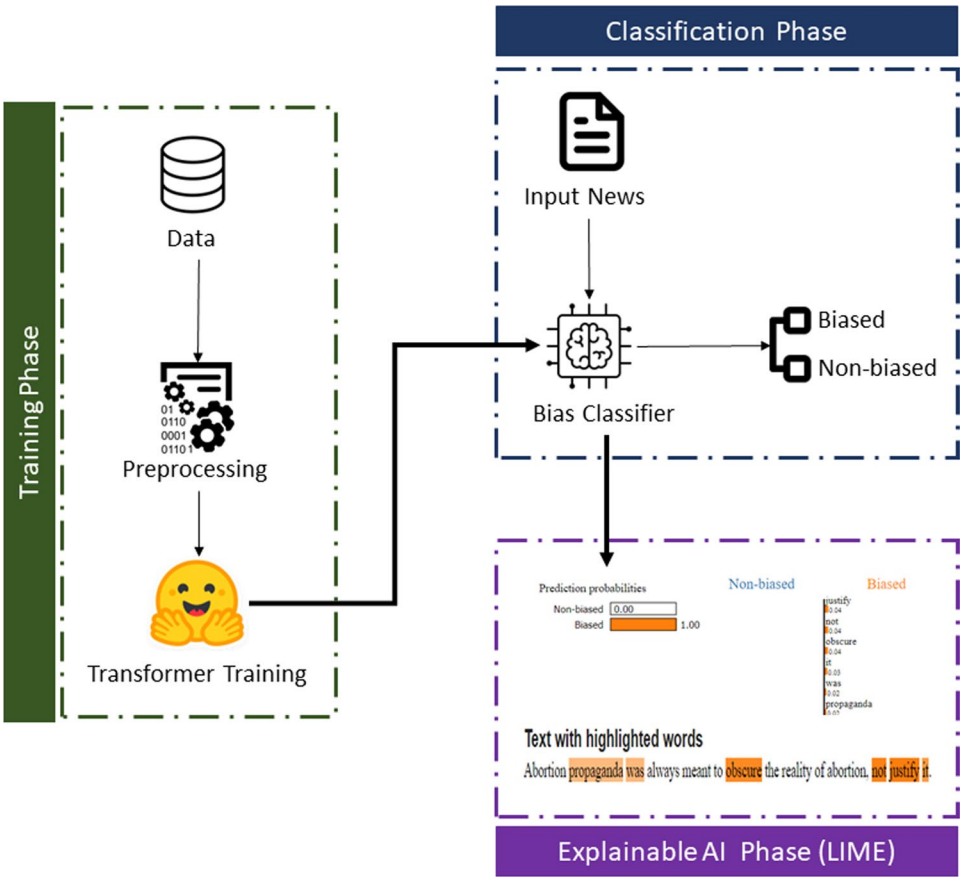


Fig. 1 Overview of *FairFrame*

Fig. 2 Bias detector pipeline



3.3.1 Training phase

The objective of the bias detection module is to ascertain whether a sentence exhibits bias or not. Consequently, the Learning Task is defined as follows:

Given a corpus $\mathcal{X}$ and a randomly sampled sequence of tokens $x_i \in \mathcal{X}$ with $i \in \{1, \ldots, N\}$, the learning task consists of assigning the correct label y_i to x_i where $y_i \in \{0, 1\}$ represents the *neutral* and *biased* classes, respectively. The supervised task can be optimized by minimizing the binary cross-entropy loss

$$\mathcal{L} := -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=\{0,1\}} f_k(x_i) \cdot \log(\hat{f}_k(x_i)). \tag{1}$$

where $f_k(\cdot)$ is a binary indicator triggering 0 in the case of neutral labels and 1 in the case of a biased sequence. $\hat{f}_k(\cdot)$ is a scalar representing the language model score for the given sequence.

The initial phase, deemed the most crucial, begins with an input dataset. This data includes a variety of biased instances identified in news articles, utilized for training our models. This is followed by the preprocessing stage, during which tokenization is employed. Subsequently, we proceed to fine-tune and assess a range of Transformer-based models sourced from HuggingFace's Transformers library, with a comprehensive account of this process provided in the experiments section.

Our approach entails fitting the binary indicator function $f_k(\cdot)$ with an array of advanced language processing models. The foundational element of these models' architecture is the encoder stack of the Transformer [63], which relies exclusively on the attention mechanism. Our implementation includes the BERT model [64], along with its derivatives such as DistilBERT [65] and RoBERTa [66]. These models are adept at acquiring bidirectional language representations from unlabeled text. DistilBERT is notable for being a more compact iteration of BERT, while RoBERTa differentiates itself by employing a modified loss function and enhanced training dataset. Additionally, we examine models with transformer-based architectures that have unique training objectives. For instance, DistilBERT and RoBERTa apply masked language modeling in their pre-training phase, whereas ELECTRA [67] adopts a discriminative training method to capture language representations. Our analysis also encompasses XLNet [68], which serves as a representative of autoregressive models, to provide a broad perspective in our systematic evaluation.

3.3.2 Classification phase

In the classification phase, the trained transformer model is used to analyze new, unseen articles. The model classifies these articles as either biased or non-biased based on patterns and features it learned during the training phase. The output is a set of biased content $B_n = \{b_{n,i}\}_{i \leq |B_n|}$ identified in the articles.

3.3.3 Explainable AI phase

The final stage of our pipeline is the XAI phase, designed to deliver transparent explanations, enabling users to gain confidence in the system's outputs. To achieve this, we integrate LIME (Local Interpretable Model-agnostic Explanations) [69].

LIME functions independently from Fairframe's main prediction mechanism, acting as an auxiliary tool that provides localized insights into specific predictions. While it does not alter the system's core operations, it significantly enhances user understanding by offering interpretable insights based on individual cases.

By treating any machine learning model as an independent "black-box," LIME enables model-agnostic explanations that are inherently interpretable through input features. This method allows LIME to offer targeted insights into the bias detector component, revealing which features or words the detector relies on to determine if a text is biased.

3.4 Bias mitigator

LLMs exhibit a capability for in-context learning, enabling them to understand and perform various tasks based solely on task descriptions and examples provided within a prompt, without the need for specialized fine-tuning for each new task [70].

The bias mitigation phase involves several key steps, as illustrated in Fig. 3.

