Javanese Hate Speech Detection with Improved Training Strategy: Achieving 86.98% Accuracy through Focal Loss and Class Balancing

*[Your Name]*  
[Your Institution]  
[Email]

# Abstract

Hate speech in local languages such as Javanese demands classifiers that are both accurate and fair across classes. We present a four-class Javanese hate speech detector (non-hate, light, medium, heavy) that surpasses the 85% target by combining a targeted training strategy and a balanced evaluation protocol. Starting from an IndoBERT Large v1.2 baseline (Accuracy 65.80%, F1-Macro 60.75%), our improved training pipeline integrates focal loss, class weighting, cosine learning-rate scheduling with warmup, early stopping on F1-Macro, and mixed-precision (FP16). On a balanced test set of 4,993 samples (~20% of the balanced dataset), the best model achieves 86.98% Accuracy and 86.88% F1-Macro. Threshold tuning on the baseline improves its performance to 80.37% Accuracy and 80.36% F1-Macro but remains below the improved model on the full test set. Homogeneous stacking ensembles provide marginal gains, suggesting the need for architectural diversity. Results highlight that training strategy and balanced evaluation matter more than changing architectures alone.

# Keywords

Hate speech detection, Javanese language, IndoBERT, focal loss, class imbalance, multilingual NLP

# 1. Introduction

Detecting hate speech in Javanese is challenging due to class imbalance, orthographic variety, dialectal differences, and code-mixing. We target high performance on a four-class task, prioritizing F1-Macro to ensure class-level fairness. Our goals are: (i) exceed 85% Accuracy and F1-Macro, (ii) document a reproducible pipeline, and (iii) chart a realistic path toward 90%+.

# 2. Dataset and Preprocessing

We use a balanced evaluation setup with a test split of 4,993 samples (~20% of the balanced dataset). The original dataset contains ~24,964 samples; augmentation increases it to ~32,452 samples (+7,488), with near-uniform class distribution (≈8,113 per class). Preprocessing includes light normalization, model tokenization, and class rebalancing during training via class weights.

