Bias Detection and Optimization of Recruitment Texts Based on BERT

Liping Feng, Na Feng, Qiaojuan Hui
School of Digital Economy and Management, Yinchuan University of Science and Technology, Yinchuan 750001, Ningxia, China
Corresponding author: Na Feng (e-mail: fengrxz@163.com)

ABSTRACT This study proposes an end-to-end framework for automatically detecting and mitigating implicit social bias in recruitment texts, providing an effective technical solution for fair and intelligent information processing in digital recruitment systems. As trustworthy semantic analysis and automated decision support become increasingly important for intelligent communication environments and information transmission applications, improving the neutrality and interpretability of textual content is also valuable for broader engineering-oriented information systems. Based on BERT, a bias detection model integrating BiLSTM and a hierarchical attention mechanism is developed to achieve fine-grained bias classification through word-level and sentence-level attention analysis. Furthermore, a T5-based bias optimization model is fine-tuned using biased–unbiased parallel corpora to accomplish semantic-preserving text rewriting. Experimental results demonstrate that the proposed detection model achieves an accuracy of 0.882 and an F1 score of 0.842 on the test set, outperforming the strongest baseline DeBERTa-v3 by 0.021 in F1 score. Ablation studies further verify that the BiLSTM and hierarchical attention modules improve the F1 score by 0.023 and 0.017, respectively. The optimization model obtains a neutrality score of 4.7 and an information retention score of 4.5 in human evaluation, while achieving BLEU and ROUGE-L scores of 0.65 and 0.71, respectively. The proposed framework provides a reliable and interpretable approach for automated fairness correction of recruitment texts and offers practical guidance for intelligent text processing and trustworthy information management in engineering applications.

INDEX TERMS bias detection, BERT, recruitment text optimization, intelligent information processing, trustworthy communication

I. INTRODUCTION

As artificial intelligence technologies pervade recruitment processes, automated resume screening and job recommendation systems are quickly gaining popularity[1,2]. However, the fairness of such systems encounters significant challenges, often originating from the training datasets themselves, that is, the social biases contained within the recruitment text. Implicit biases exist in the wording and sentence structure that becomes a subtle reinforcement of stereotypes based on sex, age, and physical ability (for example, " Courage," indicates a preferred male candidate; "Physical fitness," may obscure people with disabilities) [3,4]. These biases, which are quite common, pose challenges to fairness within the employment market and impede equal employment opportunity rights [5,6]. Furthermore, employers face potential legal risks and damage to their brand reputation. Therefore, automated detection of bias in recruitment texts and optimization of biased recruitment texts has become an important area of inquiry with implications for responsible artificial intelligence development and building an inclusive digital society. It also has both theoretical and practical significance.

In summary, existing research still faces significant limitations in fine-grained classification of bias types and high-quality automated optimization of biased text. To systematically address these issues, this study proposes an end-to-end framework for addressing bias in recruitment text. The main contributions of this paper are as follows: First, to address the limitations of fine-grained classification and interpretability, a detection model based on BERT, integrating a bidirectional long short-term memory network with a hierarchical attention mechanism, is constructed. This model not only accurately identifies bias across multiple categories, such as gender, age, and physical ability, but also clearly demonstrates the key sources of bias through visualization of attention weights. Second, in order to address the absence of optimization steps and quality constraints, this research proposes the novel use of the T5 (Text-To-Text Transfer Transformer) text-to-text generation model as a biased text optimizer. It fine-tuned this biased text optimizer using specially trained biased-unbiased parallel corpus pairs to generate neutral and to-standard job descriptions while maximizing the preservation of the original meaning. Finally, this study is systematically validated on public recruitment

data, providing a reliable technical path and empirical evidence for achieving a complete closed-loop approach to recruitment fairness, which is essential for ensuring equal opportunities across all roles in the industry,from detection to intervention.

II. RESEARCH METHODS

A. OVERALL DESIGN OF THE RESEARCH FRAMEWORK

This study constructed a systematic end-to-end bias processing framework. Its core process is shown in Figure 1, which covers three modules: data preprocessing, bias detection, and bias optimization, aiming to achieve the automatic generation of fair text from raw text.

The core of this research framework is a serialized, task-driven pipeline system. The first stage, data preprocessing, is the foundation of the entire research. Key text fields are extracted from a recruitment dataset on the Kaggle platform and rigorously cleansed and normalized. Crucially, not only is a baseline corpus constructed for model training, but high-quality biased-unbiased parallel sentence pairs are also created through manual annotation. This tailored corpus serves as the foundation for subsequent supervised learning of the bias detection model and sequence-to-sequence training of the bias optimization model. Its quality directly determines the performance ceiling of the entire system. This stage transforms the messy, unstructured raw data into structured, normalized data that can be consumed by the model, providing reliable input for subsequent analysis.

In the second stage, the bias detection model serves as the perception and judgment hub of the entire system. This paper designs a hybrid deep learning architecture based on BERT. The text to be tested is first deeply semantically encoded by the BERT encoder, generating a feature vector rich in contextual information. Next, a BiLSTM (Bidirectional Long Short-Term Memory) layer helps to capture long-range dependencies and sequential structure in the text. This article introduces a hierarchical attention mechanism: word-level attention weighs the contribution of each word in a sentence toward bias judgment, while sentence-level attention is implemented specifically to process longer job descriptions, capturing the most biased sentences. Ultimately, through the combination of both mechanisms, fine-sliced bias classification results are output, along with a visual heatmap explanation of the model's reasoning.

The bias reduction model is the third and final stage of the research's function. This model actually implements interventions, going from diagnosis to therapy. This module is based on the T5 text-to-text generation model. When the detection model detects biased text, it passes the biased text into the T5 model, now concatenated with specific instructions for mitigation of the bias. The T5 encoder understands the layered semantic meaning and the intent of the original text, while the decoder samples new text that represents a neutral and appropriate word-by-word-formulation, based on the encoder representations and bias mitigation instructions.

B. DATA PREPROCESSING AND CORPUS CONSTRUCTION

The experimental data for this study comes from the "New York City Current Job Postings" dataset, available on Kaggle (https://www.kaggle.com/datasets/new-york-city/new-york-city-current-job-postings). The dataset, which includes data on both internal and external job postings from the city's recruitment website, is authoritative, well-structured, and complete and provides the perfect foundation for a study of bias in recruiting texts.

Key text semantics-related fields were extracted from the original dataset. The extracted text was then carefully cleaned and standardized, which included removing HTML (HyperText Markup Language) tags, URLs (Uniform Resource Locator), and special characters; lowercasing the data to remove capitalization-related differences and performing tokenization using the Natural Language Toolkit (NLTK). Finally, stop words were removed, and invalid short sentences that were less than five characters were filtered out, resulting in a textual dataset with sufficient information for analysis. In the preprocessing stage, stop word removal and other operations aimed to initially reduce data noise and serve as part of the exploratory analysis. However, during the model building stage, this paper recognized that preserving complete syntactic structure is crucial for sequence models to capture long-range dependencies. Therefore, when constructing the final corpus for training the BERT-BiLSTM-Hierarchical-Attention model, the stop word filtering step was removed to maintain the original word order and grammatical integrity of the text, ensuring that the input to the BiLSTM layer is a coherent sequence of sentences.