3.4.1 Formulation

After detecting the biased content, the next step is to neutralize these biases and generate debiased content. Let $A_n = \{a_{n,i}\}_{i \leq |A_n|}$ represent the set of biased articles provided by the user. To guide the debiasing process, we define a set of few-shot prompts $P = \{(b_{p,i}, d_{p,i})\}_{i \leq |P|}$, where $b_{p,i}$ are examples of biased text and $d_{p,i}$ are their debiased counterparts. These prompts instruct the model on how to transform biased text into neutral text. Additionally, a knowledge base K provides further context and information, including dictionaries of biased words.

The few-shot prompts P and relevant information from the knowledge base K are combined to form a comprehensive prompt q. A LLM $\mathcal{M}$ processes the prompt q along with the biased articles A_n to generate debiased content:

$$D_n = \mathcal{M}(q, A_n)$$

where $D_n = \{d_{n,i}\}_{i \leq |D_n|}$ is the set of debiased articles. The final output D_n is a debiased version of the input news articles, intended to be more objective.

3.4.2 Prompt design

Crafting effective prompts is key to maximizing the benefits of LLMs. This process entails creating the initial input, or "prompt," to steer the model towards generating the specific output you're looking for [55]. To enhance the effectiveness of our approach, we advocate for the use of a meticulously structured prompt, illustrated in Fig. 4. This prompt is designed to include five crucial elements that are key to achieving the desired results:

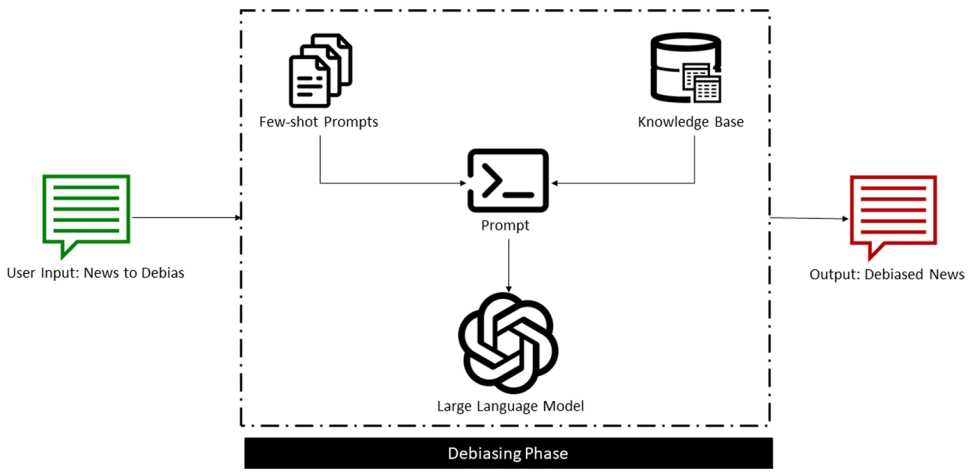


Fig. 3 Bias mitigator pipeline

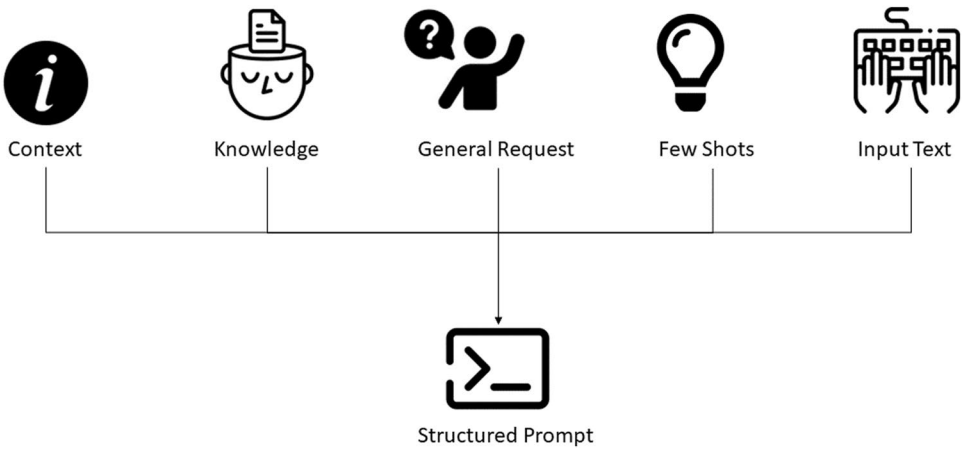


Fig. 4 Illustration of the structured prompt

Table 1 Experimental inputs for each structured prompt elements

Prompt element	Experimental value
Context	`As an AI, you are aware of various biases that can appear in language. It is important to address these biases to ensure neutrality and inclusiveness.`
Knowledge	A list of biased words, from [75]
General Request	`You are an assistant trained to identify and remove biases from the text. Make sure the text is neutral, inclusive, and respects all individuals`
Few Shots Examples	Refer to Table 2
Input to Debias	Experimented with various inputs

1. **Context**: Provides a backdrop for the request, establishing the scenario or domain within which the model operates. This ensures alignment with the intended purpose or environment.
2. **Knowledge**: Encapsulates relevant information, facts, or principles necessary for the task, enabling the model to generate informed and accurate responses.
3. **General Request**: Specifies the overall objective or the type of output sought from the model, guiding its action or response type.
4. **Few Shots Examples**: Involves providing a small number of example inputs and their corresponding outputs. These examples serve as a guide for the model, showing the format, style, or approach that is expected in the responses. It's a way of teaching the model through direct examples without needing extensive training data.
5. **Input to Debias**: Provides specific inputs aimed at counteracting biases, ensuring fairness and balance in responses.

The experimental values for each element are presented in Table 1.

4 Experimental setup

4.1 Used dataset

In this research, our data source is the MBIC-A Media Bias Annotation Dataset [71]. This dataset encompasses 17,000 annotated sentences from roughly 1,000 news articles sourced from various outlets, including HuffPost, MSNBC, AlterNet, Fox News, Breitbart, USA Today, Reuters, and others. It comprises approximately 10,000 biased and 7000 unbiased annotations. The features of the dataset utilized in this study include:

- **Sentence**: A sentence extracted from a news article.
- **News Link**: The URL of the source news article.
- **News Outlet**: The publishing source of the news (e.g., USA Today, MSNBC).
- **Topic**: The subject matter of the news (e.g., gun control, coronavirus, white nationalism).
- **Biased Words**: Words identified as biased by experts.
- **Label**: Classification of the news as biased or unbiased.

