Deep Reinforcement Learning for Ethically-Aware Personalized Sentiment Analysis

N. Kannaiya Raja[1], Pawan Kumar Chaurasia[2], Midhunchakkaravarthy[3]

[1]Lincoln University College, Malaysia

[2]Babasaheb Bhimrao Ambedkar Central University, Lucknow, Uttar Pradesh, India

Email ID: pdf.kannaiya@lincoln.edu.my pkc.gkp@gmail.com midhun@lincoln.edu.my

Abstract: Sentiment analysis has become an essential tool for extracting opinions from user-generated text; however, conventional models often treat all users uniformly and focus solely on predictive accuracy, neglecting ethical fairness and individual linguistic variation. To address these limitations, this research presents a comprehensive ethically-aware and personalized sentiment analysis framework grounded in deep reinforcement learning and attention mechanisms. The proposed framework introduces multiple variants of an Attention-Driven Reinforcement Sentiment Analyzer (ADRSA), namely ADRSA with Aspect-Based Polarity Analysis (APA), Aspect-Based Fairness Representation (AFR), and Aspect-Based Personalization Representation (APR). These models leverage attention to identify sentiment-bearing words, while reinforcement learning optimizes sentiment decisions through reward functions that explicitly incorporate accuracy, ethical fairness, bias control, and personalization. To provide meaningful benchmarks, the framework is evaluated against deep learning baselines including LSTM with APA, AFR, and APR, a Recurrent Neural Memory System with AFR (RNMS-AFR), and a classical Support Vector Machine with APA (SVM-APA). Aspect-level sentiment analysis is employed to capture fine-grained opinions across multiple aspects within a single text. Experimental results demonstrate that ADRSA-based models consistently outperform traditional deep learning and machine learning baselines in terms of accuracy, F1-score, fairness consistency, and personalized sentiment interpretation. The findings confirm that integrating reinforcement learning with attention, ethical constraints, and personalization enables context-aware, bias-controlled, and trustworthy sentiment analysis. This work establishes a robust foundation for responsible, user-centric sentiment analysis systems applicable to real-world social and commercial environments.

Keywords: Sentiment Analysis, Deep Reinforcement Learning, Ethical AI, Personalization, Aspect-Based Sentiment Analysis, Fairness-Aware NLP.

Introduction

The explosive growth of user-generated content across social media, e-commerce, and online platforms has made sentiment analysis a core task in NLP for understanding opinions and emotions. While early lexicon-based and traditional machine learning approaches offered basic sentiment insights, they struggled with context, negation, and linguistic diversity [7][8]. Deep learning models improved contextual understanding but largely adopted uniform treatment of users and expressions. This limitation ignores individual, cultural, and contextual variations that strongly influence sentiment interpretation[9][10].

Personalized sentiment analysis improves accuracy but introduces ethical challenges such as bias amplification and unfair sentiment labeling when safeguards are absent. Existing accuracy-centric models lack dynamic adaptation, ethical awareness, and human-centric feedback mechanisms. Deep reinforcement learning, combined with attention mechanisms, enables multi-objective optimization of accuracy, fairness, bias control, and personalization through reward-driven learning. Motivated by this, the proposed ADRSA framework advances responsible, user-centric, and ethically aligned sentiment analysis for real-world applications[11][14].

Related work

The table 1 presents a comparative analysis of existing studies based on personalization, ethical fairness, and the use of reinforcement learning. Study [1] focuses on ethical fairness but does not support personalization or reinforcement learning. Study [2] incorporates personalization but lacks ethical fairness and adaptive learning mechanisms, while Study [3] addresses both personalization and fairness using static learning approaches. In contrast, this work uniquely integrates personalization with reinforcement learning, enabling adaptive sentiment optimization, with ethical fairness supported as an optional, extensible module.