Key data fields utilized in the study, and specific preprocessing methods are presented in Table 1.

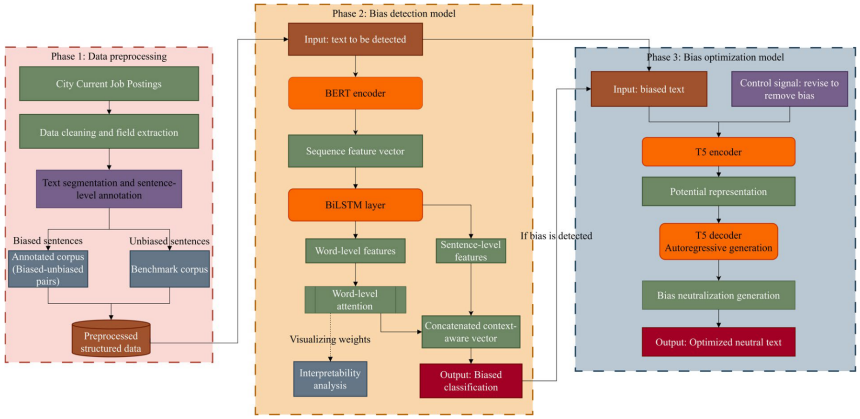

FIGURE 1. Overall research framework

TABLE 1. Position data field description and preprocessing strategy

Serial number	Field	Description	Preprocessing strategy and objectives
1	Business title	The business title of a position is usually an informal and descriptive title.	Convert to lowercase; Remove spaces on both sides of hyphens to maintain lexical integrity; Direct segmentation.
2	Civil service title	The official civil service title is standardized and formal in format.	Convert to lowercase; Perform phrase segmentation on fixed job title combinations to preserve key semantic units.
3	Job description	Core responsibilities and job description of the position.	Remove HTML tags; Convert to lowercase and participle; Key steps: Break down sentences by periods and semicolons, and break down long descriptions into independent sentences for sentence level bias analysis.
4	Minimum qual requirements	The minimum qualification requirements that applicants must meet.	Convert to lowercase and participle; Break down sentences; This field is a high-risk area for age, physical ability, and educational bias, and is a key focus for cleaning and labeling.
5	Preferred skills	The skills or experience that are prioritized often contain implicit biases.	Convert to lowercase and participle; Break down sentences; This field is a high-risk area for age and gender bias.

C. DATA ANNOTATION SYSTEM AND STATISTICAL ANALYSIS

To construct a high-quality supervised learning dataset, this paper established a fine-grained bias classification system and annotation specifications. Three annotators annotated the preprocessed sentences back-to-back according to established guidelines, resolving disagreements through discussion. The final inter-annotator reliability (Cohen's Kappa) was 0.81, indicating high consistency in the annotation results. A statistical analysis of the annotated corpus is shown in Table 2.

TABLE 2. Statistical analysis of the annotated corpus

Bias category	Proportion (%)	Example sentence
No Bias	71.2	"Proficiency in Microsoft Office Suite is required."
Gender	10.5	"Seeking a forceful and assertive leader for our sales team."
Age	8.2	"Ideal for recent college graduates or digital natives."
Physical	7.1	"Candidates must be able to lift 50 pounds unassisted."
Other	3.0	"Must be available to work nights and weekends."

Note: The annotation criteria in this study focus on identifying implicit language patterns in the text that may constitute an unjustified exclusion of a specific age group,rather than judging the ultimate reasonableness of such a requirement in actual recruitment scenarios.

Automatic detection of 'age bias' aims to identify expressions that may implicitly imply age bias; this is not equivalent to determining that such expressions necessarily constitute unreasonable discrimination in a specific recruitment context. In practical applications, optimizing the fairness of recruitment texts requires a comprehensive consideration of legal provisions, business necessity, and actual job requirements, striking a balance between eliminating arbitrary biases and retaining reasonable and relevant qualifications. The distribution of bias category data is shown in Figure 2.

In terms of quantity, non-biased examples predominate, consistent with the expectation that bias expressions in real-world scenarios are relatively subtle. Among the bias categories, gender bias is the most prevalent, while sentences expressing physical ability bias not only dominate in number but also have a higher average length and standard deviation than other categories. This means that descriptions of this type of bias are usually more specific and lengthy, and generally entail the detailed enumeration of physical tasks. This can make for a more elaborate sentence structure and increases the length of the text, which create unique challenges for models to correctly identify this type of bias.

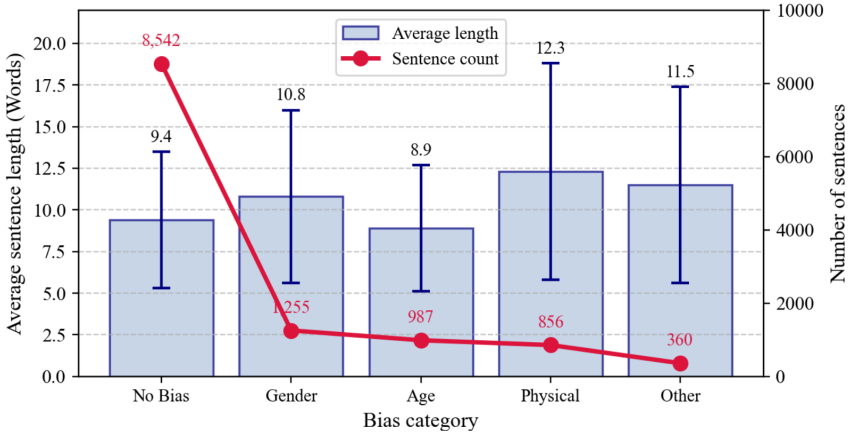

FIGURE 2. Biased category data distribution

D. DESIGN OF BIAS DETECTION MODEL BASED ON BERT

E. BERT-BILSTM-HIERARCHICAL-ATTENTION MODEL STRUCTURE

This study created a hybrid deep learning model built on BERT [7,8] to effectively classify implicit social bias in recruitment texts. This model builds on the solid foundation of understanding semantics offered by BERT while adding sequence encoding and a hierarchical attention layer to extract bias characteristics and improve transparency in model decisions. Figure 3 shows the BERT-BiLSTM-Hierarchical-Attention model structure.

