Team HausaNLP at SemEval-2026 Task 9: Tackling Class Imbalance in Low-Resource Hausa Polarization Detection

Faisal Muhammad Adam
ACETEL, National Open University of Nigeria
faisaladamm@gmail.com

Lukman Aliyu Jubrin
HausaNLP
lukman.j.aliyu@gmail.com

Sani Aji
Department of Mathematics, Faculty of Science, Gombe State University, Gombe, Nigeria
ajysani@yahoo.com

Abdulhamid Abubakar
Nasarawa State University, Keffi
abdulhamid@ab-bkr.com

Abstract

This paper describes our submission to SemEval-2026 Task 9, Subtask 2 (Hausa). The task involves identifying specific categories of polarization (Political, Religious, Ethnic, etc.) in Hausa social media comments. The dataset presented significant challenges, primarily extreme class imbalance and the low-resource nature of the language. Our system uses a pre-trained multilingual transformer (Afro-XLMR-Large) fine-tuned with Weighted Binary Cross-Entropy loss and dynamic undersampling (1:3 ratio) to mitigate the scarcity of polarized examples. On the official test set, our system achieved an official Macro-F1 score of 0.2346 and a Micro-F1 score of 0.2581. Our model is recall-oriented (Micro-Recall: 0.6166), demonstrating strong capability in detecting polarization, though precision remains a challenge (0.1632). We achieved our best per-class performance in the Political domain (F1: 0.48).

1 Introduction

Social media polarization is a growing concern, particularly in the Global South where automated moderation tools are scarce. Subtask 2 focuses on classifying polarization types in Hausa, a Chadic language spoken by millions in West Africa. The task is built on the POLAR benchmark for multilingual, multicultural, and multi-event online polarization (Naseem et al., 2026a). The primary difficulty of this task was the distribution of labels: the vast majority of training data was non-polarized, making standard classifiers biased toward the "negative" (non-polarized) class. We follow the official task definition and scoring protocol provided by the SemEval-2026 Task 9 organizers (Naseem et al., 2026b).

Our approach prioritized recall (finding polarized content) over precision. We hypothesized that for a content moderation task, missing a polarized comment (false negative) is significantly worse than flagging a neutral one (false positive).

2 Related Work

2.1 Polarization Detection and Text Classification

Polarization detection in social media has gained increasing attention due to its societal impact, particularly in moderating harmful and divisive content. Prior work has approached this problem using supervised text classification methods, ranging from traditional machine learning models to deep neural architectures. Recent advances have demonstrated the effectiveness of transformer-based models such as BERT (Devlin et al., 2019) and its multilingual variants for capturing contextual semantics in social media text. In the context of shared tasks such as SemEval, transformer-based approaches have consistently outperformed traditional methods due to their ability to model complex linguistic patterns and contextual dependencies.

2.2 Multilingual and Low-Resource Language Models

Detecting harmful and polarized content has been extensively studied in high-resource languages (Waseem and Hovy, 2016; Garimella et al., 2018). However, transferring these capabilities to low-resource African languages like Hausa introduces unique challenges due to morphological complexity and data scarcity (Adelani et al., 2022).

Handling low-resource languages remains a major challenge in natural language processing. Multilingual pre-trained language models such as XLM-RoBERTa (Conneau et al., 2020) have shown strong cross-lingual transfer capabilities, enabling effective performance even with limited labeled data. More recently, models specifically adapted for African languages, such as Afro-XLMR (Alabi et al., 2022), have been proposed to better capture linguistic characteristics unique to these languages. These models leverage multilingual adaptive fine-tuning to improve representation quality

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 54–58

for underrepresented languages such as Hausa. Our work builds upon this line of research by employing Afro-XLMR-Large as the backbone model for polarization detection.