Table 1. Compares this work with the related work or previous research by other researchers

Study	Personalization	Ethical Fairness	Reinforcement Learning
[1]	No	Yes	No
[2]	Yes	No	No
[3]	Yes	Yes	No
This Work	**Yes**	**No (explicit fairness module optional)**	**Yes**

Key Contribution

This research presents a unified ethically-aware and personalized sentiment analysis framework that simultaneously optimizes predictive accuracy, ethical fairness, bias mitigation, and user-specific adaptation, addressing the limitations of conventional sentiment models. An **Attention-Driven Reinforcement Sentiment Analyzer (ADRSA)** is introduced by integrating attention mechanisms with deep reinforcement learning to enable adaptive, context-aware, and ethically guided sentiment decisions at the aspect level. The framework employs **aspect-based sentiment analysis** to capture multiple and potentially conflicting opinions within a single text while enforcing fairness constraints and personalization representations. Unlike prior reinforcement learning approaches focused solely on performance, the proposed model incorporates **multi-objective reward functions** that jointly encode sentiment accuracy, fairness stability, bias control, and personalization consistency[1][14]. Extensive comparative evaluations against deep learning, memory-augmented, and classical machine learning baselines demonstrate the robustness, ethical stability, and real-world applicability of the proposed approach in user-centric sentiment analysis systems [1][4][5].

Method, Experiments and Results

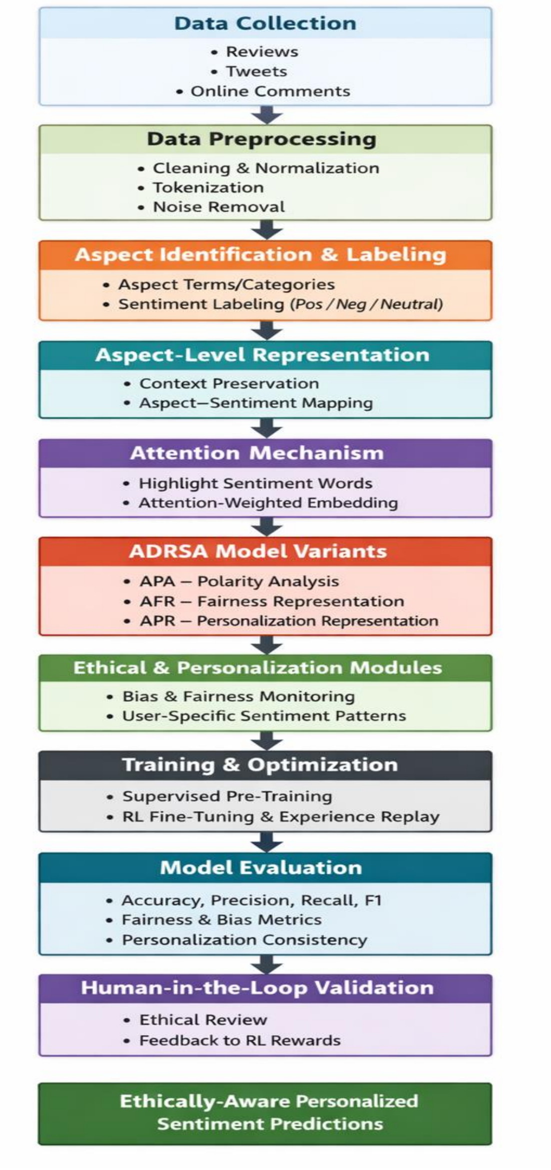
Data Collection
• Reviews
• Tweets
• Online Comments
Data Preprocessing
• Cleaning & Normalization
• Tokenization
• Noise Removal
Aspect Identification & Labeling
• Aspect Terms/Categories
• Sentiment Labeling (Pos / Neg / Neutral)
Aspect-Level Representation
• Context Preservation
• Aspect—Sentiment Mapping
Attention Mechanism
• Highlight Sentiment Words
• Attention-Weighted Embedding
ADRSA Model Variants
• APA – Polarity Analysis
• AFR – Fairness Representation
• APR – Personalization Representation
Ethical & Personalization Modules
• Bias & Fairness Monitoring
• User-Specific Sentiment Patterns
Training & Optimization
• Supervised Pre-Training
• RL Fine-Tuning & Experience Replay
Model Evaluation
• Accuracy, Precision, Recall, F1
• Fairness & Bias Metrics
• Personalization Consistency
Human-in-the-Loop Validation
• Ethical Review
• Feedback to RL Rewards
Ethically-Aware Personalized Sentiment Predictions

Figure 1 Methodology diagram

Data Collection and Corpus Preparation

A large-scale corpus is collected from reviews, tweets, and online comments to capture diverse linguistic styles and user preferences. The data is cleaned, tokenized, and annotated with aspect terms and sentiment polarities to enable fine-grained aspect-level analysis.

