SemEval 2026 Task 9

Detecting Multilingual, Multicultural and Multievent Online Polarization (POLAR)

Zeynep Şahin 30553,
Suat Emre Karabıçak 30649,
Sıla Horozoğlu 30916,
Korhan Erdoğdu 30838

Introduction- What is Polarization Detection?

- Social media amplifies divisive language and "us vs. them" narratives
    - Polarization ≠ Hate Speech ≠ SentimentHate speech: Targets individuals/groups with harmful content
    - Sentiment: Positive or negative emotion
    - Polarization: Language that divides audiences into opposing camps

- Polarized: "Those politicians are destroying OUR country!"
- Non-polarized: "I disagree with this policy decision"

Our Task - SemEval-2026 Task 9: Multilingual Challenge

- **Goal: Binary classification (polarized vs. non-polarized). The output label is either polarized (True=1) or non-polarized (False=0).**
- **Data: Social media posts in 22 languages (websites, Reddit, blogs, Bluesky, and regional forums covering topics like elections, conflicts, gender rights and migration.)**
- **Challenge: One model must work across all 22 languages**
- **22 Languages: Amharic, Arabic, Chinese, English, German, Hausa, Hindi, Italian, Nepali, Persian, Spanish, Turkish, Urdu**

Naseem et al. (2025) - POLAR dataset

Dataset Overview

```
================================================================
SUMMARY
================================================================
Total languages: 22
Total training samples: 73,681
Total dev samples: 3,687
Total samples: 77,368
Average samples per language: 3349
```

```
================================================================
CLASS BALANCE PER LANGUAGE
================================================================
amh:  2518 polarized (75.6%),   814 non-polarized (24.4%) - HEAVILY POLARIZED
arb:  1512 polarized (44.7%),  1868 non-polarized (55.3%) - BALANCED
ben:  1424 polarized (42.7%),  1909 non-polarized (57.3%) - BALANCED
deu:  1512 polarized (47.5%),  1668 non-polarized (52.5%) - BALANCED
eng:  1175 polarized (36.5%),  2047 non-polarized (63.5%) - SLIGHTLY IMBALANCED
fas:  2440 polarized (74.1%),   855 non-polarized (25.9%) - HEAVILY POLARIZED
hau:   392 polarized (10.7%),  3259 non-polarized (89.3%) - HEAVILY NON-POLARIZED
hin:  2346 polarized (85.5%),   398 non-polarized (14.5%) - HEAVILY POLARIZED
ita:  1368 polarized (41.0%),  1966 non-polarized (59.0%) - BALANCED
khm:  6029 polarized (90.8%),   611 non-polarized ( 9.2%) - HEAVILY POLARIZED
mya:  1682 polarized (58.2%),  1207 non-polarized (41.8%) - BALANCED
nep:  1008 polarized (50.3%),   997 non-polarized (49.7%) - BALANCED
ori:   683 polarized (28.8%),  1685 non-polarized (71.2%) - HEAVILY NON-POLARIZED
pan:   840 polarized (49.4%),   860 non-polarized (50.6%) - BALANCED
pol:  1003 polarized (41.9%),  1388 non-polarized (58.1%) - BALANCED
rus:  1023 polarized (30.6%),  2325 non-polarized (69.4%) - SLIGHTLY IMBALANCED
spa:  1660 polarized (50.2%),  1645 non-polarized (49.8%) - BALANCED
swa:  3504 polarized (50.1%),  3487 non-polarized (49.9%) - BALANCED
tel:  1274 polarized (53.8%),  1092 non-polarized (46.2%) - BALANCED
tur:  1155 polarized (48.9%),  1209 non-polarized (51.1%) - BALANCED
urd:  2476 polarized (69.5%),  1087 non-polarized (30.5%) - SLIGHTLY IMBALANCED
zho:  2121 polarized (49.6%),  2159 non-polarized (50.4%) - BALANCED
```

Test set: Provided by SemEval-2025 Task 9 organizers (held out for final evaluation)

