IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 15, Issue 5, May 2025, ||Series -1|| PP 08-14

Explainable AI for Trustworthy recruitment: Recognizing Bias in Job Descriptions Using RoBERTa

Mr. Ayush Warudkar
Department of Artificial Intelligence
GHRCEM, Nagpur

Mr. Hrushikesh Shenwai
Department of Artificial Intelligence
GHRCEM, Nagpur

Mr. Vedant Mankar
Department of Artificial Intelligence
GHRCEM, Nagpur

Mr. Lokesh Kamble
Department of Artificial Intelligence
GHRCEM, Nagpur

Miss. Mrudula Nanoti
Department of Artificial Intelligence
GHRCEM, Nagpur

Mr. Abhijeet Pardhi
Department of Artificial Intelligence
GHRCEM, Nagpur

Prof. Sweta Bokade
Department of Artificial Intelligence
GHRCEM, Nagpur

Received 01 May 2025; Accepted 11 May 2025

Abstract: *The general public seldom acknowledges job description bias because it remains widely unrecognized as it seriously affects both candidate variety and hiring procedure inclusivity. The research has developed Explainable AI for Trustworthy recruitment as an Artificial Intelligence system based on modern Natural Language Processing approaches to detect unintended biases in job listings. Various traditional tools based on keyword matching differ from the Explainable AI for Trustworthy recruitment system because it uses RoBERTa transformer model with contextual understanding to discover subtle intersectional biases which include gender and racial as well as age and disability aspects. The system identifies discriminatory wording to enable recruiters in writing job descriptions free of prejudice. The Explainable AI system for Trustworthy recruitment enhances both recruitment system effectiveness and universal hiring reach as well as working toward diverse workplace diversity.*

Keywords: *Bias Detection, Job Descriptions, RoBERTa, BERT, Multi-label Classification, Explainable AI (XAI), SHAP, Circumstantial Embeddings, Intersectional Bias, Trustworthy recruitment Practices, NLP, Transformer Models.*

I. Introduction:

The initial meeting between employers and candidates is job descriptions that possess critical significance during this evolving phase of inclusive hiring. These descriptions contain numerous unperceived biases which residents often fail to identify. Such biases caused by gender and race and age and disability and sexuality affect candidate validation and sustain devaluation of marginalized employee groups [2][3][4].

The traditional method of keyword detection performed poorly in measuring job listing bias since it disregards the overall meaning within the text. The detection system suffers from insufficient generalization together with inability to identify intersectional bias which covers multiple simultaneous bias types found in single sentences.

Almost all current systems function with no transparency regarding their prediction processes. We introduce a context-sensitive along with explainable bias detection system using transformer-derived NLP models to address this gap. RoBERTa model underwent multi-label classification training to detect seven specific bias categories including feminine, masculine, general, racial, age, disability and sexuality [7].

The system performs sentence-level evaluation of job descriptions to identify bias cases. Our system includes SHAP (SHapley Additive exPlanations) to explain bias predictions through the identification of key words that drive each prediction result. The calculation of Shapley values depends on the SHapley Additive exPlanations (SHAP) algorithm because it represents an excellent method to explain machine learning models' predictions. The system allows users to identify bias classifications alongside details explaining the factors that triggered detection while dealing with problems that deep learning classifiers have regarding interpretation. [8]

We trained and assessed the model using the Hugging Face Transformers library and PyTorch. The system achieved strong performance, with RoBERTa reaching 69.86% Accuracy, 91.48% Precision, 76.06% F1-score, 11.6031 Log Loss, 82.37% AUC, 0.7526 MCC, and 0.0543 Hamming Loss, outperforming the BERT baseline

[7]. A Streamlit interface was created to ease real-time interaction, enabling users to input job descriptions, receive bias analysis, and visually interpret flagged language.

Briefly, our work fills the drawbacks of earlier systems by combining circumstantial understanding [1][3][7], multi-label identification, and explainability [8] in a single, available solution — serving organizations recognize and turn down bias in job descriptions and upgrade the broadness of hiring practices [2][4][5].

