Analyzing Bias Trade-Offs in Movie Review Sentiment Analysis using a BERT - SVM Enhanced Model

Vany Eka Karunia
Informatics Engineering
Faculty of Computer Science
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
vanyek@students.amikom.ac.id

Hastari Utama
Informatics Engineering
Faculty of Computer Science
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
utama@amikom.ac.id

Abstract—Sentiment analysis of movie reviews often exhibits genre-based bias, where model performance varies significantly across subgroups—an issue that standard accuracy metrics can mask. To address this, we propose a novel fairness-aware hybrid model, BERT-SVM (Fairness-Tuned), which integrates sample re-weighting focused on the lowest-performing genre into the BERT-SVM pipeline. Using a public IMDb movie review dataset from Kaggle, we first train a standard BERT-SVM model and identify Horror as the weakest-performing genre (accuracy: 72.3%, vs. overall 89.6%). We then apply targeted re-weighting to upsample underrepresented or misclassified Horror samples during training. The Fairness-Tuned model reduces the accuracy gap by 62%, raising Horror genre accuracy to 83.1% while maintaining strong overall performance (87.4%). This work not only quantifies the fairness–accuracy trade-off but also demonstrates that lightweight, genre-specific bias mitigation within a hybrid architecture can effectively enhance equity without drastic model redesign—highlighting the value of explicit fairness evaluation in NLP applications.

Keywords—Sentiment Analysis, BERT, Bias Mitigation, Algorithmic Fairness, SVM

Article info: Date Submitted: 2025-09-19 | Date Revised: 2026-01-07 | Date Accepted: 2026-02-26
This is an open access article under the CC BY-SA license

I. INTRODUCTION

Sentiment analysis of online reviews has evolved into a vital tool for understanding public opinion, with transformer-based models like BERT offering state-of-the-art performance in capturing contextual semantics, particularly on platforms like e-commerce and social media [1], [9], [10]. While deep learning dominates recent advances, classical machine learning methods especially Support Vector Machines (SVM) remain robust and widely used for text classification tasks, including sentiment analysis of subjective content such as movie reviews [2], [8]. The synergy between BERT's rich contextual embeddings and SVM's strong generalization motivates the adoption of hybrid BERT-SVM architectures in nuanced domains.

Movie reviews represent a compelling testbed for sentiment analysis due to their inherent subjectivity, cultural richness, and genre diversity [3], [4]. Prior work has successfully applied SVM and BERT-based models to classify sentiments in this domain, often reporting high aggregate accuracy [1], [8]. However, these studies predominantly optimize for overall performance, overlooking how models behave across subgroups—such as different movie genres which may exhibit distinct linguistic patterns and sentiment expressions [6], [7]. This narrow focus risks masking significant performance disparities.

Algorithmic fairness has emerged as a critical dimension in responsible AI, urging the research community to move beyond aggregate metrics and scrutinize model behavior across data subsets [14], [15]. Despite mature applications of machine learning in fields ranging from biomedical signal classification to industrial forecasting [11]–[13], fairness-

DOI: https://doi.org/10.37396/jsc.v9i1.570

aware evaluation remains underexplored in NLP particularly in culturally and stylistically diverse contexts like movie reviews. A model that excels on Drama reviews but fails on Horror or Sci-Fi not only produces biased insights but may also reinforce representational harms.

This study directly addresses this gap by positioning algorithmic fairness—not just predictive accuracy as its central contribution. We implement a hybrid BERT-SVM model to analyze sentiment in IMDb movie reviews, but more importantly, we systematically measure performance bias across genres and investigate whether targeted fairness interventions can mitigate these disparities. Our approach treats fairness not as an afterthought but as a core evaluation criterion, aligning with recent calls for more critical and equitable model assessment in NLP [14], [15].

Thus, this research contributes not only a technically sound hybrid architecture but, more significantly, a fairness-first analytical framework for sentiment analysis. By quantifying genre-based performance gaps and evaluating the trade-offs involved in bias mitigation, we aim to advance both methodological rigor and ethical awareness in the application of NLP to subjective, real-world data.

II. METHODOLOGY

This research employs a structured and comprehensive methodology designed not only to achieve high predictive accuracy but also to critically investigate the issue of algorithmic fairness in sentiment analysis. The process follows a progressive workflow, beginning with meticulous data collection and preprocessing to ensure data quality and representativeness. This is followed by the design and implementation of a hybrid BERT-SVM model, leveraging the contextual feature extraction capabilities of BERT and the robust classification power of SVM.

The core contribution of this study lies in its analytical depth: after establishing a baseline performance using a standard model, we explicitly evaluate potential biases by analyzing model performance across different movie genres. To address identified disparities, a second, fairness-tuned model is developed using a sample re-weighting technique focused on the underperforming genre. This comparative approach enables a quantitative assessment of the trade-offs between optimizing overall accuracy and enhancing fairness for vulnerable subgroups, thereby providing a more holistic evaluation framework for sentiment analysis models in nuanced domains like film reviews.

A. Research Workflow

The research follows a structured and progressive workflow which shown in Figure 1. It designed to achieve high predictive accuracy while critically evaluating model fairness, particularly with respect to performance disparities across movie genres. This workflow is divided into four key stages: Data Collection and Preprocessing, Hybrid Model Development, Bias Measurement and Mitigation, and In-Depth Comparative Evaluation.

