Detecting Multilingual, Multicultural and Multievent Online Polarization

Zeynep Şahin 30553, Suat Emre Karabıçak 30649, Alpay Naçar 31133 ,Sıla Horozoğlu 30916, Korhan Erdoğdu 30838

1. Introduction

Over the last decade, social media has become one of the main places where people argue about politics, identity, religion, migration and other sensitive topics. These discussions often shift from simple disagreement to a much sharper pattern. People strongly align with an “us” and speak about a “them” in hostile, stereotyping or even dehumanizing ways. Automatically identifying such polarized content is important for understanding how online discussions become fragmented and how attitudes differ across languages and communities. This is the type of attitude polarization that SemEval 2026 Task 9 – *Detecting Multilingual, Multicultural and Multievent Online Polarization (POLAR)* – is designed to capture.

In this project, we work on SemEval-2026 Task 9: Multilingual Polarization Detection, focusing on Subtask 1, which frames the problem as a binary classification task. Given a short social media post, the goal is to determine whether it contains polarized language. The dataset provided for this task supports 13 languages, covering a wide geographic and linguistic area.

Polarization detection differs from related tasks such as hate speech or offensive language detection. While those tasks identify harmful content, polarization detection focuses on identifying language that divides audiences into opposing groups, often through us-versus-them framing, moral absolutism, or emotional intensity. As AlDayel and Magdy (2021) note in their comprehensive survey, stance detection and polarization detection share similar characteristics: both require understanding not just sentiment, but the positioning of text relative to controversial topics.

So far, we have completed the initial stages of the project. We downloaded and explored the full SemEval dataset, examining label distributions and language-specific characteristics. As baselines, we implemented a TF–IDF Logistic Regression model, followed by a multilingual transformer-based model. Our current multilingual XLM-RoBERTa model trained jointly on all languages achieves a noticeable improvement over the baseline (7.58%), indicating that transformer-based methods are promising for this task.

2. Dataset Selection

For this project, we use the dataset released for **SemEval-2025 Task 9**, which is based on the POLAR benchmark introduced by Naseem et al. (2025). The shared-task version

provides labeled social media posts for 13 languages, each annotated as either polarized or non-polarized. The dataset contains roughly 40000 training instances and around 6000 development examples. Our initial analysis shows that the proportion of polarized posts differs noticeably between languages. For example, Hindi has a high polarization rate, whereas languages like Turkish are more balanced. Texts are generally short and informal, reflecting the characteristics of social media communication.

The dataset used in this study is designed specifically for polarization detection rather than for sentiment, toxicity, or stance classification. Its multilingual nature aligns with the goals of our project, which include evaluating how well a single model can generalize across languages with different resources and linguistic properties. The presence of both high-resource and low-resource languages makes the dataset particularly suitable for studying cross-lingual transfer. Additionally, because POLAR is part of an official SemEval shared task, it provides standardized splits and a clear evaluation setup, which supports reproducible experiments and allows straightforward comparison with other systems.

We conducted a comprehensive literature review to identify external datasets to augment the provided dataset and potentially improve the accuracy of the model. While various studies related to harmful content detection (Founta et al., 2018; Katsarou et al., 2021) were identified, a significant challenge was encountered in label alignment. These studies utilize fine-grained or multi-label classification (classifying text as positive, neutral, offensive or specific type of hate speech) which do not align precisely with the binary classification of polarization as defined in our task. Integrating them would undermine the consistency, therefore, we proceeded exclusively with the competition dataset.

Before model training, we apply minimal preprocessing. URLs are removed, user mentions are anonymized, and the remaining content is kept largely intact to preserve meaningful linguistic cues. Texts are tokenized with the XLM-R tokenizer and padded or truncated to a standard sequence length. For our internal development, we generate stratified validation splits using a fixed random seed to maintain consistent class distribution. These steps prepare the dataset for transformer-based models without altering its original character.

3. Approach Plan

Our methodology is built upon recent advances in multilingual text classification and stance detection. We adopt a three-stage approach that combines multilingual transformers, cross-lingual transfer learning, and linguistic feature engineering.

3.1 Baseline: TF-IDF with Logistic Regression

We began with a traditional machine learning baseline using TF-IDF features with logistic regression, a well-established approach for text classification tasks. As demonstrated by Mohammad et al. (2016) in their SemEval-2016 stance detection shared task, TF-IDF representations combined with linear classifiers provide strong baselines for social media text classification. Our implementation achieved 67.2% macro F1-score across all languages,

establishing a performance floor for comparison. This baseline helps us quantify the improvement gained from more sophisticated approaches.