The base of this model is a BERT encoder. The contextually-informed word vectors produced contain a wealth of semantic information that determines the perceptual bias in the system. A bidirectional LSTM network subsequently maps the sequence to a deep encoding. Its gating structure allows for effective capture of long-range dependencies and non-linear sequential patterns in the text by combining semantic clues from disparate sources into a cohesive contextual representation that serves as the basis for the next level of analysis. The model's key innovation is its hierarchical attention mechanism built to operate on both the word and sentence levels. At the word level, the model automatically attends to important words that may have bias in them by evaluating the attention weight corresponding to each word. As an example, with job descriptions, there could be cases where the job description is lengthy and contains several sentences, and in these situations the sentences themselves can be evaluated for importance, then identifying the source of bias in each of those sentences. The hierarchical structure takes the model from a local perspective to an overall perspective, which allows for a fine-grain bias determination. At the end of the process the generated weighted vector is inputted into the classifier output, and also the attention weights can be indicated and visualized as heat maps, which provides evidence for each model decision, as the model's."

F. BERT FEATURE EXTRACTION

This study uses BERT-base-uncased as the base feature extractor for the model [9,10]. For an input sequence $X = x_1, x_2, \ldots, x_L$ consisting of L tokens. BERT outputs a sequence of hidden states rich in contextual information through its deep Transformer encoder:

$$H = h_1, h_2, \ldots, h_L = \mathrm{BERT}(X) \tag{1}$$

In formula 1, each $h_i \in R^{d_h}$, $d_h = 768$ is the hidden dimension of the BERT-base model. Unlike static word vectors, BERT's output h_i dynamically incorporates information from all other tokens in the sequence, enabling accurate understanding of the true meaning of a word in a specific context. This is key to detecting whether words like "strong" constitute gender bias in different contexts.

G. BILSTM LAYER

Although BERT can capture contextual information, its main advantage lies in short-range dependencies. To more effectively model the long-range semantic dependencies that may exist in recruitment texts, a BiLSTM is superimposed on the output of BERT [11,12].

BiLSTM processes sequences in both forward and backward directions through its internal gating mechanism:

$$\overrightarrow{h_i} = \overrightarrow{\mathrm{LSTM}}(h_i, \overrightarrow{h}_{i-1}) \tag{2}$$

$$\overleftarrow{h_i} = \overleftarrow{\mathrm{LSTM}}(h_i, \overleftarrow{h}_{i+1}) \tag{3}$$

Finally, the BiLSTM output corresponding to each token is the concatenation of its forward and backward hidden states:

$$h_i^{(lstm)} = [\overrightarrow{h_i}; \overleftarrow{h_i}] \in R^{2d_l} \tag{4}$$

In formula 4, d_l is the hidden dimension of LSTM. This step generates an enhanced sequence representation $H^{(lstm)} = h_1^{(lstm)}, \ldots, h_L^{(lstm)}$, which is better at capturing the global pattern of the entire sequence.

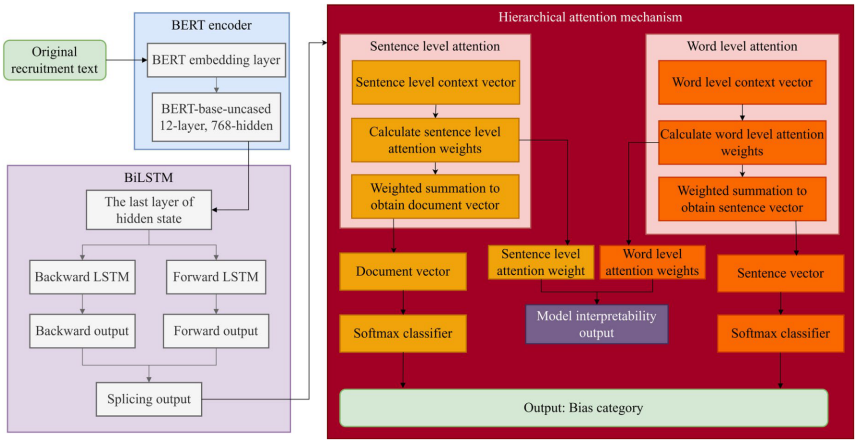

FIGURE 3. BERT-BiLSTM-Hierarchical-Attention model structure

H. HIERARCHICAL ATTENTION MECHANISM

The hierarchical attention mechanism is the core of the model to achieve interpretability and fine-grained analysis. This paper designs a two-layer attention structure [13]. The word-level attention mechanism aims to identify which specific words in a sentence contribute the most to bias.

Pass each word representation $h_i^{(lstm)}$ output by BiLSTM through a single-layer perceptron and map it to a scalar energy value e_i^w, which represents the 'importance' of the word. The energy value formula is:

$$e_i^w = u_w^\top \tanh(W_w h_i^{(lstm)} + b_w) \tag{5}$$

In Formula 5, $W_w \in R^{d_a \times 2d_l}$ is the weight matrix, $u_w \in R^{d_a}$ is the word-level context vector, and b_w is the bias term. u_w is treated as a learnable parameter during training, which can be interpreted as a "bias feature" of the query vector.

The Softmax function is used to normalize the energy values of all words to obtain the final word-level attention weights α_i.

$$\alpha_i = \frac{\exp(e_i^w)}{\sum_{j=1}^{L} \exp(e_j^w)} \tag{6}$$

The weight α_i satisfies $\sum_{i=1}^{L} \alpha_i = 1$, and its size directly reflects the importance of the word x_i.

Generate sentence vector: Through weighted summation of attention weights, the information of the entire sequence is summarized into a sentence vector of fixed dimension.

$$s = \sum_{i=1}^{L} \alpha_i h_i^{(lstm)} \tag{7}$$

This vector s is the representative feature used for the final sentence-level bias classification.

For a complete job description consisting of M sentences $S_1, S_2, \ldots, S_M$, first use the word level attention mechanism mentioned above to generate a sentence vector s_j for each sentence S_j. Subsequently, this paper repeats a similar attention process at the sentence level to identify which sentence in the entire document is the most biased.

Calculate the sentence energy value:

$$e_j^s = u_s^\top \tanh(W_s s_j + b_s) \tag{8}$$

In Formula 8, $W_s \in R^{d_a \times 2d_l}$, $u_s \in R^{d_a}$ are the sentence-level context vectors.

Normalized sentence weights:

$$\beta_j = \frac{\exp(e_j^s)}{\sum_{k=1}^{M} \exp(e_k^s)} \tag{9}$$

Generate document vectors:

$$v = \sum_{j=1}^{M} \beta_j s_j \tag{10}$$

This document vector v is used to perform overall bias classification at the document level. Finally, the sentence vector s and document vector v are fed into separate Softmax classifiers for bias category prediction.

I. DESIGN OF BIASED TEXT OPTIMIZER BASED ON T5

J. T5 MODEL SELECTION AND PRINCIPLE

This study uses the T5-base model as the core architecture of the text optimizer [14,15]. T5 uniformly transforms all natural language processing tasks into a "text-to-text" format, making it naturally suitable for text replication and style transfer tasks. Its core is a Transformer-based encoder-decoder model [16,17]. The encoder is responsible for mapping the input sequence into a continuous, context-aware latent representation, while the decoder generates the target sequence one by one in an autoregressive manner based on this representation and previously generated tokens.