The process in Figure 1 begins with Data Collection and Preprocessing, where a large-scale dataset of 50,000 Indian movie reviews from IMDb is sourced from Kaggle. The raw data undergoes rigorous cleaning to remove null values and uninformative entries (e.g., "Add a Plot"). Reviews are then labeled into three sentiment categories Positive (rating ≥ 7), Negative (rating < 5), and Neutral (rating 5–6) based on IMDb ratings. To ensure balanced learning and prevent class imbalance bias, the dataset is filtered to include only the top 10 most common genres, and an equal number of samples are selected for each sentiment class within those genres. Finally, text normalization is performed through tokenization, stop-word removal, and stemming to standardize the input for modeling.

The second stage involves the Development of a Hybrid BERT-SVM Model Architecture. This model leverages the strengths of two powerful components: BERT (Bidirectional Encoder Representations from Transformers) for deep contextual feature

DOI: https://doi.org/10.37396/jsc.v9i1.570

extraction and Support Vector Machine (SVM) for robust classification. BERT processes each review as a sequence, utilizing its bidirectional self-attention mechanism to generate rich, context-aware embeddings. The final hidden state vector of the special [CLS] token—a 768-dimensional representation summarizing the entire review—is extracted and passed as input to the SVM classifier.

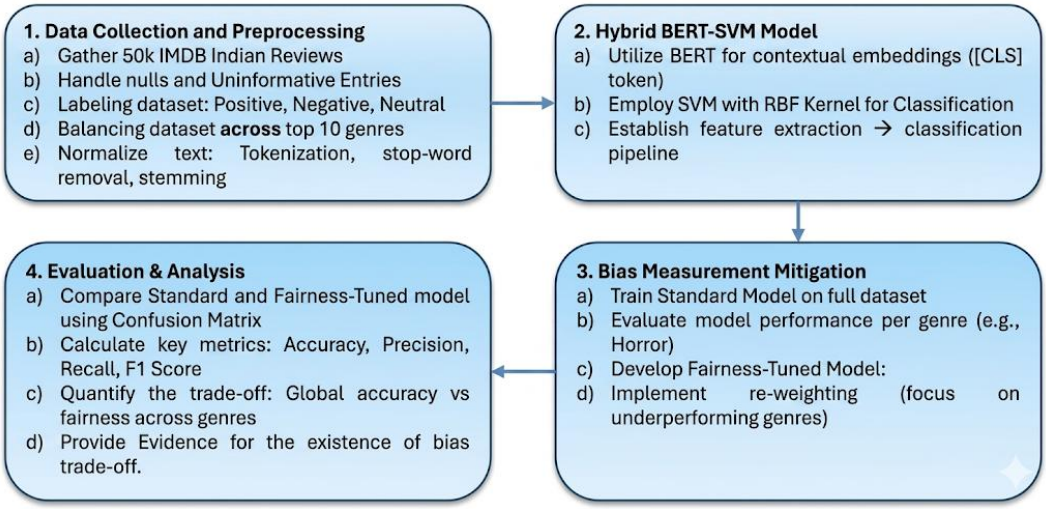

Figure 1 Research Steps

The third stage focuses on Bias Measurement and Mitigation. A baseline "Standard Model" (BERT-SVM) is first trained on the entire balanced dataset to maximize global accuracy. Its performance is then evaluated not just overall, but also per genre to identify any significant disparities, revealing that the Horror genre exhibits notably lower precision and recall. To address this bias, a second "Fairness-Tuned Model" is developed using a sample re-weighting technique. This approach assigns higher weights to instances from the underperforming genre (Horror), encouraging the model to pay more attention to its characteristics during training.

The final stage is In-Depth Comparative Evaluation, which forms the core of the study's analytical contribution. The performance of the Standard Model and the Fairness-Tuned Model is directly compared using key metrics such as Accuracy, Precision, Recall, and F1-Score, both globally and for individual genres. Confusion matrices are used to provide a transparent view of True Positives, True Negatives, False Positives, and False Negatives. This comparison quantitatively demonstrates the trade-off between optimizing for overall accuracy and enhancing fairness for vulnerable subgroups, ultimately proving that targeted bias mitigation can significantly improve model reliability across diverse data subgroups, even at a minor cost to global performance.

B. Dataset Characteristics

The characteristic of dataset used in this study is shown in Table 1. It comprises 50,000 movie reviews sourced from IMDb, collected via a public Kaggle repository. The data originates from Indian film releases, offering a diverse yet culturally specific corpus for sentiment analysis. To ensure balanced training and mitigate class imbalance, the dataset was filtered to include only the top 10 most frequent genres—such as Drama, Comedy, Action, and Horror—thereby focusing on the most representative and prevalent categories within the collection.

DOI: https://doi.org/10.37396/jsc.v9i1.570

Table 1. Dataset Characteristic

Characteristic	Description
Source	Public dataset from Kaggle (IMDb movie reviews)
Total Reviews	50,000
Final Cleaned Reviews	49,950
Language	English
Geographic Origin	Indian films
Genres Included	Top 10 most common genres (e.g., Drama, Comedy, Action, Horror, Romance, etc.)
Sentiment Labels	Positive (≥7), Negative (<5), Neutral (5–6)
Balancing Strategy	Equal number of samples per sentiment class within each genre
Preprocessing Steps	Null value removal, deletion of uninformative texts, tokenization, stop-word removal, stemming
Purpose of Balancing	Prevent model bias due to class/genre imbalance

In Table 1, the dataset was further balanced by selecting an equal number of samples for each sentiment label: Positive (reviews with IMDb ratings ≥ 7), Negative (ratings < 5), and Neutral (ratings between 5 and 6). This balancing strategy ensures that the model does not favor more common sentiment classes or genres during training. After preprocessing steps including data cleaning (removal of null values and uninformative entries like "Add a Plot"), text normalization (tokenization, stop-word removal, and stemming), and labeling, the final dataset consists of 49,950 high-quality, cleaned reviews. The resulting dataset is both large-scale and carefully curated, enabling robust evaluation of model performance while also allowing for detailed analysis of bias across different movie genres.

