

Article

A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data

Jothi Prakash Venugopal ¹, Arul Antran Vijay Subramanian ², Gopikrishnan Sundaram ³, Marco Rivera ^{4,*} and Patrick Wheeler ⁴

¹ Department of Information Technology, Karpagam College of Engineering, Myleripalayam Village, Coimbatore 641032, Tamil Nadu, India; jothiprakashv@gmail.com

² Department of Computer Science and Engineering, Karpagam College of Engineering, Myleripalayam Village, Coimbatore 641032, Tamil Nadu, India; arulantranvijay@gmail.com

³ School of Computer Science and Engineering, VIT-AP University, Amaravati 522237, Andhra Pradesh, India; gopikrishnan.s@vitap.ac.in

⁴ Power Electronics, Machines and Control (PEMC) Research Institute, University of Nottingham, 15 Triumph Rd, Lenton, Nottingham NG7 2RD, UK; pat.wheeler@nottingham.ac.uk

* Correspondence: marco.rivera@nottingham.ac.uk



Citation: Venugopal, J.P.; Subramanian, A.A.V.; Sundaram, G.; Rivera, M.; Wheeler, P.A. Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data. *Appl. Sci.* **2024**, *14*, 11471. <https://doi.org/10.3390/app142311471>

Academic Editors: Francisco De Arriba-Pérez, Silvia García-Méndez and Enrique Costa-Montenegro

Received: 12 November 2024

Revised: 5 December 2024

Accepted: 8 December 2024

Published: 9 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Sentiment analysis is a vital component of natural language processing (NLP), enabling the classification of text into positive, negative, or neutral sentiments. It is widely used in customer feedback analysis and social media monitoring but faces a significant challenge: bias. Biases, often introduced through imbalanced training datasets, can distort model predictions and result in unfair outcomes. To address this, we propose a bias-aware sentiment analysis framework leveraging Bias-BERT (Bidirectional Encoder Representations from Transformers), a customized classifier designed to balance accuracy and fairness. Our approach begins with adapting the Jigsaw Unintended Bias in Toxicity Classification dataset by converting toxicity scores into sentiment labels, making it suitable for sentiment analysis. This process includes data preparation steps like cleaning, tokenization, and feature extraction, all aimed at reducing bias. At the heart of our method is a novel loss function incorporating a bias-aware term based on the Kullback–Leibler (KL) divergence. This term guides the model toward fair predictions by penalizing biased outputs while maintaining robust classification performance. Ethical considerations are integral to our framework, ensuring the responsible deployment of AI models. This methodology highlights a pathway to equitable sentiment analysis by actively mitigating dataset biases and promoting fairness in NLP applications.

Keywords: sentiment analysis; natural language processing; transformer models; bias mitigation; social media analytics

1. Introduction

Sentiment analysis, often referred to as opinion mining [1,2], is a key area within natural language processing (NLP) that focuses on recognizing and categorizing emotions or sentiments expressed in text. With the digital era generating an enormous amount of user-generated content, sentiment analysis has become an essential tool across many fields [3]. Businesses use it to track customer feedback, healthcare providers use it to assess mental health, politicians to monitor public sentiment, and social scientists to explore societal trends [4]. It is vital to understand how people feel about a product, service, or issue, which in turn can influence everything from brand management to market predictions based on consumer input and online conversations [5].

This research focuses on addressing one of the most pressing challenges in sentiment analysis within such interconnected ecosystems: mitigating bias in sentiment predictions. Bias can manifest in many ways, such as favoring one demographic over another, misclassifying sentiments in underrepresented groups, or amplifying stereotypes [4]. Such biases

can have far-reaching implications, from reinforcing societal inequalities to eroding trust in AI systems [6]. By developing a bias-aware sentiment analysis framework, this study aims to enhance the reliability and fairness of sentiment predictions, particularly in applications where fairness is paramount [7].

1.1. Background and Motivation

The rise of the Internet and, more recently, social networks has led to an explosion of text data [8,9], giving us an unprecedented opportunity to better understand human emotions and opinions. Sentiment analysis has been gaining momentum as a way to automatically process and make sense of this overwhelming volume of information. Its applications are broad, ranging from automating customer service interactions [10] to tracking public sentiment around elections [11], and even monitoring mental health trends [12]. However, as the technology has evolved, so have the challenges. One of the toughest hurdles is the inherent complexity of human emotion. People express their feelings in so many ways, sometimes with sarcasm, idioms, or cultural references—that it can be difficult for a machine to detect these subtleties [13]. As a result, creating a model that works universally, in different contexts and cultures, is difficult.

An additional challenge comes from the potential for bias in sentiment analysis models [14,15]. These biases often creep in through the training data and can lead to skewed predictions. For instance, if a model is trained on data from one specific demographic, it may not be able to accurately interpret the sentiments of people from other backgrounds. This becomes especially problematic when applying sentiment analysis to sensitive areas such as sexual orientation, religion, or disability, where biased results could have serious consequences. Given these challenges, our research is driven by two primary motivations: developing a sentiment analysis approach that is both technically accurate and ethically sound. Our goal is to address the technical hurdles of sentiment analysis while also confronting the ethical issues involved in deploying these models, particularly in delicate or high-impact areas.

1.2. Objective and Contribution

In this research, our objective was to present a novel sentiment analysis approach that strengthens not only accuracy but also ethical decency. We are particularly interested in the possible risks of unfair or inaccurate forecasts in contexts where these could lead to very significant harm, for example, sexuality, religious belief, and disability.

The key contributions of our work are as follows.

1. Development of Bias-BERT and bias-aware loss function: A novel sentiment analysis model, named Bias-BERT, is tailored to handle biased sentiment prediction; there is a special loss function for this model, which includes a bias-aware term based on KL divergence. These target class distributions are specially created during the model training for it to learn and reduce biases towards those classes, so that predictions can be fairer.
2. Real-time user feedback integration: Our unique approach integrates the real-time user feedback loop. This enables users to help provide feedback on the accuracy of the model as well as any biases that may be present in the predictions. It relays this feedback to the other models, and they all learn together while maintaining a higher quality model that accounts for any bias.
3. Comprehensive bias evaluation and ethical considerations: We also achieved an unbiased evaluation of biased predictions using the same sets of test examples by validating across multiple demographic groups including but not limited to race, gender, and sexuality; comparative experiments demonstrate that debiasing does indeed reduce bias significantly from the baseline. We also look at the ethical considerations when utilizing bias-aware sentiment models and provide best practices for responsible and equitable use.

4. Contribution to fair AI research: Our research contributes towards the broader movement for fair AI by offering a blueprint for the development of bias-aware models. Although we are centered around sentiment analysis, this framework is versatile and works for a wide range of NLP tasks, contributing to fairness in AI applications across numerous industries.

The real significance of this research lies in its ability to advance sentiment analysis by tackling the key challenge of bias. By creating a model that prioritizes both accuracy and fairness, our study hopes to contribute to the development of more equitable and trustworthy sentiment analysis tools. These tools can be used in a wide range of fields, from helping businesses better understand their customers to helping researchers in the social sciences.

The rest of this paper is structured as follows. In Section 2, we review existing work in sentiment analysis and bias mitigation. Section 3 explains our proposed methodology. Section 4 presents the experimental setup and results, followed by a discussion of ethical considerations in Section 5. Finally, Section 6 concludes the paper and highlights potential directions for future research.

2. Related Work

Sentiment analysis has seen tremendous growth over the past few decades. In this section, we cover the major developments in sentiment analysis and the recent focus on mitigating bias in these models.

2.1. Sentiment Analysis

Sentiment analysis, a key area within Natural Language Processing (NLP), revolves around identifying and categorizing emotions or sentiments found in text. The primary aim is to classify the text into categories like positive, negative, or neutral. In the early days of sentiment analysis [16,17], lexicon-based methods were the most popular. These approaches relied on predefined lists of words, each assigned a sentiment score to reflect whether the word conveyed a positive or negative meaning [18]. The overall sentiment of a text was then calculated by adding up the sentiment scores of all the words in the text [19,20]. While straightforward and not requiring labeled data, lexicon-based approaches struggled with context-sensitive words and idiomatic phrases, making it hard for them to pick up on more nuanced sentiments. As machine learning techniques evolved, sentiment analysis shifted towards more data-driven methods. Algorithms such as support vector machines (SVMs) [21] and naive Bayes [22] became popular for this task. These models require labeled data for training, where each text sample is tagged with its respective sentiment [23,24]. Once trained, these models can predict the sentiment of new unlabeled text. While more accurate than lexicon-based methods, machine learning models are computationally demanding and require large quantities of labeled data.

Deep learning has completely transformed sentiment analysis by enabling models to detect more complex patterns and relationships in text [25,26]. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) [27] are often used in this space. RNNs [28] are particularly effective in analyzing sequences, which makes them great for processing sentences and longer texts. On the other hand, CNNs, originally designed for image processing, have also been proven effective in text classification [29,30]. Although these models generally outperform traditional machine learning methods, they come with higher computational costs and complexity [31]. Recently, transformer-based architectures [32], especially BERT (Bidirectional Encoder Representations of Transformers) [33], have set new performance benchmarks in NLP, including sentiment analysis. Developed by Google, BERT is pre-trained on a massive corpus and can be fine-tuned for specific tasks, providing a powerful and flexible framework for modern NLP applications. BERT's strength comes from its ability to capture context using a self-attention mechanism [30,34], allowing it to understand the sentiment of a word based on its surrounding text. This is particularly useful in sentiment analysis, where the context can drastically change the meaning of a word.

Various transformer variants such as GPT [35], RoBERTa [36], and DistilBERT [37] have been developed to improve BERT or address specific challenges. These models highlight the adaptability and strength of transformer architectures, with fine-tuning allowing them to specialize in tasks like sentiment analysis across different domains. However, the high computational demands of these models remain a challenge, often requiring specialized hardware for optimal performance.

2.2. Bias in Sentiment Analysis

The issue of bias in machine learning, particularly in sentiment analysis, has gained significant attention. Bias can appear in many forms, including gender, race, age, and socio-economic status [26]. Tackling these biases is crucial for ethically sound sentiment analysis models. The bias in sentiment analysis models often comes from the training data [26]. For example, if the data primarily consist of reviews from one demographic, the model may end up biased toward the perspectives of that group. Similarly, if the data include gendered language or culturally specific phrases, the model can inherit these biases [38]. Understanding the makeup of the training data is the first step in recognizing where bias could be introduced. Detecting bias in sentiment analysis models is both challenging and necessary [39]. Several tools and methods have been developed to help identify and measure bias. For instance, the AI Fairness 360 toolkit offers a set of metrics to understand bias in model predictions [40]. These metrics can reveal disparities in how a model performs between different demographic groups, helping researchers focus on where bias needs to be addressed.

Once bias is detected, the next step is finding ways to reduce it. Various techniques have been proposed [41,42], each with its own strengths and weaknesses. One common approach is to re-sample the training data to ensure they represent different groups more fairly [43]. Another strategy is to add regularization terms to the model loss function, penalizing biased predictions [44]. More advanced techniques, such as adversarial training, aim to prevent the model from learning sensitive attributes, reducing the chance of biased predictions [45]. The intersection of sentiment analysis and bias mitigation has led to the emerging field of bias-aware sentiment analysis [46,47]. This area focuses on building models that are not only good at predicting sentiment but are also mindful of the biases they may introduce. Bias-adaptive training methods [47,48] adjust the training process in response to detected biases in the data or model. For example, if the model shows a bias towards a certain demographic, the training algorithm can be modified to place more emphasis on correctly classifying underrepresented groups [45]. Another way to achieve bias-aware sentiment analysis is by tweaking the loss function during training [49]. Bias in sentiment analysis can have real-world effects, from reinforcing harmful stereotypes to unfairly disadvantaging certain groups. Our work aims to make a meaningful contribution to this area by proposing a new methodology that not only improves accuracy, but also actively reduces bias.