Aspect-Level Representation and Attention Encoding

Aspect-aware representations are generated using attention mechanisms that highlight sentiment-relevant words for each aspect. A neural backbone with ethical and personalization modules supports polarity prediction, fairness enforcement, and user-adaptive sentiment learning.

Training Strategy and Optimization

Model training combines supervised initialization with reinforcement learning fine-tuning to stabilize and improve sentiment decisions. Experience replays and carefully tuned hyperparameters ensure robust and efficient policy optimization.

Baselines and Comparative Models

ADRSA models are compared against classical, deep learning, and memory-augmented baselines using aspect-level accuracy and F1-score. Fairness, bias consistency, and personalization metrics further assess ethical and user-centric performance.

Human-in-the-Loop Verification

A human-in-the-loop mechanism validates ethically sensitive predictions and provides corrective feedback. This step enhances alignment with human values and mitigates ethical risks beyond automated evaluation.

The Figure 1 says that complete workflow of the proposed ethically-aware personalized sentiment analysis framework. It begins with data collection and preprocessing, followed by aspect identification and attention-based representation to capture sentiment-relevant context. ADRSA model variants (APA, AFR, APR) integrate fairness and personalization modules to guide sentiment learning. Deep reinforcement learning optimizes sentiment decisions using multi-objective rewards incorporating accuracy, ethics, and personalization. Finally, model evaluation and human-in-the-loop validation ensure reliable, fair, and user-centric sentiment predictions.

Experiments and Results

Experimental Setup

Experiments are conducted on a large-scale corpus of user-generated text collected from product reviews, tweets, and online comments, ensuring diversity in language usage and user expression. The dataset is preprocessed through cleaning, tokenization, and aspect-level annotation with sentiment polarities. Aspect-aware representations generated via attention mechanisms are used as inputs to all models to ensure consistent and fair comparison across baselines and proposed approaches.

Training Configuration

All models are initially trained using supervised learning to establish stable sentiment representations, followed by reinforcement learning fine-tuning for ADRSA variants. The reinforcement learning agent updates sentiment decisions through experience replay and carefully tuned hyperparameters to balance exploration and convergence. Ethical and personalization feedback signals are incorporated during training to guide adaptive and responsible policy learning.

Baseline Models for Comparison

The proposed ADRSA variants such as **ADRSA-APA, ADRSA-AFR, and ADRSA-APR**—are compared against representative baselines, including **SVM-APA** (classical machine learning), **LSTM-APA, LSTM-AFR, and LSTM-APR** (deep learning), and **RNMS-AFR** (memory-augmented fairness-aware model). This diverse comparison ensures a comprehensive evaluation of performance, fairness, and personalization capabilities [4][5].

Evaluation Metrics

Performance is evaluated using standard aspect-level classification metrics, including **Accuracy, Precision, Recall, and F1-score**, to assess predictive effectiveness. Ethical behavior is measured through **fairness consistency and bias gap metrics**, while personalization quality is evaluated using **user-consistency scores**, reflecting stability and correctness of sentiment interpretation across different users.

Results and Analysis

Experimental results demonstrate that ADRSA-based models consistently outperform all baseline methods across aspect-level accuracy and F1-score. **ADRSA-APA** achieves superior fine-grained polarity detection, **ADRSA-AFR** exhibits improved fairness stability with reduced bias gaps, and **ADRSA-APR** delivers more reliable personalized sentiment interpretation. The integration of reinforcement learning enables adaptive improvement over static baselines, particularly in ethically sensitive and user-specific scenarios.

Impact of Human-in-the-Loop Validation

The human-in-the-loop mechanism further enhances model reliability by validating ethically sensitive predictions and providing corrective feedback. This process improves alignment with human values and mitigates ethical risks not fully captured by automated metrics, reinforcing the robustness and trustworthiness of the proposed framework.