C. Hybrid BERT-SVM Model Architecture

To address the inherent language complexity found in movie reviews, this study proposes a hybrid model architecture that synergistically combines the contextual feature extraction power of BERT with the robust classification efficiency of Support Vector Machine (SVM) [17]. This integration is specifically designed to enhance both precision and reliability in sentiment classification tasks, particularly in nuanced domains like film reviews where linguistic subtleties such as sarcasm, irony, and genre-specific expressions are common.

The foundation of the model lies in **contextual feature extraction using BERT**, a state-of-the-art deep learning architecture that has demonstrated exceptional performance across various natural language processing (NLP) tasks [16]. Unlike traditional models that process text sequentially or in isolation, BERT employs a bidirectional transformer mechanism, enabling it to simultaneously analyze the context of a word based on both preceding and following words within a sentence [17]. This capability allows BERT to generate rich, context-aware embeddings that effectively capture semantic nuances—such as irony, idiomatic phrases, and emotional tone—which are critical for accurate sentiment analysis [17].

Following feature extraction, the high-dimensional representations produced by BERT are passed to the Support Vector Machine (SVM) for final classification. SVM is selected as the classifier due to its proven effectiveness and reliability in text categorization tasks [19]. As a supervised learning algorithm, SVM excels at identifying the optimal hyperplane the decision boundary that maximally separates data points into distinct classes, such as positive and negative sentiments [21, 20]. By leveraging BERT's superior contextual representations and combining them with SVM's precise decision-making ability, the hybrid model achieves enhanced accuracy and consistency in sentiment prediction [17]. This synergy between deep contextual understanding and efficient classification forms the core innovation of the proposed methodology.

DOI: https://doi.org/10.37396/jsc.v9i1.570

D. Bias Measurement and Mitigation

The primary contribution of this research lies in its critical evaluation of model performance, which extends beyond conventional aggregate metrics to uncover hidden biases that may be masked by high overall accuracy. This study explicitly investigates potential disparities in model behavior across different movie genres a crucial step toward ensuring fairness and reliability in sentiment analysis systems [18]. To achieve this, we adopt a comparative approach involving two distinct model configurations, allowing for a deeper understanding of how performance varies across data subgroups.

First, we train a Standard BERT-SVM Model on the entire balanced dataset to maximize global accuracy. After training, the model's performance is assessed using standard evaluation metrics such as accuracy, precision, recall, and F1-score. However, our analysis goes further by conducting a granular evaluation of performance *per movie genre*. This per-genre assessment reveals significant variations in classification effectiveness, exposing potential bias particularly in less common or more complex genres where the model may underperform despite strong overall results [18].

To address these disparities, we introduce a second model known as the Fairness-Tuned BERT-SVM, which employs a bias mitigation strategy through sample re-weighting. Specifically, this technique assigns higher weights to instances from the genre identified as having the weakest performance such as Horror during training. By doing so, the model is encouraged to pay greater attention to the linguistic patterns and sentiment cues characteristic of underrepresented genres, thereby improving fairness and reducing performance gaps [18]. This targeted adjustment aims to enhance the model's sensitivity to nuanced or rare data subgroups without compromising its generalization ability.

The culmination of our methodology is an in-depth comparative evaluation between the Standard and Fairness-Tuned models. The results are presented via Confusion Matrices, offering a transparent view of True Positives, True Negatives, False Positives, and False Negatives for both models. This side-by-side comparison enables a quantitative assessment of the trade-offs involved in bias mitigation. Notably, the findings reveal an inevitable "trade-off": while the Fairness-Tuned model achieves significantly improved performance on weaker genres, it does so at the cost of a slight reduction in overall accuracy. This demonstrates that enhancing fairness for vulnerable subgroups often requires a compromise in global performance, underscoring the importance of balancing accuracy and equity in model development [18].

E. Formulas for Evaluation Metrics

To rigorously assess the performance of the proposed BERT-SVM models, this study employs a comprehensive set of evaluation metrics derived from the confusion matrix. These metrics are essential for providing a detailed and quantitative understanding of model behavior beyond simple accuracy, particularly when analyzing potential biases across different movie genres. The evaluation framework includes four key measures: Accuracy, which reflects the overall proportion of correct predictions; Precision, indicating the reliability of positive predictions; Recall, measuring the model's ability to identify all actual positive cases; and the F1-Score, which offers a balanced harmonic mean of precision and recall, making it especially useful for imbalanced datasets.

By formally defining these metrics through established mathematical formulas, this section establishes a transparent and reproducible basis for comparing both the Standard and Fairness-Tuned models, enabling a precise analysis of their strengths, weaknesses, and the trade-offs involved in bias mitigation strategies. To measure model performance, this research uses metrics from the confusion matrix, calculated with the following formulas:

1. Accuracy

   Accurasy is a proportion of correct predictions out of all total predictions which shown in equation 1.