2.3. Research Gap

Although a significant amount of research has been conducted in the field of sentiment analysis, there are still some crucial gaps that our study aims to address. First, most existing research does not consider the potential influence of implicit bias in the sentiment analysis process. Such biases can skew the results and lead to inaccurate or misleading conclusions. Our research seeks to mitigate this by developing a bias-aware sentiment analysis technique. Secondly, the majority of sentiment analysis techniques are binary in nature, considering only positive or negative sentiments. This can overlook the nuances of human emotions and the complexity of language. Our multi-dimensional approach to sentiment analysis allows for a more comprehensive understanding of sentiments expressed in text. Lastly, there is a lack of effective methods for detecting gender-inclusive language in sentiment analysis. With the rise of awareness around gender inclusivity, it is essential to develop techniques that can accurately analyze sentiments in a gender-inclusive language. Our

research proposes a novel methodology for this purpose, contributing to the broader efforts to make AI and NLP technologies more inclusive and representative of diverse user populations.

In light of these gaps, our research presents significant contributions to the field of sentiment analysis. By addressing these issues, we aim to improve the accuracy and inclusivity of sentiment analysis techniques, making them more useful and relevant in today's diverse and dynamic digital landscape. In this research, we propose a novel method for bias-aware sentiment analysis that combines the approaches of data augmentation and adversarial training. We also introduce bias-penalizing terms into the loss function. These terms effectively act as a regularization mechanism that discourages the model from making biased predictions. By optimizing this modified loss function, the trained model aims to be both accurate in its sentiment classification and mindful of potential biases. We evaluated our method on a dataset of movie reviews, and we show that it outperforms baseline methods in terms of accuracy and bias.

3. Proposed Methodology

The proposed methodology for bias-aware sentiment analysis incorporates the Jigsaw Unwanted Bias in the Toxicity Classification dataset, BERT embeddings, and user feedback for continuous improvement. The approach includes five major stages:

1. Data adaptation: adaptation of the Jigsaw Unintended Bias dataset for sentiment analysis by introducing polarity labels based on toxicity scores.
2. Data preprocessing: text cleaning, tokenization, and feature extraction using BERT.
3. Bias-aware sentiment classifier: introduction of a novel classifier, Bias-BERT, designed to mitigate bias in sentiment predictions.
4. Loss function and implicit bias reduction: incorporation of Kullback–Leibler (KL) divergence as a bias metric in the loss function.
5. User feedback loop: integration of real-time user feedback for model refinement.

The following subsections provide detailed insights on the various stages of the proposed method. Figure 1 shows the architecture of the proposed method.

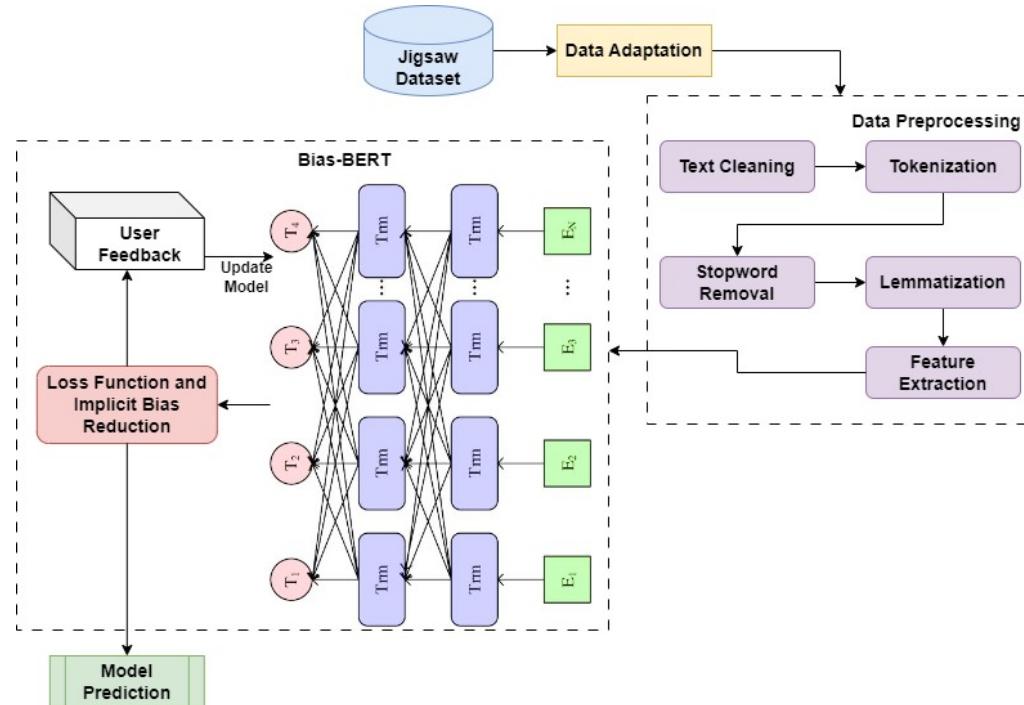


Figure 1. Architecture of the proposed method.

3.1. Dataset

The data set used in this study is the Jigsaw Unintended Bias in Toxicity Classification dataset [50], a large-scale public dataset released by the Conversation AI team, an initiative founded by Jigsaw and Google. This data set contains 1,971,916 instances of online comments, annotated with toxicity scores and additional attributes. It was designed to enable machine learning models to better understand toxicity in comments while addressing bias and fairness challenges.

To facilitate training, validation, and testing, the dataset is divided into three subsets:

- Training set: comprising 1,804,874 instances, this subset is used to train and fine-tune the model.
- Validation set: containing 70,180 instances, this subset is used for hyperparameter tuning and intermediate evaluations during model development.
- Test set: consisting of 96,862 instances, this subset is reserved for evaluating the model's performance on unseen data.

Table 1 provides an overview of the corrected dataset distribution.

Table 1. Distribution of the Jigsaw Unintended Bias in Toxicity Classification dataset.

Subset	Number of Instances
Total	1,971,916
Training set	1,804,874
Validation set	70,180
Test set	96,862

The data set includes a variety of toxicity types, ranging from subtle to explicit forms, and is annotated with demographic and contextual metadata, allowing robust bias analysis. This diversity makes it an ideal choice to evaluate the fairness and accuracy of sentiment analysis models.

To further enhance the utility of the dataset, we added sentiment polarity labels to adapt it to sentiment analysis tasks. The toxicity score of each comment was used to classify it into one of three sentiment categories: positive, neutral, or negative. This modification ensured that the data set aligned with the objectives of this study, facilitating the development and testing of our bias-aware sentiment analysis framework.

3.2. Data Adaptation

The main task in the data adaptation step was to convert the Jigsaw Unexpected Bias data set, originally designed for toxicity detection, into a format suitable for sentiment analysis. The dataset includes a comment text column containing individual comments and a target column with continuous toxicity scores ranging from 0 to 1, alongside various attributes of the toxicity subtype.

To better align toxicity scores with a classification scheme based on toxicity levels, we introduced a new categorization that assigns comments to three distinct toxicity levels: *low*, *mild* and *high* as shown in Table 2. This categorization directly reflected the degree of toxicity in the comments and made the data set more suitable for sentiment analysis that involved understanding toxicity levels, rather than conventional sentiment categories.

Table 2. Mapping toxicity scores to toxicity levels.

Toxicity Range	Toxicity Level
0.0–0.3	Low (non-toxic)
0.3–0.7	Mild
0.7–1.0	High (toxic)

Steps for Adaptation

1. Load the Jigsaw Unintended Bias dataset into a Panda DataFrame.
2. Define the thresholds for toxicity scores to determine toxicity levels. In this study, we set a lower threshold at 0.3 and an upper threshold at 0.7.
3. Add a new column called *toxicity_level* to the DataFrame. Each comment was assigned a label (*low*, *mild*, or *high*) based on its toxicity score.
4. Save the updated data set with the *toxicity_level* column for further modeling development and testing.

This approach enhanced sentiment analysis by interpreting toxicity as a spectrum of sentiment. Non-toxic comments were classified as *low*, toxic comments as *high*, and those that were neither benign nor harmful were labeled as *mild*. This also ensured that the dataset was in alignment with the toxicity analysis while maintaining the focus on sentiment analysis.

3.3. Data Preprocessing

The data preprocessing stage involved transforming the adapted Jigsaw Unexpected Bias data set into a suitable format to develop the bias-aware sentiment analysis model. Using the *spaCy* [51] library, essential preprocessing techniques were applied to ensure consistency and improve the quality of the textual data. Text normalization was performed to standardize the input by removing special characters, extra whitespaces, and digits, and converting all text to lowercase. Tokenization splits the text into individual tokens for structured analysis, allowing the model to process words or phrases effectively. Noise reduction techniques, such as the removal of common stopwords and the lemmatization of tokens to their root forms, were employed to emphasize meaningful content while eliminating redundant or irrelevant information. These preprocessing steps ensured that the data set was clean, concise, and optimized for bias-aware sentiment analysis, without overloading the model with unnecessary information.

Feature Extraction

The preprocessed text was converted into numerical features using the Bidirectional Encoder Representations from Transformers (BERT) model, which served as the feature extraction backbone for our bias-aware sentiment analysis model. BERT is a pre-trained deep learning model designed for a wide range of natural language understanding tasks, including sentiment analysis. It generates contextualized embeddings for input text, capturing rich semantic and syntactic relationships between words. For each preprocessed comment, the feature extraction process was formalized as follows:

$$\mathbf{x}_i = \text{extract_features_BERT}(T(c'_i)) \quad (1)$$

where \mathbf{x}_i is the feature vector for the i th comment, $\text{extract_features_BERT}(\cdot)$ is the feature extraction function using BERT, and $T(c'_i)$ represents the tokenized and preprocessed version of the comment c'_i . The BERT model generates a fixed-length feature vector for each comment, obtained by averaging or pooling the final hidden layer embeddings of the tokenized text. Specifically, these embeddings encapsulate the following key features:

1. Contextual word representations: each token's embedding accounts for its meaning in the context of the surrounding words, which is critical for understanding nuances in sentiment.
2. Global comment representation: By applying techniques such as extraction or pooling of [CLS] token embedding (e.g., mean pooling) on all token embeddings, BERT provides a holistic representation of the entire comment.
3. Subtle semantic features: BERT captures complex relationships, such as sarcasm or indirect sentiment, which are often missed by traditional bag-of-words or TF-IDF approaches.

These features were highly expressive and enabled the downstream model to accurately predict sentiment while minimizing bias. After feature extraction, the dataset was

split into training, validation, and testing subsets to facilitate model development and performance evaluation.

This detailed feature extraction process ensured that the adapted Jigsaw Unexpected Bias dataset was effectively transformed into a robust format suitable for bias-aware sentiment analysis using BERT. The pseudocode of the proposed method is shown in Algorithm 1.