Given an input sequence $x = (x_1, x_2, \ldots, x_m)$ and a target sequence $y = (y_1, y_2, \ldots, y_n)$, the goal of the T5 model is to maximize the conditional probability $P(y \mid x)$. This probability is autoregressively decomposed into:

$$P(y \mid x) = \prod_{t=1}^{n} P(y_t \mid y_{<t}, x) \tag{11}$$

In Formula 11, $y_{<t} = (y_1, \ldots, y_{t-1})$ represents all tokens generated before time t. At each layer of the decoder, the cross attention mechanism allows the model to dynamically focus on the most relevant part of the input sequence x when generating each new word y_t, which is the key to achieving high-quality semantic preservation rewriting.

K. TASK FORMALIZATION AND INPUT FORMAT

This paper formalizes the biased text optimization task as a controlled text replication task. To explicitly instruct the model to execute the specific instruction of "eliminating bias," a task prefix is added before the original biased sentence to form a complete input sequence.

The input sequence is constructed as : "revise to remove bias: [Original Biased Sentence]"

The target sequence is: "[Revised Neutral Sentence]"

This formal approach defines the task as conditional text generation, where the prefix "revise to remove bias:" serves as a control signal to guide the model to generate text with "neutral" properties.

L. MODEL FINE-TUNING AND INFERENCE

The pre-trained T5-base model is fine-tuned using standard maximum likelihood estimation. The model parameters are optimized by minimizing the negative log-likelihood loss function θ:

$$L(\theta) = -\sum_{(x,y)\in D} \sum_{t=1}^{n} \log P(y_t \mid y_{<t}, x; \theta) \tag{12}$$

In formula 12, D is the training dataset. This process enables the model to learn to map biased text with a specific prefix to the desired unbiased text distribution.

When the bias detection model identifies a biased sentence x_{biased}, the task prefix is appended to it, forming the complete input. This sequence is then fed into the fine-tuned T5 model. The decoder uses a beam search strategy to generate optimized text $y_{neutral}$. Beam search effectively balances generation quality and computational efficiency by retaining the k candidate sequences with the highest probability at each step (k being the beam width), ultimately outputting the sequence with the highest overall probability:

$$y_{neutral} = \arg\max_{y} P\big(y \mid \text{“revise to remove bias: ”} + x_{biased}\big) \tag{13}$$

Through this design, this study achieved an automated closed loop from "bias identification" to "bias elimination", providing a practical driving solution for creating fair recruitment copy.

III. EXPERIMENTAL SETUP

A. EXPERIMENTAL ENVIRONMENT AND PARAMETER CONFIGURATION

The experiments of this study took place on a high-performance computing server with dual NVIDIA A100-PCIE-40GB graphics cards, under an Ubuntu 20.04 operating system, and utilizing the PyTorch 2.0.1 deep learning framework. In terms of the methodology, and to ensure reproducibility, all the random seeds were fixed, and pre-trained model weights were loaded using the Transformers 4.28.1 library.

For setting model hyperparameters, the bias detection model utilized the AdamW optimizer, with a learning rate of 2×10^{-5}, implementing a linear warmup for the learning rate and a cosine annealing strategy, with a global batch size of 32. The model was trained for 100 epochs, while the performance on the validation set was observed, and an early stopping strategy was implemented. The hidden dimension of the BiLSTM layer was set at 256, the attention mechanism dimension was set at 200, and all layers utilized a dropout of 0.1 to prevent overfitting.

For the text optimization task, this article fine-tuned the T5-base model using the AdaFactor optimizer [18] at a constant learning rate of 1×10^{-4} and using a global batch size of 16. During inference, it used beam search of width 4, and set the length penalty factor to 0.6 to manage the generated text quality. The maximum length of the input sequence was limited to 128 tokens for all models. This configuration fully takes into account the trade-off between computational efficiency and text quality.

B. BASELINE MODEL

To comprehensively evaluate the performance of the BERT-BiLSTM-Hierarchical-Attention model proposed in this study, this paper selected advanced models covering different technical routes as baselines for comparison. BERT-base was included as a representative of the basic pre-training model, and its improved versions RoBERTa-large (A Robustly Optimized BERT Pretraining Approach), DeBERTa-v3 (Decoding-enhanced BERT with Disentangled Attention) [19,20], and ALBERT (A Lite BERT for Self-supervised Learning of Language Representations) [21] were also selected. These models have demonstrated excellent performance in multiple natural language understanding tasks by improving pre-training tasks, introducing decoupled attention mechanisms, and using parameter sharing strategies. The ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) model, which is based on token replacement detection pre-training, was selected. Its efficient pre-training method provides a unique comparative perspective. In addition, BART-large (Bidirectional and Auto-Regressive Transformers) [22] is a powerful sequence-to-sequence model whose encoder part is also suitable for classification tasks. To cover traditional methods, a support vector machine model based on TF-IDF (Term Frequency-Inverse Document Frequency) [23]

features is used as a non-neural network baseline. Finally, the FAST-BERT architecture, which is specifically designed for efficient text classification, is included. Its complex dense encoder design provides another valuable comparison benchmark, verifying the effectiveness and advancement of the model proposed in this study from multiple dimensions.

C. EVALUATION METRICS

For the classification task of bias detection, this paper uses standard classification evaluation metrics in machine learning. Accuracy measures the proportion of correct predictions overall, but its value is limited in datasets with imbalanced classes. This article focuses on both precision and recall calculated per class. Precision measures the proportion of samples predicted to be of a certain category that are actually of that category, and the formula is:

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

The recall rate measures the proportion of samples that are correctly predicted among samples that are actually of a certain category. The formula is:

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

The F1 formula is:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{16}$$

For text optimization tasks, this paper uses a combination of automatic and manual evaluation. The automatic evaluation uses two classic metrics from the text generation field: BLEU (Bilingual Evaluation Understudy) and ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence).

D. TRAINING LOSS CONVERGENCE

To provide an intuitive comparison of the training dynamics and learning efficiency of the proposed model and BERT-base, this article plots the loss convergence curves of both models during the training cycle in Figure 4.

From the training loss curves, one can see the systematic difference in convergence, stability, and generalization potential of the two models. The proposed model has a steeper decline early in training (around epoch 20), suggesting that this model's architecture enables faster convergence to the core patterns in the data. BERT-base shows an especially slow convergence rate in the beginning. By the mid and later training period, after a brief fluctuation (for example, around epoch 50, most likely due to the adaptations of a dynamic optimizer), the proposed model stabilizes at a much lower plateau with much less fluctuation than BERT-base.

IV. RESULTS AND ANALYSIS

A. COMPARATIVE ANALYSIS OF BIAS DETECTION MODEL PERFORMANCE

In order to systematically assess the bias detection ability of the proposed model, an exhaustive performance comparative analysis on the test set with a number of advanced baseline models was performed. The details of results are represented in Table 3.