Training/Validation split:
80% training, 20% validation from available data

Exploratory Data Analysis

- Sample size varies in training set : Swahili-swa (6991) vs. Punjabi - pan (1700)
- Class imbalance in some languages
- Text length differs by language
- Average: 109 characters per post

Related Work

- AlDayel & Magdy (2021) - Stance detection survey; transformers outperform traditional ML
  - Motivated our transformer-based approach

- Mohammad et al. (2016) - SemEval-2016 stance detection; TF-IDF baseline
  - We adopted similar baseline approach

- Zhou et al. (2015, 2016) - C-LSTM and BiLSTM+Attention for sequence classification

- Conneau et al. (2020) - XLM-RoBERTa; cross-lingual transfer learning
  - Pre-trained on 100 languages, 2.5TB data
  - Our primary model choice

- Naseem et al. (2025) - POLAR benchmark (our dataset source)

Methodology - Our Approach

TF-IDF + Logistic Regression

Neural Baselines
(CLSTM, BiLSTM)

Transformers

Preprocessing Strategy

TF-IDF:

- Remove URLs, anonymize mentions, lowercase, remove emojis

Transformers:

- Preserve case/emojis, only remove URLs/mentions
- Emojis carry sentiment (from literature)

- 80/20 stratified train/val split
- Seed=42
- Max sequence length: 128 tokens

TF-IDF + Logistic Regression Baseline

TF-IDF + Logistic Regression

Approach: Traditional machine learning

- TF-IDF: Term Frequency-Inverse Document Frequency (converts text to numerical features)
- Logistic Regression: Simple classification algorithm

Mohammad et al. (2016)

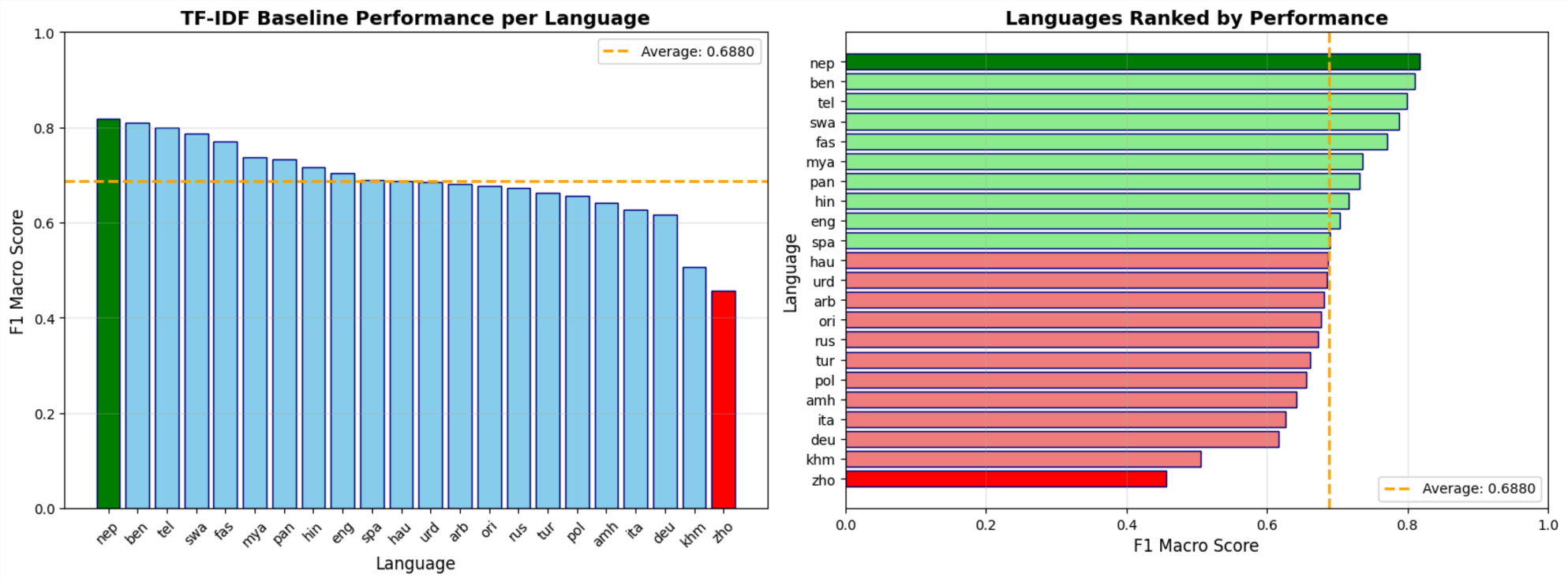
Parameters:

- Extract 1-grams and 2-grams (single words and word pairs)
- Maximum 5,000 features (most important words/phrases)
- Min document frequency: 2
- Trained separately for each language (not joint training)
- C=1.0 (regularization), Class-balanced weights

Result: 68.8% macro F1

TF-IDF + Logistic Regression Baseline

TF-IDF + Logistic Regression



Result: 68.8% macro F1

Neural Sequence Models

CLSTM

- Combines convolutional and recurrent neural networks
- CNN-->local phrase patterns (3-grams)
- LSTM-->long range context
- F1-Macro: 0.7471

BiLSTM

- With attention
- Reads sentence in both directions
- Attention learns token importance
- F1-Macro: 0.7578

BiLSTM + Lang.

- With attention and language embedding
- Adds a short learned language embedding to input
- Helps model utilize language info
- F1-Macro: 0.7650

Decoder-Only Transformer: mGPT

- Multilingual Generative Pre-Trained Transformer (GPT)
- Efficient fine-tuning using parameter-efficient adapters (LoRA) + 4-bit quantization (QLoRA)

Training Configuration:

- MAX_LENGTH=128
- 90/10 train/val split
- 1 epoch
- accuracy vs computing time trade-off

Results:

- F1-Macro: 0.7527
- Precision: 0.7562
- Recall: 0.7518
- Time: 2.5 hours

Intro to Transformers - XLM-RoBERTa Base

XLM-RoBERTa (Multilingual)

- Multilingual transformer model (Conneau et al., 2020)
- Pre-trained on 100 languages using 2.5TB of web text
- 270 million parameters
- Understands semantic meaning, not just word frequencies

- Cross-lingual transfer learning
- Semantic understanding through attention mechanisms
- Context-aware representations
- Single model for all 22 languages

Training Configuration:

- Joint training on all languages (not separate models!)
- AdamW optimizer, learning rate 2e-5
- Batch size 64, 10 epochs(with early stopping, patience 3 epochs)
- Max sequence length: 128 tokens
- Mixed precision (FP16) training

Intro to Transformers - XLM-RoBERTa Base

XLM-RoBERTa (Multilingual)

- All 22 languages trained together
- **Result: 76.07% F1**
- +7.18% improvement!

AlDayel & Magdy (2021), Lai et al. (2020)