II. Literature Survey:

The attention towards bias that affects hiring technologies has markedly increased in recent times particularly when focusing on Large Language Models (LLMs). Research investigations highlighted in this section have motivated our work approach and methodology.

2.1 Bias in Resume Screening:

Wilson and Caliskan [1] researched resume screening systems for intersectional bias using Massive Text Embedding (MTE) models. According to their research white males consistently received beneficial treatment yet black males received the most substantial unfavorable treatment throughout every examination process. Several document features together with resume length along with name repetition influenced the model outcomes according to the study. The current imperfect retrieval-based models show how systems which adapt through multiple context considerations remain important.

2.2 Gender Bias in Embeddings:

Nomelini and Marcolin [2] conducted a study which used Word2Vec embeddings to examine gender bias in job postings. The study revealed that job descriptions contained numerous feminine phrases which matched each other differently based on the selected embedding methods and dimensional structure. The investigation demonstrated that gender bias appears frequently in representing gender within corpora so we chose to use transformer-based model training for addressing this issue.

2.3 Circumstantial Classification in STEM Ads:

A semantic clustering method made by Dikshit et al. [3] implemented contrastive learning to evaluate academic job postings for their agentic, communal, and balanced profiles. Statistical analysis revealed that agentic language occurred mainly in lower parts of job description texts yet this practice seemed to deter female candidates from STEM related fields. The research investigation significantly influenced our methods for detecting sentence-level bias through the use of circumstantial encoders including RoBERTa.

2.4 Perceived Bias in IT/SE Job Ads:

content. Kanij et al. [4] compiled survey results from hiring professionals to determine the actual existence of bias in job advertisement text. A majority of people (56%) took deliberate action to combat bias in their work while the rest acknowledged its presence. The study revealed that female candidates value job advertisement language which includes basic dictionary terms in a direct format with adaptable expressions and provides concrete explanations instead of only pointing out issues.

2.5 Wider Societal-Technical Viewpoint:

Zang evaluated algorithmic systemic bias from the framework of public interest technology in his work [5]. The research based on Facebook advertising tools revealed how market and legal frameworks may create discriminatory results. Using Explainable AI (SHAP) allows our system to integrate while maintaining full accountability in automated hiring procedures as per a broader comprehension of societal-technical issues.

III. Methodology:

The system detects biases within job descriptions through the combination of circumstantial Natural Language Processing methods and multi-label classification features and Explainable AI frameworks. The segment presents an all-encompassing look at the complete processing pipeline which contains steps for both data preprocessing and model design as well as training approach and performance assessment with user interface development.

3.1 Data Cleaning:

The collection includes over 4,000 job descriptions that have received classifications into seven possible bias segments.

- **Gender-related:** feminine, masculine
- **Identity-related:** racial, age, disability, sexuality
- **General bias**

The data was categorized into:

- **Training set:** 3,000+ sentences.
- **Validation set:** ~600 sentences.
- **Test set**: ~600 sentences.

The preprocessing step removed special characters and standardized the text before splitting the sentences into tokens through RoBERTa tokenizer operation. All inputs received adjustment to reach a maximum length of 512 tokens. The Hugging Face Dataset API provided an organizational system for data storage.

3.2 Model Architecture:

The examination of bias detection in job descriptions used BERT-base and RoBERTa-base transformer architectures during experimental testing. These text classification systems employ different strategies for pretraining although they share similarities when processing text documents into many categories**.**

BERT-based Model

The BERT-base (uncased) model functions with a bidirectional transformer encoder after its training for masked language modeling and next sentence prediction. In our design:

- Bert encoder splits word tokens into pieces before creating embedding information that contains contextual content.
- The representation produced by [CLS] token goes through a dense classification head to produce prediction results.
- Operation of sigmoid functions enables the generation of separate probabilities for each of the seven bias types.
- During processing each sentence becomes a 7-dimensional output vector y ^∈[0,1]^7 while the individual vector elements indicate specific bias category probabilities..

Roberta-based Model

The chosen model for final implementation became RoBERTa-base by getting rid of sentence prediction while adopting byte-level BPE tokenization and extensive large corpus pretraining.