3.2 Foundation: Multilingual Transformers

Our primary approach builds on the findings of AlDayel and Magdy (2021), whose comprehensive survey of stance detection methods demonstrates that transformer-based models substantially outperform traditional machine learning approaches on social media text. We have already implemented this stage by training XLM-RoBERTa (Conneau et al., 2020) on the combined dataset of all 13 languages. This multilingual training strategy is motivated by the work of Lai et al. (2020), who show that joint training across multiple languages improves performance compared to language-specific models, particularly for low-resource languages. Our current implementation achieves 74.73% macro F1-score, representing a 7.58 percentage point improvement over the TF-IDF baseline and validating AlDayel and Magdy's findings regarding transformer superiority. However, we observe substantial performance variation across languages. While Nepali achieves 87.3% F1, Amharic reaches only 56.3%. This gap suggests that different languages may benefit from different training strategies, which motivates our planned extensions.

3.3 Stage 2: Cross-Lingual Transfer Learning

Our second approach addresses the weaker-performing languages through targeted cross-lingual transfer. Lai et al. (2020) demonstrate in their MultiTACOS system that training on carefully selected source languages can improve performance on target languages through knowledge transfer. We will systematically investigate which source-target language pairs produce effective transfer. Specifically, we plan to train models on subsets of high-performing languages (such as Nepali, Persian, and Chinese) and evaluate their zero-shot performance on low-performing targets (such as Amharic and German). We will also experiment with few-shot fine-tuning, where we adapt the source-trained model using small amounts of target language data. Our research questions include: which linguistic factors (language family, script, or topic similarity) predict successful transfer, and can strategic transfer outperform joint multilingual training for specific languages?

3.4 Stage 3: Linguistic Feature Engineering

Our third approach incorporates explicit linguistic features that capture polarization signals. Hofmann et al. (2022) analyze polarized political discourse and identify distinctive patterns including moral language, ideological framing, and emotional intensity markers. While transformers learn implicit representations, we hypothesize that augmenting them with explicit features may improve performance. We plan to extract features such as exclamation and capitalization ratios (emotional intensity), us-versus-them pronoun patterns (group identity), sentiment polarity scores, and counts of moral or absolutist terms. These features will be combined with XLM-RoBERTa embeddings in a multi-input neural architecture. This approach is expected to benefit languages where cultural context plays a strong role in polarization.

4. Next Steps

We have finalized our TF-IDF baseline and multilingual transformer baseline. We are beginning the transfer learning experiments, and the feature engineering approach is in the design phase. We will decide whether to pursue ensemble methods based on whether the three approaches show complementary strengths.

4.1 Potential Improvements

With more time and resources, we could improve our system in several ways. First, we need to tackle the class imbalance problem we observed across different languages. For example, Hindi has 85.5% polarized posts while Hausa only has 10.7%, which makes training difficult. We could try focal loss or weighted sampling to handle this better. Second, we should focus on languages that performed poorly, like Amharic and Italian. Applying cross-lingual transfer learning strategies could help boost their performance. Third, combining multiple models through ensemble methods might give us more reliable predictions than relying on XLM-R alone. Fourth, we'd like to test other multilingual transformers like RemBERT or the larger XLM-RoBERTa version. RemBERT consistently achieved the highest scores for most languages in the benchmark studies (Naseem et al., 2025), thereby establishing a strong foundation from which we can explore further improvements. To compare our results with the benchmark, we will conduct tests using the exact monolingual setup they employed, which corresponds to training and evaluating the model on each language separately.

4.2 Distribution of Work

Although the distribution of work is subject to change according to our findings with further improvements, the planned distribution is as follows.

Sıla will focus on fine-tuning RemBERT, incorporating solutions for the predescribed challenges such as class imbalance and cross-lingual transfer developed by the group. She will also be responsible for testing the model using the benchmark study's exact setup. Suat and Zeynep will address weak language performance and class imbalance by implementing weighted loss, weighted sampling, and zero-shot cross-lingual transfer strategies to improve results for Amharic and Italian. Korhan and Alpay will develop ensemble methods to combine predictions from different models and identify the optimal approach for our final system.

5. 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).