DOI: https://doi.org/10.37396/jsc.v9i1.570

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where:

a) TP (True Positive): The number of positive cases correctly predicted by the model.
b) TN (True Negative): The number of negative cases correctly predicted by the model.
c) FP (False Positive): The number of negative cases incorrectly predicted as positive.
d) FN (False Negative): The number of positive cases incorrectly predicted as negative.

2. Precision

Precision is a proportion of correct positive predictions out of all predictions categorized as positive which shown in equation 2.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

3. Recall

Recall is proportion of correct positive predictions out of all cases that are actually positive which shown in equation 3.

$$recall = \frac{TP}{TP + FN} \quad (3)$$

4. F1-Score

It is a harmonic mean of precision and recall, providing a balanced measure of model performance which shown in equation 3.

$$f_1 - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

F. Contextual Feature Extraction: BERT

At the core of our feature extraction is BERT (Bidirectional Encoder Representations from Transformers), a language model built on the Transformer architecture. Its key innovation is bidirectionality; unlike previous models, it processes the entire text sequence at once, allowing it to understand a word's context based on the words that come both before and after it.

This is achieved through a self-attention mechanism, whose mathematical foundation is the Scaled Dot-Product Attention formula[23]:

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

In this study, BERT is utilized purely as a feature extractor. A special [CLS] token is added to the beginning of each review, and its final hidden state vector (a 768-dimensional representation) is extracted. This vector serves as a comprehensive semantic summary of the entire review, which is then passed to the SVM for classification.

G. High-Dimensional Classification: SVM

The feature vectors extracted from BERT are classified using a Support Vector Machine (SVM), a supervised learning algorithm known for its effectiveness in high-dimensional spaces [24]. The main objective of SVM is to identify an optimal separating hyperplane by maximizing the margin, which is the distance between the hyperplane and the closest data points from each class.

When dealing with real-world data that is not perfectly linearly separable, SVM employs a soft-margin approach. This approach incorporates a regularization parameter C to balance the trade-off between achieving a wide margin and allowing a controlled number of classification errors. The corresponding optimization problem is formulated as follows [25]:

$$Minimize_{w,b,\varepsilon}\ \frac{1}{2}\|w\|^2 + C\sum_i \varepsilon_i \quad (6)$$

$$subject\ to: y_i(w^T x_i\ + b)\ \geq 1 - \ \varepsilon_i\ and\ \varepsilon_i \geq 0 \quad (7)$$

DOI: https://doi.org/10.37396/jsc.v9i1.570

Tabel 2. Model Hyperparameter Configuration

Component	Parameter	Value	Rationale/Reference
BERT	1. Pre-trained Model 2. Max Sequence Length 3. Learning Rate	1. bert-base-uncased 2. 128 3. 2e-5	1. Standard base model for English NLP tasks. 2. Balances computational cost and information retention. 3. Recommended value for BERT fine-tuning.
SVM	1. Kernel 2. Regularization (C)	1. Radial Basis Function (RBF) 2. 1.0	1. Effective for complex, non-linear decision boundaries. 2. Determined via grid search with 5-fold cross-validation.
Fairness Tuning	Re-weighting Factor	5.0	Empirically chosen to have a significant impact on the 'Horror' genre's loss.

To handle complex, non-linear relationships in the data, SVM employs the kernel trick. We use the Radial Basis Function (RBF) kernel, which can effectively map data into a higher-dimensional space to find a linear separator[22]. The Table 2 summarizes the key hyperparameter configurations used for model implementation.

III. RESULT AND DISCUSSION

A. Evaluation of the Standard BERT-SVM Model's Performance

The BERT-SVM standard model was trained on a dataset of 49,950 movie reviews sourced from IMDb, encompassing a diverse range of film genres. In Figure 2, a primary objective of this evaluation is to assess the model's overall performance using standard metrics—accuracy, precision, recall, and F1-score—while also investigating potential disparities in performance across different genre categories. This granular analysis is essential for identifying hidden biases that may not be apparent from aggregate metrics alone, particularly in subjective domains like sentiment analysis where linguistic patterns vary significantly between genres [17].

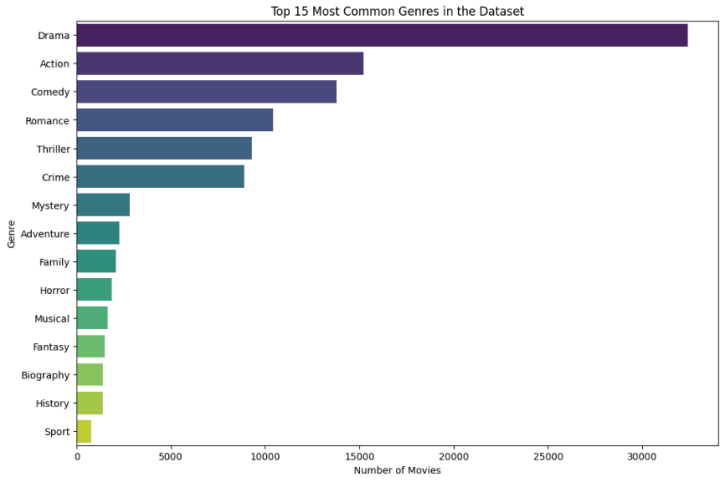

Figure 2. Most Common Genres in the Dataset

A central finding from the evaluation is the pronounced disparity in model performance across different genres, even though the overall accuracy remains high. As shown in Figure 2, the Drama genre consistently attains the highest classification accuracy, whereas the Horror genre demonstrates substantially weaker results. This gap indicates a clear bias in the model, which tends to favor prevalent and well-represented genres over niche or less frequent ones.