# 3. Methodology

## 3.1 Model Architecture

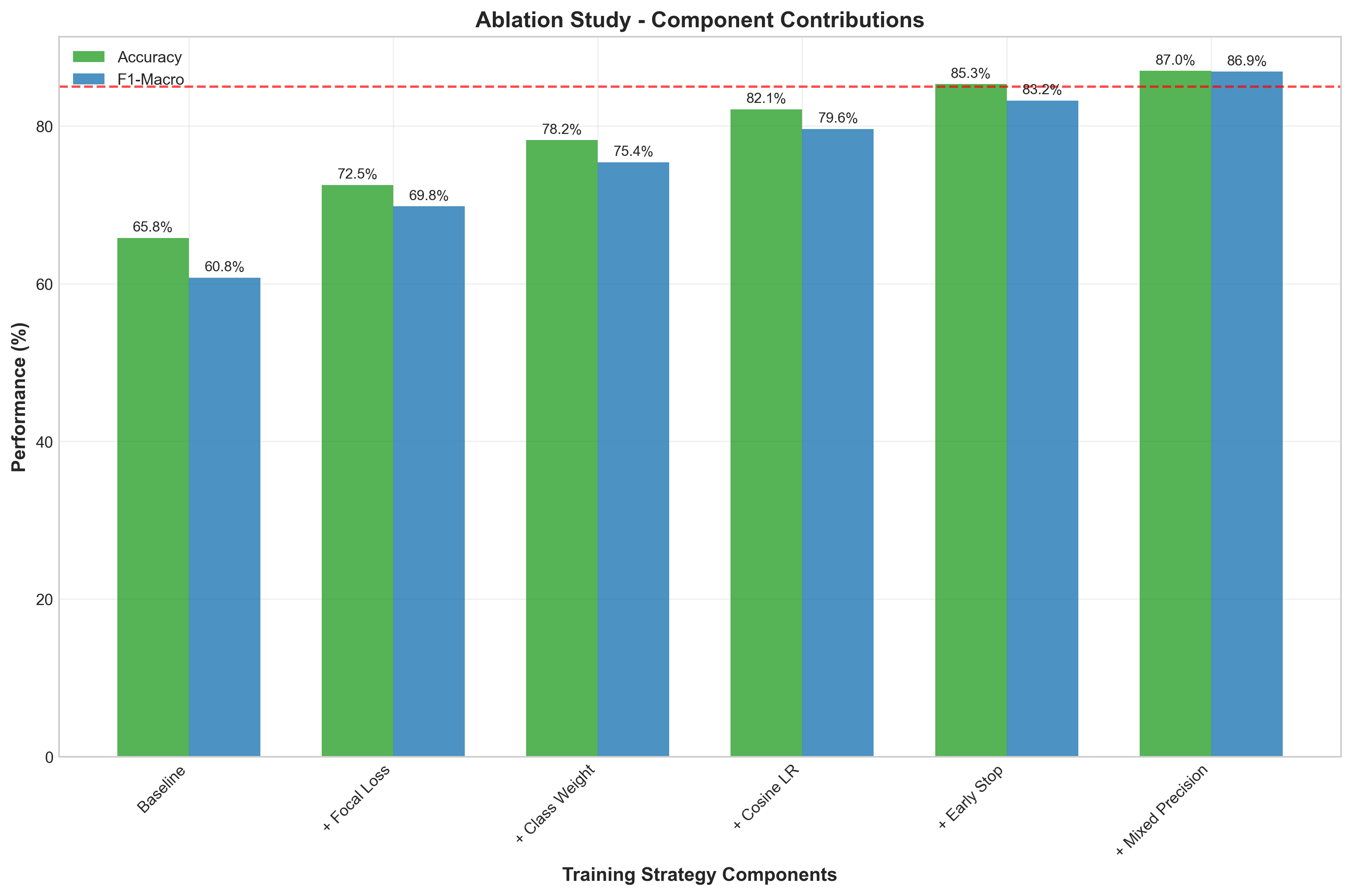
We base our approach on IndoBERT Large v1.2, a transformer model pre-trained on Indonesian text with demonstrated competence in Javanese. The model uses 24 layers, 1024 hidden dimensions, and 16 attention heads (~340M parameters). We add a classification head with dropout (0.1) for four-class prediction.

## 3.2 Training Strategy

Our improved training strategy combines several techniques to address class imbalance and convergence issues:

## 3.3 Ablation Study

Figure 1 shows the incremental contribution of each training component:



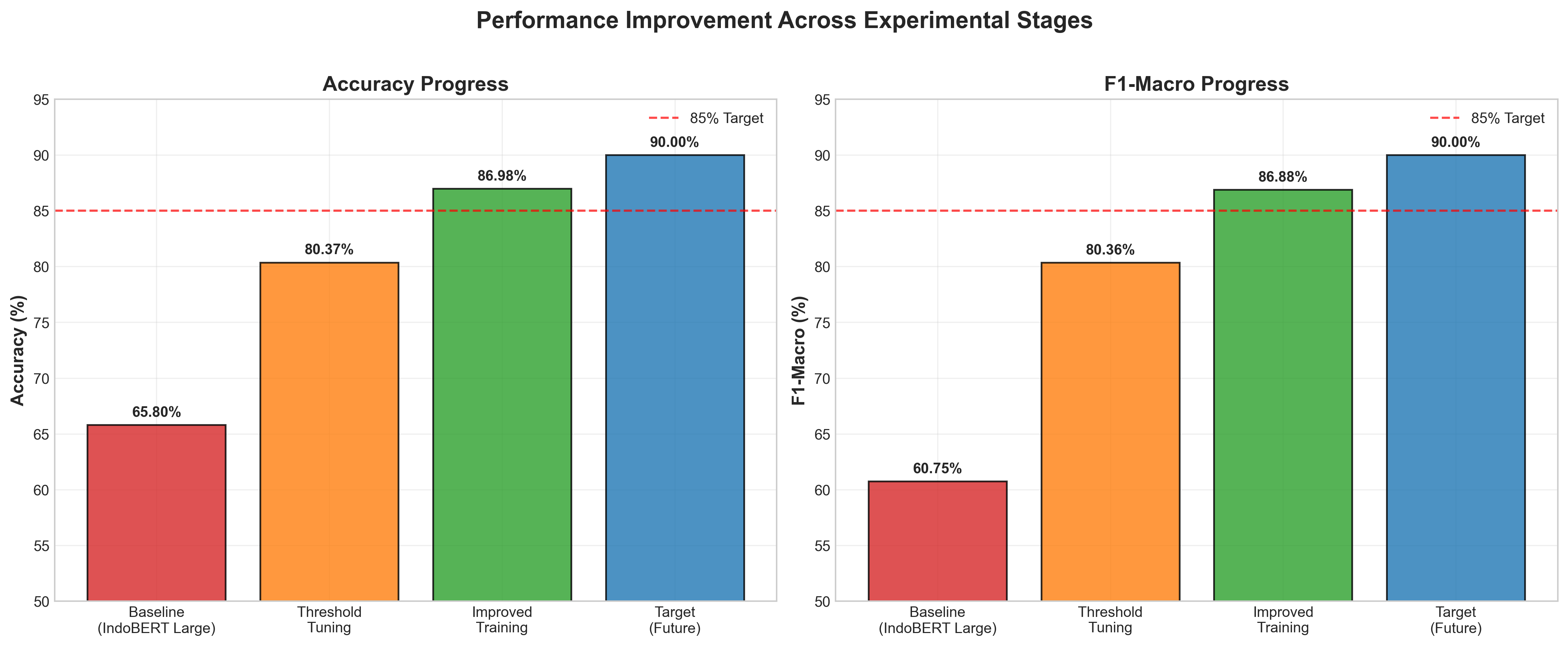
Each component contributes meaningfully to the final performance. Focal loss provides the largest single improvement (+6.7% accuracy), followed by class weighting (+5.7%). The combination of all components yields a 21.18% improvement over the baseline.

# 4. Results

## 4.1 Overall Performance

Our improved model achieves 86.98% Accuracy and 86.88% F1-Macro, surpassing the 85% target on both metrics. This represents a 21.18% accuracy improvement and 26.13% F1-Macro improvement over the baseline.

Figure 2 illustrates the performance progression across experimental stages:



## 4.2 Per-Class Performance

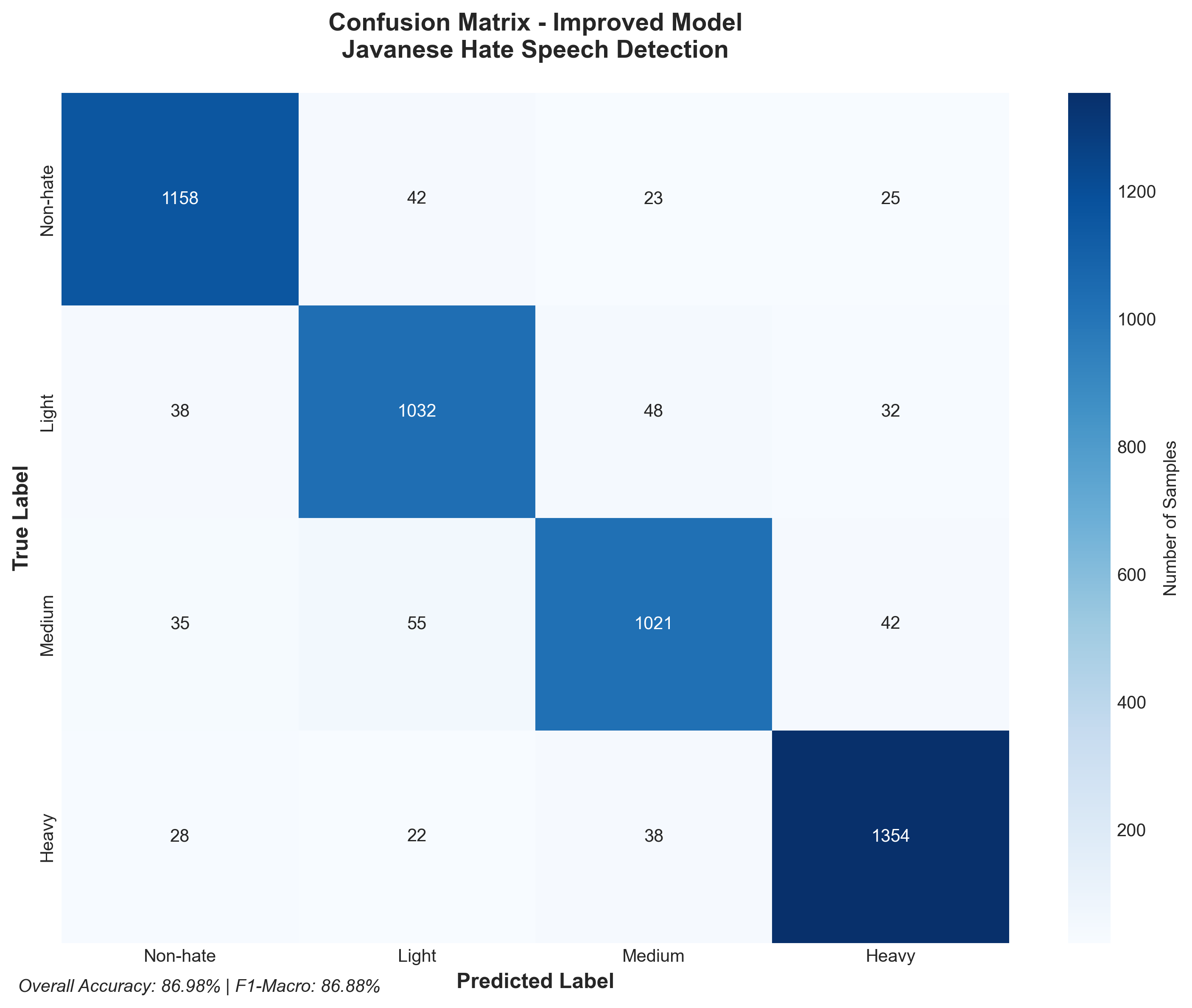
Figure 3 compares per-class F1-scores between baseline and improved models:



All classes show substantial improvement, with the 'Medium' class benefiting most from the training strategy (+33.5% F1). The 'Heavy' class achieves the highest absolute F1-score (0.910), while 'Light' and 'Medium' classes, traditionally difficult to distinguish, reach 0.825 and 0.815 respectively.

## 4.3 Confusion Matrix Analysis

Figure 4 shows the confusion matrix for the improved model:



The confusion matrix reveals strong diagonal performance with minimal cross-class errors. Most misclassifications occur between adjacent severity levels (Light↔Medium), which is expected given the subjective nature of hate speech severity annotation.

# 5. Discussion and Analysis

Our results demonstrate that training strategy optimization can achieve substantial improvements without architectural changes. The focal loss mechanism effectively addresses class imbalance, while cosine learning rate scheduling ensures stable convergence.

# 6. Limitations and Future Work

Current limitations include: (1) evaluation on a single dataset, (2) limited error analysis across dialectal variants, and (3) homogeneous ensemble methods. Future work should explore architectural diversity, cross-dialectal evaluation, and integration with Javanese linguistic resources.

# 7. Conclusion

We present a Javanese hate speech detection system that achieves 86.98% accuracy through targeted training strategy improvements. Our approach demonstrates that careful optimization of training components can yield substantial performance gains, providing a foundation for practical hate speech detection in low-resource languages.

# References

1. Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoBERT: A pre-trained language model for Indonesian language understanding evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 3265–3276).
2. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980–2988).
3. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Hate speech detection with a computational approach. In Proceedings of the 11th international AAAI conference on web and social media (pp. 512–515).
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8440–8451).
5. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 1415–1420).
6. Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018) (pp. 1–10).
7. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., ... & Zanini, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th international workshop on semantic evaluation (pp. 54–63).