TABLE 3. Performance comparison of bias detection binary classification task

Model	Accuracy	Precision	Recall	F1
BERT-base	0.854	0.812	0.793	0.802
TF-IDF + SVM (Support Vector Machine)	0.783	0.745	0.698	0.721
RoBERTa-large	0.862	0.826	0.805	0.815
BART-large	0.859	0.821	0.802	0.811
DeBERTa-v3	0.868	0.831	0.812	0.821
ELECTRA	0.858	0.819	0.798	0.808
ALBERT	0.847	0.804	0.786	0.795
FAST-BERT	0.851	0.808	0.791	0.799
Proposed	0.882	0.849	0.835	0.842

A review of overall performance trends indicates that deep learning methods based on pre-trained language models, far surpassed any performance provided by traditional TF-IDF + SVM methods, indicating the importance of deep semantic understanding in detecting implicit bias. In the pre-trained models, models with larger parameters and more advanced architectures (for example, RoBERTa-large and DeBERTa-v3) indicated performance benefit across all baseline models they were tested against. DeBERTa-v3, which leverages its decoupled attention mechanism and improved mask decoder to achieve the best performance among all baseline models, had an F1 score of 0.821. Traditional BERT-base and its variants (namely, ALBERT and FAST-BERT) faced considerable performance bottlenecks for bias detection, where neither recall nor precision was greater than 0.815. The data suggests that a single pre-trained model suffers a performance limit, even with a pre-trained language model, without a focused building to the model architecture for bias detection. The BERT-BiLSTM-Hierarchical-Attention model presented in this study achieved the highest performance on all evaluation metrics, obtaining an F1 score of 0.842, an improvement of 0.021 over the strongest baseline, DeBERTa-v3. This advantage in performance is the result of the model's targeted architecture: the BiLSTM layer effectively enhances the model's ability to capture long-term dependencies in text sequences, allowing it to better understand the distribution of bias expressions in context. The hierarchical attention mechanism assigns importance weights to words and sentences dynamically and concentrates on the relevant bias cues. This helps improve the model's sensitivity to subtle bias expressions (as indicated by improved recall) and accuracy in ambiguous cases (as indicated by improved precision). Thus, the model's architecture allows for a more complete and accurate identification of different

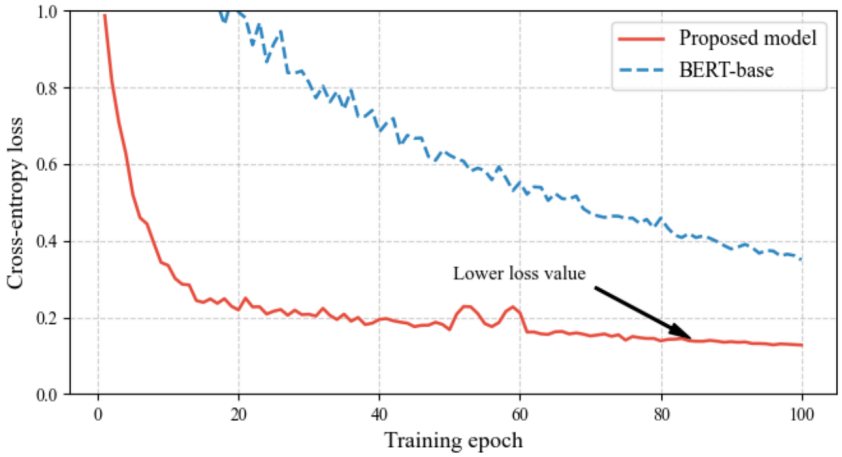

FIGURE 4. Training loss convergence curve

categories of bias in text, supporting discoveries on the model's consistently superior overall performance to current state-of-the-art baselines.

While advanced pre-trained models like DeBERTa-v3 are powerful in general semantic understanding and handling contextual dependencies, their architectures are primarily designed for broad natural language understanding tasks. This study proposes a BERT-BiLSTM-Hierarchical-Attention model, a specialized architectural enhancement for the task of bias detection. The addition of the BiLSTM layer is not intended to replace the Transformer's long-range dependency capture capabilities, but rather to specifically enhance the sequential modeling of bias expression patterns in text, such as identifying biased descriptions with specific structures spanning multiple clauses. Simultaneously, the hierarchical attention mechanism provides targeted focus at both the word and sentence levels, which is crucial for locating key semantic units of bias. This combination of a pre-trained general semantic foundation (BERT) and task-oriented sequence and attention enhancements (BiLSTM + hierarchical attention) allows the model to focus more on learning the specific contextual patterns and feature distributions required for bias detection, thus outperforming general strong baseline models that have not undergone specific structural adjustments for this task.

Confusion matrices on fine-grained bias classification is shown in Figure 5.

The comparative analysis of confusion matrices (Figure 5) clearly reveals distinct performance patterns across models. The traditional TF-IDF + SVM method (Figure 5a) exhibits significant classification weaknesses, as indicated by its attenuated diagonal elements and widely scattered off-diagonal misclassifications. The misclassification rate in the "unbiased" category is especially high (1,666 true unbiased are classified correctly but 470 are classified as biased). This underscores the limited ability of a bag-of-words based method to infer semantic context and results in many neutral expressions being misclassified. DeBERTa-v3 (Figure 5b) and FAST-BERT (Figure 5c), state-of-the-art pre-trained options, perform better, with much stronger diagonal elements, particularly in the "gender bias" and "age bias" categories. Notably, both of these models still misclassify a considerable amount of "unbiased" vs "other bias." This shows that even the best general-purpose pre-trained models have limitations in their architecture when they are being used for the very specific task of bias detection that requires fine-grained semantic understanding.

The suggested framework (Figure 5d) demonstrates excellent confusion matrix all classes, with over the diagonal being the most noticeable and sparse off-diagonal. In the case of the "unbiased" category, the suggested framework enhanced the accuracy for that class from DeBERTa-v3's 1778 count to 1880 and decreased the false positives from 358 to 265. The enhanced accuracy of the suggested framework's prediction of "unbiased" categories can mostly be attributed to the layered attention mechanism of the BiLSTM, which allows for deep comprehension of textual context, enhancing the BiLSTM's ability to differentiate between truly biased expressions and seemingly similar neutral expressions. In the case of minority category recognition, the suggested framework achieved a larger number of correct classifications than other models suggest, conceptualizing the enhancement of the BiLSTM layer to capture class related long-range dependencies and classify physical ability requirement descriptions throughout the text. Ultimately, one of the most significant findings of this study is that the proposed framework achieved 79 correct classifications in the "Other Biases" category with the smallest sample size, demonstrating the strong generalization capability of the BiLSTM output to identify rare bias patterns.

B. ABLATION EXPERIMENT ANALYSIS

To assess the influence of each element of the model, this paper sequentially eliminated the BiLSTM layer and hierarchical attention mechanism, and devised three ablation experiment models for comparison. The outcomes are displayed in Table 4.

