DeepSemantics at SemEval-2026 Task 9: Label-Wise Optimization with Adaptive Focal Loss for Polarization Manifestation Identification

Eliasse Tiao, Josue R. Edou, Mahugnon A. L. Gohouede
African Institute for Mathematical Sciences (AIMS), South Africa
eliasse@aims.ac.za, josue@aims.ac.za , aimeloick@aims.ac.za

Abstract

In this paper, we present our system for SemEval-2026 Task 9, which focuses on the fine-grained identification of polarization manifestations in multilingual social media content. Our approach combines transformer-based encoders (RoBERTa-base for English and Afro-XLM-R-small for Hausa) within a One-vs-Rest (OvR) framework, complemented by controlled oversampling, Adaptive Focal Loss, and label-wise threshold optimization. To mitigate severe class imbalance and label sparsity, we adopt language-specific optimization strategies supported by pairwise χ^2 independence analysis. Our system achieves macro-F1 scores of 0.464 in English and 0.192 in Hausa on the official test sets, ranking 5th in Hausa and 14th in English on the official leaderboard. Our code is publicly available [1].

1 Introduction

Social media messages often reflect strong and opposing attitudes related to social, political, or identity divisions, which contribute to online polarization (Conover et al., 2011; Garimella et al., 2018; Barberá, 2020; Kubin and von Sikorski, 2021). Subtask 3 of the POLAR 2026 shared task(Naseem et al., 2026a), Manifestation Identification, focuses on classifying messages according to multiple manifestations of polarization, such as Vilification, Dehumanization, Extreme Language, Lack of Empathy, Invalidation and Stereotype.

This task is particularly challenging due to severe class imbalance, sparse annotations, and low-resource languages such as Hausa, where training data is limited. Traditional joint multi-label approaches can often underperform in this setting, as rare labels generate sparse gradients that are dominated by frequent manifestations, resulting in poor recall for minority classes.

To address these challenges, we adopt a One-vs-Rest (OvR) strategy that models each manifestation independently. This label-wise formulation is complemented by language-specific optimization: Adaptive Focal Loss for English dynamically reweights gradients to emphasize rare classes, and weighted Binary Cross-Entropy with controlled oversampling for Hausa compensates for extreme imbalance. Thresholds for each label are optimized via stratified K-fold validation to improve decision calibration. Transformer encoders (RoBERTa-base for English and Davlan Afro-XLM-R-small for Hausa) provide contextual embeddings that enhance multilingual representation learning.

Experimental comparisons confirm that OvR consistently outperforms joint multi-label models, particularly for rare manifestations where joint modeling often fails. Our system achieves macro-F1 scores of 0.464 in English and 0.192 in Hausa on the official test sets, ranking 5th in Hausa and 14th in English. These results demonstrate robust performance under label sparsity and low-resource conditions, and motivate our label-wise modeling approach, which effectively captures fine-grained polarization manifestations while mitigating gradient interference in highly imbalanced multi-label settings.

2 Related Work

Fine-grained polarization analysis extends beyond binary detection by identifying specific manifestations expressed in text. This formulation naturally leads to a multi-label classification setting, where instances may exhibit multiple manifestations simultaneously. Prior work in multi-label learning has shown that joint modeling approaches often struggle under severe class imbalance and sparse label co-occurrence (Tsoumakas and Katakis, 2007; Read et al., 2011).

Transformer-based models such as BERT and

[1] https://github.com/TIAO-Eliasse/DeepSemantics-SemEval

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1632–1640

multilingual variants like XLM-R have become the dominant backbone for text classification tasks (Devlin, 2019; Conneau, 2020). While these models provide strong contextual representations, their effectiveness remains sensitive to imbalance in fine-grained classification scenarios.

To mitigate imbalance, previous studies have explored data-level strategies such as oversampling and text augmentation (Wei and Zou, 2019a; Feng et al., 2021), as well as loss-level approaches including Focal Loss (Lin et al., 2017). Problem-transformation methods such as One-vs-Rest classifiers have also been shown to improve robustness in highly imbalanced multi-label settings (Tsoumakas and Katakis, 2007).