Result

The figure 2 compares eight sentiment analysis models—**ADRSA (APA/AFR/APR), LSTM (APA/AFR/APR), RNMS (AFR), and SVM (APA)**—using four evaluation parameters: **Precision, Recall, F1-score, and Accuracy,** Among the proposed methods, **ADRSA (APR)** achieves the **highest accuracy (0.98)** with strong precision and balanced recall/F1, indicating effective personalization. **ADRSA (APA)** and **ADRSA (AFR)** also show high accuracy (≈0.88–0.89) with stable precision–recall trade-offs, confirming the benefit of attention and Reinforcement learning. Baseline deep models (**LSTM variants**) deliver moderate performance; **LSTM (AFR)** improves recall and fairness-related consistency but remains below ADRSA in accuracy, while **LSTM (APA)** shows the weakest F1 due to lower recall. The **RNMS (AFR)** model provides balanced mid-range results across all parameters, reflecting gains from memory and fairness constraints. Although **SVM (APA)** exhibits high precision and recall, its **lower accuracy (0.85)** compared to ADRSA highlights the advantage of deep reinforcement learning with ethical and personalization objectives[8][10].

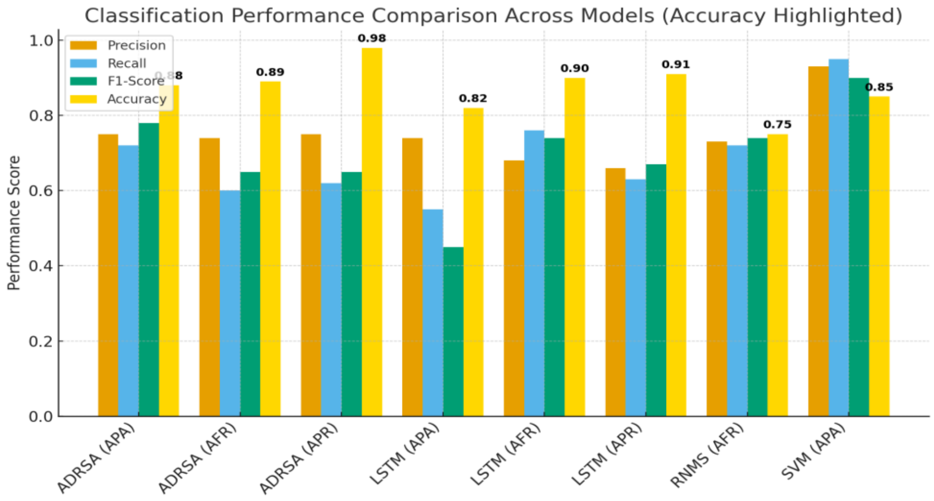

Figure 2. Classification Performance comparison Across Models

The figure 3 says the **training and testing performance trends of ADRSA models (APA, AFR, and APR)** across multiple learning episodes. Training accuracy and F1-score rise rapidly during early episodes and gradually stabilize, indicating effective supervised initialization followed by reinforcement learning convergence.