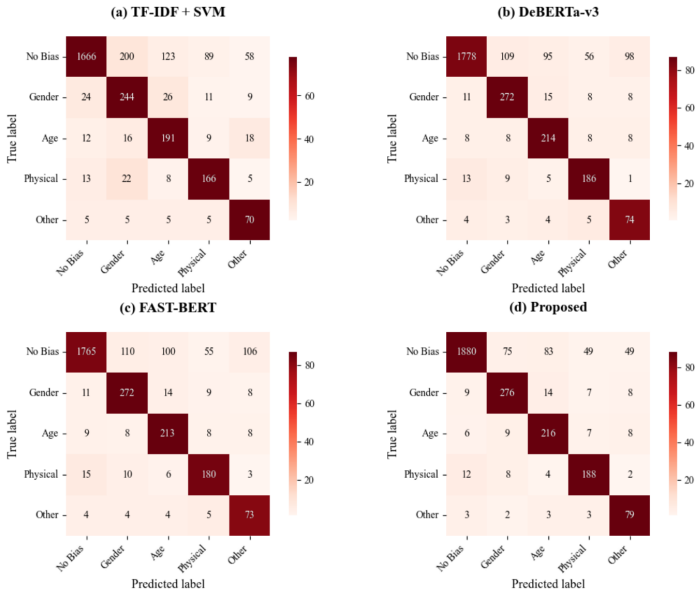

FIGURE 5. Confusion matrix comparison: (a) TF-IDF + SVM; (b) DeBERTa-v3; (c) FAST-BERT; (d) Proposed

TABLE 4. Ablation experiment performance comparison

Model	Accuracy	Precision	Recall	F1
BERT	0.854	0.812	0.793	0.802
BERT-BiLSTM	0.869	0.832	0.818	0.825
BERT-BiLSTM-Hierarchical-Attention	0.882	0.849	0.835	0.842

The ablation experiment results indicate that there was a constant improvement in performance with the price of adding each component. The basic BERT model performed consistently, as found in numerous metrics, but suffered significant limitations. The F1 score was only 0.802 and revealed that the pre-trained model did not understand the bias detection task without informed architectural and subsequent tuning. The addition of the BiLSTM layer provided improvements across all metrics, however; the F1 score was improved from 0.802 to 0.825 while accuracy improved from 0.854 to 0.869. This represents a significant improvement and evidence of the importance of sequence modeling for the bias detection task. BiLSTM model long-range dependencies in the text to better represent the distribution patterns of bias expressions in context. This results in a more precise identification of underlying implicit biases that need contextual justification to be informative.

Once a hierarchical attention mechanism was incorporated, model performance improved even further, achieving an F1 score of 0.842, which represents an additional 0.017 improvement over the BERT-BiLSTM version. This improvement manifests primarily in a balanced increase in precision and recall, demonstrating that the attention mechanism enhances the model's ability to identify salient information significantly while improving its accuracy of distinguishing ambiguous cases. The hierarchical attention mechanism dynamically assigns importance weights, or attending weights, at both the word and sentence level of the model, allowing the model to automatically attend to the most discriminative semantic units of the text while filtering irrelevant information.

C. MODEL INTERPRETABILITY ANALYSIS

To further investigate the model's decision-making process, this paper visualized word-level attention weights and carried out an evaluation regarding the model's identification of biases. The results are presented in Figure 6.

In the visualization of the physical ability bias condition (Figure 6a), the model shows an explicit focus to the key phrase "able to lift 50 lbs." The attention weight distribution shows that the three tokens of "lift" (0.25), "able" (0.15), and "50" (0.18) received the highest attention, completely corresponding to each component of the physical ability bias. This attention demonstrates that the model is aware of the action "lift" and its link to physical ability, while also recognizing the significance of the number "50" in providing a specified amount of physical requirement. Notably, the model does not just attend to all verbs or nouns but attends to semantic units that truly are aspects of ability requirements. This indicates the model learned the essential nature of physical ability bias itself-that is, explicit requirements for specific physical tasks may mean unreasonable exclusion of some applicants for that job.

In the age bias case analysis (Figure 6b), it is seen that the model's attention mechanism demonstrates a nuanced ability to identify words with implicit age signals. "recent" (0.28), "college" (0.22), and "graduates" (0.18) dominated the attention weights. Collectively, these three words are likely to describe the group of young job seekers. Notably, the model's highest attention weight was given to "recent," indicating its comprehension of the descriptive importance

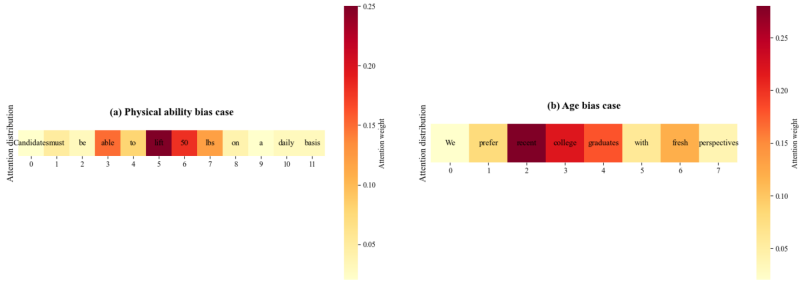

FIGURE 6. Attention weight visualization heat map: (a) Physical ability bias case; (b) Age bias case

of time and age bias. The model was also able to give moderate attention to "fresh" (0.12), demonstrating its ability to capture metaphorical language of youth. This subtle attention allocation, demonstrates the model's competence to identify not only explicit age-restricted language, but also implicit age preference, as evidenced by recognition of descriptors of one's educational history, temporal experience in relation to the specific job, and descriptors of age, demonstrating the model's ability to mature in complex semantic understanding. However, the visualization of attention weights essentially reveals an allocation of feature importance within the model, showing which input information the model relies on when making classification decisions. The model assigns high weights to words like 'lift' and 'recent,' aligning with manually labeled bias cues, confirming that the model has effectively learned surface or deep language patterns related to data labels. It's important to clarify that this consistency alone does not rigorously prove that the model 'understands' the social or ethical reasons these words constitute bias, such as understanding the age-rejection effect potentially implied by 'recent graduate.' It primarily indicates that the model has learned to use these lexical patterns as strongly correlated features for predicting bias.

D. BIASED TEXT OPTIMIZATION RESULTS

In assessing the appropriateness of the text optimizer, this paper examined some representative bias scenarios to show the model's capability to eliminate biased language while safeguarding core job duties. Table 5 presents a side-by-side comparison of the text with the initial and optimized recommendations along with a description of the modifications that occurred.