The design incorporates:

- A Every token within a Roberta encoder produces outstanding circumstantial embedding outputs.
- The [CLS] representation enters a classification layer that produces output nodes using seven categories for bias recognition.
- The final outputs from Sigmoid activation represent probabilities between 0 and 1 which correspond to the different bias categories.
- BERT generates a vector y ^∈[0,1]7 as an output equivalent to its own operations.

The above combination of being implemented in the same multi-label classification framework led to RoBERTa's selection as the final model because it presented superior Accuracy, F1-score and narrower miscategorization rates (refer to section 4).

3.3 Training Configuration:

The Hugging Face Trainer API enabled the model optimization with parameters set as follows:

- **Learning rate:** 3×10^{-5}
- **Batch size:** 12 (train), 14 (eval)
- **Epochs:** 5
- **Weight decay:** 0.01
- **Loss function:** Binary Cross-Entropy (BCE) for multi-label classification

A checkpoint with the minimum validation loss was selected as the most suitable one.

3.4 Evaluation Metrics:

Our performance assessment included measuring accuracy, precision, recall and F1-Score as well as AUC and Matthews Correlation Coefficient and Log Loss with Hamming Loss calculations.

- **Accuracy, Precision, Recall, F1-Score**
- **AUC (Area Under the Curve)**
- **Matthews Correlation Coefficient (MCC)**
- **Log Loss**
- **Hamming Loss**

Classification reports together with confusion matrices allowed us to evaluate the prediction accuracy for individual bias categories made by our model.

3.5 Explainability with SHAP:

The SHAP analysis (SHapley Additive exPlanations) served to improve openness capabilities in the model. SHAP calculates value contributions of all words present in the input sentence that ultimately produces the predictive output. Users can use this capability to identify bias-related words that the model detected in order to understand its predictive selections.

3.6 Streamlit-Based Interface:

The Streamlit framework enabled design of a web interface which provided easy accessibility to users who lacked technical knowledge. Key features include:

- Input Users can input job descriptions through the provided dialogue box.
- Running an analysis requires users to click the 'Analyze Bias' button which starts the model inference process.
- The tool presents detected bias types and confidence rating measurements to its users.
- Color codes through highlight allow users to identify biased words and phrases in their SHAP value ranges. Red highlights demonstrate high bias levels.

The platform provides an easy-to-use interface that helps human resource experts and recruiters make real-time inclusive job post creations.

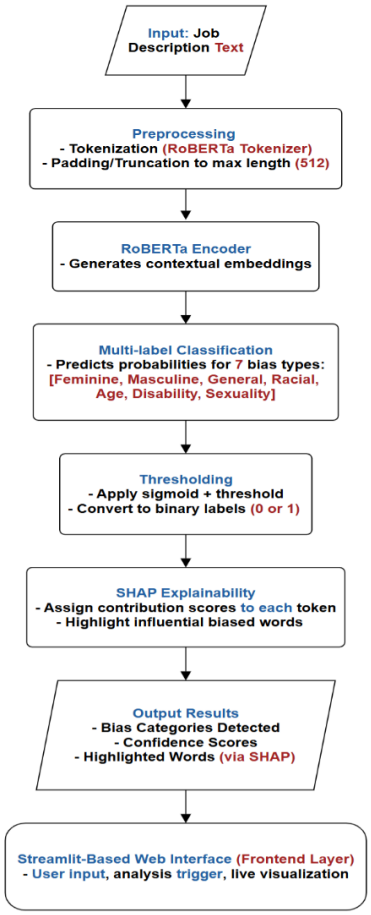

Figure 1: Flowchart of the Bias Detection Process.

A Streamlit interface presents the system architecture which begins with job description entry followed by multistage analysis of biases and SHAP-based visual representation.

IV. Experiments & Result:

Our bias detection system achieved evaluation through multiple tests that used both BERT and RoBERTa models for analysis on more than 4,000 labeled job description sentences.