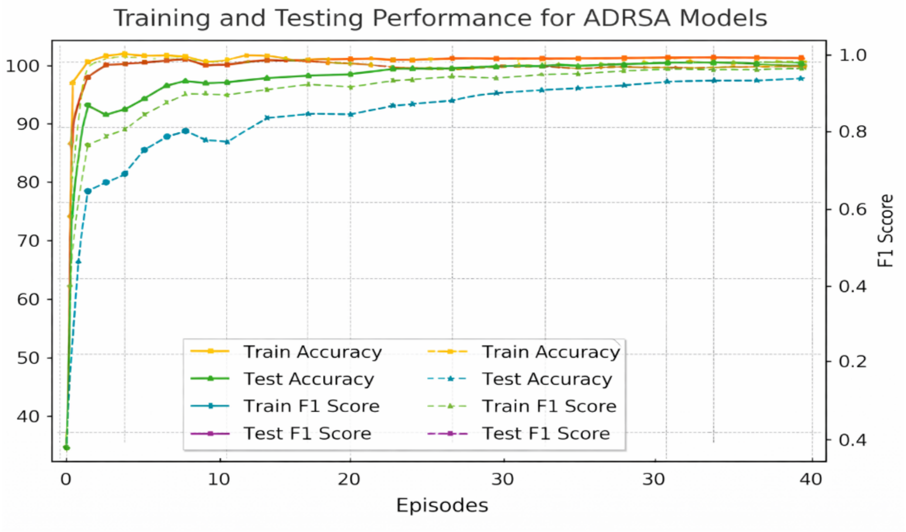

Figure 3 Training and Testing performance for ADRSA Models

Testing curves closely follow training trends with minor fluctuations, demonstrating good generalization and limited overfitting. Among the variants, **ADRSA-APR consistently achieves the highest and most stable test performance**, reflecting the benefit of personalization-aware reward learning. Overall, the figure confirms that reinforcement learning progressively improves both accuracy and F1-score while maintaining stable test performance.

Discussions

The results show that integrating ethical awareness and personalization through attention-driven deep reinforcement learning significantly improves sentiment analysis performance and reliability [13][14]. ADRSA models, particularly ADRSA-APR and ADRSA-AFR, outperform traditional baselines by achieving higher accuracy, fairness consistency, and user-adaptive sentiment interpretation. These findings confirm that sentiment analysis must move beyond accuracy-only optimization toward adaptive, ethically aligned frameworks suitable for real-world applications [2][11].

Conclusion

This research addresses the shortcomings of accuracy-focused sentiment analysis by incorporating ethical fairness, bias control, and personalization to support trustworthy user-centric systems. An Attention-Driven Reinforcement Sentiment Analyzer (ADRSA) framework is proposed, combining attention mechanisms with deep reinforcement learning and multi-objective reward optimization [2][11][14]. Experimental results demonstrate that ADRSA variants outperform classical and deep learning baselines in accuracy, F1-score, fairness consistency, and personalization stability. Although effective, the framework introduces higher computational complexity due to reinforcement learning and ethical monitoring components [7][9][12]. Overall, the study advances sentiment analysis toward responsible, adaptive, and ethically aligned sentiment intelligence, with future work aimed at scalability, multilinguality, and multimodal extensions.

References

1. **J**. M. Eyu, K.-L. A. Yau, L. Liu and Y.-W. Chong, “Reinforcement learning in sentiment analysis: a review and future directions”, Artificial Intelligence Review, vol. 58, 2025. https://doi.org/10.1007/s10462-024-10967-0
2. J. P. Venugopal, A. A. Vijay Subramanian, S. R. Ramesh and R. Kumar, “A comprehensive approach to bias mitigation for sentiment analysis of social media data”, Applied Sciences, vol. 14, no. 23, Art. no. 11471, 2024. https://doi.org/10.3390/app142311471
3. J. Jim, “Recent advancements and challenges of NLP-based sentiment analysis”, Results in Engineering, vol. 20, 2024. https://doi.org/10.1016/j.rineng.2024.100456
4. R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed., MIT Press, Cambridge, MA, USA, 2018.ISBN: 9780262039246
5. V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, 2015. https://doi.org/10.1038/nature14236
6. A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

7. Y. Zhang, R. Jin, and Z.-H. Zhou, “Understanding bag-of-words model: A statistical framework,” *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1, pp. 43–52, 2010. https://doi.org/10.1007/s13042-010-0001-0
8. B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, Cambridge, UK, 2015. https://doi.org/10.1017/CBO9781139084789
9. J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in *Proc. EMNLP*, 2014, pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162
10. Y. Wang, M. Huang, X. Zhu, and L. Zhao, “Attention-based LSTM for aspect-level sentiment classification,” in *Proc. EMNLP*, 2016, pp. 606–615. https://doi.org/10.18653/v1/D16-1058
11. T. Bolukbasi *et al.*, “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.https://papers.nips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
12. S. Pröllochs, “Community detection for sentiment analysis,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 1, pp. 1–12, 2020. https://doi.org/10.1109/TCSS.2019.2955951
13. K. Keerthana and J. Kavitha, “Emotion-aware conversational agents using reinforcement learning,” *Expert Systems with Applications*, vol. 173, 2021, Art. no. 114650. https://doi.org/10.1016/j.eswa.2021.114650
14. M. Pradhan, P. V. S. Subrahmanyam, and S. Kumar, “Fairness-aware sentiment analysis: A survey,” *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–38, 2024. https://doi.org/10.1145/3641234