Building on these findings, our approach integrates transformer representations with controlled oversampling, data augmentation, Adaptive Focal Loss, and a One-vs-Rest formulation, explicitly targeting the extreme sparsity and imbalance observed in manifestation identification.

3 System Overview

For manifestation identification, the focus shifts to label heterogeneity and rare classes. Traditional monolithic multi-label frameworks, which predict all labels jointly, often underperform on rare labels and fail to capture label-specific decision boundaries (Read et al., 2011; Tsoumakas and Katakis, 2007).

To address these limitations, we adopt a One-vs-Rest (OvR) strategy, modeling each manifestation independently.

To further address class imbalance, we combine controlled minority oversampling with loss-based reweighting strategies. For English, Adaptive Focal Loss dynamically down-weights easy examples and emphasizes rare manifestations. For Hausa, because of severe imbalance, we employ cost-sensitive learning via weighted binary cross-entropy combined with controlled oversampling.

3.1 Controlled Oversampling and Augmentation

To address the severe class imbalance, we implement a label-wise controlled oversampling strategy within our OvR framework. For each binary classifier, oversampling is triggered only when the imbalance ratio (negative-to-positive) exceeds a threshold of 5:1.

Specifically, the number of duplications R is dynamically calculated as: $R = \lceil (N_{neg}/N_{pos})/5 \rceil - 1$, where N_{neg} and N_{pos} are the counts of negative and positive instances for the target label, respectively. Positive instances are then duplicated R times to ensure the final training ratio does not exceed 5:1. To prevent the model from overfitting on identical samples, we follow duplication with Easy Data Augmentation (EDA) (Wei and Zou, 2019b), applying synonym replacement and random insertion to the oversampled set. This approach ensures that sampling is controlled independently for each manifestation, providing the necessary gradient signal for rare labels like lack_of_empathy without introducing the noise associated with full class balancing.

3.2 Dynamic Gradient Optimization via Adaptive Focal Loss

As introduced in the previous section, data augmentation and controlled oversampling mitigate class imbalance at the data level. However, these techniques alone are insufficient to address imbalance during optimization, particularly in One-vs-Rest settings where rare polarization labels generate sparse and noisy gradients. We therefore complement data-level balancing with a loss-level strategy based on Adaptive Focal Loss, which dynamically reweights gradient contributions according to prediction difficulty and class distribution (Lin et al., 2017).

Let x_i denote an input text with binary label $y_i \in \{0, 1\}$, and let $\hat{y}_i = \sigma(z_i)$ be the predicted probability. Unlike standard Binary Cross-Entropy (BCE), which assigns equal importance to all samples, focal loss down-weights well-classified instances and emphasizes hard examples. The loss for a single instance is defined as:

$$\mathcal{L}_{\mathrm{FL}}(x_i) = -\alpha_{y_i} \, (1 - p_i)^{\gamma} \, \log(p_i), \qquad (1)$$

where

$$p_i = \begin{cases} \hat{y}_i & \text{if} \quad y_i = 1, \\ 1 - \hat{y}_i & \text{if} \quad y_i = 0. \end{cases} \qquad (2)$$

The focusing parameter $\gamma > 0$ controls the degree to which easy examples are down-weighted, thereby preventing dominant negative samples from driving the optimization. In addition, class-dependent weights α_{y_i} are used to explicitly compensate for label imbalance:

$$\alpha_{y_i} = \begin{cases} \alpha \cdot w_+ & \text{if } y_i = 1, \\ (1-\alpha) \cdot w_- & \text{if } y_i = 0 \end{cases} \quad (3)$$

where w_+ and w_- are dynamically computed from inverse class frequencies.

From an optimization perspective, this formulation induces a form of dynamic gradient reweighting. For confidently predicted samples ($p_i \to 1$), the gradient magnitude decreases proportionally to:

$$\frac{\partial \mathcal{L}_{\text{FL}}}{\partial z_i} \propto (1-p_i)^\gamma,$$

which stabilizes training and prevents gradient domination by frequent classes. Consequently, rare polarization manifestations exert a stronger influence during learning.

Within our One-vs-Rest architecture, Adaptive Focal Loss is applied independently to each binary classifier, enabling label-specific optimization dynamics. In combination with data augmentation and oversampling, this strategy substantially improves recall for rare labels while maintaining precision on frequent ones.

