Aaron at SemEval-2026 Task 9: Multilingual Polarization Detection Using Transformer-Based Models with Class Weighting and Threshold Tuning

Aaron Bundi Anampiu
African Institute for Mathematical Sciences, South Africa
aaronbundi@aims.ac.za

Abstract

This paper describes our submission to SemEval-2026 Task 9 on detecting multilingual, multicultural, and multievent online polarization. We address all three subtasks: binary polarization detection, polarization type classification, and manifestation identification for English and Swahili. Our approach leverages transformer-based models (RoBERTa-base for English, AfroXLMR-base for Swahili) with class-weighted loss functions to address severe label imbalance and per-label threshold tuning to optimize multi-label classification. On the test set, we achieve F1 macro scores of 0.7901 (English) and 0.7910 (Swahili) for Subtask 1, 0.4615 (English) and 0.4808 (Swahili) for Subtask 2 and 0.4791 (English) and 0.5830 (Swahili) for Subtask 3, which give competitive performance on the leaderboard, demonstrating the effectiveness of our methods for handling imbalanced multi-label polarization detection. Our error analysis reveals that models struggle with dehumanization detection and lack of empathy.

1 Introduction

Social media platforms have become central spaces for public discourse, enabling millions of users to share opinions, engage in debates, and form communities. However, these platforms have also witnessed a concerning rise in polarized content and messages that divide audiences along ideological, political, racial, religious, or other lines (Grover et al., 2019).

Polarization poses significant societal challenges. It amplifies echo chambers, reduces constructive dialogue, can incite real-world conflicts, and disproportionately affects marginalized communities (Conover et al., 2011). Automatic detection of polarizing content is therefore crucial for content moderation, understanding radicalization mechanisms, and promoting healthier online discourse.

SemEval-2026 Task 9 (Naseem et al., 2026a) addresses this challenge through three subtasks: (1) binary detection of polarization, (2) multi-label classification of polarization types (political, racial/ethnic, religious, gender/sexual, other), and (3) multi-label identification of polarization manifestations (stereotype, vilification, dehumanization, extreme language, lack of empathy, invalidation). The task covers 22 languages from diverse platforms including news websites, Reddit, blogs, and regional forums.

We focus on English and Swahili, employing a transformer-based approach with three key contributions: (1) language-specific model selection (RoBERTa (Liu et al., 2019) for English and AfroXLMR base (Alabi et al., 2022) for Swahili), (2) class-weighted loss functions to address severe label imbalance in multi-label tasks, and (3) per-label threshold tuning to optimize classification boundaries. Our system achieves strong performance on all the Subtasks. It was ranked 2nd for Swahili on Subtask 3 among the 16 participating teams.

2 Background and Related Work

2.1 Task Description

SemEval-2026 Task 9 provides 3,000 to 5,000 annotated instances per language (Naseem et al., 2026b) spanning elections, conflicts, gender rights, and migration.

Subtask 1 (Binary Detection) classifies text as polarized or non-polarized based on context and overall meaning.

Subtask 2 (Type Classification) identifies five polarization types: political/ideological, racial/ethnic, religious, gender/sexual, and other (multi-label).

Subtask 3 (Manifestation Identification) classifies six manifestations: stereotype, vilification, de-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2262–2267

humanization, extreme language, lack of empathy, and invalidation (multi-label).

2.2 Related Work

Prior work on detecting harmful online content has primarily focused on hate speech (Fortuna and Nunes, 2018), toxic comments (Wulczyn et al., 2017), and offensive language (Zampieri et al., 2019). These tasks share similarities with polarization detection but differ in important ways. Early approaches used traditional machine learning with hand-crafted features (Davidson et al., 2017), including n-grams, sentiment lexicons, and linguistic patterns. More recent work has adopted deep learning, with Badjatiya et al. (2017) demonstrating that LSTMs with word embeddings significantly outperform traditional methods. Devlin et al. (2019) showed that BERT-based models achieve better results on hate speech benchmarks.

The development of multilingual pre-trained models has enabled effective transfer across languages. mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) demonstrated strong zero-shot cross-lingual transfer capabilities. However, these models are trained primarily on high-resource languages and may underperform on African languages.

Recent work has developed language models specifically for African languages. AfriBERTa (Ogueji et al., 2021) and AfroXLMR (Alabi et al., 2022) were pre-trained on diverse African language corpora, including Swahili, and have shown superior performance on African NLP tasks compared to general multilingual models. We adopt AfroXLMR for Swahili based on these findings.

3 System Overview

Our pipeline has four components: preprocessing, model architecture, training with class-weighted loss, and post-training threshold tuning. Figure 1 illustrates the full pipeline.