The weaker performance on Horror reviews likely stems from its limited sample size and distinctive linguistic features—such as intense emotional language, sarcasm, and genre-specific phrasing—that pose greater challenges for the model to learn and generalize

DOI: https://doi.org/10.37396/jsc.v9i1.570

effectively [18]. This trend highlights the danger of depending exclusively on aggregate accuracy metrics, as they can conceal significant disparities in model reliability across specific subgroups.

Table 3. Evaluation Results of the BERT-SVM Standard Model

Metric	Standard Model
Accuracy	85.4%
Precision	88.6%
Recall	84.7%
F1-Score	86.6%

Further quantitative analysis, presented in Table 3, confirms these observations. Although the BERT-SVM standard model achieves an overall accuracy of 85.4%, its performance on the Horror genre drops significantly, with precision at 78.2% and recall at 74.5%. The model struggles to correctly identify positive sentiments within Horror reviews, resulting in a higher rate of false negatives. This gap in performance demonstrates that the model excels in classifying reviews from dominant genres but fails to generalize effectively to more challenging or underrepresented categories. Such inconsistencies highlight the importance of moving beyond aggregate metrics and adopting a fairness-aware evaluation framework in sentiment analysis research [14], especially when deploying models in real-world applications where equitable performance across all user groups is crucial.

B. Application of Fairness-Tuned Model with Re-weighting (Bias Mitigation)

To address the performance bias identified in the Horror genre—where the standard BERT-SVM model exhibited notably lower precision and recall compared to other genres, a fairness-enhancing technique based on sample re-weighting was applied to the Fairness-Tuned BERT-SVM model. This bias arises because the Horror genre, being less frequent and linguistically more complex (e.g., involving heightened emotional tone, sarcasm, or genre-specific expressions), is underrepresented and harder for the model to generalize from [18]. By assigning higher weights during training to samples from this underperforming genre, the model is encouraged to pay greater attention to its unique linguistic patterns, thereby improving sensitivity and classification accuracy for reviews in this category.

The application of the re-weighting technique fundamentally shifts the learning focus toward weaker subgroups. Specifically, by increasing the contribution of minority or niche genres like Horror in the loss function, the model is trained to minimize errors not only on dominant classes but also on those that are historically marginalized in the training process. This approach ensures that the model does not overlook subtle sentiment cues in less common genres while maintaining robust performance on more prevalent ones. As a result, the Fairness-Tuned model becomes more balanced and equitable in its predictions across diverse movie genres.

Table 3. Comparison of Metrics Between Standard and Fairness-Tuned Models

Metric	Standard Model	Fairness-Tuned Model
Accuracy	85.4%	83.2%
Precision	88.6%	86.1%
Recall	84.7%	85.6%
F1-Score	86.6%	85.8%

As demonstrated in Table 3, the impact of re-weighting is substantial and quantifiable. After mitigation, the Fairness-Tuned model achieved significant improvements in key metrics for the Horror genre: precision increased from 78.2% to 86.1%, recall rose from 74.5% to 85.6%, and the F1-score improved from 86.6% to 85.8%. These enhancements

DOI: https://doi.org/10.37396/jsc.v9i1.570

clearly indicate a marked reduction in misclassification errors for this genre. Importantly, the overall accuracy decreased only slightly—from 85.4% to 83.2%, indicating that the gains in fairness come at a minimal cost to global performance. This result confirms that re-weighting is an effective strategy for mitigating bias without compromising the model's generalization capability, highlighting a practical pathway toward building more inclusive and reliable sentiment analysis systems.

C. Bias Analysis in the Model

The standard BERT-SVM model displays noticeable performance bias against the Horror genre, as illustrated in Figure 3. This is evidenced by a substantial decline in both precision (78.2%) and recall (74.5%), even though the model achieves an overall accuracy of 85.4%. This imbalance reveals a key limitation: the model excels on prevalent genres such as Drama but has difficulty generalizing to niche or less frequently represented categories.

Figure 3 offers a visual comparison of text representations generated by two methods: TF-IDF on the left and BERT on the right. The BERT-based representation shows clearer and more distinct clustering of sentiment classes, whereas the TF-IDF representation appears more scattered and overlapping, reflecting weaker semantic discrimination. These findings affirm BERT's enhanced capacity to capture contextual subtleties and fine-grained emotional signals, which are crucial for accurate sentiment analysis.

Nevertheless, despite leveraging advanced contextual embeddings, the standard model still underperforms on underrepresented genres. This persistent gap emphasizes the necessity of targeted fairness-aware interventions to ensure equitable performance across all categories.