The optimization cases illustrate a methodical rewriting strategy to address different types of bias. With regard to age bias (Cases 1, 2, and 5), the model successfully reconstructed age-based descriptions to be more competence- and skill-based. For example, "young and dynamic" changed to "energetic and proactive," maintaining the necessary positive disposition a candidate might have, while removing an age bias. In the same way "recent college graduate" became "relevant education background," effectively avoiding a bias to willful recent graduates, while maintaining reasonably appropriate educational expectations. Both changes reflected the model's understanding of both the nature of age bias; it did not merely delete sensitive language, but changed it with appropriate language in years of functional description in the type of requirements the text was suggesting. Similarly in Case 5, "digital native" - a very specific generational bias - was solidly and confidently changed into "technical proficiency and adaptability," which conveyed the models ability to easily and confidently transformed even implicit age bias.

The model provided a broader range of optimization techniques in the areas of gender bias and physical ability. The model implemented a "functionalized" rewording technique to address physical ability bias (Cases 3 and 6). It shifted requirement descriptions that specified particular physical abilities to merely functional skills to accomplish the task. For example, "physically strong" was reformulated as "able to safely operate," thus avoiding the exclusionary requirement of particular physical conditions yet still fulfilling the functional requirements of the job. The model could simply employ inclusive language and equal qualifications to ameliorate discrimination regarding gender bias (Cases 4 and 8). Case 7 was particularly valuable considering the complexity of social biases, as the model successfully repositioned the wording "Native English speakers" which may involve biases related to nationality and race to a language proficiency requirement. All the optimizations adhered to the same basic rule: equity and utility must be achieved in the optimized text while reducing discriminatory language and needing to hold constant as much as possible the functional needs of the role.

This study combined manual evaluation with automated metrics to thoroughly assess the performance of the suggested text optimizer by comparing the T5 model with sophisticated baseline techniques in a number of dimensions, such as text similarity, neutrality, fluency, and information retention. A comparison of the outcomes of biased text optimization is displayed in Figure 7.

The models' variations in performance along three important dimensions are evident in Figure 7a. Because of their simple word replacement approach, rule-based methods do fairly well in information retention (3.8 points), but they have serious problems with neutrality (2.8 points) and flu-

TABLE 5. Text optimization case display table

Serial number	Original text	Optimized Text	Optimization explanation
1	We are looking for a young and dynamic sales manager.	We are looking for an energetic and proactive sales manager.	Replace age-related vocabulary with ability oriented descriptions.
2	The ideal candidate will be a recent college graduate.	The ideal candidate will have relevant educational background.	Eliminate age preference for fresh graduates and focus on the essence of educational background.
3	Applicants must be physically strong to handle heavy equipment.	Applicants must be able to safely operate heavy equipment.	Transforming physical ability requirements into task execution ability.
4	We prefer female candidates for this customer service role.	We welcome all qualified candidates for this customer service role.	Eliminating gender preferences and emphasizing equal qualifications.
5	The position requires a digital native with tech savvy.	The position requires strong technical proficiency and adaptability.	Transforming intergenerational characteristics into specific skill requirements.
6	This job is suitable for able-bodied individuals only.	This job requires the physical capacity to perform essential functions.	Replace discriminatory expressions with functional descriptions.
7	Native English speakers are preferred.	Strong communication skills in English are required.	Translate language background preferences into skill requirements.
8	Male workers are needed for this construction role.	All qualified workers are encouraged to apply for this construction role.	Eliminating gender restrictions and emphasizing equal opportunities.

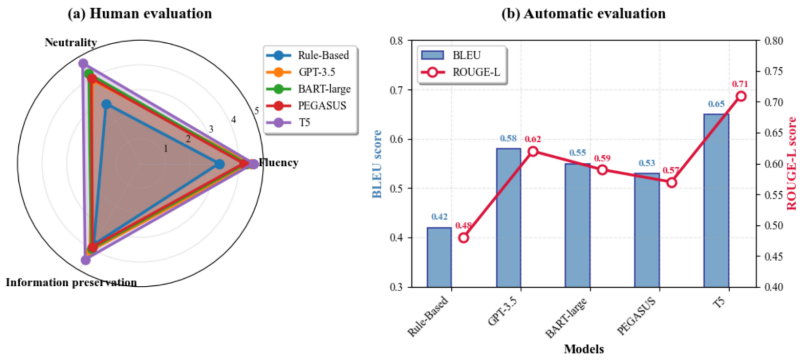

FIGURE 7. Biased text optimization results: (a) Manual evaluation; (b) Automatic evaluation

ency (3.2 points). This implies that in complex semantic scenarios, simple string replacements are unable to adequately address bias and struggle to preserve natural text flow. GPT-3.5 (Generative Pre-trained Transformer 3.5) demonstrates strong language generation capabilities, achieving a high score in fluency (4.1 points), but its neutrality (3.9 points) in bias elimination remains insufficient, reflecting the limitations of large, general-purpose language models without targeted bias training. BART-large and PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence models) performed well across all dimensions but were not particularly outstanding. T5 model achieved optimal or near-optimal performance across all three dimensions, particularly in neutrality (4.7 points) and information retention (4.5 points). This demonstrates that the T5 model, fine-tuned based on targeted parallel corpus, can better balance the core contradictions of bias elimination and information retention.

The results of the human evaluation are further supported by Figure 7b, which also offers a quantitative performance comparison. The T5 model outperforms the other models on the BLEU metric, scoring 0.65, which shows that the accuracy and standardization of the optimization are reflected in the generated text's higher lexical and phrasal similarity to the human-written reference text. It is noteworthy that the rule-based approach scored the lowest on the BLEU metric (0.42), consistent with what was seen from the assessment of its fluency in the human evaluation. This suggests that it is difficult to generate idiomatic text through simple dictionary-based substitution. The T5 model's ranking did reflect that it had the highest automatic metric (ROUGE-L) as well (0.71), indicating that the semantic content and overall structure of the generated text are consistent with the reference text. Overall it is reasonable to conclude that T5 model performed better than the GPT-3.5, achieving well on the automatic metrics, including the degree of semantic importance being extracted or retained. To be specific the ROUGE-L gap (0.62 vs. 0.71) does imply that there is still future improvement, intuitively, to be made regarding giving a full fidelity to the original semantic confidence suggested by the source material. The extensive automatic evaluation discovered that the T5 model has the most reasonable balance of the quality of text (including confidence of meaning and semantics being retained) and minimizing bias after fine-tuning its work.

V. CONCLUSION

A thorough framework for addressing bias in recruitment texts was effectively developed by this study, which holds significant value for promoting equitable hiring practices in

the industry. The detection level was capable of accurately identifying multiple bias types, such as gender, age, and physical ability through the bimodal deployment of BERT feature extraction, BiLSTM sequence encoding, and a hierarchical attention mechanism. The analysis of the model's interpretability indicated that it could focus on important biased terms. Finally, the optimization level was capable of achieving semantic-level transformation from biased text to neutral text through a T5-based text-generation model, using conditional sequence generation work that removed discriminating language while retaining important job-related information. This paper proposes a bias detection model, underpinned by deep semantic understanding and a mechanism of attention, to successfully respond to the challenge of implicit bias detection. Furthermore, the paper develops a text generation based bias optimization strategy, to enclose a full closed-loop, end-to-end detection to intervention process. The validation of the proposed model is examined using experiments on recruitment data, which provides a reliable and effective method of automation to address bias removal. This paper provides valuable technological support to improve transparency of model decisions and trustworthiness of findings, which enhances employment equity within the industry and the responsible development of artificial intelligence through an interpretable analysis.