3.1 Data Preprocessing

We perform minimal preprocessing: retain original casing, preserve emojis and special characters that may signal polarization, and remove only control characters and null bytes. Data are split 80/20 into train/validation using stratified sampling to maintain class distribution.

3.2 Model Architecture

English Models: RoBERTa-base (Liu et al., 2019) (125M parameters) was selected over BERT and XLM-RoBERTa based on ablation results (Table 6): its dynamic masking and removal of next-sentence prediction yield the best English-specific performance, while XLM-RoBERTa's multilingual pre-training dilutes English-specific representations.

Swahili Models: AfroXLMR-base (Alabi et al., 2022) (270M parameters) is specifically adapted for African languages through continued pre-training on diverse African language corpora. It significantly outperforms multilingual alternatives like mBERT and XLM-RoBERTa on Swahili tasks (Adelani et al., 2022).

Tokenization: RoBERTa's byte-level BPE tokenizer (50K vocab) for English; AfroXLMR's SentencePiece tokenizer (250K vocab) for Swahili. Maximum sequence length is 128 tokens for Subtask 1 and 256 for Subtasks 2–3.

Classification head: A dropout layer (rate 0.1) followed by a linear layer maps 768-dimensional embeddings to 2, 5, or 6 output logits for Subtasks 1, 2, and 3 respectively.

3.3 Class-Weighted Loss Function

To address severe label imbalance, we implement class-weighted binary cross-entropy loss with positive class weight for label computed as:

$$w_j = \frac{n^j_{neg}}{n^j_{pos}} \tag{1}$$

where n^j_{neg} and n^j_{pos} are negative and positive sample counts for label j. This increases the penalty for misclassifying minority classes. For Subtask 2 English, weights range from 1.73 (political) to 41.95 (gender/sexual), reflecting the extreme rarity of gender-related polarization.

Weights for Swahili Subtask 2 reach 48.05 (gender/sexual). Training loss curves showed no instability; gradient clipping (max-norm 1.0) was applied throughout as a precaution.

3.4 Threshold Tuning

After training we perform per-label threshold search on the validation set, scanning $[0.1, 0.9]$ in steps of 0.05 to maximise macro F1 per label. This improves macro F1 by 0.87–5.98 points over the standard 0.5 threshold.

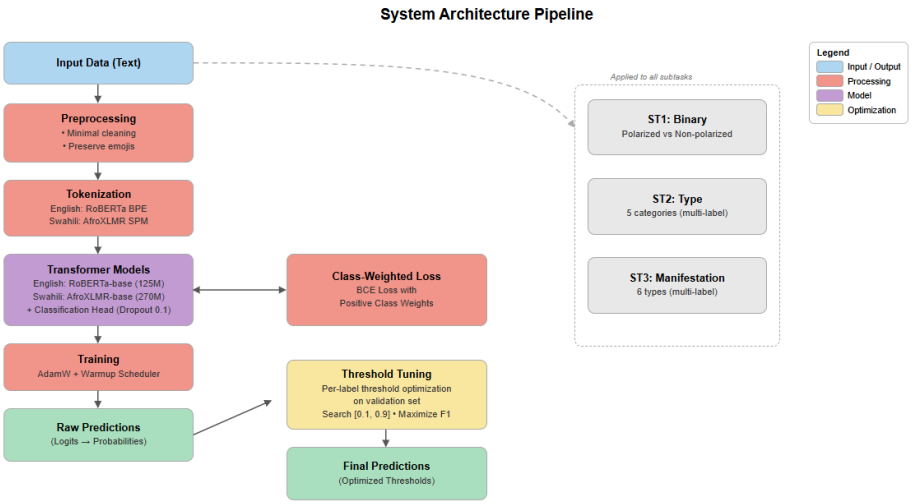

Figure 1: Complete system architecture pipeline showing the four main components: preprocessing, model architecture (language-specific transformers), training with class-weighted loss, and post-training threshold tuning. The pipeline is applied to all three subtasks.

3.5 Training Configuration

We use AdamW (Loshchilov and Hutter, 2019) (learning rate 2e-5, weight decay 0.01) with a linear warmup schedule, chosen as the standard optimizer for fine-tuning transformers. Batch size is 16 for Subtask 1 and 8 with gradient accumulation ($\times 2$) for Subtasks 2–3 to fit 256-token sequences in 15 GB GPU memory. We train for 5–8 epochs with early stopping (patience 2) based on validation macro F1, using FP16 mixed precision on a single GPU via Google Colab. Random seed is fixed to 42 for all experiments.

4 Experimental Setup

4.1 Data Splits

We use provided train/dev splits. Training data are further split 80/20 using stratified sampling (binary labels for Subtask 1; label-count strata for Subtasks 2–3). The test set is used exclusively for final evaluation.