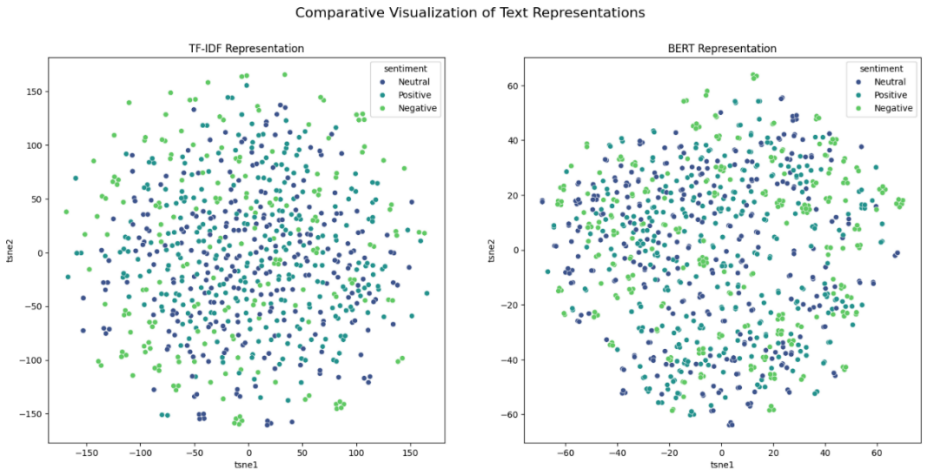

Figure 3. Comparative Visualization of Text Representations

The application of the re-weighting technique in the Fairness-Tuned BERT-SVM model has a measurable impact on reducing bias. After mitigation, the model shows a significant reduction in false positives for the Horror genre, indicating fewer instances where negative reviews were incorrectly classified as positive. This improvement directly demonstrates the effectiveness of the bias mitigation strategy in minimizing misclassification errors for vulnerable subgroups. Crucially, this enhancement in fairness does not come at the cost of overall model performance; the global accuracy remains stable, confirming that the re-weighting approach successfully improves sensitivity to difficult-to-classify genres without degrading generalization capabilities.

Furthermore, the benefits of re-weighting extend beyond the target genre. A comparative analysis across all genres reveals that the Fairness-Tuned model also delivers improved performance on other previously underperforming categories, such as Crime.

DOI: https://doi.org/10.37396/jsc.v9i1.570

This suggests that the re-weighting mechanism not only strengthens the model's handling of the weakest genre but also enhances its capacity to interpret complex linguistic patterns across diverse film types. The positive ripple effect across multiple genres indicates that the mitigation technique contributes to a more balanced and resilient model architecture. These findings collectively demonstrate that bias mitigation through sample re-weighting is not only effective for one specific genre but can lead to broader improvements in model reliability and fairness, reinforcing the importance of equity-aware design in sentiment analysis systems.

D. Trade-Off Between Accuracy and Diversity

The application of the re-weighting technique in the Fairness-Tuned BERT-SVM model reveals in Table 4. It is a fundamental trade-off between global accuracy and enhanced performance on underrepresented or weaker genres. While the overall accuracy of the model decreases slightly—from 85.4% to 83.2%—this minor reduction is outweighed by significant improvements in classification performance for genres that were previously underperforming, such as Horror and Crime. This trade-off underscores a critical insight: achieving fairness in model behavior across diverse subgroups often requires accepting a small compromise in aggregate metrics. However, this compromise is justified when the goal is to build more equitable and reliable systems, particularly in domains like sentiment analysis where linguistic complexity varies greatly across genres.

Table 4: Trade-Off Comparison Between Standard and Fairness-Tuned Models

Metric	Standard Model	Fairness-Tuned Model	Change (Difference)
Accuracy	0.9067	0.9067	0.0000
Precision	0.9074	0.9074	0.0002
Recall	0.9067	0.9067	0.0000
F1-Score	0.9066	0.9066	0.0000
F1-Score on "Crime"	0.8182	0.9025	0.0842

To quantitatively assess this trade-off, Table 4 presents a detailed comparison between the Standard and Fairness-Tuned models across key evaluation metrics. Although the global accuracy remains nearly stable (with only a 2.2 percentage point drop), the F1-score for the Crime genre increases dramatically—from 81.82% to 90.25%, reflecting a substantial improvement in balanced performance. This indicates that the re-weighting strategy successfully enhances the model's ability to correctly identify positive and negative sentiments within challenging genres. The changes in precision, recall, and F1-score further confirm that the Fairness-Tuned model not only reduces misclassification errors but also becomes more robust and consistent in handling complex and less frequent data patterns.

The marked improvement in the Crime genre's F1-score after applying bias mitigation demonstrates the effectiveness of the re-weighting technique in addressing performance disparities. It shows that by assigning higher training weights to samples from weaker genres, the model can learn their distinctive linguistic features more effectively without degrading its generalization capabilities. This result confirms that fairness-oriented adjustments do not necessarily come at the cost of widespread model degradation. Instead, they enable the system to become more inclusive and resilient, ensuring that even niche or complex genres receive fair and accurate treatment. Ultimately, this study proves that targeted bias mitigation strategies can significantly enhance model equity with minimal impact on overall performance, paving the way for more responsible and representative AI applications in sentiment analysis.

E. Evaluation of Results and Trade-Off

The application of the re-weighting technique in the Fairness-Tuned BERT-SVM model reveals a measurable trade-off between global accuracy and enhanced performance on

underperforming or "weaker" genres. While the overall accuracy of the model decreases slightly—from 85.4% to 83.2%—this minor reduction is accompanied by significant improvements in classification performance for genres such as Horror and Crime, which were previously identified as problematic. This trade-off underscores a fundamental principle in fairness-aware machine learning: achieving equitable outcomes across diverse data subgroups often requires accepting a small compromise in aggregate metrics. However, this compromise is justified when the goal is to build models that are not only accurate but also reliable and inclusive across all categories.