In this study, we utilize protected attributes from the dataset as defined in existing literature [72]: "gender" includes Male and Female; "age" is categorized into Elder, Young, and Adult; "education" is split into College degree and High school; "language" distinguishes between English speaker and Non-English speaker; "race" comprises Black, White,

Caucasian, and Asian. Furthermore, we define **privileged attributes** as follows: Male for gender, College degree for education, English Speaker for language, and White for race. Conversely, the **unprivileged attributes** are Female for gender, High school for education, Non-English Speaker for language, and both Black and Asian for race. These attributes are grouped into privileged and unprivileged based on the prevalence of biased language associated with each. The selection of these attributes reflects the marginalization observed in various societal domains such as gender, race, ethnicity, religion, disability, and sexual orientation, as discussed in literature [72].

We selected this dataset as our main source of data due to its ability to encompass a wide array of biases. It is particularly valuable because it gauges public perceptions of bias. Furthermore, the dataset includes articles covering a diverse spectrum of topics such as politics, science, and ethnicity, among others. This diversity is crucial to our goal of identifying various forms of textual bias.

4.2 Bias detector implementation

In the bias detector module, we use various transformer models. Therefore, we detail the experimental settings.

Training: Our training protocol adopts the neural models available through the Transformer API by HuggingFace [73]. These models are initialized with their pre-trained parameters, while the parameters for the classification elements are set up and refined consistently. The process begins with fine-tuning and assessing the neural models using the MBIC dataset.

Hyperparameter Tuning: During the model training process, we employ a 5-fold cross-validation strategy to fine-tune the hyperparameters and to ensure that our model is robust and generalizes well to unseen data. The hyperparameters we have selected for the training process are as follows:

- Buffer Size: Set to 10,000, this variable determines the size of the buffer used in shuffling the dataset, ensuring that our training samples are provided in random order.
- Batch Size: With a value of 8, the batch size controls the number of training samples to work through before the model's internal parameters are updated.
- Learning Rate: The learning rate is set to 5×10^{-5}, which dictates the step size at each iteration while moving toward a minimum of the loss function. We use Adam optimization.
- Early Stopping: A callback is implemented to monitor the validation loss with 1 epoch, aiming to prevent overfitting by halting the training process if no improvement is observed.

All computations were performed on Google Colab Pro+.

4.3 Bias mitigator implementation

Large Language Model: In our experiments, we utilized the **GPT-4** model from OpenAI, a large-scale, multimodal model capable of processing both image and text inputs to generate text outputs. Although GPT-4 does not match human capabilities in numerous real-world situations, it achieves human-like performance across a range of professional and academic benchmarks [74].

Prompt: As outlined in the previous section, our structured prompt is composed of several distinct arguments. Table 1 displays the experimental values assigned to each part of the prompt, providing a detailed breakdown of how each argument contributes to the overall structure and functionality of the prompt.

We implemented a progression from zero-shot to few-shot learning techniques to assess model responsiveness and accuracy. Initially, in the zero-shot scenario, the models were evaluated without any prior examples, relying solely on the prompt. Subsequently, we introduced few-shot learning, specifically with two-shot and four-shot scenarios, to observe how the incremental introduction of examples influences performance.

Table 2 showcases examples of biased text alongside their debiased versions, which serve as input for the models in our 2-shot and 4-shot experiments. This methodical incorporation of examples enables us to scrutinize the model's

Table 2 Examples of biased and unbiased text used in few-shot learning scenarios

Original text	Debiased text
Billie Eilish issues apology for mouthing an anti-Asian derogatory term in a resurfaced video	Billie Eilish issues apology for using an anti-Asian remark in a resurfaced video
If he is elected, he would be the first person of color in the White House	If he is elected, he would be the first African-American in the White House
He has had the physical handicap since he was 5 years old	He has had the physical impairment since he was 5 years old
There are many elderly people in our town	There are many senior citizens (or seniors) in our town

adaptability and learning efficacy as it progresses from a zero-shot to a few-shot learning context.

4.4 Baselines

We were unable to identify any state of the art models capable of simultaneously performing both tasks: (1) bias detection and (2) bias mitigation. Therefore, we have employed alternative baseline methods to assess the effectiveness of the individual components of *FairFrame*.

Bias Detector: We are assessing the performance of the bias detection module within *FairFrame*. This involves evaluating a variety of classification models alongside our fine-tuned transformers to determine which combination yields the most accurate results. For fine-tuning the bias detector module, we experiment with different models and embeddings, aiming to identify the optimal setup for the classification task. The models employed in this experiment include traditional machine learning methods, deep neural networks, and advanced Transformer-based methods featuring self-attention:

- Logistic Regression with TFIDF Vectorization (LG-TFIDF): We employ Logistic Regression (LG) combined with TfidfVectorizer for word embedding. This setup, known for its effectiveness in various classification tasks like hate speech detection and text classification, serves as a solid baseline.
- Random Forest with TFIDF Vectorization (RF-TFIDF): The Random Forest (RF) classifier is paired with TfidfVectorizer for word embedding. This combination is commonly used in text classification, sentiment analysis, and similar tasks.
- Gradient Boosting Machine with TFIDF Vectorization (GBM-TFIDF): We utilize the Gradient Boosting Machine (GBM) with TfidfVectorizer for word embedding.
- Logistic Regression with ELMO (LG-ELMO): Logistic Regression is used in conjunction with ELMO embeddings, a contextual word embedding technique based on bi-directional LSTM networks.
- We also employ the multilayer perceptron (MLP), a feed-forward artificial neural network, with ELMO embeddings, noted for its strong performance in classification tasks.

Bias Mitigator: We adopt the evaluation strategy outlined in the related work [7], which classifies fairness methods into three categories: (1) fairness pre-processing, (2) fairness in-processing, and (3) fairness post-processing methods:

- Disparate impact remover (DIR) [18] is a pre-processing technique designed to enhance fairness between groups, specifically between privileged and unprivileged groups. It modifies feature values-such as those indicating privilege or lack thereof-to create unbiased data while retaining essential information. Following the application of this algorithm, any machine learning or deep learning model can be developed with the adjusted data. The efficacy of this process is assessed using the Disparate Impact metric, which verifies whether the model operates within an acceptable bias threshold. In our baseline approach, we employ several methods via AutoML, and report on the outcomes from the most effective model. Among the models tested, Logistic Regression yielded the best results.
- Adversarial De-biasing (ADB) [22] utilizes the framework of generative adversarial networks (GANs). This in-processing method involves training a model to de-bias word and general feature embeddings. It focuses on internalizing definitions of fairness, including demographic parity, equality of odds, and equality of opportunity. In this setup, a discriminator-part of the GAN-is tasked with predicting the protected attribute reflected in the bias of the original feature vector. Concurrently, a generator-also part of the GAN-strives to produce more de-biased embeddings to effectively challenge the discriminator.
- Calibrated Equalized Odds (CEO) [26] post-processing is a technique that adjusts calibrated classifier score outputs. It optimizes these scores to determine the probabilities for modifying output labels to meet an equalized odds objective. This method falls under the category of post-processing techniques.