3.3 Cost-Sensitive Optimization for Hausa

Preliminary experiments applying Adaptive Focal Loss to Hausa resulted in gradient vanishing and unstable convergence. We attribute this instability to the extreme label sparsity of the Hausa dataset ($< 1\%$ positive instances for several classes) — a regime in which the focal modulation term $(1-p_i)^\gamma$ suppresses gradients for the dominant negative class, leaving insufficient signal for minority class learning. Weighted BCE, by directly scaling the loss via $w_+ = N_{neg}/N_{pos}$, provides more stable gradient updates under such extreme imbalance. For an instance x_i, the loss is defined as:

$$\mathcal{L}_{\text{BCE}}(x_i) = -w_+ y_i \log(\hat{y}_i) - w_-(1-y_i)\log(1-\hat{y}_i) \quad (4)$$

where $w_+ = N_{neg}/N_{pos}$ scales the contribution of the minority class and $w_- = 1$. This cost-sensitive approach compensates for the imbalance by directly inflating the loss of rare manifestations, preventing the model from stagnating on the dominant negative class.

3.4 Threshold Optimization

Even with data-level balancing and loss reweighting, using a fixed decision threshold (e.g., 0.5) can be suboptimal, particularly for rare polarization labels. To address this, we perform a complementary threshold optimization step at inference time.

Thresholds are selected independently for each One-vs-Rest classifier by maximizing the F1-score on the validation set. To improve robustness, the validation data is split into stratified $K = 5$folds, thresholds are computed per fold, and the final threshold is obtained by averaging across folds.

It is applied uniformly across English and Hausa classifiers.

4 Data Exploration

In this section we begin by analyzing the distributional properties of the datasets , with a particular focus on class imbalance and label co-occurrence patterns. These aspects are critical, as they directly influence model design choices and evaluation strategies.

4.1 Subtask 3: Manifestation Identification

4.1.1 Imbalance analysis in dataset

We analyze the manifestation-level annotations, which further increase the granularity and sparsity of the task. Figures 1 and 2 show a highly imbalanced label distribution in both languages.

Most manifestations are extremely rare, reinforcing the need for label-specific learning strategies.

4.2 Independence Analysis of Manifestation Identifications

To examine label interactions, we conduct pairwise χ^2 independence tests on manifestation labels. As shown in Table 3, the English dataset exhibits statistically significant dependencies across nearly all label pairs, indicating frequent co-occurrence of manifestations. In contrast, the Hausa dataset shows significant dependencies for only a subset of pairs, while several others display high p-values, suggesting weaker or approximate independence. Overall, the dependency structure is heterogeneous and language-dependent.

Importantly, statistical dependence between labels does not necessarily imply that joint multi-label modeling is optimal. Prior work in multi-label learning highlights that explicitly modeling label correlations is not universally beneficial and may depend on data characteristics (Tsoumakas and Katakis, 2007). Under severe class imbalance, joint optimization can amplify gradient dominance

of frequent labels, potentially limiting the learning of rare manifestations.

Given the imbalance-sensitive nature of the POLAR task and the uneven dependency patterns observed across languages, we adopt a One-vs-Rest formulation. This approach enables label-specific optimization, stabilizes training for minority classes, and remains robust to varying levels of inter-label dependence.

5 Experimental Setup

For Subtask 3 (manifestation identification), we split the initial training data into training and validation sets using a Multilabel Stratified Shuffle Split (80/20), preserving the distribution of all polarization labels. The validation set was used for model selection, hyperparameter tuning, and threshold optimization.

Hyperparameters such as learning rate, batch size, number of epochs, and loss-specific parameters were tuned on the validation set. Thresholds for each One-vs-Rest classifier were optimized per label via grid search over [0.01, 0.99] using stratified K-fold splits, with final thresholds averaged across folds to improve robustness.

To address class imbalance, different strategies were adopted for each language. For English, Adaptive Focal Loss dynamically reweighted gradients to improve recall for rare manifestations while preserving precision. For Hausa, cost-sensitive learning with weighted BCE combined with controlled oversampling (triggered when imbalance ratio exceeded 5:1) was applied.