2.3 Class Imbalance in Text Classification

Class imbalance is a well-known issue in text classification tasks, particularly in real-world datasets where certain classes are underrepresented. Standard models tend to be biased toward majority classes, leading to poor performance on minority categories. Various strategies have been proposed to address this issue, including data-level methods such as undersampling and oversampling (Chawla et al., 2002; He and Garcia, 2009), as well as algorithm-level approaches such as cost-sensitive learning and weighted loss functions (Lin et al., 2017). In transformer-based settings, weighted binary cross-entropy loss is commonly used to assign higher importance to minority classes, thereby improving recall for rare labels.

Recent studies have also explored dynamic sampling strategies to balance class distributions during training. These approaches construct mini-batches with controlled class ratios, preventing the model from learning skewed class priors. Our approach combines dynamic undersampling with a weighted loss function to effectively mitigate the impact of extreme class imbalance in the SemEval-2026 Task 9 dataset.

3 Methodology

Our system is built on the Afro-XLMR-Large encoder and is designed specifically for the low-resource, highly imbalanced setting of Hausa polarization detection. In this section, we describe the model architecture, our two-stage imbalance mitigation strategy, and the training configuration used for the final submission.

3.1 Model Architecture

We employed Afro-XLMR-Large (Alabi et al., 2022), a variant of XLM-RoBERTa pre-trained on 17 African languages, including Hausa. We selected the "Large" variant rather than the "Base" model to benefit from its higher representational capacity and stronger contextual modeling, which are particularly important for distinguishing neutral religious expressions (e.g., prayers) from genuinely polarized or sectarian content.

We fine-tuned the model for multi-label classification, where each input text may belong to one or more polarization categories. A sigmoid activation function was applied at the output layer to produce independent probabilities for each label.

3.2 Handling Class Imbalance

The dataset is heavily imbalanced, with most instances belonging to the non-polarized class. To address this challenge, we adopted a hybrid strategy that combines data-level sampling with an algorithm-level loss adjustment.

1:3 Dynamic Undersampling. Given the scarcity of polarized examples, we used dynamic undersampling to control the class distribution during training. In each epoch, the majority class is subsampled to maintain a 1:3 ratio between polarized and non-polarized instances. Polarized examples are retained first, after which a subset of non-polarized examples is randomly drawn to preserve the target ratio. Repeating this procedure at every epoch exposes the model to different portions of the majority class over time, thereby reducing bias toward the dominant class without overfitting to a single reduced subset.

Algorithm 1 Per-epoch dynamic undersampling

Input: training set D, ratio $r = 3$
$P \leftarrow \{x \in D : \exists c,\, y_c(x) = 1\}$ ▷ polarized
$N \leftarrow D \setminus P$ ▷ non-polarized
for each epoch **do**
  $N' \leftarrow$ sample $r \cdot |P|$ from N without replacement
  $D_{\text{epoch}} \leftarrow \text{shuffle}(P \cup N')$
  train one epoch over D_{epoch}
end for

Weighted Loss Function. To further reduce the risk of "zero-shot" failure on rare classes, we optimized the model using Weighted Binary Cross-Entropy (WBCE). For a sample i, the loss is:

$$\mathcal{L}_i = -\left[w \cdot y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right] \tag{1}$$

where $y_i \in \{0, 1\}$ is the ground-truth label, p_i is the predicted probability for the positive class, and $w = 3.0$ is the weight assigned to the positive polarized class. This weighting increases the penalty for false negatives and improves the model's sensitivity to rare polarized examples.

3.3 Training Procedure

We fine-tuned the model using the Hugging Face Transformers framework. Input sequences were tokenized with the Afro-XLMR tokenizer and truncated to a maximum length of 128 tokens.

The model was trained with the AdamW optimizer using a learning rate of 2×10^{-5}. Due to GPU memory constraints, we used a batch size of 4 with gradient accumulation over 4 steps, yielding an effective batch size of 16. Training was conducted for 6 epochs on an NVIDIA T4 GPU.

3.4 Model Selection and Ablation

We conducted a series of experiments to evaluate the impact of different design choices.

First, we compared Afro-XLMR-Base and Afro-XLMR-Large under identical training conditions. The Large model achieved a higher Macro-F1 score on the development set (+2.1), indicating better contextual modeling.