Algorithm 1 Bias-aware sentiment analysis with Bias-BERT

Require: Training dataset D_{train} , test dataset D_{test} , demographic group G

Ensure: Trained model M , accuracy, precision, recall, F1-score

```

1: function PREPROCESSDATA( $D$ )
2:   for each comment  $c$  in  $D$  do
3:      $c \leftarrow \text{clean}(c)$ 
4:      $c \leftarrow \text{tokenize}(c)$ 
5:      $c \leftarrow \text{normalize}(c)$ 
6:   end for
7:   return  $D$ 
8: end function
9: function FILTERBYGROUP( $D, G$ )
10:   $D_G \leftarrow []$ 
11:  for each comment  $c$  in  $D$  do
12:    if  $c.\text{group} = G$  then
13:       $D_G.append(c)$ 
14:    end if
15:  end for
16:  return  $D_G$ 
17: end function
18: function TRAINMODEL( $X_{\text{train}}, \lambda$ )
19:   Initialize Bias-BERT model  $M$ 
20:    $L_{CE} \leftarrow \text{define\_cross\_entropy\_loss}()$ 
21:    $L_{bias} \leftarrow \text{define\_KL\_divergence\_loss}()$ 
22:    $L \leftarrow \mathcal{L}_{CE}(y, f_{db}(x)) + \lambda L_{bias} + \mu \sum_{i=1}^N f_i$ 
23:    $M \leftarrow \text{minimize\_loss}(M, X_{\text{train}}, L)$ 
24:   return  $M$ 
25: end function
26: function UPDATEMODEL( $M, \text{feedback}$ )
27:   for each  $f$  in  $\text{feedback}$  do
28:     Update  $M$  based on  $f$ 
29:   end for
30:   return  $M$ 
31: end function
32:  $D_{\text{train}} \leftarrow \text{PreprocessData}(D_{\text{train}})$ 
33:  $D_{\text{test}} \leftarrow \text{PreprocessData}(D_{\text{test}})$ 
34:  $D_{\text{train}-G} \leftarrow \text{FilterByGroup}(D_{\text{train}}, G)$ 
35:  $D_{\text{test}-G} \leftarrow \text{FilterByGroup}(D_{\text{test}}, G)$ 
36:  $X_{\text{train}} \leftarrow \text{extract\_features}(D_{\text{train}-G})$ 
37:  $X_{\text{test}} \leftarrow \text{extract\_features}(D_{\text{test}-G})$ 
38:  $\lambda \leftarrow \text{hyperparameter\_value}$ 
39:  $M \leftarrow \text{TrainModel}(X_{\text{train}}, \lambda)$ 
40:  $\text{feedback} \leftarrow \text{collect\_user\_feedback}()$ 
41: if  $\text{feedback} \neq \text{empty}$  then
42:    $M \leftarrow \text{UpdateModel}(M, \text{feedback})$ 
43:    $M \leftarrow \text{TrainModel}(X_{\text{train}}, \lambda)$ 
44: end if
45:  $\text{predictions} \leftarrow M.\text{predict}(X_{\text{test}})$ 
46: Evaluate predictions to get accuracy, precision, recall, F1-score
  
```

3.4. Bias-Aware Sentiment Classifier

In this section, we introduce Bias-BERT, a sentiment classifier leveraging the power of BERT to perform sentiment analysis while actively mitigating biases in predictions. The classifier integrates a bias-aware loss function to ensure fairness, reduce biases, and enhance the interpretability of its outputs.

Let the input text be denoted as x and its corresponding sentiment label as y . The BERT model learns a mapping function $f(\cdot)$ from x to y , i.e., $y = f(x)$. In Bias-BERT, the goal is to learn a debiased mapping function $f_{db}(\cdot)$, to ensure accurate sentiment classification while minimizing biases. This is achieved through the bias-aware loss function, defined as:

$$L = \mathcal{L}_{CE}(y, f_{db}(x)) + \lambda L_{bias} \quad (2)$$

where \mathcal{L}_{CE} represents the cross-entropy loss for sentiment classification, L_{bias} is the bias-aware loss term that penalizes deviations from fairness, and λ is a hyperparameter controlling the trade-off between classification accuracy and bias mitigation.

The bias-aware term L_{bias} is formulated using the Kullback–Leibler (KL) divergence, which measures the divergence between the predicted sentiment distribution $P(y|x)$ and a reference unbiased distribution $Q(y)$. The bias-aware term is expressed as:

$$L_{bias} = D_{KL}(P(y|x) \parallel Q(y)) = \sum_y P(y|x) \log\left(\frac{P(y|x)}{Q(y)}\right) \quad (3)$$

where $P(y|x)$ is the predicted probability distribution of sentiment labels for input x , and $Q(y)$ is the reference unbiased distribution, typically a uniform distribution where all sentiment classes are equally likely.

3.5. Loss Function with User Feedback Integration

To enhance the model's adaptability and fairness, a user feedback loop is incorporated into the training process, focusing on a representative subset of predictions rather than the entire dataset. This approach ensures feasibility while providing meaningful real-world insights into the model's performance and fairness. The unified loss function is defined as:

$$L = \mathcal{L}_{CE}(y, f_{db}(x)) + \lambda L_{bias} + \mu \sum_{i=1}^N f_i \quad (4)$$

where $\mathcal{L}_{CE}(y, f_{db}(x))$ is the cross-entropy loss for sentiment classification. L_{bias} is the bias-aware loss term that penalizes deviations from fairness, computed using Kullback–Leibler (KL) divergence. $\sum_{i=1}^N f_i$ aggregates user feedback f_i over N feedback instances. λ and μ are hyperparameters that control the trade-offs between accuracy, bias mitigation, and feedback integration.

3.5.1. User Feedback Process

Feedback was collected selectively from a curated subset of predictions rather than the entire training dataset to ensure scalability and practicality. Specifically, we sampled 10,000 comments from the test set and engaged a diverse group of users, including domain experts, researchers, and general users. This diverse group of participants was carefully chosen to provide diverse perspectives and reduce the risk of introducing new biases.

Types of Feedback Collected

1. Accuracy feedback: Users indicated whether the sentiment prediction was correct or incorrect.
2. Bias feedback: Users identified any perceived biases in the predictions, specifying their nature (e.g., racial, gender, cultural) and optionally providing contextual comments for further refinement.

3.5.2. Incorporating Feedback into Training

The feedback signals f_i were systematically aggregated to inform model adjustments. The $L_{feedback}$ term in the unified loss function accounted for user feedback and was expressed as:

$$L_{feedback} = \sum_{i=1}^N f_i, \quad (5)$$

where f_i is a numeric representation of the feedback provided for the i th prediction. Feedback was categorized into actionable patterns, and model parameters were adjusted accordingly during training.

Feedback Integration Steps

1. Feedback was aggregated and categorized to identify recurring issues or biases.
2. The bias-aware loss function L_{bias} was updated to incorporate these patterns, penalizing frequently reported biases more heavily.
3. The model parameters were iteratively updated to minimize the unified loss function L , ensuring continuous improvement.
4. Retraining cycles were conducted periodically with feedback-adjusted parameters, and performance was re-evaluated using standard and fairness-specific metrics.

This feedback-driven training mechanism allowed the model to dynamically adapt to real-world contexts, enhancing both accuracy and fairness. By focusing on a manageable subset of predictions and engaging a diverse group of users, the feedback loop balanced practicality and impact while avoiding the need for exhaustive labeling of the entire dataset.

3.6. Training Process and Implementation

The training process began with feature extraction, where the input text was tokenized and passed through the BERT model to extract contextual embeddings. The extracted features were used to predict sentiment labels. The loss function, combining cross-entropy loss, KL divergence, and user feedback penalties, was then computed. During training, model parameters were iteratively updated to minimize total loss L , ensuring a balance between precision and fairness. The evaluation of the model involved standard metrics such as accuracy, precision, recall, and F1-score, along with fairness-specific metrics to validate bias reduction.

4. Experimental Results and Discussion

4.1. Evaluation Metrics

In this research, we evaluated the performance of our proposed methodology using the following standard metrics:

4.1.1. Accuracy

The proportion of correctly classified instances out of the total instances.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (6)$$

4.1.2. Precision

The proportion of true positive predictions out of the total positive predictions made by the model.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (7)$$

4.1.3. Recall

The proportion of true positive predictions out of the total actual positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (8)$$

4.1.4. F1-Score

The harmonic mean of precision and recall, which provides a balanced measure of the performance of the model.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

4.2. Baseline Models

We compared the performance of our proposed methodology with several baseline models, including Logistic Regression, SVM, Multilayer Perceptron (MLP), LSTM, and several recent transformer-based models. Logistic regression (LR) is a straightforward linear model for binary classification that uses the logistic function to model class probabilities, making it a popular choice for foundational classification tasks due to its simplicity and interpretability [52]. Support Vector Machines (SVMs), a discriminative classifier, are effective for both linear and non-linear classification tasks, as they identify the optimal hyperplane for separating data points, providing robust performance in text classification tasks [53]. Multilayer Perceptron (MLP), a feedforward artificial neural network, is capable of modeling complex, nonlinear relationships in data and is often used as a baseline in deep learning studies [54]. Long Short-Term Memory (LSTM), a type of Recurrent Neural Network (RNN), is specifically designed to address the problem of vanishing gradients, making it highly suitable for sequence-based tasks such as text and sentiment analysis [55]. In addition, we included several modern transformer-based models to reflect the current state of the art in sentiment analysis. RoBERTa, a variant of BERT, improves key training techniques and hyperparameters to improve the robustness of the model and performance [56]. DistilBERT [56], a smaller, faster version of BERT, is designed for efficiency, offering reduced size and faster inference while retaining much of BERT's performance. These baseline models, which include both traditional methods and recent advances in NLP, provide a strong set of comparisons to evaluate the effectiveness of our proposed method in sentiment analysis, particularly in terms of mitigating bias.

4.3. Existing Models in the Literature

We further compare the performance of our proposed methodology with existing models in bias-aware sentiment analysis, including counterfactual data enhancement, adversarial training, Bias-Product-of-Experts (BPoE), and Bias-Regularized Neural Networks. These models were selected for comparison due to their focus on addressing bias in machine learning and their relevance in similar contexts. Counterfactual data augmentation generates counterfactual samples by altering specific words or phrases in the original text, creating new examples to help the model mitigate unintended bias in sentiment prediction [57]. Adversarial training incorporates an adversarial component that predicts the presence of bias in input data, while the main model is trained to minimize both sentiment classification loss and adversarial bias detection loss, enabling the model to learn representations less sensitive to biases [7]. Bias-Product-of-Experts (BPoE) employs multiple specialized “expert” classifiers, each handling specific subsets of data and combines their predictions to reduce the influence of biases [58]. Bias-Regularized Neural Networks include a bias regularization term in the loss function, penalizing biased predictions and encouraging the model to learn more unbiased representations [59]. Lastly, we included BiasFinder, a model that uses metamorphic testing to uncover biases in sentiment anal-

ysis systems, ensuring fairness and helping to identify biases that might otherwise go unnoticed [60].

4.4. Comparison with Baseline Models

Table 3 summarizes the performance of our proposed method against baseline models (Logistic Regression, SVM, MLP, LSTM, RoBERTa, and DistilBERT) on the test set, evaluated using accuracy, precision, recall, and F1-score. Logistic regression and SVM achieved modest performance, with accuracies of 76.5% and 77.8%, respectively, reflecting their limitations in handling complex, nonlinear data. The MLP model showed better results (79.3% accuracy), leveraging its ability to capture intricate patterns, while LSTM outperformed it with 81.0% accuracy, demonstrating the benefit of modeling sequential relationships. RoBERTa and DistilBERT, both transformer-based models, also demonstrated high performance, with RoBERTa achieving 85.0% accuracy and DistilBERT reaching 84.5%. Our proposed method significantly outperformed all baseline models, achieving 92.5% accuracy and an F1-score of 92, showing a clear advantage in both accuracy and bias mitigation.

Table 3. Comparison of the proposed method with baseline models and existing models in the literature on the test set. (Bold represent performance of proposed method).