RoBERTa-base (Liu et al., 2019) (English) and Afro-XLM-R-small (Hausa) (Alabi et al., 2022) encoders were fine-tuned for Subtask 3.

6 Results

6.1 Validation results and model selection

We first compare the One-vs-Rest (OvR) and Multi-label (ML) approaches on the English validation set. As shown in Table 1, OvR consistently outperforms the joint Multi-label formulation across most manifestations.

In particular, substantial gains are observed for Extreme Language (+0.27), Invalidation (+0.24), and Lack of Empathy (+0.36). The latter is especially notable, as the Multi-label model completely fails to detect this rare class (F1 = 0.00), while OvR achieves 0.36.

	English		Hausa
Label	**OvR**	**Multi-label**	**OvR**
Stereotype	0.560	0.420	0.286
Vilification	0.590	0.580	0.111
Dehumanization	0.430	0.380	0.364
Extreme Language	0.570	0.300	0.375
Lack of Empathy	0.360	0.000	0.057
Invalidation	0.530	0.290	0.000
Macro-F1	**0.500**	0.330	**0.200**

Table 1: Validation set F1-scores for Subtask 3. The One-vs-Rest (OvR) approach consistently outperforms the joint Multi-label baseline across both languages.

Overall, OvR reaches a macro-F1 of 0.50, compared to 0.33 for the Multi-label approach. This 17-point improvement confirms that independent label modeling is better suited to the sparse and weakly co-occurring manifestation structure of the dataset.

These findings support our hypothesis that joint modeling struggles under extreme label imbalance and limited co-occurrence patterns, as minority-class gradients become dominated by frequent manifestations.

For Hausa, only the OvR framework was adopted based on the strong validation results obtained in English. Using Davlan Afro-XLM-R-small, the model achieves a macro-F1 of 0.20 on the validation set (Table 1). Performance varies considerably across labels. Extreme Language (0.375) and Dehumanization (0.364) show relatively stronger detection capability, whereas Lack of Empathy (0.057) and Invalidation (0.00) remain highly challenging.

6.2 Test Results

Based on validation performance, the OvR approach was selected for official submission in both languages. Table 2 reports the test results.

In English, the system achieves a macro-F1 of 0.464. Strongest performance is observed for Vilification (0.625) and Extreme Language (0.564), while Dehumanization (0.378) and Lack of Empathy (0.356) remain comparatively difficult.

For Hausa, the macro-F1 reaches 0.192. Although lower than English, the model achieves non-trivial detection across most manifestations, though performance remains limited throughout, with Stereotype (0.297) and Dehumanization (0.252) showing relatively stable performance.

Label	English F1	Hausa F1
Stereotype	0.420	0.297
Vilification	0.625	0.165
Dehumanization	0.378	0.252
Extreme Language	0.564	0.185
Lack of Empathy	0.356	0.102
Invalidation	0.442	0.154
Macro-F1	0.464	0.192

Table 2: F1-scores for Subtask 3 (Manifestation Identification) on the test set: English vs Hausa (One-vs-Rest).

7 Error Analysis

Analysis of model predictions reveals distinct structural challenges across both languages.

English System As shown in Figure 9, closely related categories such as dehumanization and stereotype are frequently absorbed into the broader vilification class. Furthermore, invalidation and lack_of_empathy act as attractor classes, being consistently over-predicted. Error rate analysis (see Fig 6 in Appendix) indicates that performance degradation is triggered by the mere presence of polarization (error rates $> 92\%$) rather than multi-label complexity. Sensitivity to contextual length is also high, with error rates increasing by +1.18% per additional word(Figure 7).

Hausa System The Hausa system exhibits a prediction collapse into a limited subset of classes. Figure 10 shows that stereotype and dehumanization are frequently misclassified as invalidation. Unlike the English case, the primary bottleneck in Hausa is positive-class sparsity rather than semantic boundary modeling; 92.5% of polarized texts contain only a single label (Figure 8a). Contextual length plays a secondary role (+0.45% error increase per word, Figure 8b) compared to the extreme data imbalance documented in Section 4. To address potential concerns regarding the highly optimized thresholds (0.990) for certain Hausa labels like Vilification and Extreme Language, our error analysis confirms the model does not degenerate into a trivial majority-class (all-negative) classifier. As shown in the FP Co-confusion matrix (Figure 10), the model actively triggers positive predictions that cross this strict threshold. For instance, it generates at least 45 positive predictions for Vilification and 7 for Extreme Language just among its inter-class confusions alone. This confirms that the threshold effectively isolates the strongest signals in a severely imbalanced setting (e.g., 1.3% prevalence for Vilification) rather than suppressing predictions entirely, which is further validated by the non-zero macro-F1 scores.