Performance by Language Current Results

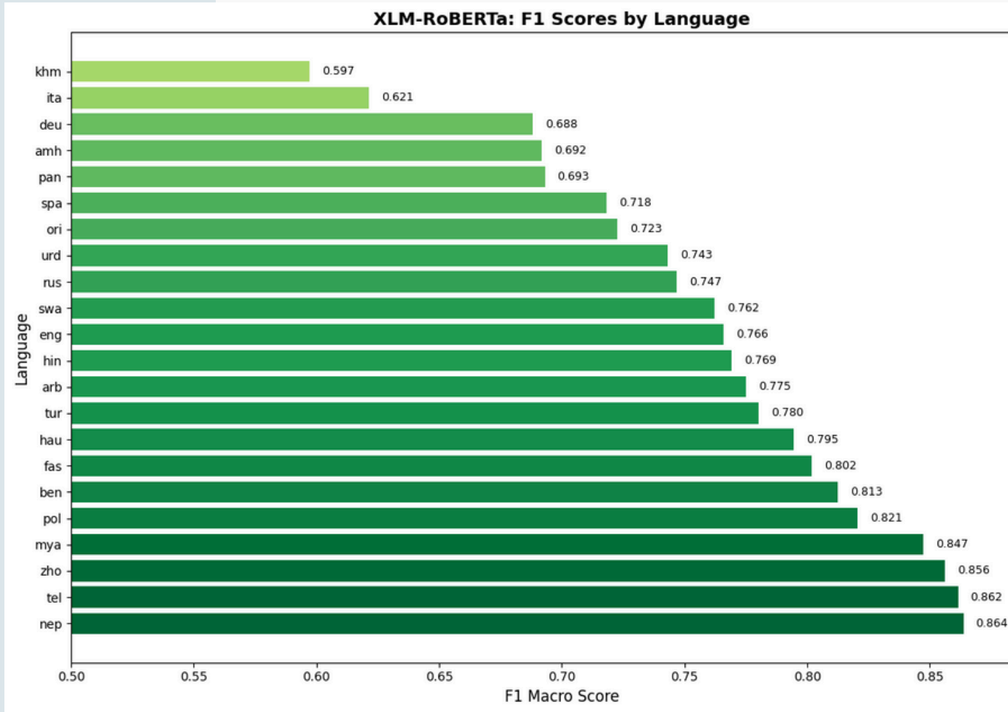
🏆 Top Performers:

- Nepali: 86.4%
- Telugu: 86.2%
- Chinese: 85.6%
- Burmese: 86.9%
- Bengali: 84.0%

⚠️ Need Improvement:

- Khmer: 59.7%
- Persian: 62.1%
- German: 68.8%

Why the gap 27.2 % ? → Motivates our next approaches!



Transformer Models

Training Configurations:

- Learning rate: 2e-5
- Batch size: 64
- Warmup ratio: 0.06
- Weight decay: 0.01 (L2 regularization)
- Max epochs: 10
- Early stopping: Patience=3 (on validation F1 macro)
- Precision: FP16 mixed precision (2× faster training)
- Optimization: AdamW optimizer

Model 1: XLM-RoBERTa-base

- **Source:** Conneau et al. (2020)
- **Parameters: 270M**
- **Why chosen:** Strong cross-lingual transfer baseline
- **Result: 76.07% F1**

Model 2: XLM-RoBERTa-large

- **Source:** Conneau et al. (2020)
- **Parameters: 550M**
- **Why chosen:** Test capacity vs performance tradeoff
- **Result: 78.14% F1 (best single model)**

Model 3: InfoXLM-base

- **Source:** Microsoft
- **Parameters: 270M**
- **Why chosen:** Superior cross-lingual transfer potential
- **Result: 77.43% F1**

Model 4: mDeBERTa-v3-base

- **Source:** Microsoft
- **Parameters: 278M**
- **Why chosen:** Different attention mechanism may capture different aspects of polarization
- **Result: 77.91% F1**

Model 5: RemBERT

- **Source:** Google (via Naseem et al., 2025)
- **Parameters: 580M (largest model)**
- **Why chosen:** Best performer in POLAR benchmark
- **Result: 76.98% F1**

Handling Class Imbalances

Solution 1: Focal Loss (Lin et al., 2017)

Source: "Focal Loss for Dense Object Detection" - ICCV 2017

Original use: Object detection with class imbalance

Our adaptation: Multilingual text classification

How It Works: Easy examples : $(1-p_t)^\gamma \approx 0 \rightarrow$ loss ≈ 0

Hard examples: $(1-p_t)^\gamma \approx 1 \rightarrow$ loss $\approx$ full

Intuition: "Don't waste time on easy, well-classified examples. Focus training on hard, misclassified examples."

Impact: +0.43% F1 improvement (76.07% → 76.50%)

Ensemble Method

Solution 2: Language-Aware Weighted Sampling

Why Language-Aware?