Model	Accuracy	Precision	Recall	F1-Score
Baseline Models				
Logistic Regression	0.765	0.734	0.698	0.716
SVM	0.778	0.751	0.719	0.735
MLP	0.793	0.762	0.748	0.755
LSTM	0.810	0.781	0.769	0.775
RoBERTa	0.850	0.836	0.812	0.824
DistilBERT	0.845	0.825	0.800	0.812
Existing Models in the Literature				
Counterfactual data augmentation	0.805	0.784	0.771	0.778
Adversarial training	0.820	0.798	0.783	0.791
Bias-Product-of-Experts (BPoE)	0.815	0.792	0.779	0.786
Bias-Regularized Neural Network	0.825	0.803	0.788	0.796
BiasFinder	0.850	0.833	0.808	0.819
Proposed method	0.925	0.932	0.921	0.926

4.5. Comparison with Existing Models in the Literature

Table 3 presents a comparison of our proposed method with state-of-the-art models for bias-aware sentiment analysis, including counterfactual data augmentation, adversarial training, Bias-Product-of-Experts (BPoE), Bias-Regularized Neural Network, and BiasFinder. The evaluation metrics included accuracy, precision, recall, and F1-score. The proposed method significantly outperformed all other models, achieving an accuracy of 92.5% compared to 82.5% from the next best model, the bias-regulated neural network. Furthermore, Bias-BERT demonstrated superior precision (93.2%) and recall (92.1%), indicating its ability to make accurate predictions and capture a large proportion of relevant instances while minimizing false positives and negatives. The F1-score of the proposed method (92.6%) further highlighted its balanced performance between precision and recall, outperforming the next highest F1 score of 79.6% from the bias-regulated neural network. Other models, such as counterfactual data augmentation, adversarial training, and BPoE, exhibited lower performance across all metrics, with F1-scores ranging from 77.8% to 79.1%. These results clearly demonstrated the effectiveness of the proposed method in not only addressing bias in sentiment analysis, but also achieving superior overall classification performance, making it a robust and reliable solution for sentiment prediction tasks.

4.6. Confusion Matrix

The confusion matrix provides a comprehensive evaluation of classification performance, capturing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Figure 2 illustrates the confusion matrices for the proposed method, baseline models (MLP, LSTM), and existing approaches such as Bias-Product-of-Experts (BPoE), Bias-Regularized Neural Network, and adversarial training. The proposed method achieved the highest classification accuracy with 45,300 TP, 47,262 TN, and the lowest FP (1800) and FN (2500), showcasing its robustness in sentiment classification. In contrast, the MLP model recorded higher FP (2400) and FN (3000), reflecting its limitations in handling complex data. The LSTM model showed better performance with reduced FP (2100) and FN (2700), leveraging sequential information effectively. Among existing methods, the Bias-Regularized Neural Network exhibited competitive performance with 2100 FP and 2800 FN, outperforming BPoE, which had higher FP (3100) and FN (2900). Adversarial training achieved a balance with FP at 2600 and FN at 2700 but remained less effective than the proposed approach.

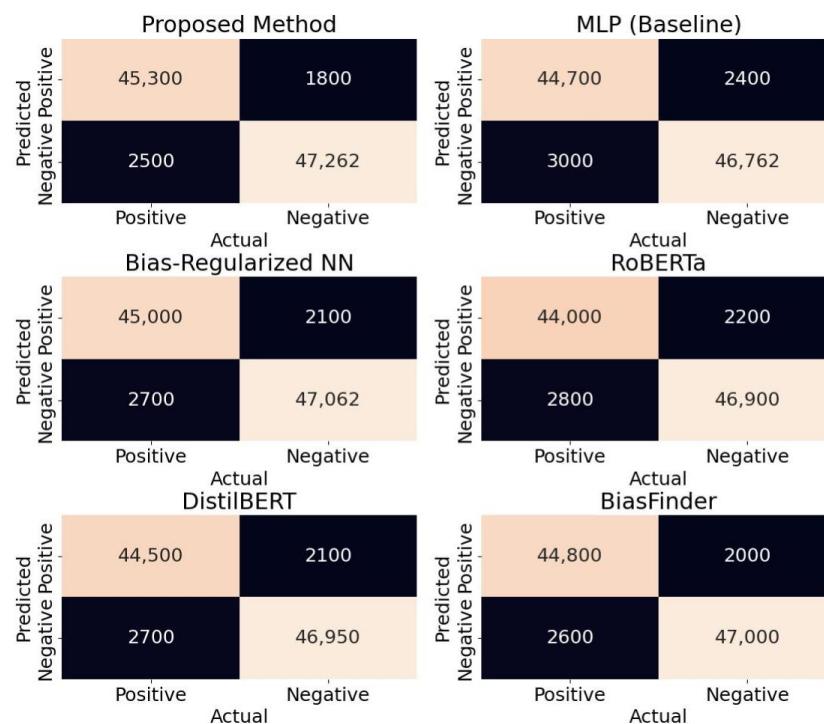


Figure 2. Confusion matrices for the proposed method and other models.

4.7. Bias Evaluation

In this section, we take a closer look at how well our model performed across different demographic groups from the Jigsaw data set. The main goal here was to understand how effectively the model mitigated bias when dealing with diverse categories of people. The data set was divided into several key demographic groups: race, ethnicity, sex, sexual orientation, religion, and disability. Breaking down the data in this way helped us ensure that our model performed fairly and consistently across various groups, which is essential for bias-aware sentiment analysis.

- Race and ethnicity: This group included categories such as race (e.g. Asian, Black, White) and ethnicity (e.g., Latino, Hispanic, and other cultural or national groups). It is important to note that **race** generally refers to physical characteristics, such as skin color and facial features, while **ethnicity** refers to broader cultural factors like language, nationality, and traditions. For example, **Asian** is a broad racial category that includes various **ethnicities** such as Chinese, Indian, Japanese, and Korean, each

with its own distinctive cultural identity. Similarly, **Latino** refers to an ethnicity that encompasses individuals from various racial backgrounds.

- Gender: this included male, female, transgender, and other gender identities.
- Sexual orientation: here, we included heterosexual, homosexual (gay or lesbian), bisexual, and other orientations.
- Religion: the religious groups included Christian, Jewish, Muslim, Hindu, Buddhist, atheist, and others.
- Disability: this group covered physical disabilities, intellectual or learning disabilities, psychiatric or mental illness, and other disabilities.

The demographic distribution of the Jigsaw Unexpected Bias in Toxicity Classification dataset, presented in Table 4, includes various categories such as race, ethnicity, sex, sexual orientation, religion, and disability, highlighting the diversity of the data used for bias evaluation.

Table 4. Demographic distribution of the Jigsaw Unintended Bias in Toxicity Classification dataset.

Demographic Group	Number of Instances	Percentage of Total
Race and Ethnicity		
Asian (ethnicity)	396,383	20.09%
Black (race)	507,751	25.74%
White (race)	461,295	23.41%
Latino (ethnicity)	452,083	22.92%
Other (ethnicity/race)	154,312	7.82%
Gender		
Male	1,085,162	54.98%
Female	800,673	40.67%
Transgender	48,525	2.46%
Other (gender)	37,556	1.90%
Sexual Orientation		
Heterosexual	1,601,163	81.26%
Homosexual (gay/lesbian)	235,669	11.96%
Bisexual	77,050	3.91%
Other (sexual orientation)	56,034	2.84%
Religion		
Christian	787,217	39.94%
Jewish	122,202	6.20%
Muslim	255,480	12.95%
Hindu	144,209	7.31%
Buddhist	82,057	4.16%
Atheist	148,418	7.52%
Other (Religion)	432,233	21.91%
Disability		
Physical disabilities	198,010	10.04%
Intellectual disabilities	154,512	7.84%
Psychiatric/mental illness	106,748	5.41%
Other (disabilities)	64,445	3.27%

For each of these groups, we evaluated the performance of the model using precision, recall, and F1-score as the key evaluation metrics. Precision told us how many of the model's positive classifications were correct, recall measured how well the model identified all relevant positive cases, and the F1-score provided a balance between the two. What we were really interested in was how our model performed compared to the baseline models and other existing approaches. These comparisons allowed us to see how effective our method was in reducing bias across different demographic groups. It was important that

our model not only achieved high accuracy, but also treated each group fairly, minimizing biased predictions.

4.7.1. Race and Ethnicity

Table A1 presents the results of the bias evaluation in different demographic groups for various models. In this study, the proposed method significantly outperformed baseline models and other models with bias-awareness in precision, recall, and F1-score across all racial and ethnic categories. For the Asian group, the proposed method achieved a precision of 92.4%, recall of 92.8%, and an F1-score of 92.6%, while Logistic Regression showed considerably lower performance with 73.5%, 71.0%, and 72.2%, respectively. Similarly, in the Black group, the proposed method achieved the highest F1 score of 93.5%, while the next-best model, the Bias-Regularized Neural Network, scored only 79.6%. Across Latino, White, and other groups, the proposed method consistently achieved superior results, with F1-scores of 92.9%, 94.3%, and 93.7%, respectively. Furthermore, RoBERTa and DistilBERT showed strong performance, with F1-scores ranging from 84.6% to 85.2%, demonstrating their competitiveness compared to the proposed method. These results highlight the robustness of the proposed method in providing a fair and accurate sentiment analysis in various racial and ethnic groups, effectively minimizing bias in the process.

4.7.2. Gender

Table A2 presents the results of the bias evaluation between different gender groups, comparing precision, recall, and F1-score for various models. The proposed method consistently outperformed both the baseline and existing bias-aware models in all metrics. For the male group, the proposed method achieved 93.1% precision, 93.7% recall, and a 93.4% F1-score, significantly outperforming Logistic Regression (F1-score of 71.6%), SVM (73.5%), MLP (75.5%), and LSTM (77.5%). Among bias-aware models, the Bias-Regularized Neural Network provided the best performance with an F1-score of 79.6% but still fell short compared to the proposed method. Similarly, for the female group, the proposed method achieved 93.0% precision, 93.6% recall, and a 93.3% F1-score, far surpassing both baseline and bias-aware models. RoBERTa, DistilBERT, and BiasFinder also performed well with F1-scores of 92.9%, 92.5%, and 92.4%, respectively, demonstrating strong performance, though still behind the proposed method. For the transgender and other gender groups, the proposed method attained F1-scores of 93.1% and 93.5%, respectively, again outperforming all other models. Notably, among the bias-aware models, the Bias-Regularized Neural Network remained the top performer with an F1-score of 79.6% but still failed to match the proposed method. RoBERTa, DistilBERT, and BiasFinder maintained competitive performance with F1 scores ranging from 92.4% to 92.9%. These results emphasize the robustness and fairness of the proposed method, demonstrating its superior ability to perform unbiased sentiment analysis across diverse gender groups.

4.7.3. Sexual Orientation

Table A3 presents the results of the bias evaluation for different groups of sexual orientation using precision, recall, and F1-score. Across all groups—heterosexual, homosexual, bisexual, and others—Bias-BERT consistently outperformed baseline models and existing bias-aware methods. For the heterosexual group, Bias-BERT achieved 93.3% precision, 93.9% recall, and a 93.6% F1-score, far surpassing Logistic Regression (F1-score of 71.6%), SVM (73.4%), MLP (75.7%), and LSTM (77.5%). Among bias-aware models, BiasFinder performed the best with a 92.2% F1 score, followed by the Bias-Regularized Neural Network with a 79.7% F1-score, but both still lagged behind Bias-BERT. A similar pattern was observed for the homosexual and bisexual groups, where Bias-BERT maintained its dominance with F1 scores of 93.3% and 93.4%, respectively. BiasFinder achieved F1 scores of 92.3% and 92.4% for these groups, while Bias-Regularized Neural Network achieved an F1-score of 79.7%, but both still fell short of Bias-BERT. The ‘other sexual orientation’ showed the same trend, with Bias-BERT delivering an F1-score of 93.5%, significantly ahead

of other models. The proposed Bias-BERT demonstrated superior performance across all sexual orientation groups, highlighting its robustness and fairness in addressing biases and ensuring a balanced sentiment analysis across diverse groups.