8 Conclusion

In this work, we presented our system for the POLAR 2026 shared task on multilingual manifestation identification of polarization in social media. Our approach integrates transformer-based encoders with controlled oversampling, Adaptive Focal Loss, and label-wise threshold optimization within a One-vs-Rest framework, explicitly designed to address severe class imbalance and label sparsity.

On the validation set, the One-vs-Rest framework achieved a macro-F1 of 0.50 for English and 0.20 for Hausa. On the official test set, the system obtained macro-F1 scores of 0.464 for English and 0.192 for Hausa, demonstrating stable behavior across both languages despite the severe class sparsity of the Hausa setting.

Future work will explore hybrid strategies combining label-wise optimization with structured modeling of label dependencies, enhanced cross-lingual transfer techniques for low-resource languages such as Hausa, access to larger models and GPUs and additional data collection.

9 Limitations

Our study has three main limitations. First, the free Google Colab environment prevented us from exploring larger transformer models or extensive hyperparameter tuning. Second, the small and highly imbalanced Hausa dataset limits generalization for rare manifestations. Finally, results rely on a single run (seed = 42), though our stratified K-fold threshold optimization demonstrated stable convergence, supporting the reliability of the reported metrics.

Acknowledgments

We would like to thank Shamsuddeen Hassan Muhammad and Idris Abdulmumin for their course on Natural Language Processing, which provided the foundation for this work, and for their encouragement to participate in the POLAR 2026 shared task competition.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pablo Barberá. 2020. Social media, echo chambers, and political polarization. *Social Media + Society*, 6(3):1–12.

Alexis Conneau. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter.

Jacob Devlin. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of ACL-IJCNLP*, pages 968–988.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. *Preprint*, arXiv:1801.01665.

Emily Kubin and Christian von Sikorski. 2021. Three dimensions of political polarization. *Political Psychology*, 42(S1):3–33.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

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, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Grigorios Tsoumakas and Ioannis Manousos Katakis. 2007. Multi-label classification: An overview. *Int. J. Data Warehous. Min.*, 3:1–13.

Jason Wei and Kai Zou. 2019a. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*, pages 6382–6388.

Jason Wei and Kai Zou. 2019b. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *Preprint*, arXiv:1901.11196.

A Appendix

A.1 Subtask 3: Manifestation Identification

The third subtask (Naseem et al., 2026a) focuses on identifying the rhetorical and linguistic manifestations of polarization in social media posts. Each instance can be associated with one or more labels, including , Vilification, Dehumanization, Extreme Language, Lack of Empathy, and Invalidation (Naseem et al., 2026b) .This subtask is formulated as a multi-label classification problem, aiming to capture how polarized opinions are expressed rather than just whether they exist.

Table 4 summarizes the final configurations for English and Hausa models.

A.2 Optimized Decision Thresholds

Table 5 reports the label-wise optimized thresholds for English and Hausa datasets.

A.3 Figures

Error analysis :Qualitative examples. Table 6 illustrates representative error cases for each system. For English, texts discussing political or geopolitical topics without explicit hostility such as analytical commentary on denazification or culture

Label 1	Label 2	EN	HA
vilification	extreme_language	3.70e-271	2.76e-04
stereotype	vilification	1.04e-168	5.59e-02
vilification	dehumanization	8.80e-147	4.44e-01
vilification	invalidation	4.91e-143	1.00e+00†
extreme_language	invalidation	2.44e-124	1.64e-02
vilification	lack_of_empathy	6.90e-124	1.00e+00†
stereotype	extreme_language	1.57e-122	8.82e-16
extreme_language	lack_of_empathy	2.85e-120	1.11e-01
stereotype	dehumanization	1.29e-113	6.89e-02
dehumanization	extreme_language	2.61e-100	5.81e-23
stereotype	invalidation	3.06e-76	6.57e-02
dehumanization	lack_of_empathy	1.05e-71	1.03e-03
lack_of_empathy	invalidation	8.57e-68	1.00e+00†
dehumanization	invalidation	1.22e-56	1.00e+00†
stereotype	lack_of_empathy	6.46e-56	2.85e-04

Table 3: P-values of pairwise χ^2 independence tests for manifestation labels in English (EN) and Hausa (HA). †Values of 1.00 indicate zero observed co-occurrence in the Hausa dataset; the χ^2 statistic is undefined for these pairs, which are treated as independent.