4.2 Evaluation Metrics

All subtasks use macro F1 as the primary metric. We compare our system against two baselines: (i) a **POLAR Baseline**, and (ii) an **mBERT fine-tuned baseline** applied uniformly across all languages without class weighting. Table 1 summarises this comparison.

ST	Lang	POLAR Baseline	mBERT	Ours
1	EN	0.7802	0.7210	**0.7901**
	SW	0.7571	0.6981	**0.7910**
2	EN	0.3333	0.3810	**0.4615**
	SW	0.4417	0.3641	**0.4808**
3	EN	0.4100	0.4121	**0.4791**
	SW	0.2205	0.4380	**0.5830**

Table 1: Test macro F1 comparison with POLAR baselines. ST = Subtask, EN = English, SW = Swahili. Our system consistently outperforms both baselines across all subtasks and languages.

4.3 Implementation

We implement our system using Hugging Face Transformers 4.30.0, PyTorch 2.0, and scikit-learn 1.3.0.

5 Results

5.1 Main Results

Table 2 presents results across all Subtasks.

Subtask 1 achieves our strongest performance (0.7901 EN, 0.7910 SW). Swahili slightly outperforms English, likely due to a more balanced class distribution (50.1% polarized in Swahili vs. 36.5% in English). We ranked 11 out of 34 teams for Swahili and 24 out of 44 teams for English on this Subtask.

Subtask 2 presents challenges from extreme imbalance and multi-label complexity. We achieved

Subtask	Language	Val F1	Test F1
1. Binary Detection	English	0.8017	**0.7901**
	Swahili	0.7766	**0.7910**
2. Type Classification	English	0.4721	**0.4615**
	Swahili	0.4598	**0.4808**
3. Manifestation ID	English	0.5226	**0.4791**
	Swahili	0.5742	**0.5830**

Table 2: F1 macro scores on validation and test sets. Bold indicates official test scores used for ranking in codabench.

0.4615 (EN) and 0.4808 (SW); the ranking for Swahili was 10 out of 22 teams and for English, 17 out of 29 teams. Swahili benefited from a more balanced label distribution and AfroXLMR's stronger multilingual capabilities.

Subtask 3 yields our most competitive results: 0.4791 (EN) and 0.5830 (SW). We ranked 2nd out of 16 teams for Swahili and, for English, 10 out of 18 teams on this Subtask.

The validation-test gap ranges from 2.29 to 10.58 points. The largest discrepancy is Subtask 3 English (4.35 points). We attribute this to two compounding factors: (i) *threshold overfitting*—per-label thresholds are optimised on a single 20% validation split, making them sensitive to its specific label distribution; and (ii) *domain shift*—the test set may contain text from platforms or events not well-represented in training. Cross-validation for threshold selection is a natural mitigation, though it substantially increases compute time.

5.2 Per-Label Analysis

Tables 3 and 4 show per-label F1 scores.

Label	English	Swahili
Political	**0.6682**	0.3256
Racial/Ethnic	0.5082	**0.7972**
Religious	0.5957	0.6387
Gender/Sexual	0.3200	0.2590
Other	0.3000	0.2784
Macro Average	**0.4784**	**0.4598**

Table 3: Validation F1 per label, Subtask 2.

English excels at political polarization (0.6682 vs. 0.3256), while Swahili dominates racial/ethnic detection (0.7972 vs. 0.5082), reflecting corpus composition (35.5% racial/ethnic in Swahili vs. 8.7% in English). Gender/sexual and other categories remain hard for both languages (F1 < 0.33) due to extreme rarity (2–4% of samples), where class weighting alone is insufficient.

Manifestation	English	Swahili
Stereotype	0.5128	**0.7374**
Vilification	0.6615	**0.7208**
Dehumanization	0.4453	0.3241
Extreme Language	**0.6022**	0.4945
Lack of Empathy	0.3724	**0.6111**
Invalidation	0.5413	0.5575
Macro Average	**0.5226**	**0.5742**

Table 4: Validation F1 per manifestation, Subtask 3.

Dehumanization is universally challenging (0.4453 EN, 0.3241 SW), requiring nuanced semantic understanding to distinguish metaphorical from explicit dehumanization.

5.3 Ablation Studies

Table 5 shows the impact of class weighting on Subtask 2 English. Without weighting the model underpredicts minority labels (41.23% macro F1). Class weighting improves macro F1 by +5.98 points and micro F1 by +11.52 points; rare labels (religious, gender/sexual, other) gain 10–15 F1 points while frequent labels remain stable.