4.1 Model Performance Comparison:

The evaluation of models took place through standard assessment methods for multi-label classification. Throughout the evaluation tests the RoBERTa model demonstrated superior performance than the BERT Model in every scoring category:

Metric	BERT	RoBERTa
Accuracy (%)	63.01	69.86
Precision (%)	81.25	91.48
Recall (%)	62.97	66.43
F1-Score (%)	67.43	76.06
AUC (%)	73.24	82.37
Log Loss	14.63	11.60
MCC	0.6080	0.7526
Hamming Loss	0.0895	0.0543

The research findings validate the capacity of RoBERTa's circumstantial model to detect different and connected types of bias that exist in written materials.

4.2 Confusion Matrix Analysis:

Evaluation of classification precision relied on studying the test set confusion matrices from BERT and RoBERTa. The seven different bias categories show correct grading results through these data matrices.

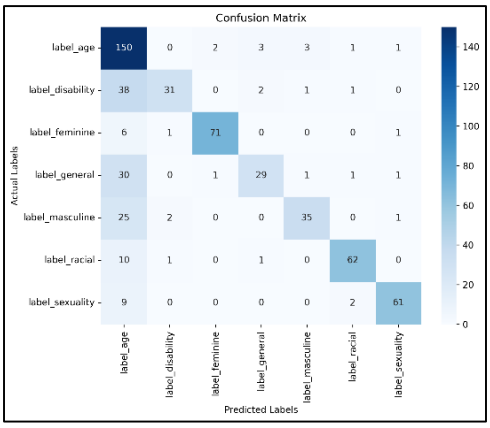

Figure 2: BERT Confusion Matrix

The predictions from BERT contained errors because the label_masculine and label_general categories often matched incorrect predictions to label_age. When the model scored feminine and racial labels accurately it had average confusion levels throughout the process.

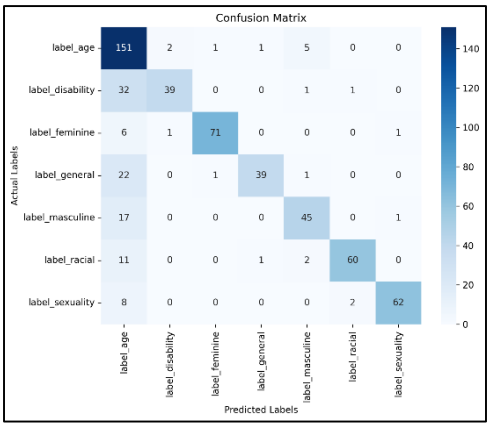

Figure 3: RoBERTa Confusion Matrix

The accuracy levels of RoBERTa exceeded those of BERT while producing fewer incorrect matches. Label_masculine and label_general underwent major advancements in RoBERTa which produced comprehensive and precise predictions for each of the seven bias categories when compared to BERT's performance.

V. Discussion:

The research results demonstrate that Roberta achieves superior performance compared to BERT when detecting bias patterns in job advertising texts. Through maximal accuracy and f1-score and minimal error rates RoBERTa proves better than other models at detecting circumstantial and intersectional bias. Context-aware models prove superior to classic keyword-based systems as well as embedding-only approaches for handling bias detection tasks.

Model decisions and their bias detection methods became clearer through SHAP explainability because the system identified the precise words that created bias. Such enhancements provide multiple benefits including better user confidence while also making the tool practical for HR professionals. The system continued to display overlapping labels which mainly affected the differentiation of age bias from disability bias categories because of their comparable linguistic features. Despite positive outcomes from the model across multiple labels additional development opportunities exist to enhance performance when working with large diverse datasets in particular concerning sexual orientation and other protected groups' bias types.

This system delivers three key features through its integrated precision, explainability and high user-friendliness. Non-technical users can navigate the Streamlit interface to get instant feedback on their job descriptions without any technical difficulties which helps them create more inclusive job content.

VI. Conclusion:

The project developed a system which examined job description bias by analyzing the surrounding text of each statement. The system based on highly refined RoBERTa model achieved better results by using multi-label classification to detect seven definite bias types while surpassing BERT performance in essential metrics. Through SHAP implementation the system provided sentence-level explanations to users about flagging reasons but delivered it via the easy-to-use Streamlit interface.

The research demonstrates how transformer-based models can detect hidden biases which conventional inspection methods miss. The system operates to detect biased language while assisting organizations with diverse job ad development in order to achieve unbiased hiring.