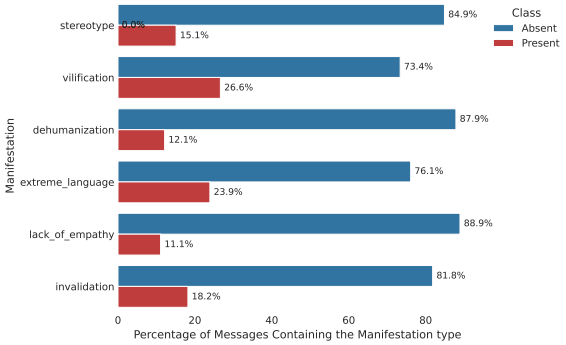


Figure 1: Distribution of manifestation labels in the English dataset.

wars — are systematically over-predicted across all six labels, revealing that the model conflates politically charged discourse with polarization manifestations. Conversely, highly polarized short texts such as "Kamala Harris has the easiest job ever" are entirely missed, suggesting that implicit or ironic framing falls below the model's detection threshold. For Hausa, the dominant error pattern is label substitution: the model predicts plausible but incorrect labels, reflecting unstable decision boundaries under low-resource conditions.

B Supplemental Figures and Qualitative Analysis

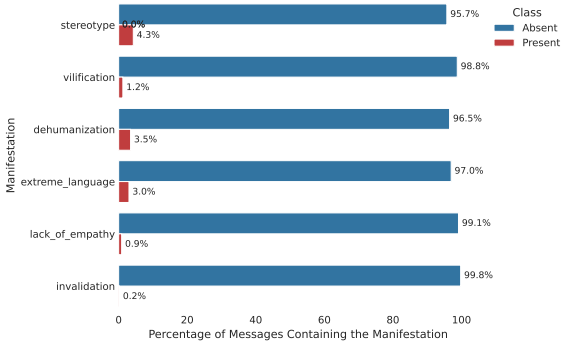


Figure 2: Distribution of manifestation labels in the Hausa dataset.

	EN	**HA**
Model	RoBERTa-base	Afro-XLM-R-small
LR	1e−5	1e−5
Batch size	4	4
Grad. Accum.	1	2
Epochs	5	3
Weight decay	0.01	0.01
Max length	512	512
Loss	Focal	Weighted BCE
γ	2.0	–
α	0.25	–
Class weight	–	$N_{\text{neg}}/N_{\text{pos}}$
Oversampling	–	ratio > 5
Threshold search	[0.01–0.99]	[0.01–0.99]

Table 4: Final training configurations (EN = English, HA = Hausa).

Manifestation	English	Hausa
Vilification	0.18	0.990
Extreme Language	0.41	0.990
Stereotype	0.57	0.900
Invalidation	0.21	0.500
Lack of Empathy	0.07	0.010
Dehumanization	0.31	0.090

Table 5: Label-wise optimized thresholds.

Figure 3: Thematic Analysis of Bigrams by Hausa Discourse Category

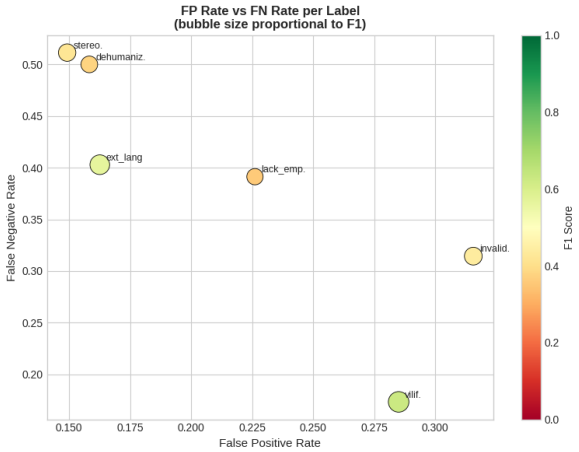


Figure 4: FP vs FN rate scatter (English)