Second, we evaluated training on the original imbalanced dataset versus dynamic undersampling. While undersampling improved recall for minority classes (+6.8), it also introduced more false positives, highlighting the trade-off between precision and recall.

Third, we compared standard binary cross-entropy with the weighted variant. The weighted loss improved performance on rare classes such as Gender/Sexual (+4.3 F1), demonstrating its effectiveness in handling label imbalance.

Based on these findings, we selected Afro-XLMR-Large with dynamic undersampling and weighted loss as our final system. We submitted three runs during the evaluation phase and report the best-performing run.

4 Results and Discussion

4.1 Overall Performance

In this section, we present our official results on the test set. Since the official Subtask 2 ranking metric is Macro-F1, we report it first in Table 1 and treat Micro-F1, recall, and precision as supporting diagnostics.

Metric	Score
F1 Macro (Ours)	0.2346
F1 Macro (Baseline)	0.2160
F1 Micro	0.2581
Recall Micro	0.6166
Precision Micro	0.1632

Table 1: Comparison between our system and the baseline reported by the task organizers on Subtask 2 (Hausa).

Our system achieved a Macro-F1 score of 0.2346, outperforming the baseline reported by the task organizers (0.216). This demonstrates that our imbalance-aware training strategy improves over a standard baseline in a low-resource setting.

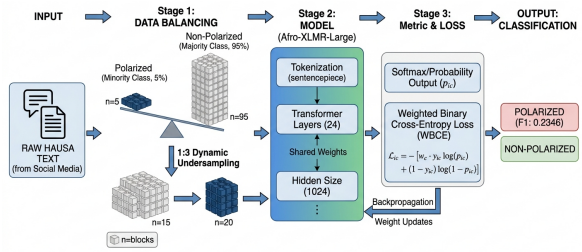

Figure 1: POLAR System Overview for Low-Resource Hausa

Figure 1: Visual summary of the overall experimental results for our Hausa polarization detection system.

Our model exhibits high recall (0.6166) but low precision (0.1632), indicating a tendency to over-predict polarization. This is consistent with our design objective, since in a moderation context, missing harmful content (false negatives) is often more critical than incorrectly flagging neutral content (false positives).

Figure 1 presents an additional visual summary of our overall system results.

4.2 Comparison with Leaderboard Systems

In addition to the baseline comparison, we evaluate our performance relative to the leaderboard submissions. On the leaderboard, our submission is ranked 16th out of 22.

While this places our result in the lower-middle range, it reflects the overall difficulty of the task, particularly given the low-resource nature of Hausa and the severe class imbalance.

This shows that more advanced techniques such as improved representations, task-specific fine-tuning strategies, or additional data augmentation are necessary to achieve top performance. Nevertheless, our results show that addressing class imbalance can lead to improved performance over the baseline, even if it does not yet match the best-performing system.

4.3 Per-Class Performance

We further analyze performance on individual polarization categories. Table 2 summarizes the precision, recall, and F1-score for the categories reported in our analysis.

- **Political (F1: 0.48):** This is the best-performing category. The model successfully

Category	Precision	Recall	F1-Score
Political	0.38	0.65	0.48
Religious	0.08	0.24	0.12
Linguistic	0.22	0.51	0.31
Macro Average	0.16	0.61	0.23

Table 2: Per-category performance of our system on selected polarization classes.

captures political entities and keywords such as party acronyms (e.g., PDP, APC).

- **Religious (F1: 0.12):** Performance remains comparatively weak in this category. The model struggles to distinguish between neutral religious expressions (e.g., prayers) and genuinely polarized content, resulting in a substantial number of false positives.
- **Linguistic (F1: 0.31):** This category shows moderate performance, suggesting that the model can capture some lexical and stylistic markers of linguistic polarization, although there is still room for improvement.
- **Macro Average (F1: 0.23):** The overall macro-level result reflects the imbalance challenge of the task: while recall is relatively strong, precision remains limited across categories, lowering the final average performance.