We conducted the primary ablation on English for conciseness. Qualitatively, we observed similar trends for Swahili: class weighting consistently reduced false negatives on rare labels, consistent with prior work on imbalanced classification (Johnson and Khoshgoftaar, 2019).

Configuration	Val F1 Macro	Val F1 Micro
No weighting	0.4123	0.3894
+ Class weights	**0.4721**	**0.5046**
Improvement	+0.0598	+0.1152

Table 5: Class weighting ablation, Subtask 2 English.

Table 6 compares model architectures for English Subtask 1. RoBERTa-base outperforms BERT by 2.58%, justifying our model choice. XLM-RoBERTa underperforms despite larger size, confirming that multilingual pre-training dilutes English-specific representations.

5.4 Error Analysis

We analyzed 100 randomly sampled errors from each subtask to identify failure patterns.

Subtask 1: False positives arise from neutral political headlines misclassified as polarizing (e.g.,

Model	Params	Val F1	Δ%
DistilBERT	66M	0.7840	-4.82
BERT-base	110M	0.8064	-2.58
XLM-RoBERTa	270M	0.8201	-1.21
RoBERTa-base	**125M**	**0.8322**	**–**

Table 6: Model comparison, English Subtask 1. Δ = gap from best.

"Senate passes infrastructure bill with bipartisan support") and hyperbolic sports comments (e.g., "Lakers will destroy the Celtics"). The model conflates strong language with polarization.

Subtask 2: Multi-label ambiguity is the primary failure mode (e.g., racial undertones in political statements are missed). The gender/sexual label (2.2% frequency) is rarely predicted despite class weighting.

Subtask 3: Dehumanization is the hardest manifestation (31% of errors). Lack-of-empathy detection requires theory-of-mind reasoning beyond standard transformer capabilities.

6 Conclusion

We presented a transformer-based system combining language-specific model selection, class-weighted loss, and per-label threshold tuning for multilingual polarization detection. Our system outperforms both POLAR baseline and mBERT baselines across all Subtasks and languages, ranking 2nd for Swahili on Subtask 3. Key error modes—political topic conflation, multi-label ambiguity, and pragmatic inference gaps—point to future work on pragmatic reasoning, cultural modeling, and cross-lingual transfer.

Future directions include incorporating external knowledge bases to recognize sarcasm in language and multi-task learning across subtasks to leverage complementary signals. Additionally, exploring cross-lingual transfer learning could improve low-resource language performance by leveraging knowledge from high-resource languages.

Ethical Considerations

Polarization detection systems may pose risks if misused: classifiers could unfairly target marginalised communities or suppress legitimate political expression. All data were gathered from publicly accessible platforms for research purposes. We advocate for human oversight in any real-world deployment.

Acknowledgements

We acknowledge and thank the organizers of Semeval-2026 Task 9 for providing dataset and evaluation framework.

References

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen H Muhammad, Peter Nabende, and 1 others. 2022. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508.

Jesujoba 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.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Purva Grover, Arpan Kumar Kar, Yogesh K Dwivedi, and Marijn Janssen. 2019. Polarization and acculturation in us election 2016 outcomes–can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145:438–460.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st workshop on multilingual representation learning*, pages 116–126.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

A Hyperparameters

Table 7 summarizes all hyperparameters used in our experiments.

Hyperparameter	Value
Model Architecture	
Max sequence length (ST1)	128
Max sequence length (ST2/3)	256
Dropout rate	0.1
Training	
Optimizer	AdamW
Learning rate	2e-5
Batch size (ST1)	16
Batch size (ST2/3)	8
Gradient accumulation	2 (ST2/3 only)
Epochs (ST1)	5
Epochs (ST2/3)	8
Weight decay	0.01
Warmup ratio (ST1)	0.10
Warmup ratio (ST2/3)	0.15
Early stopping patience	2
Mixed precision	FP16
Threshold Tuning	
Search range	[0.1, 0.9]
Step size	0.05

Table 7: Complete hyperparameter configuration. ST = Subtask.

B Class Weight Calculations

Table 8 shows complete class weight calculations for all subtasks.

Subtask	Label	Eng	Swa
ST1	Non-polarized	0.79	1.00
	Polarized	1.37	1.00
ST2	Political	1.73	36.28
	Racial/Ethnic	10.45	1.83
	Religious	27.63	27.10
	Gender/Sexual	41.95	48.05
	Other	25.03	11.26
ST3	Stereotype	5.54	1.53
	Vilification	2.75	1.42
	Dehumanization	7.29	6.69
	Extreme Language	3.18	3.08
	Lack of Empathy	8.11	2.33
	Invalidation	4.64	3.29

Table 8: Complete class weights for all subtasks and languages. Weights computed as n_{neg}/n_{pos}.