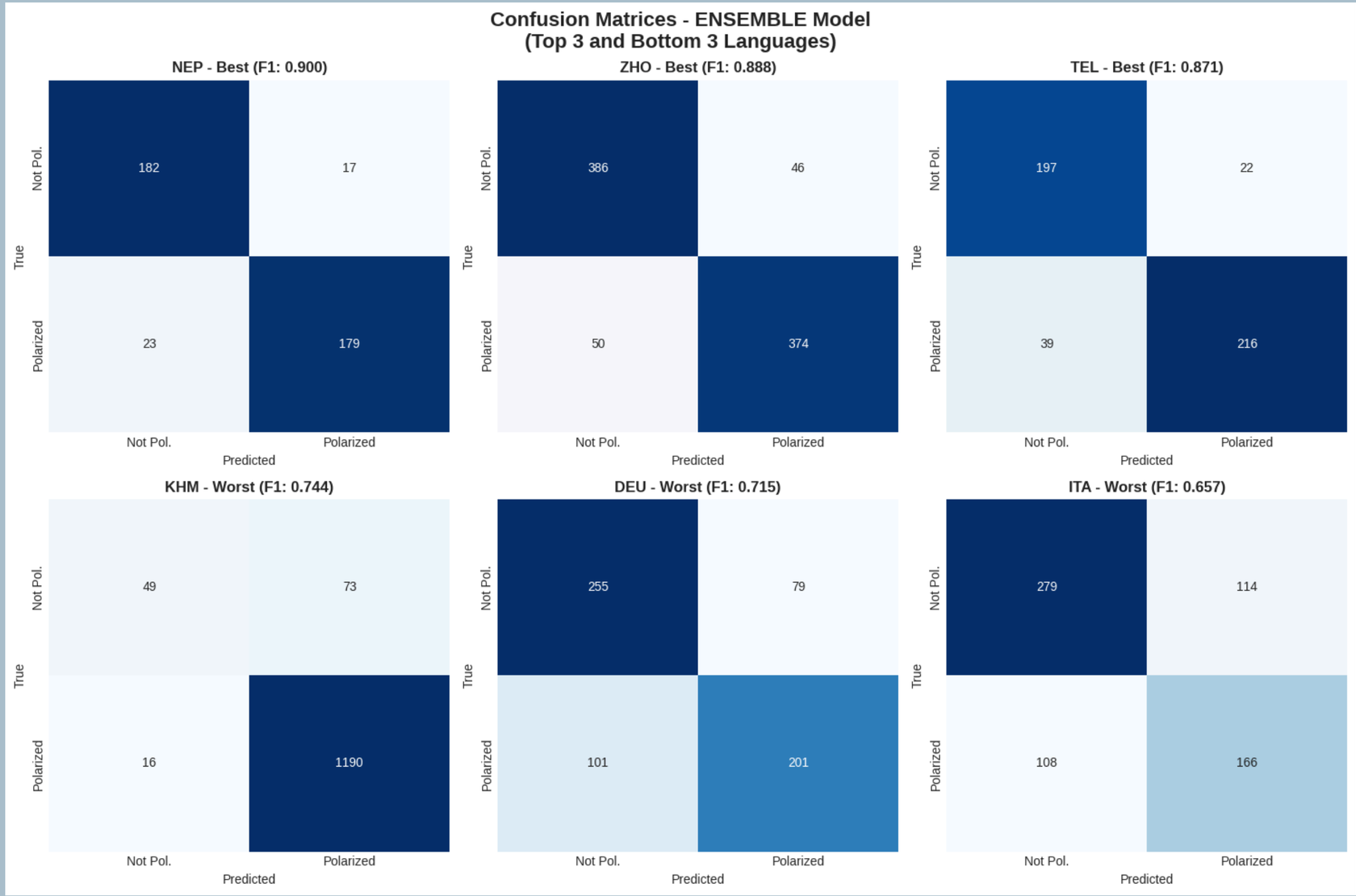
• Imbalance varies dramatically by language