Founta, A. M., et al. (2018). Large scale crowdsourcing and characterization of Twitter abusive behavior. *arXiv preprint arXiv:1802.00393*.

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).

K. Katsarou, S. Sunder, V. Woloszyn, and K. Semertzidis, "Sentiment Polarization in Online Social Networks: The Flow of Hate Speech," in *2021 Eighth International Conference on Social Network Analysis, Management and Security (SNAMS)*, 2021, pp. 1–6.

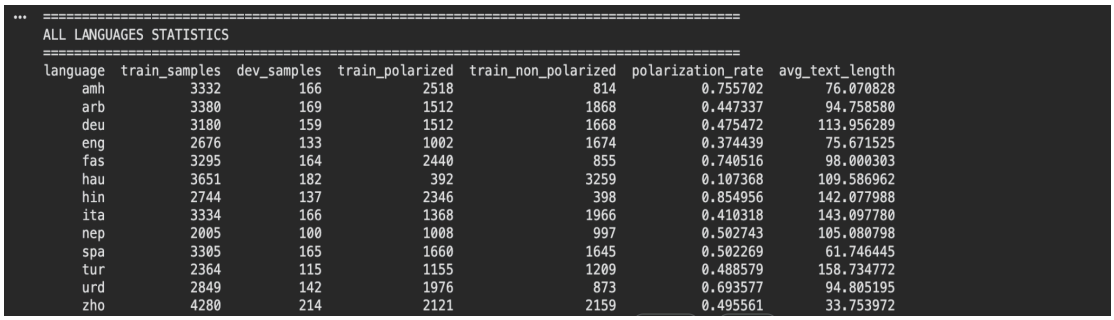
Lai, M., Cignarella, A. T., Hernández Farías, D. I., Bosco, C., Patti, V., & Rosso, P. (2020). Multilingual stance detection in social media political debates. Expert Systems with Applications, 143, 113045.

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).

Naseem, U., et al. (2025). POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. arXiv preprint arXiv:2505.20624.

6. Appendix

Figure 1: All Languages Statistics



```
================================================================================
ALL LANGUAGES STATISTICS
================================================================================
language  train_samples  dev_samples  train_polarized  train_non_polarized  polarization_rate  avg_text_length
     amh           3332          166             2518                  814           0.755702        76.070828
     arb           3380          169             1512                 1868           0.447337        94.758580
     deu           3180          159             1512                 1668           0.475472       113.956289
     eng           2676          133             1002                 1674           0.374439        75.671525
     fas           3295          164             2440                  855           0.740516        98.000303
     hau           3651          182              392                 3259           0.107368       109.586962
     hin           2744          137             2346                  398           0.854956       142.077988
     ita           3334          166             1368                 1966           0.410318       143.097780
     nep           2005          100             1008                  997           0.502743       105.080798
     spa           3305          165             1660                 1645           0.502269        61.746445
     tur           2364          115             1155                 1209           0.488579       158.734772
     urd           2849          142             1976                  873           0.693577        94.805195
     zho           4280          214             2121                 2159           0.495561        33.753972
```

Figure 2: Summary of the Dataset

```
================================================================================
SUMMARY
================================================================================
Total languages: 13
Total training samples: 40,395
Total dev samples: 2,012
Total samples: 42,407
Average samples per language: 3107
```