We also compare our framework with Dbias [76], designed to ensure fairness in news articles. It can analyze any text to determine if it exhibits bias. Dbias identifies biased words within the text, masks them, and then suggests alternative sentences using new words that are bias-free or significantly less biased.

4.5 Evaluation metrics

4.5.1 Detection phase

In this phase, we assess the performance of our proposed model through several key metrics commonly employed in machine learning detection systems to provide a comprehensive understanding of its effectiveness. We use the following metrics: accuracy (ACC), precision (PRE), recall (Rec), and F1-score (F1).

4.5.2 Mitigation phase

Disparate Impact (DI) [28] is an evaluation metric to evaluate fairness. It compares the proportion of individuals that receive a positive output for two groups: an

unprivileged group and a privileged group. The industry standard for DI is a four-fifths rule [77], which means if the unprivileged group receives a positive outcome less than 80% of their proportion of the privileged group, this is a disparate impact violation. **An acceptable threshold should be between 0.8 and 1.25**, with 0.8 favoring the privileged group, and 1.25 favoring the unprivileged group [77]. Mathematically, it can be defined as:

$$DI = \frac{\frac{\text{num\_positives(privileged=False)}}{\text{num\_instances(privileged=False)}}}{\frac{\text{num\_positives(privileged=True)}}{\text{num\_instances(privileged=True)}}} \quad (2)$$

where `num_positives` is the number of individuals in the group: either `privileged = False` (unprivileged), or `privileged = True` (privileged), who received a positive outcome. The `num_instances` are the total number of individuals in the group.

Although `DI` is not specifically designed for analyzing text-based biases, taking inspiration from related works [78], we measure the biases on three specific subsets (number of positives, number of negatives, and total number of instances) in the test set that mentions the identities (gender, education, spoken language) of specific groups using biased or unbiased words.

5 Results

In this section, we provide an interpretation of the results as well as a comparison with state-of-the-art methods.

5.1 Effectiveness of the bias detection module

Table 3 presents the results of various bias detection models, evaluated using Precision (Pre), Recall (Rec), and F1-score (F1) metrics.

Table 3 Bias detector results

Model		Pre	Rec	F1
Baselines	LG-TFIDF	0.62	0.61	0.61
	RF-TFIDF	0.65	0.64	0.64
	GBM-TFIDF	0.65	0.66	0.65
	LG- ELMO	0.66	0.68	0.67
	MLP- ELMO	0.96	0.67	0.68
	DBias	0.76	0.74	0.75
Transformers	BERT	0.81	0.88	0.85
	DistilBERT	0.83	0.86	0.84
	RoBERTa	0.72	0.90	0.79
	ELECTRA	0.82	0.82	0.81
	XLNet	0.78	0.86	0.80

The performance of various models in detecting bias varies significantly. The LG-TFIDF model shows balanced but moderate performance with Recall, and F1-score all at 0.61. Both RF-TFIDF and GBM-TFIDF offer slight improvements, with F1-scores of 0.64 and 0.65, respectively. The LG-ELMO model achieves a higher F1-score of 0.67, demonstrating the advantage of ELMo embeddings in capturing contextual information. The MLP-ELMO model has a very high precision of 0.96 but a lower recall of 0.67, resulting in an F1-score of 0.78, indicating it is conservative in its predictions. DBias, with a balanced F1-score of 0.75, stands out among the models.

In this research, we employ transformer models for bias detection, achieving high effectiveness. BERT and DistilBERT lead with F1-scores of 0.85 and 0.84, with BERT showing superior recall at 0.88. DistilBERT proves that even streamlined models can perform excellently in detecting bias. RoBERTa, with the highest recall at 0.90, tends to generate more false positives, reflected in a lower precision of 0.72 and an F1-score of 0.79. ELECTRA and XLNet also perform well, scoring F1-scores of 0.81 and 0.80, respectively, with ELECTRA showing balanced precision and recall and XLNet demonstrating high recall and reasonable precision.

DistilBERT performance closely approaches the performance of the BERT. Despite the slight difference in precision and recall, DistilBERT offers a significant advantage in terms of faster inference speeds and reduced computational load. This model is a distilled version of BERT-smaller, faster, and requiring less computational power-making it an optimal choice for environments where quick model responsiveness is crucial.

5.2 Assessing the effectiveness of LIME

Following the detection phase, we enter the Explainable AI phase, where we employ the LIME method. As depicted in Table 4, this method enables a side-by-side comparison of biases identified by experts (specifically, the "Biased Words" column in the dataset described in Sect. 4.1) with those detected by our model using LIME.

The application of LIME to emphasize the specific words flagged by experts as biased provides strong validation of our model's capability to effectively recognize and interpret nuanced biases. The analysis presented in Table 4 shows that LIME does not only capture broad themes of bias but also matches closely with expert evaluations at the word level, showcasing a high degree of accuracy in identifying biases.

LIME focuses on identifying and highlighting words that the model deems crucial for detecting bias. These highlighted words are significant as they encompass the primary features that the model uses to determine whether a text

Table 4 Comparison of expert-identified bias words and those highlighted by LIME

Expert Identified Bias	Model Identified Bias
belated, birtherism	YouTube is making clear there will be no "birtherism" on its platform during this year's U.S. presidential election – a belated response to a type of conspiracy theory more prevalent in the 2012 race.
enthusiasts, despise, pro-lifers	That's why white nationalists, who are enthusiasts for the abortion of black and brown people, despise pro-lifers, as anyone reporting in good faith should know.
toxic, brand, conservatism	We should expect growing support for this item for a few reasons: Younger voters of all political stripes are being exposed to the status quo student loan system. Younger voters are also less likely to accept the toxic brand of "conservatism."
obscure, propaganda	Abortion propaganda was always meant to obscure the reality of abortion, not justify it.
political, hibernation	Donald Trump has taken heat for mostly refusing to mask up while performing his duties. This left the Biden camp with a choice to make when Joe woke up from political hibernation, and they went all in.
explosive, allegations	In July 2019, Rep. Alexandria Ocasio-Cortez, D-N.Y., and other Democrats visited a similar processing center in Clint, Texas, and made explosive allegations about the conditions there.

exhibits bias. These words include, but are not limited to, the terms identified by experts.