• Prevents language-specific bias

Impact: Works synergistically with focal loss

Combined improvement: +0.43% F1

Ensemble Method



Results

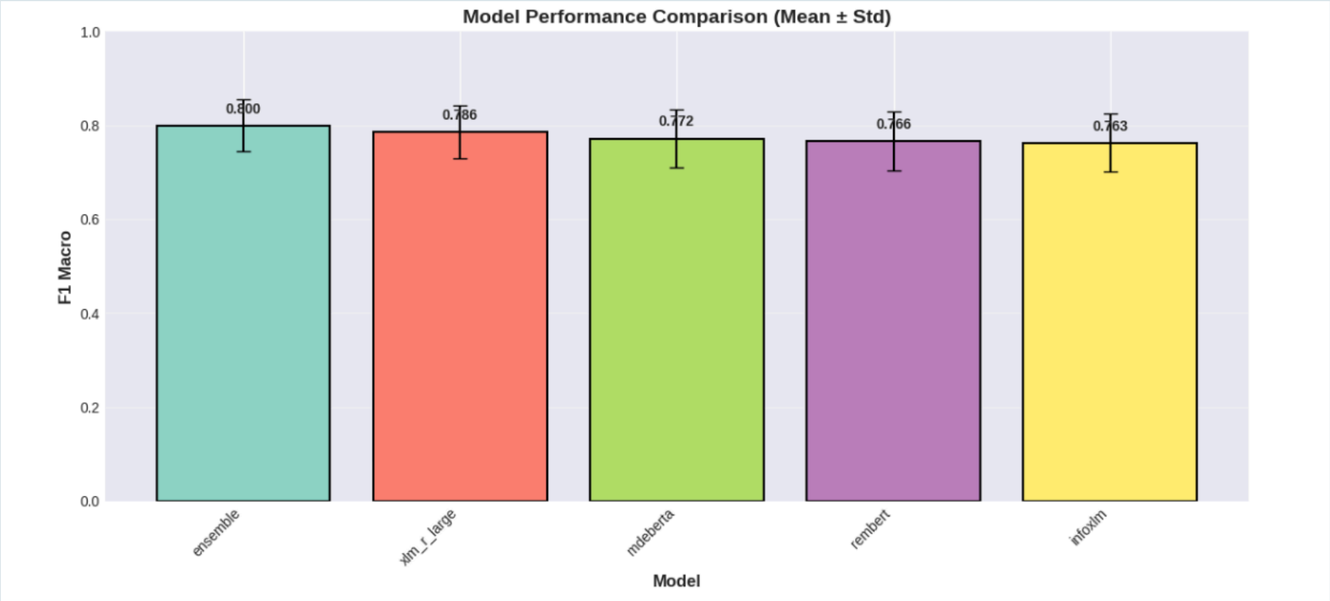
Table 1: Overall Model Performance

Model	F1 Macro	Macro Precision	Macro Recall
Traditional baseline			
TF-IDF + Logistic Regression	68.80	69.59	69.74
Neural Sequence Models			
CLSTM	0.7471	0.7471	0.7470
BiLSTM + Attention	0.7578	0.7577	0.7587
BiLSTM + Attention + Language Emb.	0.7650	0.7648	0.7655
Decoder-based transformers			
mGPT	0.7527	0.7562	0.7518
Encoder-based transformers			
XLM-RoBERTa-base (Baseline Tra)	76.07		
XLM-RoBERTa-base + Focal Loss	76.50		
mDeBERTa-v3-base	77.91		
InfoXLM-base	77.43		
XLM-RoBERTa-large	78.14		
RemBERT	76.98		
Ensemble (5 models)	79.69		

- Metric: Macro F1
- TF-IDF baseline: 68.8%
- Best single model (XLM-R-Large): 78.1%
- Final ensemble (5 models): 79.7%
- +12.5 F1 improvement over baseline