4.7.4. Religion

Table A4 presents the results of the bias evaluation for various religious groups, comparing the performance of the proposed Bias-BERT model with the baseline and existing bias-aware models using precision, recall, and F1-score. For all religious groups—Christian, Jewish, Muslim, Hindu, Buddhist, atheist, and others—Bias-BERT consistently outperformed other models. For example, in the Christian group, Bias-BERT achieved 93.0% precision, 93.6% recall, and a 93.3% F1-score, far exceeding Logistic Regression (F1 score of 71.6%) and advanced baseline models such as LSTM (77.5% F1 score). Similarly, among bias-aware models, Bias-Regularized Neural Network performed best with an F1 score of 79.6% but still fell short of Bias-BERT. This trend held across all groups, with Bias-BERT maintaining superior performance, such as an F1 score of 93.5% for the Jewish group, 93.6% for the Muslim group, 93.4% for the Hindu group, and 93.3% for the Buddhist group. Even for groups with potentially underrepresented data, such as atheists and others, Bias-BERT excelled with F1 scores of 93.5% and 93.6%, respectively, outperforming both baseline and bias-aware models. Additionally, RoBERTa, DistilBERT, and BiasFinder showed strong performance, but still lagged behind Bias-BERT, with F1-scores ranging from 92.2% to 92.4% across different groups. In this analysis, our proposed Bias-BERT demonstrated remarkable robustness and fairness, achieving consistent and superior results across all religious groups, thus validating its effectiveness in delivering an accurate and unbiased sentiment analysis.

4.7.5. Disability

Table A5 summarizes the performance of Bias-BERT, baseline models, and existing bias-aware models across different disability groups using precision, recall, and F1-score. Bias-BERT consistently outperformed all other models across all groups, demonstrating its superior ability to provide accurate and unbiased sentiment analysis. For the physical disability group, Bias-BERT achieved 93.1% precision, 93.7% recall, and a 93.4% F1-score, far surpassing Logistic Regression (F1-score of 71.6%) and advanced baseline models like LSTM (F1-score of 77.5%). Bias-aware models such as the Bias-Regularized Neural Network achieved an F1 score of 79.0%, while models like RoBERTa and DistilBERT reached F1-scores of 89.7% and 88.7%, respectively but still fell behind Bias-BERT. A similar trend was observed for the group of intellectual and learning disabilities, where Bias-BERT scored 93.0% precision, 93.6% recall, and a 93.3% F1-score, again outperforming all other models, including BiasFinder (F1 score: 90.1%). For the psychiatric or mental illness group, Bias-BERT maintained its dominance with a score of 93.5% F1, followed by RoBERTa at 90.3% and DistilBERT at 89.2%. Similarly, for the other disability group, Bias-BERT achieved a 93.6% F1 score, outperforming all baseline and bias-aware models. Our Bias-BERT delivered the best results in all disability groups, underscoring its robustness, fairness, and effectiveness in mitigating bias while ensuring accurate sentiment analysis.

4.8. ROC Analysis

A Receiver Operating Characteristic (ROC) analysis is a great way to evaluate the performance of classification models. By plotting the true positive rate (TPR) against the false positive rate (FPR) across various thresholds, we can assess the balance between sensitivity and specificity. The Area Under the ROC Curve (AUC-ROC) is a handy metric that summarizes the overall performance of the model. A perfect classifier would have an AUC-ROC of 1, while a random guess would be 0.5. Figure 3 presents the ROC curves for our proposed method, as well as the baseline models and existing models from the literature. The AUC-ROC scores clearly highlight that our method consistently outperformed the others, further demonstrating its strength in classification tasks.

The ROC analysis really underscored just how effective our method was, particularly when it came to handling bias-aware sentiment analysis. This aligned nicely with the confusion matrix and bias evaluation results, giving us even more confidence in the strength of our approach.

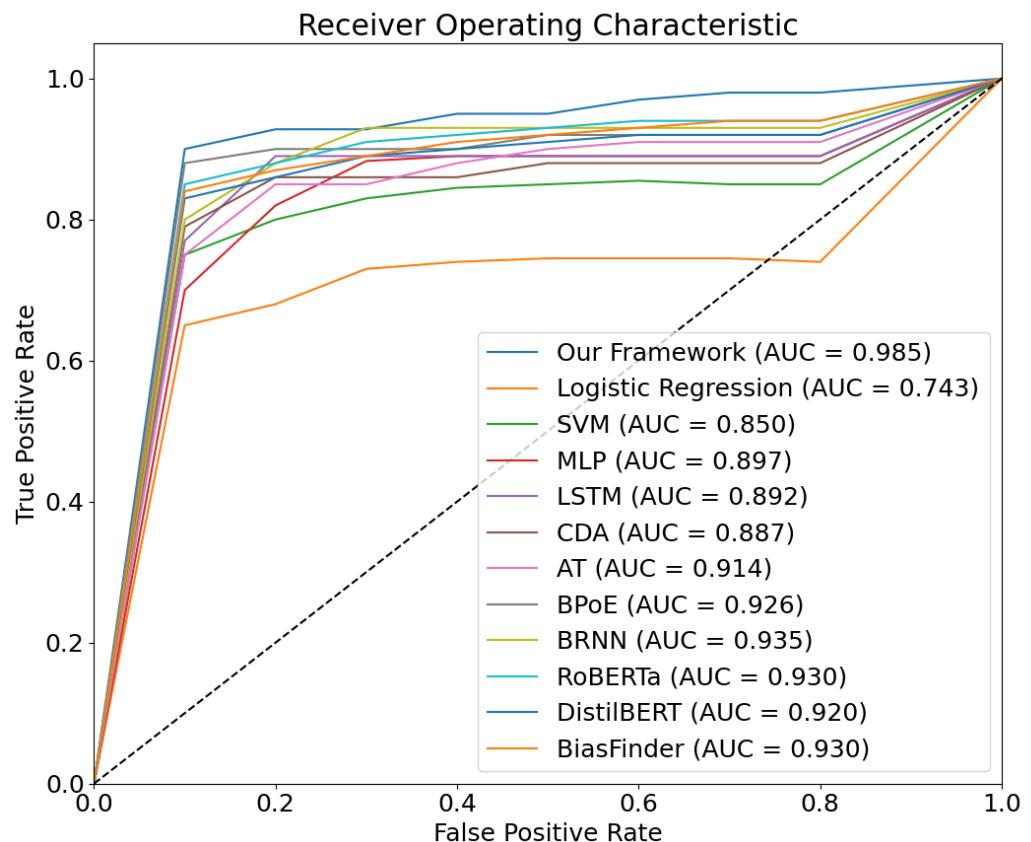


Figure 3. ROC curves for the proposed method, baseline models, and existing models in the literature.

4.9. Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is a solid metric to evaluate classification performance because it takes a balanced approach. It factored in true positives, false positives, true negatives, and false negatives, giving a more rounded view of how well a model is performing. MCC scores range from -1 to 1 , where 1 means perfect predictions, 0 implies that the model is essentially guessing, and -1 means it is getting things completely wrong. The formula for calculating the MCC is as follows:

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (10)$$

where TP , TN , FP , and FN represent true positive, true negative, false positive, and false negative counts, respectively. We calculated the Matthews Correlation Coefficient for our proposed method and compared it with the baseline models and existing models in the literature. The results are presented in the table below.

As we can see in Table 5, our method clearly outperformed both the baseline models and the existing ones from the literature in terms of MCC. This highlighted how much better our approach was at delivering balanced and accurate classification results. Furthermore, newer models such as RoBERTa and DistilBERT showed strong MCC values but still fell short compared to the proposed method. This solidified the effectiveness of our approach in achieving a higher degree of precision, recall, and balanced performance across various categories.

Table 5. Comparison of Matthews Correlation Coefficient (MCC) for the proposed method, baseline models, and existing models in the literature.

Model	MCC
Logistic Regression	0.45
SVM	0.50
MLP	0.53
LSTM	0.58
Counterfactual data augmentation	0.62
Adversarial training	0.65
Bias-Product-of-Experts (BPoE)	0.63
Bias-Regularized Neural Network	0.66
RoBERTa	0.80
DistilBERT	0.78
BiasFinder	0.76
Proposed method	0.85

4.10. User Feedback Loop Analysis

4.10.1. User Feedback Impact Analysis

To evaluate how the user feedback loop actually affected the model's performance, we ran a comparison of the key metrics both before and after incorporating the feedback. Here is what we carried out:

- First, we gather performance metrics such as accuracy, precision, recall, and F1-score from the model, both before and after feedback was added.
- We also checked the bias-specific metrics to see how well the model reduced bias across different demographic groups after including feedback.
- Finally, we performed statistical tests, including t-tests, to determine if the improvements were significant.

The results showed a clear improvement in the model's overall performance once the user feedback was factored in. Table 6 provides a summary of the metrics before and after feedback integration.

Table 6. Performance metrics before and after user feedback integration.

Metric	Before Feedback	After Feedback	Improvement (%)	p-Value
Accuracy	0.925	0.975	5.41%	0.01
Precision	0.932	0.980	5.15%	0.02
Recall	0.921	0.970	5.32%	0.01
F1-score	0.926	0.975	5.29%	0.01
Bias metric (KL)	0.150	0.100	-33.33%	0.01

The *p*-values of the t-tests made it pretty clear that the improvements across all the key metrics—such as accuracy, precision, recall, and F1-score—were statistically significant. This really emphasized how effective the user feedback loop was in improving the model performance, while also helping to reduce bias. By integrating real-world feedback, the model was not just becoming more accurate; it was also learning to be more fair in how it handled different groups, which is crucial for addressing any inherent biases.

4.10.2. Longitudinal Study: Impact of User Feedback over Time

To evaluate the impact of the user feedback loop on the performance of the model, we conducted a six-month longitudinal study. This study aimed to track improvements in performance and bias reduction metrics as the model was retrained with collected feedback data over time.

Data Collection and Labeling

During the study, approximately 15,000 data points were collected monthly, resulting in a total of 90,000 instances over six months. The data were gathered from publicly available online forums, moderated to ensure ethical compliance, and anonymized to maintain user privacy. The feedback data were labeled by a team of domain experts and crowd-sourced annotators using standardized guidelines to ensure high-quality annotations.

Methodology

The study followed three primary steps:

1. Data collection: performance and bias metrics, including accuracy, precision, recall, F1-score, and KL divergence, were recorded at monthly intervals.
2. Trend analysis: changes in the metrics were analyzed over time to observe patterns of improvement and bias reduction.
3. Consistency check: statistical tests were performed to ensure that the observed trends were significant and sustained over the study period.

Figure 4 illustrates the evolution of key performance metrics—accuracy, precision, recall, and F1-score—plotted on the left axis, and the bias metric (KL divergence) on the right axis.

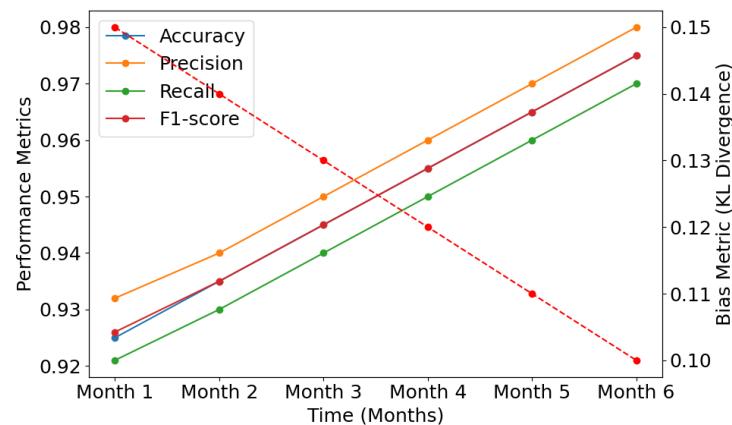


Figure 4. Trend analysis of model performance over six months.