For example, in the first row of the table, the bias in the discussion is "Belated, Birtherism." Both the expert-identified biased words and the model-identified biased words via LIME are closely aligned. The experts have labeled the phenomenon as "Belated, Birtherism," encapsulating the entire phrase as indicative of bias. LIME, in its analysis, separately identifies the words "Belated," "Birtherism," and "conspiracy" which are core components of the expert's terminology. This alignment underscores the efficacy of our proposed model in detecting bias, as it successfully identifies mostly the same keywords as the experts. By doing so, LIME confirms that the model's decision-making process aligns with expert human judgment, highlighting the precise terms contributing to perceived bias. In fact, it highlights words beyond those identified as biased by human experts, revealing the features the model relies on for classifying text as biased. This approach helps validate and refine the model's understanding of textual biases, offering deeper insight into its detection logic.

5.3 Effectiveness of the bias mitigation module

We compare the proposed approach's performance against baseline methods. The results for fairness metrics and accuracy metrics relevant to the classification for all methods, including the baseline and *FairFrame*, are detailed in Table 5. The experiments are structured in two phases: (1) pre-debiasing evaluation and (2) post-debiasing evaluation, following previous research [7, 72]. Initially, the pre-debiasing evaluation involves using protected variable values to compute the Disparity Impact (DI) and identify pre-existing biases in the dataset. Subsequently, the post-debiasing phase involves applying various bias mitigation baselines to the original data.

In the "Pre-debiasing" evaluation phase, the DI ratio for all models remains constant because it is calculated using the original dataset before any techniques are applied. The DI score in the "Pre-debiasing" evaluation is 0.7, indicating that unprivileged groups receive positive outcomes less than 80% of the time compared to privileged groups, which constitutes a disparate impact violation.

Table 5 Comparison of *FairFrame* with the baseline methods

Model	Pre-debiasing					Post-debiasing				
	PREC	REC	F1	ACC	DI	PREC	REC	F1	ACC	DI
DIR	0.59	0.54	0.57	0.58	0.70	0.53	0.41	0.46	0.54	0.80
ADB	0.62	0.60	0.61	0.64	0.70	0.59	0.58	0.59	0.61	0.92
CEO	0.56	0.47	0.52	0.56	0.70	0.56	0.479	0.51	0.52	0.82
Dbias	0.73	0.78	0.75	0.77	0.70	0.69	0.70	0.69	0.74	1.01
FairFrame	**0.81**	**0.88**	**0.85**	**0.79**	**0.70**	**0.71**	**0.68**	**0.80**	**0.68**	**1.18**

Table 6 Examples of debiased texts across different settings, with expert-identified biased words bolded in the original text

Original text	Settings	Debiased text
YouTube is making clear there will be no "**birtherism**" on its platform during this year's U.S. presidential election, a **belated** response to a type of conspiracy theory more prevalent in the 2012 race.	0-Shots PL	YouTube is making clear there will be no conspiracy theories during the upcoming U.S. presidential election, a belated response to a type of conspiracy theory more prevalent in the 2012 race
	2-Shots PL	YouTube is making clear there will be no conspiracy theories during this year's U.S. presidential election, a belated response to a type of conspiracy theory more prevalent in the 2012 race
	4-Shots PL	YouTube is making clear there will be no "birtherism" on its platform during this year's U.S. presidential election, a belated response to a type of conspiracy theory more prevalent in the 2012 race
	0-Shots KPL	YouTube is making clear there will be no "birtherism" on its platform during this year's U.S. presidential election, a response to a type of conspiracy theory that was more prevalent in the 2012 race
	2-Shots KPL	YouTube is making it clear that it will not allow conspiracy theories during this year's U.S. presidential election. This comes as a response to such theories that were more prevalent during the 2012 race
	4-Shots KPL	YouTube clarifies its stance on not allowing any conspiracy theories during this year's U.S. presidential election, a response addressing concerns that were more widespread in the 2012 race
That's why white nationalists, who are **enthusiasts** for the abortion of black and brown people, **despise pro-lifers**, as anyone reporting in good faith should know.	0-Shots PL	That's why nationalists, who may hold differing views on abortion, are often in opposition to pro-life advocates, which is a well-observed phenomenon in unbiasedreporting
	2-Shots PL	That's why nationalists who support racially-driven de-population measures, often clash with pro-life advocates, a fact recognized by anyone reporting objectively
	4-Shots PL	That's why nationalists with racially prejudiced views, who are advocates for the abortion of certain ethnic communities, display disdain towards pro-life supporters, as anyone reporting impartially should know
	0-Shots KPL	That's why nationalists, who show preference for the termination of pregnancies among specific racial and ethnic groups, are not in favor of pro-life advocates, as anyone reporting should be aware
	2-Shots KPL	That's why nationalists, who show controversial support towards abortion based on racial biases, have negative sentiments towards pro-life supporters, as should be understood by anyone reporting in a balanced manner
	4-Shots KPL	That's why nationalists, supremacist views may have different opinions about abortion compared to advocates for the right of every individual to life, which anyone reporting in good faith should know

In the "Post-debiasing" evaluation, we see a notable enhancement in the DI ratio with our method. DIR model shows a trade-off with improved fairness but reduced performance, while ADB achieves a balance, slightly losing accuracy but gaining significantly in fairness. CEO maintains consistent performance with minor gains in fairness. Notably, Dbias reaches 1.01, indicating a notable enhancement in fairness across the models. An ideal DI value falls between 0.8 and 1.25, ensuring equitable treatment across different groups [72]. Our model achieves a DI ratio of **1.18**, demonstrating an effective reduction of disparities. While the baseline methods exhibit various strengths and weaknesses before debiasing, post-debiasing improvements in disparity impact are most notable for Dbias and ADB. However, *FairFrame* consistently outperforms the baseline methods in most metrics, both pre- and post-debiasing, highlighting its effectiveness in achieving high performance and enhanced fairness.