Figure 3: Results from Exploratory Data Analysis

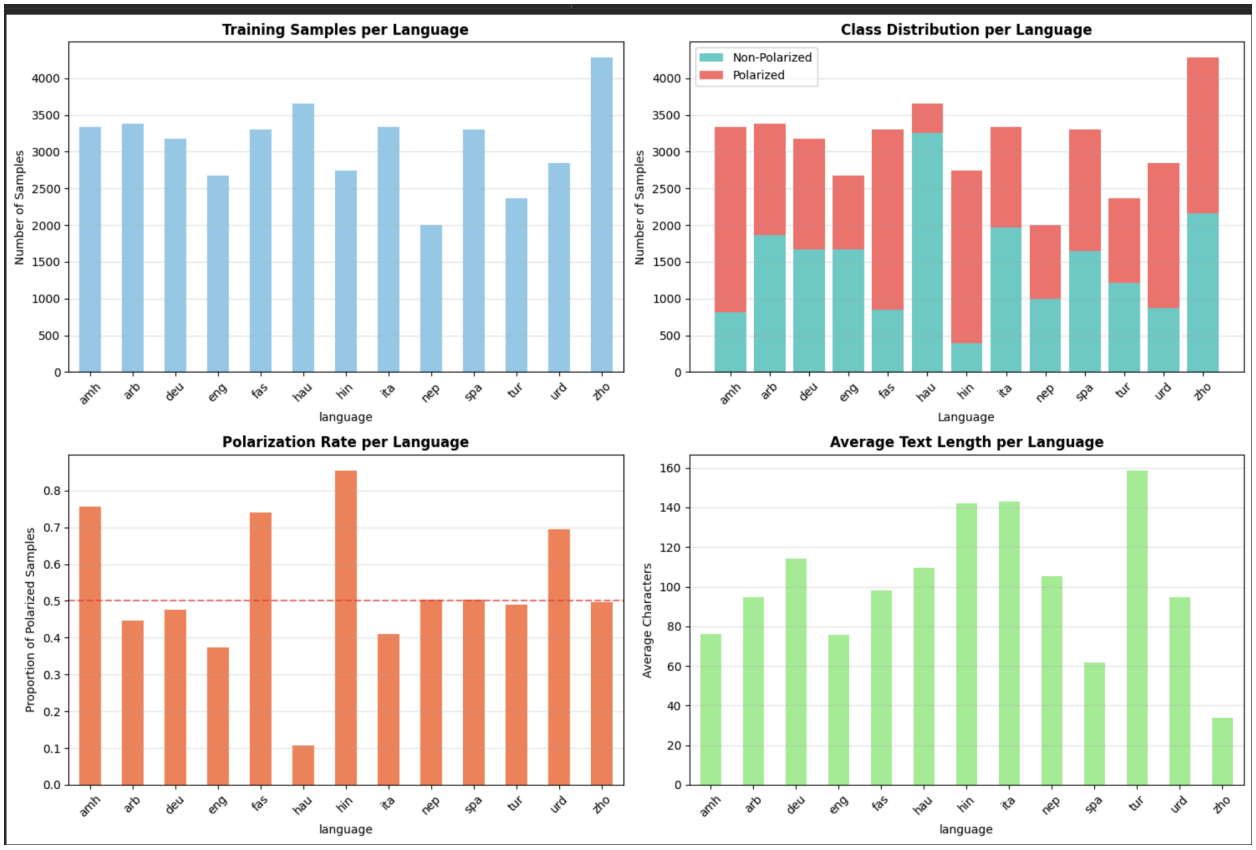
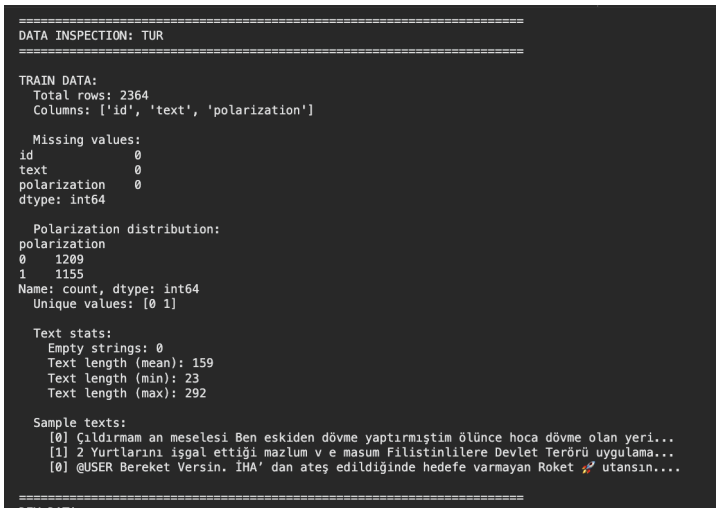