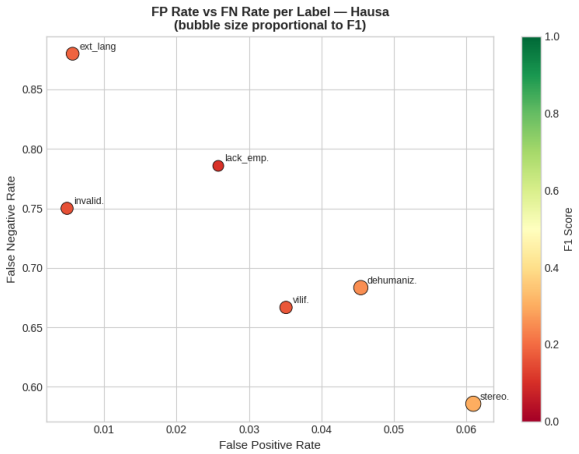


Figure 5: FP vs FN rate scatter (Hausa)

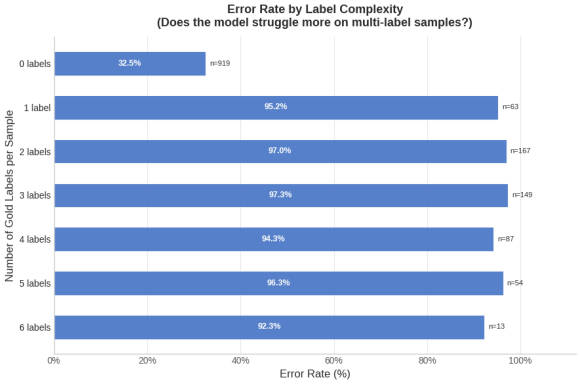


Figure 6: Model error rate (English case) by number of gold labels per text

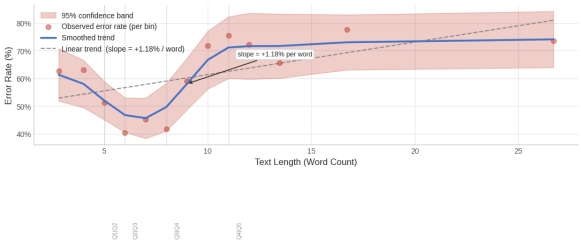
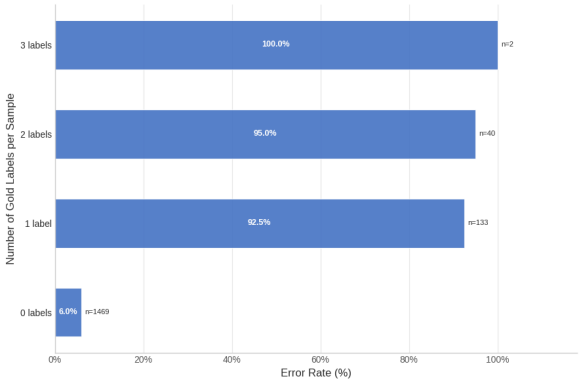
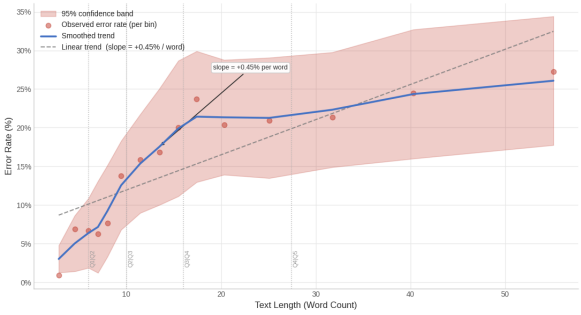


Figure 7: Error rate of RoBERTa-base (English case) as a function of text length



(a) Model Error Rate (Hausa case) by number of labels in the text



(b) Error rate of Afro-XLMR-small (Hausa case) as a function of text length

Figure 8: Error analysis for the Hausa system.