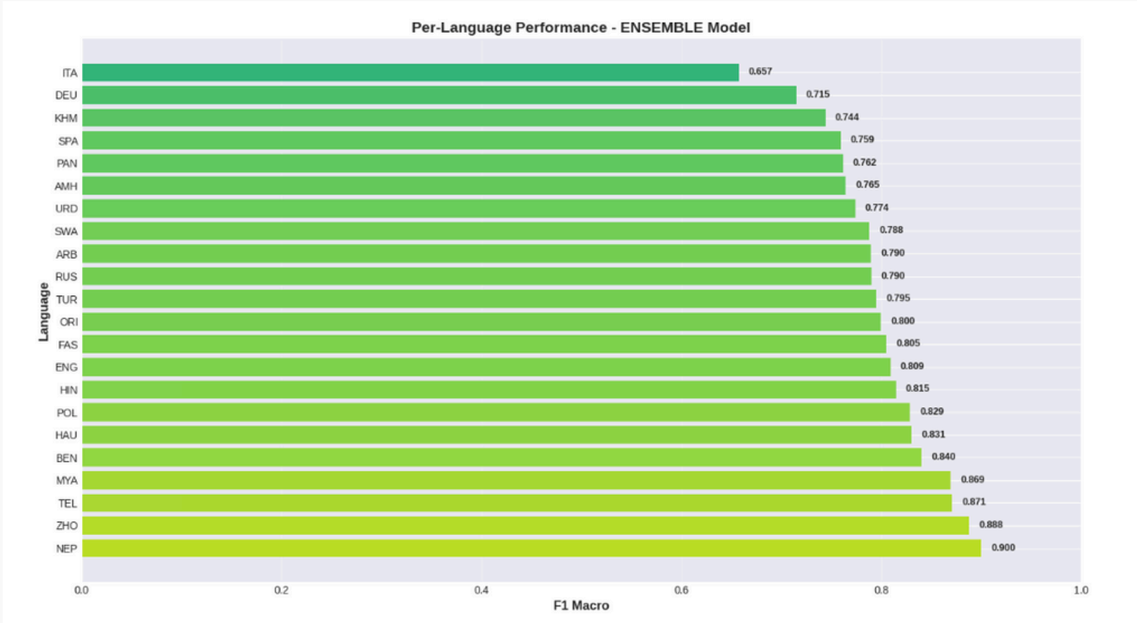
Model Comparison

- Traditional ML < Neural < Transformer models
- Best non-transformer:
- BiLSTM + Attention + Language Embedding → 76.5%
- Encoder-based transformers outperform decoder-based (mGPT)
- Ensemble consistently best



Per-Language Results

- Large variation across languages
- Best languages:
  - Nepali, Chinese, Telugu (F1 ≈ 0.88–0.90)
- Hardest languages:
  - Italian, German (F1 ≈ 0.65–0.71)
- No single model dominates all languages



Codabench Submission Results

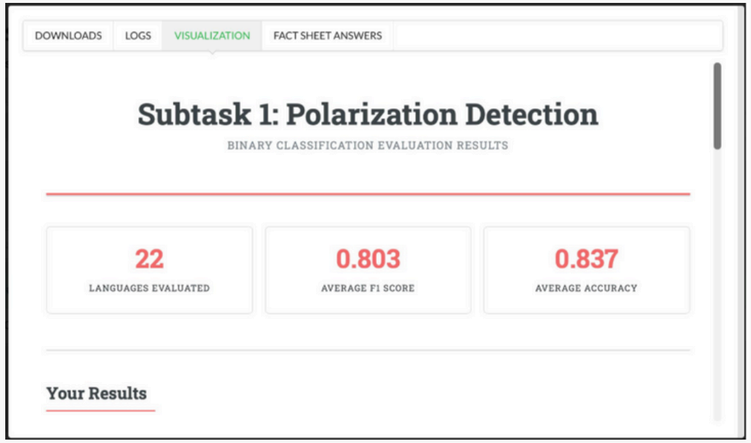
3 Individual Submissions:

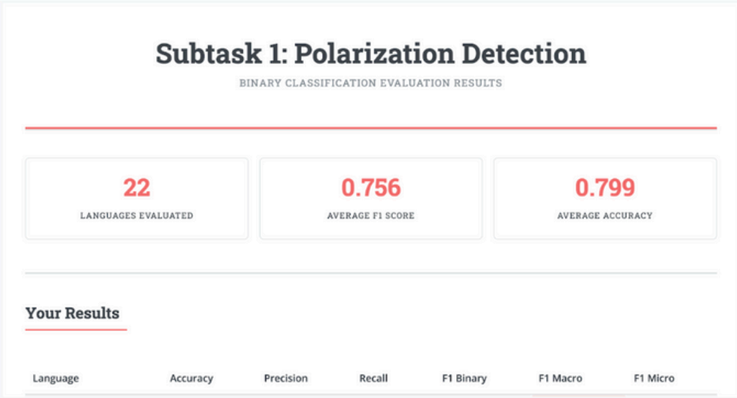
1) Transformers ensemble of 5 models with focal loss and weighted sampling.