Researchers should extend the data collection while developing algorithms that reduce human error and implement automatic text generation capabilities for live rephrasing of biased content.

References:

[1] K. Wilson and A. Caliskan, "Gender, race, and intersectional bias in resume screening via language model retrieval," *Proc. AAAI/ACM Conf. on AI, Ethics, and Society (AIES)*, Seattle, WA, USA, 2024.

[2] G. C. Nomelini and C. B. Marcolin, "Gender bias in large language models: A job postings analysis," *Revista de Administração Mackenzie*, vol. 25, no. 6, pp. 1–27, 2024.

[3] M. Dikshit, H. Bouamor, and N. Habash, "Investigating gender bias in STEM job advertisements," *Proc. Workshop on Gender Bias in NLP (GeBNLP)*, 2024.

[4] T. Kanij, J. Grundy, and J. McIntosh, "Enhancing understanding and addressing gender bias in IT/SE job advertisements," *J. of Information and Software Technology*, 2024.

[5] J. Zang, *Case Studies in Public Interest Technology*, Ph.D. dissertation, Harvard Univ., MA, USA, 2021.

[6] A. J. Koch, S. D. D'Mello, and P. R. Sackett, "A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making," *J. Appl. Psychol.*, vol. 99, no. 4, pp. 128–161, 2015.

[7] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint*, arXiv:1907.11692, 2019.

[8] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] J. Sánchez-Monedero, L. Dencik, and L. Edwards, "What does it mean to 'solve' the problem of discrimination in hiring?: Social, technical and legal perspectives from the UK on automated hiring systems," *Proc. FAT* '20, pp. 458–468, 2020.

[10] W. Chan et al., "KERMIT: Generative insertion-based modeling for sequences," *arXiv preprint*, arXiv:1906.01604, 2019.

[11] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," in *Machine Learning Challenges*, Springer, 2006.

[12] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[13] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *IEEE Symp. on Security and Privacy (SP)*, 2016, pp. 598–617.

[14] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[16] D. Sundararaman and V. Subramanian, "Debiasing gender bias in information retrieval models," *arXiv preprint*, arXiv:2208.01755, 2022.

[17] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] L. Wang et al., "Improving text embeddings with large language models," *arXiv preprint*, arXiv:2401.00368, 2023.

[19] S. Leavy, "Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning," in *Proc. Int. Conf. on Software Engineering*, 2018, pp. 14–16.

[20] M. Meyer, A. Cimpian, and S. J. Leslie, "Women are underrepresented in fields where success is believed to require brilliance," *Frontiers in Psychology*, vol. 6, 2015.

[21] T. Mikolov et al., "Efficient estimation of word representations in vector space," *Proc. Workshop at ICLR*, 2013.

[22] R. H. Bernstein et al., "Assessing gender bias in particle physics and social science recommendations for academic jobs," *Social Sciences*, vol. 11, no. 2, p. 74, 2022.

[23] M. P. Born and T. W. Taris, "The impact of the wording of employment advertisements on students' inclination to apply for a job," *The Journal of Social Psychology*, vol. 150, no. 5, pp. 485–502, 2010.
[24] B. Casad et al., "Gender inequality in academia: Problems and solutions for women faculty in STEM," *J. of Neuroscience Investigation*, vol. 99, no. 1, pp. 13–23, 2020.
[25] S. Campero, "Hiring and intra-occupational gender segregation in software engineering," *Am. Sociol. Rev.*, vol. 86, no. 1, pp. 60–92, 2021.
[26] G. Catolino et al., "Gender diversity and women in software teams: How do they affect community smells?" in *Proc. IEEE/ACM 41st Int. Conf. on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, 2019, pp. 11–20.
[27] J. Angwin and N. Scheiber, "Dozens of companies are using Facebook to exclude older workers from job ads," *ProPublica*, Dec. 2017.
[28] Applied, "Scaling fast: how to get hiring right," *Technical Report*, 2019.
[29] J. B. Merrill and A. Tobin, "Facebook is letting job advertisers target only men," *ProPublica*, Sept. 2018.