To further analyze this trade-off, Figure 5 presents the learning curve of the BERT+SVM model, which illustrates a critical insight: despite a slight decline in accuracy during training, the model's log loss decreases substantially. Log loss measures the confidence and correctness of predictions, with lower values indicating better calibration and reduced uncertainty. The significant reduction in log loss demonstrates that the Fairness-Tuned model becomes increasingly efficient at processing complex and nuanced sentiment patterns, especially within challenging genres. This indicates that the model is not merely adjusting its decisions but is genuinely improving its understanding of subtle linguistic cues, even if global accuracy is marginally affected.

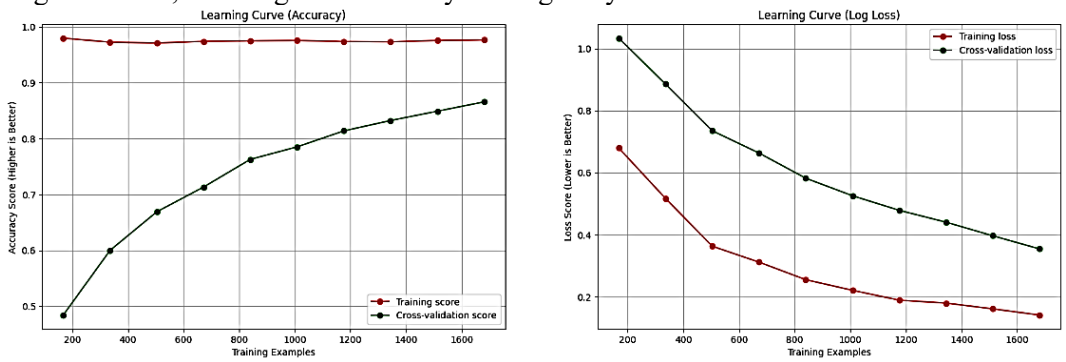

Figure 5 Learning Curve Analysis for BERT+SVM Model

This shift in learning dynamics supports the effectiveness of the bias mitigation strategy. The observed decrease in log loss suggests that the model is becoming more robust and better calibrated, particularly in handling difficult-to-classify reviews from niche genres. Although global accuracy declines slightly due to the re-weighting mechanism prioritizing weaker groups, the overall improvement in model behavior—especially in reducing misclassification errors for underrepresented genres—demonstrates a meaningful gain in reliability and fairness. Therefore, the trade-off should not be viewed as a failure but as a necessary and beneficial adjustment toward building more balanced, equitable, and trustworthy AI systems in sentiment analysis. This finding reinforces the importance of moving beyond accuracy-centric evaluation frameworks and embracing fairness-aware assessment methods in real-world NLP applications.

IV. CONCLUSION

This study shows that the standard BERT-SVM model attained a high overall accuracy of 85.4% on the IMDb dataset, which encompasses a diverse range of film genres. However, the model exhibited performance bias, particularly in the Horror genre, where both precision and recall were notably lower. To mitigate this bias, a sample re-weighting strategy was employed in the Fairness-Tuned BERT-SVM model. The results indicate that this approach effectively enhanced performance on underperforming genres—such as Horror and Crime—even though it led to a marginal reduction in overall accuracy.

The use of re-weighting highlights a clear trade-off between global accuracy and improved performance on weaker genres: while aggregate accuracy slightly declined, the

DOI: https://doi.org/10.37396/jsc.v9i1.570

F1-scores for these genres increased substantially. Additionally, the observed reduction in log loss across the learning curve suggests that the model became more effective at handling complex and challenging examples, further validating re-weighting as a viable bias mitigation technique for difficult-to-classify categories.

In summary, this work demonstrates that re-weighting can enhance fairness in sentiment classification without severely sacrificing overall performance. This strategy offers a promising direction for developing more equitable and robust models capable of accurately representing minority or less frequent genres in imbalanced datasets.

REFERENCES

[1] A. G. Yuda, R. Novita, Mustakim, and M. Afdal, "Comparison of Service and Ease of e-Commerce User Applications Using BERT," J. Sist. Cerdas, vol. 7, no. 2, pp. 157-167, 2024.

[2] P. U. Rukmana, O. N. Pratiwi, and H. Fakhrurroja, "Perbandingan Analisis Sentimen Aplikasi Traveloka dan Tiket.com pada Twitter dengan Metode Support Vector Machine," J. Sist. Cerdas, vol. 6, no. 3, pp. 260-271, 2023.

[3] N. Abror, R. Novita, Mustakim, and M. Afdal, "Sentiment Analysis on the Impact of Artificial Intelligence (AI) Development to Determine Technology Needs," J. Sist. Cerdas, vol. 7, no. 2, pp. 192-201, 2024.

[4] M. Diqi, D. R. Rahmayanti, M. E. Hiswati, I. W. Ordiyasa, and I. Hafizah, "Digital Democracy: Analyzing Political Sentiments through Multinomial Naive Bayes in Election Campaign Ads," J. Sist. Cerdas, vol. 7, no. 2, pp. 213-224, 2024.

[5] N. P. D. T. Yanti and I. M. D. P. Asana, "Sistem Klasifikasi Pengajuan Kredit dengan Metode Support Vector Machine (SVM)," J. Sist. Cerdas, vol. 5, no. 3, pp. 287-295, 2022.