Text	Gold	Pred.
EN — False Positive		
"fear and ignorance.. encouraged by populism to push for culture wars."	∅	all 6
EN — False Negative		
"Kamala Harris has the easiest job ever"	all 6	∅
HA — Label substitution		
"Mun aura amarar yai da aidun addinin Islama – Faransa"	ext_lang	vilif., stereo., invalid., lack_emp.

Table 6: Error cases. ∅ = no label predicted or expected. "all 6" = full manifestation set.

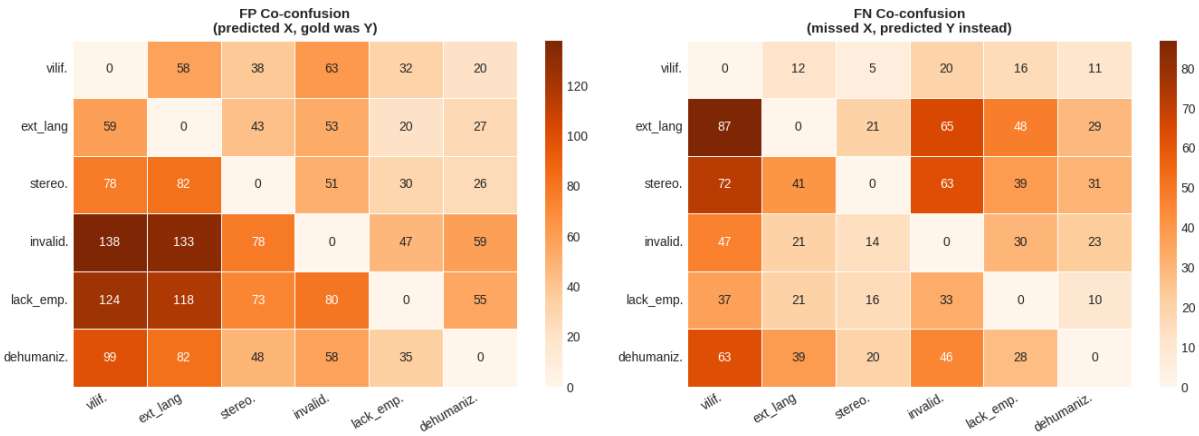


Figure 9: Confusion matrices for Subtask 3. English shows semantic overlap with vilification.

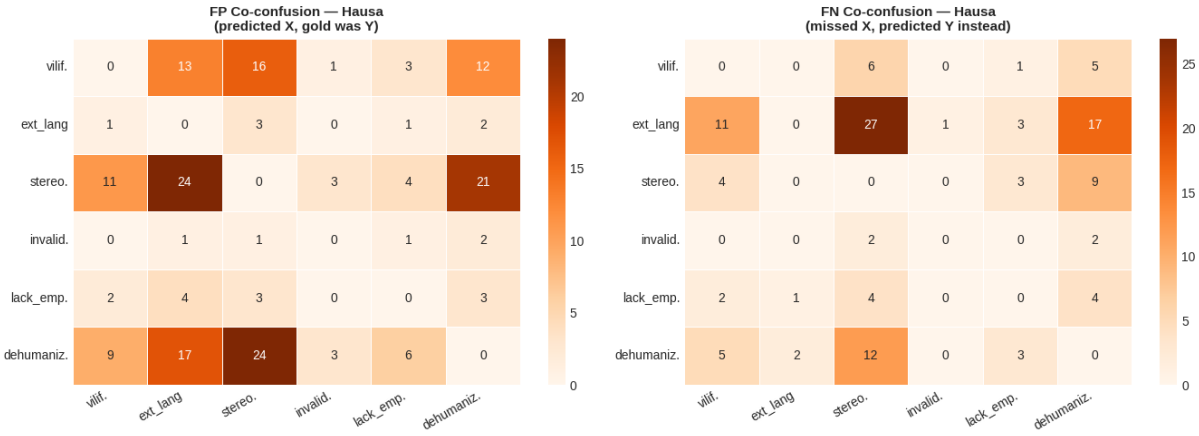


Figure 10: Confusion matrices for Subtask 3. Hausa shows prediction collapse due to sparsity.