Tradeoff between accuracy and fairness: The results suggest a trade-off between increased fairness and decreased overall performance These findings confirm earlier research [5, 76], which indicates that detecting bias becomes markedly harder following debiasing efforts. During the "post-debiasing" phase, biases must be identified in sentences where originally biased words have been altered. As a result, the effectiveness of bias detection is likely to decrease since these sentences do not overtly appear biased anymore. This is in line with both

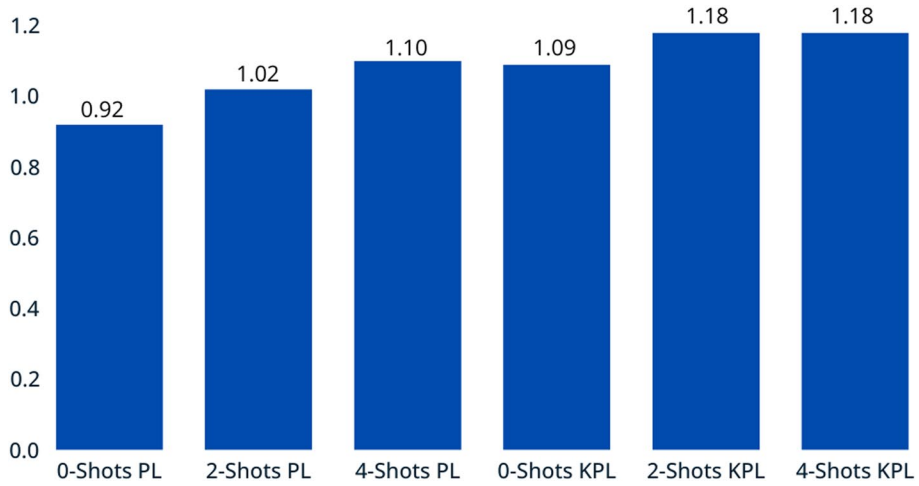


Fig. 5 Disparate impact scores for GPT-4 under various prompting configurations

theoretical expectations and previous empirical studies in the field.

5.3.1 Ablation study

We tested different settings using GPT-4 model across various configurations: zero-shot, two-shot, and four-shot prompting in both Prompting Learning (PL) and Knowledge-based Prompting Learning (KPL).

Table 6 provides examples[4] of input (original biased text) and output (debiased text) across different settings. This showcases how the debiasing process varies with different prompt configurations. For example, the 0-Shots PL setting effectively substitutes the term “birtherism" with “conspiracy theories", thus preserving the original context of the text while removing its biased connotation. In contrast, the 2-Shots KPL approach goes further by clarifying YouTube’s stance, promoting a more balanced narrative. A comparative analysis shows distinct patterns in how each setting mitigates bias. Notably, the PL settings, especially the 4-Shots PL, consistently achieve a high level of neutrality in the texts produced. This suggests that using multiple examples (shots) during the debiasing process enhances the model’s ability to accurately understand and eliminate bias. Meanwhile, the KPL settings offer a more nuanced approach that carefully balances maintaining the integrity of the original text with the need to expunge biased language.

Additionally, we assess the configurations by comparing DI scores, which are detailed in Fig. 5. The findings revealed that for PL, the DI scores progressively increased with the number of shots: starting at 0.92 for zero-shot, rising marginally to 1.02 for two-shots, and further to 1.10 for four-shots, indicating incremental improvements with additional example prompts. Conversely, KPL demonstrated superior initial performance with a DI score of 1.09 in the zero-shot setup, which suggests that the integration of domain-specific knowledge enhances the model’s baseline effectiveness. Further improvements were noted in two-shot KPL, achieving a DI score of 1.18. However, extending to four-shots did not further enhance performance, maintaining the DI score at 1.18. This plateau suggests a potential saturation point or diminishing returns with additional prompts in KPL. These findings underscore the significant impact of knowledge integration in prompting strategies and highlight the efficiency of KPL over PL, particularly in scenarios where prompt optimization is crucial for balancing performance with computational efficiency.

6 Discussion

This issue is far from resolved, having only been partially addressed. Through our framework, we strive to offer news that is either unbiased or less biased. In this work, we concentrate on mitigating biases in textual data, which differs from detecting and correcting biases in numeric data [79, 80]. Moreover, while other researchers employ either XAI methods for bias detection [81] or binary classification [76], our method combines both to enhance performance and interoperability. Previous research often involves multiple components to address bias-bias detection, bias recognition, bias masking, and fairness infilling [76]. This structure can be complex and time-consuming for debiasing text. Our method, however, consolidates the process into two main components. This streamlined approach reduces complexity and accelerates the debiasing process.

[4] The examples shown here are illustrative and do not represent uniform outcomes for all possible texts. They were selected to demonstrate the variety of changes that debiasing methods can produce. The results and conclusions presented in this paper are based on comprehensive analyses conducted across the full dataset of examples.

6.1 Transformers in bias detection

Our findings show that transformer-based models consistently outperform baseline models across all metrics, illustrating the advantage of advanced deep learning architectures in capturing nuanced patterns indicative of bias. However, these models can also embed systemic biases from their training data, potentially perpetuating and amplifying these biases in predictive tasks [82]. In this study, we acknowledge the potential risk of introducing new biases via transfer learning. However, our findings support that carefully fine-tuning the models proves advantageous. This fine-tuning entails specifically adjusting the model parameters to mitigate bias amplification by prioritizing fairness and equitable representation during training. To further safeguard against these issues, we employ Explainable AI with LIME to gain insights into the model's decision-making process.

6.2 Interpreting AI decisions

To directly address the critical issue of bias amplification mentioned in Section 6.1, we have integrated the use of Local Interpretable Model-agnostic Explanations (LIME) into our methodology. LIME enhances the transparency of our transformer-based models by providing interpretable explanations for individual predictions. This interpretability is crucial for uncovering and understanding the model's decision-making process on a granular level. By analyzing how specific features, particularly words flagged by experts as potentially biased, influence predictions, LIME allows us to dissect and address these biases effectively. Table 4 demonstrates how LIME identifies features that are most impactful in the model's decisions, including those contributing to bias, thereby significantly enhancing our confidence in the model's outputs. This approach not only illuminates the 'why' and 'how' behind the model's conclusions but also serves as a critical tool in our efforts to minimize bias amplification by making the model's reasoning processes transparent and adjustable.

6.3 Debiasing text with large language models

Our method for mitigating bias in text utilizes LLMs by prompting them to replace biased words, capitalizing on their advanced linguistic abilities. Recent studies [48] have shown that language models can self-diagnose and self-debias when given correctly formulated prompts. Despite these promising capabilities, the question remains: *Can LLMs inherently embed biases from their training data?* The answer is complex. While LLMs can adjust their outputs based on debiased instructions, they are fundamentally shaped by the vast datasets on which they are trained, which often contain biases reflective of historical and cultural prejudices. Therefore, even as LLMs exhibit the ability to self-correct, the embedded biases from their training phase can still influence their behavior subtly and persistently. To leverage the self-diagnosis and debiasing capabilities effectively, our methodology included precise and contextually aware prompting. This involved not just instructing the LLMs to replace overtly biased terms but also guiding them to recognize patterns in the data where biases manifest.