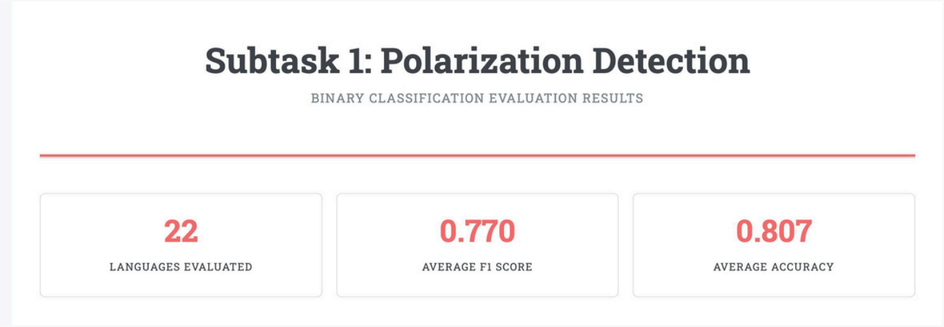
2) XLM-Roberta-Base

3) Ensemble of 3 models (miniLM, mbert, distilmbert)

4)SemEval Official Competition Organizer's Result







Discussion

Dataset Impact

- Class imbalance strongly affects performance
- Language-specific issues:
- Script differences (Chinese, Arabic)
- Low-resource languages
- Cross-lingual transfer helps but is not sufficient

Advantages & Disadvantages

Advantages

- Unified Multilingual Training (22 languages)- strong generalization
- Ensemble captures complementary model strengths
- Focal loss improves robustness under class imbalance
- Performance-Weighted Ensemble

Disadvantages

- High computational and memory cost
- Slower inference due to ensemble
- Uneven distribution of data across languages
- Limited interpretability of predictions

Limitations 

- Performance gaps remain for some languages
- GPU memory constraints limited full ensemble evaluation
- Class imbalance not fully resolved
- Interpretability: black box predictions
- Generalization to new events/topics unclear

Future Improvements 

- Language-specific fine-tuning for weak languages
- Feature extraction for interpretability
- Back-translation for data augmentation
- Model distillation for efficiency

Conclusion

- Multilingual transformers significantly outperform traditional baselines
- Joint multilingual training enables knowledge transfer
- Handling class imbalance is critical for polarization detection
- Ensemble learning provides the most robust performance across languages from our trial
- Performance gaps across languages still remain
- Future work should focus on efficiency and low-resource languages
- Low-resource languages benefit from high-resource languages
- Focal Loss Crucial for Imbalance Without imbalance handling: Model predicts majority class

References

- AlDayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. Information Processing & Management, 58(4), Article 102597. https://doi.org/10.1016/j.ipm.2021.102597
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & 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). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747
- Lai, M., Cignarella, A. T., Hernández Farías, D. I., Bosco, C., Patti, V., & Rosso, P. (2020). Multilingual stance detection in social media political debates. Expert Systems with Applications, 143, 113045. https://doi.org/10.1016/j.eswa.2019.113045
- 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 (ICCV) (pp. 2980–2988).https://doi.org/10.1109/ICCV.2017.324
- Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 Task 6: Detecting stance in tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (pp. 31–41). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1003
- Naseem, U., et al. (2025). POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. arXiv. https://arxiv.org/abs/2505.20624
- Zhou, C., Sun, C., Liu, Z., & Lau, F. C. M. (2015). A C-LSTM neural network for text classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 39–44). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1013
- Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 207–212). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1022

Thank you