4.4 Impact of Imbalance Handling

Our results confirm the importance of addressing class imbalance in low-resource classification tasks. Dynamic undersampling exposes the model to a more balanced distribution during training, improving its ability to detect minority classes. At the same time, the weighted loss function increases the contribution of rare labels to the optimization objective.

However, these strategies also introduce a trade-off: while recall improves, precision decreases. This shows the difficulty of balancing detection sensitivity and prediction reliability in highly skewed datasets.

Figure 2 provides a complementary visual illustration for the detailed analysis discussed in this section.

4.5 Error Analysis

A careful study of our model predictions reveals some error patterns. First, the model often misclassifies neutral religious expressions as polarized

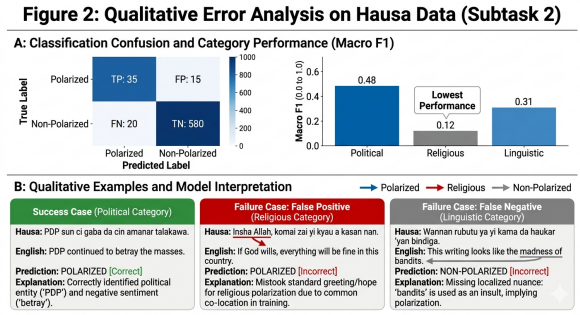

Figure 2: Visual illustration accompanying the detailed analysis of system behavior and class-level performance.

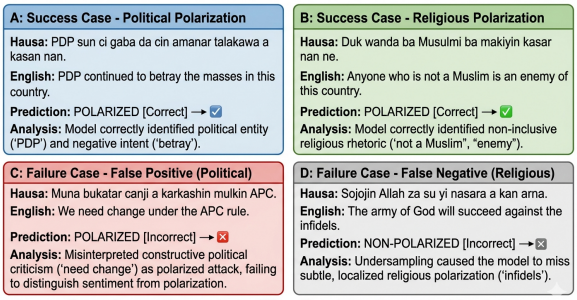

Figure 3: Qualitative Analysis of Hausa Polarization Detection (Subtask 2)

Figure 3: Additional visual summary supporting the error analysis and qualitative discussion of system predictions.

content due to lexical overlap. Second, ambiguous statements are frequently misinterpreted, showing the limitations of relying only on textual cues without broader context. Finally, the scarcity of training examples in some categories limits the model's ability to generalize effectively.

These observations suggest directions for future work, including incorporating richer contextual information, applying data augmentation techniques, and exploring calibration methods to better manage the precision–recall trade-off.

Figure 3 provides an additional visual summary related to the error patterns and qualitative analysis discussed above.

5 Conclusion

Our participation in SemEval-2026 highlights the trade-off between precision and recall in highly imbalanced datasets. By using Afro-XLMR-Large and dynamic undersampling, we built a system that is highly sensitive to polarization (recall above 60%). While this results in more false positives, it shows that low-resource models can successfully learn to flag harmful content even with limited training examples.

For reproducibility, we will release training scripts, configuration files, and inference code after the official evaluation period.

Ethical Considerations

Our system is designed as a decision-support tool for moderation, not a fully autonomous judge. Because false positives may disproportionately affect dialectal and identity-linked expressions, human review remains necessary before punitive actions are taken.

References

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026. POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. *arXiv preprint arXiv:2505.20624*. URL https://arxiv.org/abs/2505.20624.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, Dheeraj Kodati, Sahar Moradizeyveh, Firoj Alam, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Nelson Odhiambo Onyango, Clemencia Siro, Ibrahim Said Ahmad, Lilian Wanzare, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026. SemEval-2026 Task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, CA, USA. Association for Computational Linguistics.

Jacob Devlin, et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of SRW at HLT-NAACL*.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.

David Adelani, et al. 2022. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of EMNLP*.

Alexis Conneau, et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Jesujoba Alabi, Kwabena Amponsah, and David Adelani. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of COLING*.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*.

Tsung-Yi Lin, et al. 2017. Focal loss for dense object detection. In *Proceedings of ICCV*.