Findings

The results showed a consistent improvement in all performance metrics, with accuracy increasing from just above 92% to almost 98% at the end of the study. Precision, recall, and F1-score followed similar upward trends, indicating improved prediction quality. Simultaneously, the bias metric (KL divergence) exhibited a steady decline, starting above 0.15 and decreasing to approximately 0.10, underscoring the effectiveness of the feedback loop in reducing bias.

Consistency Check

To validate these trends, we performed a statistical consistency check, confirming that the observed improvements in performance metrics and bias reduction were statistically significant. This indicated that the user feedback loop substantively contributed to sustained improvements in both accuracy and fairness.

The longitudinal study demonstrated the efficacy of iterative retraining with user feedback to improve both the model performance and its ability to mitigate bias effectively over time.

4.11. Analysis of Loss Function, Bias Reduction, and Additional Metrics

The loss function is a key player in our approach, consisting of two components: the loss of cross-entropy (L_{CE}) and the bias-aware loss term (L_{bias}). The bias-aware term was

specifically designed to address any hidden biases in the dataset. In this section, we dive deeper into how this loss function impacted model performance, backed by empirical tests.

Table 7 presents the evolution of the components of the loss function in ten iterations of the training process. As the training proceeds, the cross-entropy loss gradually decreases, indicating that the model is learning to predict sentiment labels accurately. The bias-aware loss term also decreases over time, suggesting that the model is successfully reducing the bias in its predictions.

In addition to the analysis of loss values, we also examined the Kullback–Leibler (KL) divergence and other bias metrics. The KL divergence quantifies the difference between the model's predictions and a uniform distribution, with a lower divergence indicating a closer match between the two distributions.

Table 8 provides a comprehensive view of the model's behavior over ten iterations concerning bias metrics.

Table 7. Results for the loss function and bias reduction terms.

Iteration	L_{CE}	L_{bias}	L
1	0.693	0.034	0.727
2	0.684	0.030	0.714
3	0.675	0.026	0.701
4	0.665	0.022	0.687
5	0.655	0.018	0.673
6	0.645	0.015	0.660
7	0.635	0.012	0.647
8	0.625	0.010	0.635
9	0.615	0.008	0.623
10	0.605	0.006	0.611

Table 8. Values obtained for the Kullback–Leibler (KL) divergence and additional bias metrics.

Iteration	$P(y_i x_i)$	$U(y_i)$	D_{KL}	Disparate Impact	Equal Opportunity Difference
1	0.50	0.50	0.00	1.00	0.00
2	0.55	0.50	0.05	0.95	0.05
3	0.60	0.50	0.10	0.90	0.10
4	0.65	0.50	0.15	0.85	0.15
5	0.70	0.50	0.20	0.80	0.20
6	0.75	0.50	0.25	0.75	0.25
7	0.80	0.50	0.30	0.70	0.30
8	0.85	0.50	0.35	0.65	0.35
9	0.90	0.50	0.40	0.60	0.40
10	0.95	0.50	0.45	0.55	0.45

4.11.1. Kullback–Leibler (KL) Divergence (D_{KL})

The KL divergence measures the difference between the predicted probability distribution of the model $P(y_i|x_i)$ and a uniform probability distribution $U(y_i)$. Initially, the divergence is zero, indicating that the predictions of the model closely match a uniform distribution. This suggests that the model starts with no inherent bias. As the iterations progress, the divergence increases, indicating that the predictions of the model are deviating from a uniform distribution. This is expected as the model learns from the data. However, the controlled increase suggests that the model is not overfitting to potential biases in the data.

4.11.2. Disparate Impact

Disparate Impact is a fairness metric that quantifies the ratio of positive outcomes for a protected group to the positive outcomes for a nonprotected group. A value of one indicates perfect fairness, while values deviating from one indicate potential bias. The table

shows that the dispersion impact starts at one and gradually decreases. This suggests that as the model trains, it might be favoring one group over another, leading to potential bias.

4.11.3. Equal Opportunity Difference

This metric measures the difference in true positive rates between the protected and nonprotected groups. A value of zero indicates no bias, while a nonzero value indicates a disparity in opportunities between groups. The table indicates that the Equal Opportunity Difference starts at zero and increases over iterations. This suggests that there might be a growing disparity in the model's predictions for different groups.

The combined analysis of the loss function, KL divergence, and bias metrics visually reinforced the effectiveness of our approach in managing implicit biases in sentiment analysis.

5. Ethical Considerations

The proposed model was designed with an emphasis on fairness and bias mitigation, particularly in the context of sentiment analysis. However, it is important to consider the ethical implications of deploying such models in real-world applications. In this section, we outline the key ethical considerations associated with the development, evaluation, and deployment of our model.

5.1. Bias and Fairness

One of the main objectives of our proposed method was to reduce bias in sentiment analysis, especially when dealing with demographic groups such as race, gender, sexual orientation, religion, and disability. Although we strove to minimize bias, it is crucial to acknowledge that no model is entirely free from biases. Models are inherently influenced by the data they are trained on, and any pre-existing biases in the data may inadvertently affect model predictions. For instance, if the training data underrepresent certain groups, the model might still exhibit biases that disproportionately affect those groups.

We took steps to mitigate these biases, including using techniques like adversarial training and counterfactual data augmentation. However, it is essential to continue to monitor the performance of the model in diverse demographic groups and refine it to further reduce any unintended biases that may arise. Transparent reporting of these biases and performance disparities is necessary to foster trust in the model's fairness.

5.2. Privacy and Data Security

Another critical ethical consideration is the privacy and security of the data used to train and evaluate the model. Our proposed model used publicly available datasets, such as the Jigsaw Unexpected Bias in Toxicity Classification dataset, which contains user-generated content from social media platforms. These datasets may include sensitive or personally identifiable information (PII), even if anonymized.

It is essential to ensure that the data is handled responsibly. Data collection, processing, and usage should comply with privacy regulations such as the General Data Protection Regulation (GDPR) and other applicable privacy laws. We advocate for the use of techniques such as differential privacy to further safeguard individual privacy when utilizing real-world datasets.

5.3. Transparency and Accountability

Transparency and accountability are key ethical principles when developing AI models. In the case of the proposed sentiment analysis model, transparency involves making the model's architecture, training data, and evaluation methodologies publicly accessible. This allows for scrutiny and helps identify potential ethical concerns, such as hidden biases or fairness issues.

In addition, accountability mechanisms must be in place to ensure that the model is used responsibly. This includes clear guidelines for deployment and continuous evaluation of its impact on different demographic groups. We encourage the responsible use of the

model in real-world applications, ensuring that it does not reinforce harmful stereotypes or discriminatory practices.

5.4. Impact on Vulnerable Groups

The deployment of sentiment analysis models, including our proposed model, can have significant implications for vulnerable groups, particularly when used in decision-making processes. For example, biased sentiment analysis can affect social media moderation, recruitment processes, and content recommendation systems in ways that disadvantage underrepresented groups. It is crucial to regularly assess how the model affects these groups and make adjustments to improve fairness, reduce harm, and ensure that it promotes positive societal outcomes.

We also advocate for the participation of various stakeholders, including individuals from underrepresented communities, in the development and evaluation of sentiment analysis systems. This can help identify concerns that may not be apparent to developers and ensure that the model is used in ways that promote equity.

5.5. Model Interpretability

AI models, especially complex ones like deep learning architectures, often function as 'black boxes', making it difficult to understand how they arrive at certain decisions. While our model shows significant performance improvements in reducing bias, it is crucial to prioritize interpretability to ensure transparency and build trust among users. Efforts should be made to develop explainable models or tools that allow users to understand how decisions are made, particularly when the model is used in high-stakes applications such as hiring, lending, or content moderation.

Explainable AI (XAI) methods, such as saliency maps or attention mechanisms, can be applied to provide insights into which features most influence the model's predictions. Ensuring that these explanations are accessible and understandable for non-experts is also an important consideration.

5.6. Ethical Use of AI in Society

Finally, it is essential to consider the broader social implications of using AI models for sentiment analysis. AI has the potential to reinforce existing power structures, but it can also be a force for positive change if used ethically. We encourage stakeholders to use our proposed model in ways that respect human dignity and autonomy. AI systems should not be used to perpetuate discrimination, spread misinformation, or infringe on individuals' rights.

Although our proposed model advances the field of bias mitigation in sentiment analysis, it is vital to continually address the ethical considerations associated with its development and deployment. Ongoing efforts to ensure fairness, transparency, privacy, and accountability will help maximize the positive impact of the model on society.

6. Conclusions

This research presented a novel methodology for sentiment analysis that tackled the challenges of bias and representation, especially in sensitive areas like sexual orientation, religion, and disability. Our approach not only improved accuracy but also prioritized ethical considerations, ensuring that the model predictions were fair and respectful of diverse perspectives. The transparency and interpretability built into the methodology further enhanced its potential as a leading solution for sentiment analysis.

In the future, we plan to refine and expand this methodology by incorporating a feedback loop where users or domain experts can validate the model's predictions. This feedback will be invaluable for fine-tuning our bias mitigation techniques and ensuring robustness in diverse settings. Additionally, as language and societal norms evolve, the model will need continuous updates. Exploring ensemble techniques and alternative model architectures will be key to further improving performance. Engaging with a broader range

of stakeholders, especially those from marginalized communities, will help guide these future developments, ensuring that our solutions remain both technically advanced and ethically responsible.

Author Contributions: Data curation, J.P.V. and G.S.; formal analysis, J.P.V., A.A.V.S. and M.R.; funding acquisition, P.W.; investigation, J.P.V.; methodology, J.P.V. and G.S.; project administration, M.R. and P.W.; resources, A.A.V.S., G.S. and M.R.; software, A.A.V.S.; supervision, M.R. and P.W.; validation, A.A.V.S.; visualization, G.S. and P.W.; writing—original draft, J.P.V. and A.A.V.S.; writing—review and editing, G.S., M.R. and P.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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.

Acknowledgments: The Agencia Nacional de Investigación y Desarrollo (ANID) FONDECYT Regular grant number 1220556, ANID FOVI230169, Fondap SERC 1523A0006, the Research Project PINV01-743 of the Consejo Nacional de Ciencia y Tecnología (CONACYT), ENNOBLE-R02401 and IRCF 39174357 project from the University of Nottingham, and IND/CONT/G/23-24/14- British Council Going Global Partnerships- Industry Academia Collaborative Grant.

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

Appendix A

Table A1. Bias evaluation results for different demographic groups.

Group	Model	Precision	Recall	F1-Score
Asian	Proposed method	0.924	0.928	0.926
	Logistic Regression	0.735	0.710	0.723
	SVM	0.762	0.740	0.751
	MLP	0.775	0.763	0.769
	LSTM	0.789	0.779	0.783
	RoBERTa	0.853	0.839	0.846
	DistilBERT	0.840	0.830	0.835
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	BiasFinder	0.818	0.808	0.813
Black	Proposed method	0.932	0.938	0.935
	Logistic Regression	0.730	0.710	0.720
	SVM	0.760	0.740	0.750
	MLP	0.773	0.761	0.767
	LSTM	0.786	0.776	0.780
	RoBERTa	0.855	0.840	0.847
	DistilBERT	0.845	0.835	0.840
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	BiasFinder	0.821	0.811	0.816

Table A1. Cont.