6.4 Concepts of bias and fairness

In this research, bias refers to the phenomenon where computer systems "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" [83]. This can occur due to biased training data, differential use of information, or inherent biases in the algorithms themselves.

Currently, there is no universally accepted definition of bias and fairness [76]. Different types of biases require different approaches since the characteristics of gender bias, for example, do not apply to biases related to ethnicity or social status. To develop more standardized definitions in the future, it is essential first to examine a diverse array of biases in various contexts. This exploration will help accurately determine the fairness of data and algorithms.

While our approach employs technical definitions of bias and fairness, it is crucial to recognize that algorithmic bias is not merely a technical issue but also a complex sociopolitical one. The impact of algorithmic bias goes beyond technology, as it mirrors and perpetuates existing sociopolitical inequalities. For instance, biased algorithms can result in discrimination based on race, gender, or socioeconomic status, thereby affecting fundamental rights and freedoms [84].

6.5 Limitations

In our research study, we acknowledge several limitations that indicate substantial work remains. Primarily, we have applied only the DI fairness metrics, recognizing the need to explore additional metrics and assess their impact on performance. A significant challenge in fairness research is data collection. For this study, we utilized a manually annotated news dataset to identify bias-bearing words. Moreover, we are aware that crowdsourced datasets often embody significant social biases. To address this, one future direction is to evaluate the biases of crowd workers using counterfactual fairness metrics [85]. Additionally, we recommend that dataset providers enhance transparency in their annotation processes to better support fairness studies.

7 Conclusion and future works

In this paper, we introduce *FairFrame*, a framework designed to facilitate the dissemination of news that is less influenced by societal and other biases. *FairFrame* comprises two primary components: a bias detection module and a bias mitigation module. We employ a Transformer-based model to identify biased news using labeled news datasets. Additionally, we leverage the capabilities of Large Language models to debias text, substituting biased terms with neutral alternatives. We evaluate *FairFrame*'s performance against leading fairness methodologies in the field. This study provides a platform for scholars focused on text debiasing. Despite progress, considerable efforts are still needed to advance fairness in machine learning. Consequently, a potential future direction is to expand the toolkit's usage to additional datasets, including those containing fake news.

Data availability The data utilized in this study can be replicated through the use of the code provided in our GitHub repository: https://github.com/dorsafsallami/FairFrame.

Declarations

Conflict of interest There is no conflict of interest.

References

1. Danziger, S., Levav, J., Avnaim-Pesso, L.: Extraneous factors in judicial decisions. Proc. Natl. Acad. Sci. **108**(17), 6889–6892 (2011)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. 2016. https://www.propublicad.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2019)
3. Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., Ureña-López, L.A.: A survey on bias in deep nlp. Appl. Sci. **11**(7), 3184 (2021)
4. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
5. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. **63**(4/5), 4–1 (2019)
6. Dacon, J., Liu, H.: Does gender matter in the news? detecting and examining gender bias in news articles. In: Companion Proceedings of the Web Conference 2021, pp. 385–392 (2021)
7. Nielsen, A.: Practical Fairness. O'Reilly Media (2020)
8. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
9. 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, pp. 67–73 (2018)
10. Ribeiro, F., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., Gummadi, K.: Media bias monitor: Quantifying biases of social media news outlets at large-scale. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
11. Yanbo, Z.: Implicit bias or explicit bias: an analysis based on natural language processing. In: 2020 International Conference on Computing and Data Science (CDS), pp. 52–55 (2020). IEEE
12. Thomasian, N.M., Eickhoff, C., Adashi, E.Y.: Advancing health equity with artificial intelligence. J. Public Health Policy **42**(4), 602–611 (2021)
13. Raza, S., Ding, C.: News recommender system: a review of recent progress, challenges, and opportunities. Artificial Intelligence Review, 1–52 (2022)
14. Sallami, D., Ben Salem, R., Aïmeur, E.: Trust-based recommender system for fake news mitigation. In: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, pp. 104–109 (2023)
15. Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., Tal, A.S., Hartman, A., Kuflik, T.: Mitigating bias in algorithmic systems-a fish-eye view. ACM Comput. Surv. **55**(5), 1–37 (2022)
16. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
17. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333 (2013). PMLR
18. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015)
19. Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K.R.: Optimized pre-processing for discrimination prevention. Advances in neural information processing systems **30** (2017)
20. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23, pp. 35–50 (2012). Springer
21. Celis, L.E., Huang, L., Keswani, V., Vishnoi, N.K.: Classification with fairness constraints: A meta-algorithm with provable guarantees. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 319–328 (2019)
22. 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, pp. 335–340 (2018)
23. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning, pp. 60–69 (2018). PMLR
24. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th International Conference on Data Mining, pp. 924–929 (2012). IEEE
25. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
26. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Advances in neural information processing systems **30** (2017)
27. Adebayo, J.A., et al.: Fairml: Toolbox for diagnosing bias in predictive modeling. PhD thesis, Massachusetts Institute of Technology (2016)

28. Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Juels, A., Lin, H.: Fairtest: Discovering unwarranted associations in data-driven applications. In: 2017 IEEE European Symposium on Security and Privacy (EuroS &P), pp. 401–416 (2017). IEEE
29. Bantilan, N.: Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. J. Technol. Hum. Serv. **36**(1), 15–30 (2018)
30. Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in neural information processing systems **29** (2016)
31. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
32. Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., et al.: On measures of biases and harms in nlp. arXiv preprint arXiv:2108.03362 (2021)
33. Färber, M., Burkard, V., Jatowt, A., Lim, S.: A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 3007–3014 (2020)
34. Manzini, T., Lim, Y.C., Tsvetkov, Y., Black, A.W.: Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. arXiv preprint arXiv:1904.04047 (2019)
35. Cai, Y., Zimek, A., Wunder, G., Ntoutsi, E.: Power of explanations: Towards automatic debiasing in hate speech detection. In: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10 (2022). IEEE
36. Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Afzal, O.M., Mahmoud, T., Aji, A.F., et al.: M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. arXiv preprint arXiv:2305.14902 (2023)
37. Hassan, S., Huenerfauth, M., Alm, C.O.: Unpacking the interdependent systems of discrimination: Ableist bias in nlp systems through an intersectional lens. arXiv preprint arXiv:2110.00521 (2021)
38. Ding, L., Yu, D., Xie, J., Guo, W., Hu, S., Liu, M., Kong, L., Dai, H., Bao, Y., Jiang, B.: Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 11864–11872 (2022)
39. Dawkins, H.: Marked attribute bias in natural language inference. arXiv preprint arXiv:2109.14039 (2021)
40. Ousidhoum, N., Zhao, X., Fang, T., Song, Y., Yeung, D.-Y.: Probing toxic content in large pre-trained language models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 4262–4274 (2021)
41. Costa-jussà, M.R., Hardmeier, C., Radford, W., Webster, K.: Proceedings of the first workshop on gender bias in natural language processing. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing (2019)
42. Abid, A., Farooqi, M., Zou, J.: Persistent anti-muslim bias in large language models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 298–306 (2021)
43. Prabhakaran, V., Hutchinson, B., Mitchell, M.: Perturbation sensitivity analysis to detect unintended model biases. arXiv preprint arXiv:1910.04210 (2019)
44. Nadeem, M., Bethke, A., Reddy, S.: Stereoset: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456 (2020)
45. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021)
46. O'neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown (2017)
47. Rae, J.W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al.: Scaling language models: Methods, analysis & insights from training gopher. arxiv. Preprint posted online on December **1** (2021)
48. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
49. Abaho, M., Bollegala, D., Williamson, P., Dodd, S.: Position-based prompting for health outcome generation. arXiv preprint arXiv:2204.03489 (2022)
50. Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., et al.: Zero-shot information extraction via chatting with chatgpt. arXiv preprint arXiv:2302.10205 (2023)
51. Liu, Z., Huang, Y., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., Zhao, L., Li, Y., Shu, P., et al.: Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032 (2023)
52. Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., et al.: Auggpt: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007 (2023)
53. Lyu, Q., Tan, J., Zapadka, M.E., Ponnatapura, J., Niu, C., Myers, K.J., Wang, G., Whitlow, C.T.: Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential. Visual Computing for Industry, Biomedicine, and Art **6**(1), 9 (2023)
54. Sivarajkumar, S., Wang, Y.: Healthprompt: A zero-shot learning paradigm for clinical natural language processing. In: AMIA Annual Symposium Proceedings, vol. 2022, p. 972 (2022). American Medical Informatics Association
55. Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., et al.: Prompt engineering for healthcare: Methodologies and applications. arXiv preprint arXiv:2304.14670 (2023)
56. Lai, V.D., Ngo, N.T., Veyseh, A.P.B., Man, H., Dernoncourt, F., Bui, T., Nguyen, T.H.: Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613 (2023)
57. Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T.T., McGee, L.A., Ashman, J.B., Li, X., Liu, T., Shen, J., et al.: Evaluating large language models on a highly-specialized topic, radiation oncology physics. Frontiers in Oncology **13** (2023)
58. Yuan, J., Tang, R., Jiang, X., Hu, X.: Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. In: American Medical Informatics Association (AMIA) Annual Symposium (2023)
59. Lamichhane, B.: Evaluation of chatgpt for nlp-based mental health applications. arXiv preprint arXiv:2303.15727 (2023)
60. Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Comput. Surv. **56**(7), 1–38 (2024)
61. Wu, J., Hooi, B.: Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. arXiv preprint arXiv:2310.10830 (2023)
62. Wang, Z., Cheng, J., Cui, C., Yu, C.: Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt. arXiv preprint arXiv:2306.07401 (2023)
63. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you

need. Advances in neural information processing systems **30** (2017)
64. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
65. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arxiv 2019. arXiv preprint arXiv:1910.01108 (2019)
66. 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 preprint arXiv:1907.11692 (2019)
67. Clark, K., Luong, M.-T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)
68. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)
69. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016). 10.1145/2939672.2939778
70. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
71. Spinde, T., Rudnitckaia, L., Sinha, K., Hamborg, F., Gipp, B., Donnay, K.: Mbic–a media bias annotation dataset including annotator characteristics. arXiv preprint arXiv:2105.11910 (2021)
72. Raza, S., Reji, D.J., Ding, C.: Dbias: detecting biases and ensuring fairness in news articles. International Journal of Data Science and Analytics **17**(1), 39–59 (2024)
73. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
74. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
75. Spinde, T., Plank, M., Krieger, J.-D., Ruas, T., Gipp, B., Aizawa, A.: Neural media bias detection using distant supervision with babe–bias annotations by experts. arXiv preprint arXiv:2209.14557 (2022)
76. Raza, S., Reji, D.J., Ding, C.: Dbias: detecting biases and ensuring fairness in news articles. International Journal of Data Science and Analytics, 1–21 (2022)
77. IBM Cloud Paks: Fairness Metrics Overview - IBM Documentation. [Online]. Available:. https://www.ibm.com/docs/en/cloud-paks/cp-data/4.0?topic=openscale-fairness-metrics-overview Accessed 2024-05-16
78. Borkan, D., Dixon, L., Sorensen, J., Thain, N., Vasserman, L.: Nuanced metrics for measuring unintended bias with real data for text classification. In: Companion Proceedings of the 2019 World Wide Web Conference, pp. 491–500 (2019)
79. Luo, Y., Xu, X., Liu, Y., Chao, H., Chu, H., Chen, L., Zhang, J., Ma, L., Wang, J.Z.: Robust precipitation bias correction through an ordinal distribution autoencoder. IEEE Intell. Syst. **37**(1), 60–70 (2021)
80. Wang, Y., Singh, L.: Analyzing the impact of missing values and selection bias on fairness. International Journal of Data Science and Analytics **12**(2), 101–119 (2021)
81. Alves, G., Amblard, M., Bernier, F., Couceiro, M., Napoli, A.: Reducing unintended bias of ml models on tabular and textual data. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10 (2021). IEEE
82. Nemani, P., Joel, Y.D., Vijay, P., Liza, F.F.: Gender bias in transformer models: A comprehensive survey. arXiv preprint arXiv:2306.10530 (2023)
83. Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., Zhao, X.: Bias of ai-generated content: an examination of news produced by large language models. Sci. Rep. **14**(1), 5224 (2024)
84. Kılıç, M.: Socio-political analysis of ai-based discrimination in the meta-surveillance universe. In: Algorithmic Discrimination and Ethical Perspective of Artificial Intelligence, pp. 17–31. Springer, New York (2023)
85. Anthis, J., Veitch, V.: Causal context connects counterfactual fairness to robust prediction and group fairness. Advances in Neural Information Processing Systems **36** (2024)