AUTHOR CONTRIBUTIONS

Conceptualization –Liping Feng, Na Feng and Qiaojuan Hui; methodology – Na Feng and Qiaojuan Hui; investigation – Liping Feng, Na Feng and Qiaojuan Hui; writing-original draft preparation – Liping Feng, Na Feng and Qiaojuan Hui. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by,

Ningxia Hui Autonomous Region Level First-Class Human Resource Management Professional Project (Yinchuan University of Science and Technology).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] T. Sachs, A. C. Homan, and B. Lancee, "Impression formation of majority and minority applicants during resume screening—Does processing more information reduce prejudice? Journal of Applied Social Psychology," 2024; 54(7):387-404, doi: 10.1111/jasp.13047.

[2] M. Adamovic, "Analyzing discrimination in recruitment: A guide and best practices for resume studies," International Journal of Selection and Assessment, vol. 28, no. 4, pp. 445-464, 2020, doi: 10.1111/ijsa.12298.

[3] S. Celik and M. V. Turker, "Can eye movements be a predictor of implicit attitudes? Discrimination against disadvantaged individuals during the recruitment process," Istanbul Business Research, vol. 51, no. 2, pp. 459-489, 2022, doi: 10.26650/ibr.2022.51.837555.

[4] S. Hennekam, J. Peterson, L. Tahssain-Gay, and J. P. Dumazert, "Recruitment discrimination: How organizations use social power to circumvent laws and regulations," The International Journal of Human Resource Management, vol. 32, no. 10, pp. 2213-2241, 2021, doi: 10.1080/09585192.2019.1579251.

[5] D. Hangartner, D. Kopp, and M. Siegenthaler, "Monitoring hiring discrimination through online recruitment platforms," Nature, vol. 589, no. 7843, pp. 572-576, 2021, doi: 10.1038/s41586-020-03136-0.

[6] A. Waling, A. Lyons, B. Alba, V. Minichiello, C. Barrett, M. Hughes, et al., "Recruiting stigmatised populations and managing negative commentary via social media: A case study of recruiting older LGBTI research participants in Australia," International Journal of Social Research Methodology, vol. 25, no. 2, pp. 157-170, 2022, doi: 10.1080/13645579.2020.1863545.

[7] R. Bhardwaj, N. Majumder, and S. Poria, "Investigating gender bias in bert," Cognitive Computation, vol. 13, no. 4, pp. 1008-1018, 2021, doi: 10.1007/s12559-021-09881-2.

[8] J. R. Minot, N. Cheney, M. Maier, D. C. Elbers, C. M. Danforth, and P. S. Dodds, "Interpretable bias mitigation for textual data: Reducing genderization in patient notes while maintaining classification performance," ACM Transactions on Computing for Healthcare, vol. 3, no. 4, pp. 1-41, 2022, doi: 10.1145/3524887.

[9] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A fine-tuned BERT-based transfer learning approach for text classification," Journal of Healthcare Engineering, vol. 2022, no. 1, Art. no. 3498123, 2022, doi: 10.1155/2022/3498123.

[10] A. K. Nandanwar and J. Choudhary, "Contextual embeddings-based web page categorization using the fine-tune bert model," Symmetry, vol. 15, no. 2, pp. 395, 2023, doi: 10.3390/sym15020395.

[11] K. Liu, Y. Feng, L. Zhang, R. Wang, W. Wang, X. Yuan, et al., "An effective personality-based model for short text sentiment classification using BiLSTM and self-attention," Electronics, vol. 12, no. 15, pp. 3274, 2023, doi: 10.3390/electronics12153274.

[12] L. Enamoto, A. R. A. S. Santos, R. Maia, L. Weigang, and G. P. R. Filho, "Multi-label legal text classification with BiLSTM and attention," International Journal of Computer Applications in Technology, vol. 68, no. 4, pp. 369-378, 2022, doi: 10.1504/IJCAT.2022.125186.

[13] W. Wang, Y. Zhang, Y. Sui, Y. Wan, Z. Zhao, J. Wu, et al., "Reinforcement-learning-guided source code summarization using hierarchical attention," IEEE Transactions on Software Engineering, vol. 48, no. 1, pp. 102-119, 2020, doi: 10.1109/TSE.2020.2979701.

[14] M. W. B. D. Satya, A. Luthfiarta, and M. N. Althoff, "Comparative Analysis of T5 Model Performance for Indonesian Abstractive Text Summarization," SISTEMASI, vol. 14, no. 3, pp. 1092-1106, 2025, doi: 10.32520/stmsi.v14i3.4884.

[15] E. Daraghmi, L. Atwe, and A. Jaber, "A Comparative Study of PEGASUS, BART, and T5 for Text Summarization Across Diverse Datasets," Future Internet, vol. 17, no. 9, pp. 389, 2025, doi: 10.3390/fi17090389.

[16] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1-67, 2020, doi: 10.48550/arXiv.1910.10683.

[17] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," ACM Computing Surveys, vol. 56, no. 3, pp. 1-37, 2023, doi: 10.1145/3617680.

[18] R. Rao, S. Sharma, and N. Malik, "Automatic text summarization using transformer-based language models," International Journal of System Assurance Engineering and Management, vol. 15, no. 6, pp. 2599-2605, 2024, doi: 10.1007/s13198-024-02280-4.

[19] J. C. Timoneda and S. V. Vera, "BERT, RoBERTa, or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text," The Journal of Politics, vol. 87, no. 1, pp. 347-364, 2025, doi: 10.1086/730737.

[20] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa," Applied Intelligence, vol. 51, no. 6, pp. 3522-3533, 2021, doi: 10.1007/s10489-020-01964-1.

[21] W. Safira, B. Prabaswara, A. Stevens Karnyoto, and B. Pardamean, "Leveraging albert for sentiment classification of long-form chatgpt reviews on twitter," International Journal of Computing and Digital Systems, vol. 17, no. 1, pp. 1-12, 2024, doi: 10.12785/ijcds/1570999256.

[22] M. La Quatra and L. Cagliero, "Bart-it: An efficient sequence-to-sequence model for italian text summarization," Future Internet, vol. 15, no. 1, pp. 15, 2022, doi: 10.3390/fi15010015.

[23] R. K. Dey and A. K. Das, "Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis," Multimedia Tools and Applications, vol. 82, no. 21, pp. 32967-32990, 2023, doi: 10.1007/s11042-023-14653-1.