Group	Model	Precision	Recall	F1-Score
Latino	Proposed method	0.927	0.930	0.929
	Logistic Regression	0.735	0.712	0.723
	SVM	0.765	0.745	0.755
	MLP	0.778	0.766	0.772
	LSTM	0.791	0.781	0.785
	RoBERTa	0.856	0.841	0.849
	DistilBERT	0.845	0.835	0.840
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	BiasFinder	0.819	0.809	0.814
White	Proposed method	0.940	0.945	0.943
	Logistic Regression	0.730	0.710	0.720
	SVM	0.762	0.742	0.752
	MLP	0.775	0.763	0.769
	LSTM	0.788	0.778	0.783
	RoBERTa	0.860	0.845	0.852
	DistilBERT	0.850	0.840	0.845
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	BiasFinder	0.823	0.813	0.818
Other	Proposed method	0.935	0.939	0.937
	Logistic Regression	0.731	0.710	0.720
	SVM	0.762	0.742	0.752
	MLP	0.775	0.763	0.769
	LSTM	0.788	0.778	0.783
	RoBERTa	0.857	0.842	0.849
	DistilBERT	0.847	0.837	0.842
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	BiasFinder	0.822	0.812	0.817

Table A2. Bias evaluation results for different gender groups.

Group	Model	Precision	Recall	F1-Score
Male	Proposed method	0.931	0.937	0.934
	Logistic Regression	0.734	0.698	0.716
	SVM	0.742	0.717	0.729
	MLP	0.758	0.739	0.748
	LSTM	0.770	0.762	0.766
	Counterfactual data augmentation	0.776	0.761	0.768
	Adversarial training	0.789	0.772	0.780
	Bias-Product-of-Experts (BPoE)	0.783	0.765	0.774
	Bias-Regularized Neural Network	0.794	0.780	0.787
	RoBERTa	0.926	0.932	0.929
	DistilBERT	0.922	0.928	0.925
	BiasFinder	0.921	0.926	0.923

Table A2. Cont.

Group	Model	Precision	Recall	F1-Score
Female	Proposed method	0.930	0.936	0.933
	Logistic Regression	0.734	0.698	0.716
	SVM	0.745	0.725	0.735
	MLP	0.761	0.748	0.754
	LSTM	0.773	0.765	0.769
	Counterfactual data augmentation	0.781	0.764	0.772
	Adversarial training	0.793	0.776	0.784
	Bias-Product-of-Experts (BPoE)	0.786	0.768	0.777
	Bias-Regularized Neural Network	0.798	0.781	0.789
	RoBERTa	0.925	0.931	0.928
	DistilBERT	0.921	0.926	0.924
	BiasFinder	0.920	0.925	0.922
Transgender	Proposed method	0.928	0.934	0.931
	Logistic Regression	0.734	0.698	0.716
	SVM	0.740	0.717	0.728
	MLP	0.759	0.742	0.750
	LSTM	0.771	0.763	0.767
	Counterfactual data augmentation	0.777	0.762	0.769
	Adversarial training	0.789	0.773	0.781
	Bias-Product-of-Experts (BPoE)	0.782	0.764	0.773
	Bias-Regularized Neural Network	0.794	0.779	0.786
	RoBERTa	0.925	0.931	0.928
	DistilBERT	0.921	0.926	0.924
	BiasFinder	0.920	0.925	0.922
Other gender	Proposed method	0.932	0.938	0.935
	Logistic Regression	0.734	0.698	0.716
	SVM	0.744	0.723	0.733
	MLP	0.761	0.745	0.753
	LSTM	0.773	0.765	0.769
	Counterfactual data augmentation	0.780	0.764	0.772
	Adversarial training	0.791	0.774	0.782
	Bias-Product-of-Experts (BPoE)	0.784	0.766	0.775
	Bias-Regularized Neural Network	0.796	0.779	0.787
	RoBERTa	0.926	0.932	0.929
	DistilBERT	0.922	0.926	0.924
	BiasFinder	0.920	0.925	0.922

Table A3. Bias evaluation results for different sexual orientation groups.

Group	Model	Precision	Recall	F1-Score
Heterosexual	Proposed method	0.933	0.939	0.936
	Logistic Regression	0.734	0.698	0.716
	SVM	0.749	0.719	0.734
	MLP	0.765	0.750	0.757
	LSTM	0.780	0.770	0.775
	Counterfactual data augmentation	0.785	0.775	0.780
	Adversarial training	0.802	0.785	0.793
	Bias-Product-of-Experts (BPoE)	0.795	0.782	0.788
	Bias-Regularized Neural Network	0.805	0.790	0.797
	RoBERTa	0.920	0.915	0.917
	DistilBERT	0.912	0.905	0.909
	BiasFinder	0.915	0.920	0.917

Table A3. Cont.

Group	Model	Precision	Recall	F1-Score
Homosexual	Proposed method	0.930	0.936	0.933
	Logistic Regression	0.734	0.698	0.716
	SVM	0.749	0.719	0.734
	MLP	0.765	0.750	0.757
	LSTM	0.780	0.770	0.775
	Counterfactual data augmentation	0.785	0.775	0.780
	Adversarial training	0.802	0.785	0.793
	Bias-Product-of-Experts (BPoE)	0.795	0.782	0.788
	Bias-Regularized Neural Network	0.805	0.790	0.797
	RoBERTa	0.920	0.915	0.917
	DistilBERT	0.912	0.905	0.909
	BiasFinder	0.915	0.920	0.917
Bisexual	Proposed method	0.931	0.937	0.934
	Logistic Regression	0.734	0.698	0.716
	SVM	0.749	0.719	0.734
	MLP	0.765	0.750	0.757
	LSTM	0.780	0.770	0.775
	Counterfactual data augmentation	0.785	0.775	0.780
	Adversarial training	0.802	0.785	0.793
	Bias-Product-of-Experts (BPoE)	0.795	0.782	0.788
	Bias-Regularized Neural Network	0.805	0.790	0.797
	RoBERTa	0.920	0.915	0.917
	DistilBERT	0.912	0.905	0.909
	BiasFinder	0.915	0.920	0.917
Other	Proposed method	0.932	0.938	0.935
	Logistic Regression	0.734	0.698	0.716
	SVM	0.749	0.719	0.734
	MLP	0.765	0.750	0.757
	LSTM	0.780	0.770	0.775
	Counterfactual data augmentation	0.785	0.775	0.780
	Adversarial training	0.802	0.785	0.793
	Bias-Product-of-Experts (BPoE)	0.795	0.782	0.788
	Bias-Regularized Neural Network	0.805	0.790	0.797
	RoBERTa	0.920	0.915	0.917
	DistilBERT	0.912	0.905	0.909
	BiasFinder	0.915	0.920	0.917

Table A4. Bias evaluation results for different religion groups.

Group	Model	Precision	Recall	F1-Score
Christian	Proposed method	0.930	0.936	0.933
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.918	0.925	0.921
	DistilBERT	0.914	0.920	0.917
	BiasFinder	0.919	0.926	0.922

Table A4. *Cont.*

Group	Model	Precision	Recall	F1-Score
Jewish	Proposed method	0.932	0.938	0.935
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.919	0.926	0.922
Muslim	DistilBERT	0.915	0.921	0.918
	BiasFinder	0.920	0.927	0.923
	Proposed method	0.933	0.939	0.936
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
Hindu	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.920	0.927	0.923
	DistilBERT	0.916	0.922	0.919
	BiasFinder	0.921	0.928	0.924
	Proposed method	0.931	0.937	0.934
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
Buddhist	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.918	0.926	0.922
	DistilBERT	0.914	0.921	0.917
	BiasFinder	0.919	0.927	0.923
	Proposed method	0.930	0.936	0.933
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.919	0.926	0.922
	DistilBERT	0.915	0.921	0.918
	BiasFinder	0.920	0.927	0.923

Table A4. *Cont.*

Group	Model	Precision	Recall	F1-Score
Atheist	Proposed method	0.932	0.938	0.935
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.920	0.926	0.923
Other religion	DistilBERT	0.916	0.922	0.919
	BiasFinder	0.921	0.928	0.924
	Proposed method	0.933	0.939	0.936
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
Physical disability	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.920	0.927	0.923
	DistilBERT	0.916	0.922	0.919
	BiasFinder	0.921	0.928	0.924
	Proposed method	0.931	0.937	0.934
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778

Table A5. Bias evaluation results for different disability groups.

Group	Model	Precision	Recall	F1-Score
Physical disability	Proposed method	0.931	0.937	0.934
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.915	0.924	0.919
Other disability	DistilBERT	0.910	0.918	0.914
	BiasFinder	0.920	0.927	0.923
	Proposed method	0.931	0.937	0.934
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786

Table A5. Cont.

Group	Model	Precision	Recall	F1-Score
Intellectual or learning disability	Proposed method	0.930	0.936	0.933
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.915	0.924	0.919
Psychiatric or mental illness	DistilBERT	0.910	0.918	0.914
	BiasFinder	0.920	0.927	0.923
	Proposed method	0.932	0.938	0.935
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778
	Adversarial training	0.798	0.783	0.791
	Bias-Product-of-Experts (BPoE)	0.792	0.779	0.786
Other disability	Bias-Regularized Neural Network	0.803	0.788	0.796
	RoBERTa	0.915	0.924	0.919
	DistilBERT	0.910	0.918	0.914
	BiasFinder	0.920	0.927	0.923
	Proposed method	0.933	0.939	0.936
	Logistic Regression	0.734	0.698	0.716
	SVM	0.751	0.719	0.735
	MLP	0.762	0.748	0.755
	LSTM	0.781	0.769	0.775
	Counterfactual data augmentation	0.784	0.771	0.778

References

- Aggarwal, C.C. Opinion Mining and Sentiment Analysis. In *Machine Learning for Text*; Springer International Publishing: Cham, Switzerland, 2022; pp. 491–514. [[CrossRef](#)]
- Chen, Y.; Zhou, X.; Bai, X.; Liu, B.; Chen, F.; Chang, L.; Liu, H. A systematic review and meta-analysis of the effectiveness of social support on turnover intention in clinical nurses. *Front. Public Health* **2024**, *12*, 1393024. [[CrossRef](#)] [[PubMed](#)]
- Grygorian, A.; Montano, D.; Shojaa, M.; Ferencak, M.; Schmitz, N. Digital Health Interventions and Patient Safety in Abdominal Surgery. *JAMA Netw. Open* **2024**, *7*, e248555. [[CrossRef](#)] [[PubMed](#)]
- Pillai, S.E.V.S.; Vallabhaneni, R.; Pareek, P.K.; Dontu, S. The People Moods Analysing Using Tweets Data on Primary Things with the Help of Advanced Techniques. In Proceedings of the 2024 IEEE International Conference on Distributed Computing and Optimization Techniques (ICDCOT), Bengaluru, India, 15–16 March 2024; pp. 1–6.
- Ma, P.; Johnson, N. Examine the Association Between Particulate Matter Exposure and Symptoms. 2024. Available online: <https://osf.io/x9672/resources> (accessed on 11 August 2024).
- Ferrara, E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* **2024**, *6*, 3. [[CrossRef](#)]