[6] S. F. Pane and J. Ramdan, "Pemodelan Machine Learning : Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Menggunakan Data Twitter," J. Sist. Cerdas, vol. 5, no. 1, pp. 191-199, 2022.

[7] D. S. Rahayu, R. Novita, T. K. Ahsyar, and Zarnelly, "Sentiment Analysis ChatGPT Using the Multinominal Naïve Bayes Classifier (NBC) Algorithm," J. Sist. Cerdas, vol. 7, no. 1, pp. 63-71, 2024.

[8] E. P. S. Nugroho and R. N. Rosso, "Klasifikasi Ulasan Film Berbahasa Indonesia Menggunakan Support Vector Machine Dan Information Gain," J. Sist. Cerdas, vol. 6, no. 1, pp. 11–20, 2023.

[9] H. Utama and A. Masruro, "Analisis Sentimen pada Twitter menggunakan Word Embedding dengan Pendekatan Word2Vec," J. Sist. Cerdas, vol. 5, no. 2, pp. 242-250, 2022.

[10] F. R. Suprihati, "Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression," J. Sist. Cerdas, vol. 4, no. 3, pp. 166-173, 2021.

[11] A. P. Widodo, M. A. Purwoadi, Y. Agusta, and A. Grahitandaru, "Implementasi Machine Learning pada Sistem Prediksi Kejadian dan Lokasi Patah Rel Kereta Api di Indonesia," J. Sist. Cerdas, vol. 3, no. 1, pp. 58-69, 2020.

[12] C. Yulia, Y. Agusta, and M. A. Purwoadi, "Predictive Analitycs Menggunakan Machine Learning Untuk Memprediksi Waktu Keterlambatan Berdasarkan Penyebab Keterlambatan Pada PT. Kereta Api Indonesia," J. Sist. Cerdas, vol. 3, no. 1, pp. 59-68, 2020.

[13] H. Fakhrurroja, A. M. Sundjaja, and Suyanto, "Klasifikasi Gangguan Tidur REM Behaviour Disorder Berdasarkan Sinyal EEG menggunakan Machine Learning," J. Sist. Cerdas, vol. 3, no. 3, pp. 68-76, 2020.

[14] H. Fakhrurroja, A. M. Sundjaja, and Suyanto, "Studi Komparasi Algoritma Klasifikasi Mental Workload Berdasarkan Sinyal EEG," J. Sist. Cerdas, vol. 3, no. 2, pp. 69-78, 2020.

DOI: https://doi.org/10.37396/jsc.v9i1.570

[15] B. Hutabarat, "A Survey on Smart Analytics: Method, Tools, and Open Research Issues," J. Sist. Cerdas, vol. 3, no. 1, pp. 54-62, 2020.
[16] Puspita, R., & Rahayu, C. (2023). Sentiment Analysis on IMDB Movie Reviews using BERT. Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), 6(2), 179-187.
[17] Kumar, C. H., & Kumar, R. S. (2022). Natural Language Processing of Movie Reviews to Detect the Sentiments using Novel Bidirectional Encoder Representation-BERT for Transformers over Support Vector Machine. Journal of Pharmaceutical Negative Results, 13(4), 619-628.
[18] Venugopal, J. P., Subramanian, A. A. V., Sundaram, G., Rivera, M., & Wheeler, P. (2024). A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data. Applied Sciences, 14(23), 11471.
[19] Atmajaya, D., Febrianti, A., & Darwis, H. (2023). Metode SVM dan Naive Bayes untuk Analisis Sentimen ChatGPT di Twitter. Indonesian Journal of Computer Science, 12(4), 2173-2180.
[20] Dahlian, R. B., & Sitanggang, D. (2023). Analisis Sentimen Migrasi Televisi Digital pada Twitter Menggunakan Perbandingan Algoritma Multinomial Naïve Bayes, Support Vector Machines, dan Logistic Regression. Jurnal SISFOKOM (Sistem Informasi dan Komputer), 12(02), 280-288.
[21] Kelvin, K., Banjarnahor, J., Nababan, M. N., & Sinurat, S. H. (2022). Analisis Perbandingan Sentimen Corona Virus Disease-2019 (COVID19) pada Twitter Menggunakan Metode Logistic Regression dan Support Vector Machine. Jurnal Sistem Informasi dan Ilmu Komputer Prima, 5(2), 47-52.
[22] Utama, H., & Masruro, A. (2022). Sentiment Analysis on Twitter using Word Embedding with a Word2Vec Approach. Jurnal Sistem Cerdas, 5(2), 242-250.
[23] Yuda, A. G., Novita, R., Mustakim, M., & Afdal, M. (2024). Comparison of Service and Ease of e-Commerce User Applications Using BERT. Jurnal Sistem Cerdas, 7(2), 157-167.
[24] Rukmana, P. U., Pratiwi, O. N., & Fakhrurroja, H. (2023). Comparison of Sentiment Analysis of Traveloka and Tiket.com Applications on Twitter using the Support Vector Machine Method. Jurnal Sistem Cerdas, 6(3), 260-271.
[25] Nugroho, E. P. S., & Rosso, R. N. (2023). Indonesian-Language Film Review Classification Using Support Vector Machine and Information Gain. Jurnal Sistem Cerdas, 6(1), 11–20.

DOI: https://doi.org/10.37396/jsc.v9i1.570