SemEval 2026 Task 9

Detecting Multilingual, Multicultural and Multievent Online Polarization (POLAR)

Zeynep Şahin 30553,
Suat Emre Karabıçak 30649,
Alpay Naçar 31133,
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 13 languages (websites, Reddit, blogs, Bluesky, and regional forums covering topics like elections, conflicts, gender rights and migration.)
- Challenge: One model must work across all 13 languages
- 13 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: 13
Total training samples: 40,395
Total dev samples: 2,012
Total samples: 42,407
Average samples per language: 3107
```

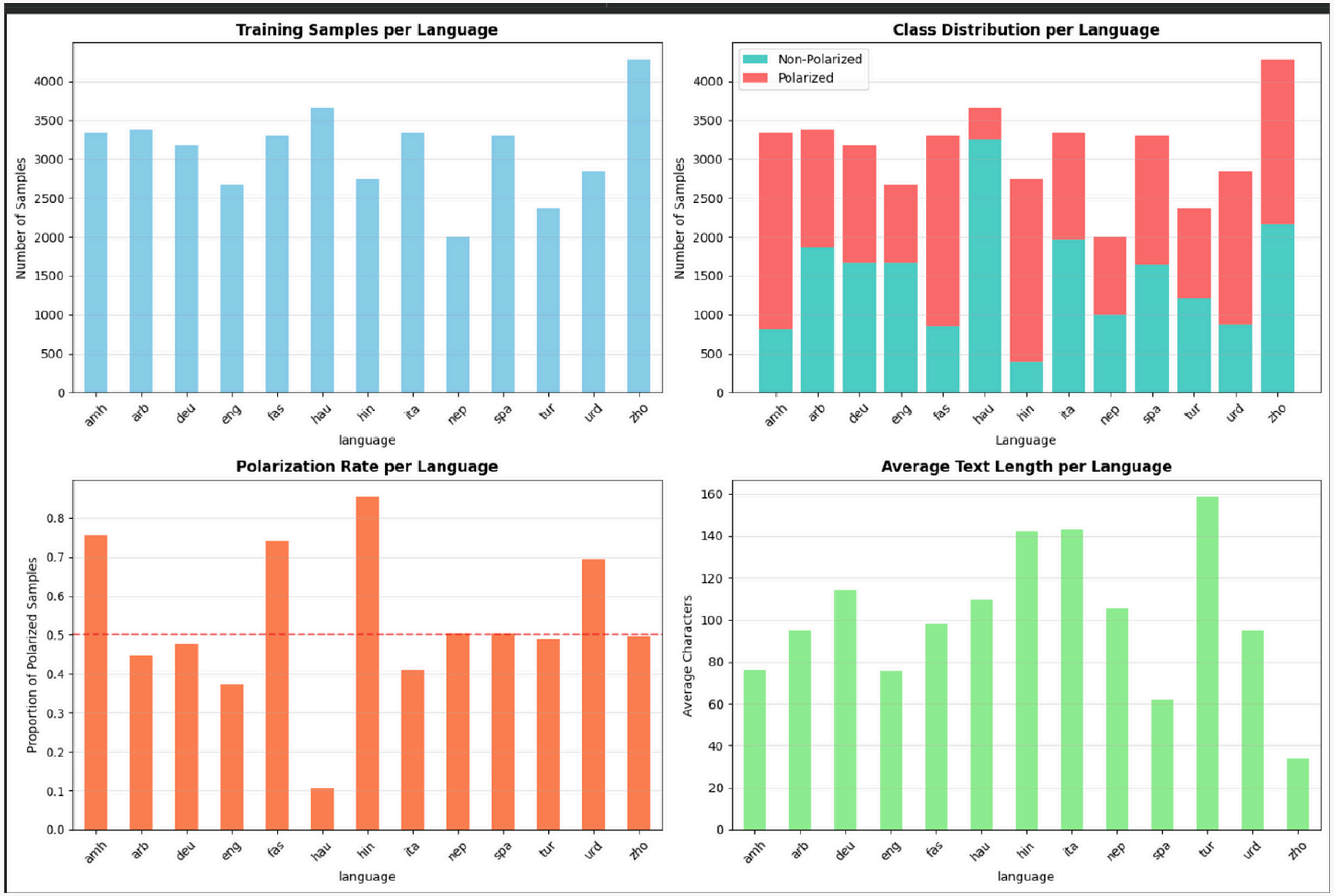
```
... ============================================================================
    CLASS BALANCE PER LANGUAGE
    ============================================================================
    amh:  2518 polarized (75.6%),   814 non-polarized (24.4%)
    arb:  1512 polarized (44.7%),  1868 non-polarized (55.3%)
    deu:  1512 polarized (47.5%),  1668 non-polarized (52.5%)
    eng:  1002 polarized (37.4%),  1674 non-polarized (62.6%)
    fas:  2440 polarized (74.1%),   855 non-polarized (25.9%)
    hau:   392 polarized (10.7%),  3259 non-polarized (89.3%)
    hin:  2346 polarized (85.5%),   398 non-polarized (14.5%)
    ita:  1368 polarized (41.0%),  1966 non-polarized (59.0%)
    nep:  1008 polarized (50.3%),   997 non-polarized (49.7%)
    spa:  1660 polarized (50.2%),  1645 non-polarized (49.8%)
    tur:  1155 polarized (48.9%),  1209 non-polarized (51.1%)
    urd:  1976 polarized (69.4%),   873 non-polarized (30.6%)
    zho:  2121 polarized (49.6%),  2159 non-polarized (50.4%)
```

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

<u>Training/Validation split:</u>

85% training, 15% validation from available data

Exploratory Data Analysis



- Sample size varies: Chinese (4,280) vs. Nepali (2,005)
- Class imbalance in some languages
- Text length differs by language
- Average: 109 characters per post

Our Approach - Baselines Complete

TF-IDF + Logistic Regression

- Traditional ML approach
- Result: 67.2% F1

Mohammad et al. (2016)

XLM-RoBERTa (Multilingual)

- All 13 languages trained together
- Result: 74.73% F1
- +7.5% improvement!

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

Performance by Language Current Results

🏆 Top Performers:

- Nepali: 87.3%
- Chinese: 86.6%
- Hausa: 82.8%
- Persian: 82.0%

⚠️ Need Improvement:

- Amharic: 56.3%
- Italian: 63.7%
- German: 64.8%

Why the gap 31 % ? → Motivates our next approaches!

Planned Approaches

Stage 2: Cross-Lingual Transfer Learning

- The Idea:
- Use knowledge from strong languages to help weak ones
- Example: Train on Nepali (87.3%) → transfer to Amharic (56.3%)
- Two Methods:
    - Zero-shot: Test on target language without any target data
    - Few-shot: Fine-tune with small amount of target language data
- Lai et al. (2020) - MultiTACOS system

Additional Planned Improvements

- **1. Handle Class Imbalance**
- Problem: Hindi has 85.5% polarized, Hausa only 10.7%
- Solution: Focal loss + weighted sampling
- These techniques help the model learn from rare examples
- **2. Test Larger Models**
- RemBERT: Winner in POLAR benchmark study - Naseem et al. (2025)
- XLM-RoBERTa-large: More parameters = better performance - (Conneau et al., 2020)
- **3. Ensemble Methods**
- Combine predictions from multiple models
- Like getting a "second opinion" from different experts
- Usually more reliable than single model
- **4. Add Extra Features**
- Punctuation patterns (!!!, ???), Hashtags and emojis, Sentence length
- **5. Language-Specific Fine-tuning**
- After training on all languages, fine-tune separately per language
- Customize for each language's unique patterns - Hofmann et al. (2022)

Team Work Distribution

RemBERT Fine-tuning

- Integrate class imbalance solutions
- Apply cross-lingual transfer
- Benchmark testing

Sıla Horozoğlu

Weak Language Performance

- Weighted loss & sampling
- Zero-shot transfer
- Target: Amharic & Italian

Zeynep Şahin & Suat Emre Karabıçak

Ensemble Methods

- Combine model predictions
- Optimize final system

Alpay Naçar & Korhan Erdoğdu

References

AlDayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. Information Processing & Management, 58(4), 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).

Hofmann, V., Pierrehumbert, J. B., & Schütze, H. (2022). Modeling ideological salience and framing in polarized online environments. In Findings of the Association for Computational Linguistics: NAACL 2022 (pp. 1934-1951).

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.

Mohammad, S., 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).