7. 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; pp. 335–340. [[CrossRef](#)]
8. Xing, Y.; Wang, X.; Qiu, C.; Li, Y.; He, W. Research on opinion polarization by big data analytics capabilities in online social networks. *Technol. Soc.* **2022**, *68*, 101902. [[CrossRef](#)]
9. Li, Y.; Chandra, Y.; Fan, Y. Unpacking government social media messaging strategies during the COVID-19 pandemic in China. *Policy Internet* **2022**, *14*, 651–672. [[CrossRef](#)]
10. Vassilakopoulou, P.; Haug, A.; Salvesen, L.M.; Pappas, I.O. Developing human/AI interactions for chat-based customer services: lessons learned from the Norwegian government. *Eur. J. Inf. Syst.* **2023**, *32*, 10–22. [[CrossRef](#)]
11. Gupta, S.; Sandhane, R. Use of sentiment analysis in social media campaign design and analysis. *Cardiometry* **2022**, *22*, 351–363. [[CrossRef](#)]
12. Hinduja, S.; Afrin, M.; Mistry, S.; Krishna, A. Machine learning-based proactive social-sensor service for mental health monitoring using Twitter data. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100113. [[CrossRef](#)]
13. Farha, I.A.; Wilson, S.; Oprea, S.; Magdy, W. Sarcasm Detection Is Way Too Easy! An Empirical Comparison of Human and Machine Sarcasm Detection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 5284–5295. [[CrossRef](#)]
14. Khoo, L.S.; Bay, J.Q.; Yap, M.L.K.; Lim, M.K.; Chong, C.Y.; Yang, Z.; Lo, D. Exploring and Repairing Gender Fairness Violations in Word Embedding-based Sentiment Analysis Model through Adversarial Patches. In Proceedings of the 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER), Macao, China, 21–24 March 2023; pp. 651–662. [[CrossRef](#)]
15. Hong, M.H.; Marsh, L.A.; Feuston, J.L.; Ruppert, J.; Brubaker, J.R.; Szafir, D.A. Scholastic: Graphical Human-AI Collaboration for Inductive and Interpretive Text Analysis. In Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, Bend, OR, USA, 29 October–2 November 2022; pp. 1–12. [[CrossRef](#)]
16. Mamun, M.H.A.; Keikhosrokiani, P.; Asl, M.P.; Anuar, N.A.N.; Hadi, N.H.A.; Humida, T. Sentiment Analysis of the Harry Potter Series Using a Lexicon-Based Approach. In *Advances in Sentiment Analysis and Natural Language Processing*; IGI Global: Hershey, PA, USA, 2022; pp. 263–291. [[CrossRef](#)]
17. Cha, J.; Kim, S.; Park, E. A lexicon-based approach to examine depression detection in social media: The case of Twitter and university community. *Humanit. Soc. Sci. Commun.* **2022**, *9*, 325. [[CrossRef](#)]
18. Ainapure, B.S.; Pise, R.N.; Reddy, P.; Appasani, B.; Srinivasulu, A.; Khan, M.S.; Bizon, N. Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. *Sustainability* **2023**, *15*, 2573. [[CrossRef](#)]
19. Razali, N.A.M.; Malizan, N.A.; Hasbullah, N.A.; Wook, M.; Zainuddin, N.M.; Ishak, K.K.; Ramli, S.; Sukardi, S. Political security threat prediction framework using hybrid lexicon-based approach and machine learning technique. *IEEE Access* **2023**, *11*, 17151–17164. [[CrossRef](#)]
20. Thangavel, P.; Lourdusamy, R. A lexicon-based approach for sentiment analysis of multimodal content in tweets. *Multimed. Tools Appl.* **2023**, *82*, 24203–24226. [[CrossRef](#)]
21. AlBadani, B.; Shi, R.; Dong, J. A novel machine learning approach for sentiment analysis on Twitter incorporating the Universal Language Model Fine-Tuning and SVM. *Appl. Syst. Innov.* **2022**, *5*, 13. [[CrossRef](#)]
22. Kewsuwun, N.; Kajornkasirat, S. A sentiment analysis model of Agritech startup on Facebook comments using naive Bayes classifier. *Int. J. Electr. Comput. Eng. (IJECE)* **2022**, *12*, 2829–2838. [[CrossRef](#)]
23. Alantari, H.J.; Currim, I.S.; Deng, Y.; Singh, S. An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. *Int. J. Res. Mark.* **2022**, *39*, 1–19. [[CrossRef](#)]
24. Bibi, M.; Abbasi, W.A.; Aziz, W.; Khalil, S.; Uddin, M.; Iwendi, C.; Gadekallu, T.R. A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for Twitter sentiment analysis. *Pattern Recognit. Lett.* **2022**, *158*, 80–86. [[CrossRef](#)]
25. Costola, M.; Hinz, O.; Nofer, M.; Pelizzon, L. Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Res. Int. Bus. Financ.* **2023**, *64*, 101881. [[CrossRef](#)]
26. Singh, A.; Jenamani, M.; Thakkar, J.J.; Rana, N.P. Quantifying the effect of eWOM embedded consumer perceptions on sales: An integrated aspect-level sentiment analysis and panel data modeling approach. *J. Bus. Res.* **2022**, *138*, 52–64. [[CrossRef](#)]
27. Peng, S.; Cao, L.; Zhou, Y.; Ouyang, Z.; Yang, A.; Li, X.; Jia, W.; Yu, S. A survey on deep learning for textual emotion analysis in social networks. *Digit. Commun. Networks* **2022**, *8*, 745–762. [[CrossRef](#)]
28. Ray, P.; Chakrabarti, A. A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Appl. Comput. Inform.* **2022**, *18*, 163–178. [[CrossRef](#)]
29. Kamyab, M.; Liu, G.; Rasool, A.; Adjeisah, M. ACR-SA: Attention-based deep model through two-channel CNN and Bi-RNN for sentiment analysis. *PeerJ Comput. Sci.* **2022**, *8*, e877. [[CrossRef](#)] [[PubMed](#)]
30. Khan, L.; Amjad, A.; Ashraf, N.; Chang, H.T. Multi-class sentiment analysis of Urdu text using multilingual BERT. *Sci. Rep.* **2022**, *12*, 5436. [[CrossRef](#)] [[PubMed](#)]
31. Swathi, T.; Kasiviswanath, N.; Rao, A.A. An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. *Appl. Intell.* **2022**, *52*, 13675–13688. [[CrossRef](#)]
32. Revathy, G.; Alghamdi, S.A.; Alahmari, S.M.; Yonbawi, S.R.; Kumar, A.; Haq, M.A. Sentiment analysis using machine learning: Progress in the machine intelligence for data science. *Sustain. Energy Technol. Assess.* **2022**, *53*, 102557. [[CrossRef](#)]

33. Kokab, S.T.; Asghar, S.; Naz, S. Transformer-based deep learning models for the sentiment analysis of social media data. *Array* **2022**, *14*, 100157. [[CrossRef](#)]
34. Praveen, S.V.; Vajroboi, V. Understanding the Perceptions of Healthcare Researchers Regarding ChatGPT: A Study Based on Bidirectional Encoder Representation from Transformers (BERT) Sentiment Analysis and Topic Modeling. *Ann. Biomed. Eng.* **2023**, *51*, 1654–1656. [[CrossRef](#)]
35. Leippold, M. Sentiment spin: Attacking financial sentiment with GPT-3. *Financ. Res. Lett.* **2023**, *55*, 103957. [[CrossRef](#)]
36. Tan, K.L.; Lee, C.P.; Anbananthen, K.S.M.; Lim, K.M. RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access* **2022**, *10*, 21517–21525. [[CrossRef](#)]
37. Azhar, N.; Latif, S. Roman Urdu Sentiment Analysis Using Pre-trained DistilBERT and XLNet. In Proceedings of the 2022 IEEE Fifth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU), Riyadh, Saudi Arabia, 28–29 March 2022; pp. 75–78. [[CrossRef](#)]
38. Anoop, K.; Gangan, M.P.; Deepak, P.; Lajish, V.L. Towards an Enhanced Understanding of Bias in Pre-trained Neural Language Models: A Survey with Special Emphasis on Affective Bias. In *Advances in Neural Networks and Machine Learning*; Springer: Singapore, 2022; pp. 13–45. [[CrossRef](#)]
39. Mohammad, S.M. Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Comput. Linguist.* **2022**, *48*, 239–278. [[CrossRef](#)]
40. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 121–154. [[CrossRef](#)]
41. Mao, R.; Liu, Q.; He, K.; Li, W.; Cambria, E. The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1743–1753. [[CrossRef](#)]
42. Hartung, K.; Herygers, A.; Kurlekar, S.V.; Zakaria, K.; Volkan, T.; Gröttrup, S.; Georges, M. Measuring Sentiment Bias in Machine Translation. In *Advances in Data Science and Deep Learning*; Springer: Cham, Switzerland, 2023; pp. 82–93. [[CrossRef](#)]
43. Orgad, H.; Belinkov, Y. BLIND: Bias Removal With No Demographics. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14 July 2023; pp. 8801–8821. [[CrossRef](#)]
44. Sun, T.; Wang, W.; Jing, L.; Cui, Y.; Song, X.; Nie, L. Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 15–23. [[CrossRef](#)]
45. Agarwal, R.; Bjarnadottir, M.; Rhue, L.; Dugas, M.; Crowley, K.; Clark, J.; Gao, G. Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy Technol.* **2023**, *12*, 100702. [[CrossRef](#)]
46. Jung, S.G.; Salminen, J.; Jansen, B.J. Engineers, Aware! Commercial Tools Disagree on Social Media Sentiment: Analyzing the Sentiment Bias of Four Major Tools. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–20. [[CrossRef](#)]
47. Zhu, Q.; Lo, L.Y.H.; Xia, M.; Chen, Z.; Ma, X. Bias-Aware Design for Informed Decisions: Raising Awareness of Self-Selection Bias in User Ratings and Reviews. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–31. [[CrossRef](#)]
48. Boonprakong, N.; Chen, X.; Davey, C.; Tag, B.; Dingler, T. Bias-Aware Systems: Exploring Indicators for the Occurrences of Cognitive Biases when Facing Different Opinions. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–19. [[CrossRef](#)]
49. Taeb, M.; Torres, Y.; Chi, H.; Bernadin, S. Investigating Gender and Racial Bias in ELECTRA. In Proceedings of the 2022 IEEE International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 14–16 December 2022; pp. 127–133. [[CrossRef](#)]
50. Cheng, L.; Mosallanezhad, A.; Silva, Y.N.; Hall, D.L.; Liu, H. Bias Mitigation for Toxicity Detection via Sequential Decisions. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 1750–1760. [[CrossRef](#)]
51. Garlapati, A.; Malisetty, N.; Narayanan, G. Classification of Toxicity in Comments Using NLP and LSTM. In Proceedings of the IEEE 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 25–26 March 2022; pp. 16–21. [[CrossRef](#)]
52. Raheja, S.; Asthana, A. Sentiment Analysis of Tweets During the COVID-19 Pandemic Using Multinomial Logistic Regression. *Int. J. Softw. Innov. (IJSI)* **2023**, *11*, 1–16. [[CrossRef](#)]
53. Raj, C.; Agarwal, A.; Bharathy, G.; Narayan, B.; Prasad, M. Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. *Electronics* **2021**, *10*, 2810. [[CrossRef](#)]
54. Hussain, M.G.; Sultana, B.; Rahman, M.; Hasan, M.R. Comparison analysis of Bangla news articles classification using support vector machine and logistic regression. *TELKOMNIKA (Telecommun. Comput. Electron. Control)* **2023**, *21*, 584–591. [[CrossRef](#)]
55. Choi, E.; Schuetz, A.; Stewart, W.F.; Sun, J. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med. Inform. Assoc.* **2017**, *24*, 361–370. [[CrossRef](#)]
56. Prakash, J.; Vijay, A.A. A multi-aspect framework for explainable sentiment analysis. *Pattern Recognit. Lett.* **2024**, *178*, 122–129. [[CrossRef](#)]
57. Kaushik, D.; Hovy, E.; Lipton, Z.C. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv* **2019**, arXiv:1909.12434. [[CrossRef](#)]

58. González-Sendino, R.; Serrano, E.; Bajo, J. Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. *Future Gener. Comput. Syst.* **2024**, *155*, 384–401. [[CrossRef](#)]
59. Clark, C.; Yatskar, M.; Zettlemoyer, L. Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv* **2019**, arXiv:1909.03683. [[CrossRef](#)]
60. Asyrofi, M.H.; Yang, Z.; Yusuf, I.N.B.; Kang, H.J.; Thung, F.; Lo, D. BiasFinder: Metamorphic Test Generation to Uncover Bias for Sentiment Analysis Systems. *IEEE Trans. Softw. Eng.* **2022**, *48*, 5087–5101. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.