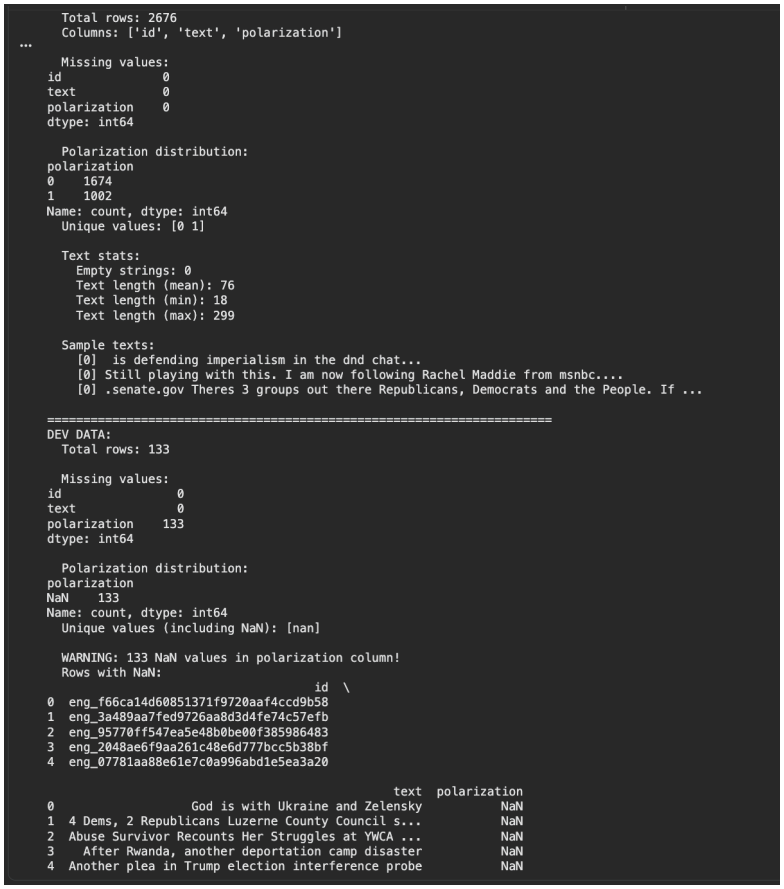
Figure 4: Comparison of Class Balance per Language

```
================================================================
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%)
```

Figure 5 and 6: Data Inspection for Turkish Language



```
================================================================
DATA INSPECTION: TUR
================================================================

TRAIN DATA:
  Total rows: 2364
  Columns: ['id', 'text', 'polarization']

  Missing values:
id              0
text            0
polarization    0
dtype: int64

  Polarization distribution:
polarization
0    1209
1    1155
Name: count, dtype: int64
  Unique values: [0 1]

  Text stats:
    Empty strings: 0
    Text length (mean): 159
    Text length (min): 23
    Text length (max): 292

  Sample texts:
    [0] Çıldırmam an meselesi Ben eskiden dövme yaptırmıştım ölünce hoca dövme olan yeri...
    [1] 2 Yurtlarını işgal ettiği mazlum v e masum Filistinlilere Devlet Terörü uygulama...
    [0] @USER Bereket Versin. İHA' dan ateş edildiğinde hedefe varmayan Roket 🚀 utansın....

================================================================
```



```
      Total rows: 2676
      Columns: ['id', 'text', 'polarization']
…
      Missing values:
    id              0
    text            0
    polarization    0
    dtype: int64

      Polarization distribution:
    polarization
    0    1674
    1    1002
    Name: count, dtype: int64
      Unique values: [0 1]

      Text stats:
        Empty strings: 0
        Text length (mean): 76
        Text length (min): 18
        Text length (max): 299

      Sample texts:
        [0]  is defending imperialism in the dnd chat...
        [0] Still playing with this. I am now following Rachel Maddie from msnbc....
        [0] .senate.gov Theres 3 groups out there Republicans, Democrats and the People. If ...

    ================================================================
    DEV DATA:
      Total rows: 133

      Missing values:
    id                0
    text              0
    polarization    133
    dtype: int64

      Polarization distribution:
    polarization
    NaN    133
    Name: count, dtype: int64
      Unique values (including NaN): [nan]

      WARNING: 133 NaN values in polarization column!
      Rows with NaN:
                                          id  \
    0  eng_f66ca14d60851371f9720aaf4ccd9b58
    1  eng_3a489aa7fed9726aa8d3d4fe74c57efb
    2  eng_95770ff547ea5e48b0be00f385986483
    3  eng_2048ae6f9aa261c48e6d777bcc5b38bf
    4  eng_07781aa88e61e7c0a996abd1e5ea3a20

                                                  text  polarization
    0                  God is with Ukraine and Zelensky           NaN
    1  4 Dems, 2 Republicans Luzerne County Council s...           NaN
    2  Abuse Survivor Recounts Her Struggles at YWCA ...           NaN
    3    After Rwanda, another deportation camp disaster           NaN
    4  Another plea in Trump election interference probe           NaN
```