Semeval task 9
Detecting Multilingual, Multicultural and Multievent Online Polarization

Gheorghiu Maria, Vlad Adriana, Iftene Tudor

Full data available only on jan 10th

1. Description

The problem addresses the automatic detection of online polarization, defined not just as disagreement, but as the sharp division of opinions into opposing groups, characterized by hostility, exclusion, and “us vs. them” narratives.

Current NLP models excel at detecting “toxicity” (slurs/insults) or “stance” (pro/con), but they struggle to detect the rhetorical strategies used to deepen social divides. This task aims to solve this across 22 languages (Amharic, Arabic, Bengali, Burmese, Chinese, English, German, Hausa, Hindi, Italian, Khmer, Nepali, Odia, Persian, Polish, Punjabi, Russian, Spanish, Swahili, Telugu, Turkish, Urdu), handling diverse cultural contexts (e.g., elections in India vs. conflicts in the Middle East).

2. Dataset and subtasks

We will be using sites that provide news (such as Reddit and blogs), while also regional forums, all of them including the coverage of politically charged subjects in order to create our database. This will range from elections to conflicts, gender rights and more. Also, per each language (22 possible in total) that will be taken in consideration we will have 3000-5000 annotated instances.
The goal of this project can be divided in 3 different subtasks:

- <u>Subtask 1: Polarization Detection</u>
  In order for a text to be classified as polarized, it should include at least one polarized specified characteristic. If a text does not include any, it will be classified as non-polarized (only 2 classes)
  The text must overall include a form of polarization to be considered as such, containing only a word of phrase that may conduce to interpretation of their validity (for example if a word that usually represents clear polarisation in a text that is part of, the same word won’t have a use if the overall context doesn’t support the polarization)

- <u>Subtask 2: Polarization Type Classification</u>

The goal of this task is to identify the specific societal dimension defining the conflict in a polarized text. The categories are as follows:

- *Political*: Disputes over governance, ideology or elections.
- *Religious*: Conflicts based on faith, scripture or religious identity.
- *Racial/Ethnic*: Divisions based on heritage, skin colour or nationality.
- *Gender/Sexual Orientation*: Conflicts regarding identity, rights and gender roles.
- *Other*: Socio-economic status, sports rivalries, etc.

- <u>Subtask 3: Manifestation Identification</u>
  The most complex aspect of the problem, it requires the model to identify the rhetorical device or psychological mechanism used to express polarization (how the writer is attacking the other side).
    - *Stereotyping*: Reducing a group to a simplified, often negative trait.
    - *Vilification*: Portraying the out-group as inherently evil, corrupt or malicious.
    - *Dehumanization*: Denying the out-group human qualities (e.g. comparing them to animals, pests or objects).
    - *Extreme Language*: Using hyperbolic adjectives or intensifiers to amplify threat or difference.
    - *Lack of Empathy*: Explicitly disregarding the suffering or perspective of the out-group.
    - *Invalidation*: Dismissing the lived experiences or legitimate concerns of the out-group as fake or irrelevant.

3. Scoring metric

The metric for all subtasks is MacroF1, meaning that all labels are weighed equally. For example, in the case of subtask 1, which required a binary classification, 50% of the score goes to tests labeled as false and 50% to those labeled as true, regardless of how many tests there are with each of the labels. Additionally, only the languages each team chooses to submit for count towards the score (and those languages, again, weigh equally on the final score, regardless of how many tests there are per language). This means two things:

- the accuracy displayed on the leaderboards may not directly correlate the actual score
- we can choose to submit for one, multiple, or all languages. we won't lose out on points for choosing not to tackle every language.

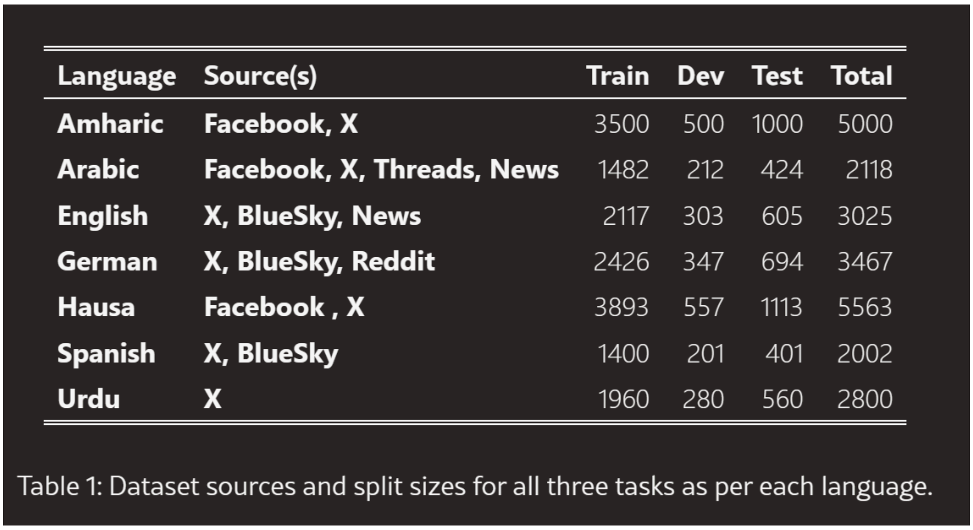
4. Dataset benchmark

The following section includes key paragraphs and tables from the following article: https://arxiv.org/html/2505.20624v1

Despite growing attention, computational approaches to polarization suffer from major limitations. First, most existing datasets focus on English or high-resource languages, reflecting a widespread trend across NLP tasks that ignores the rich diversity of linguistic and sociocultural contexts in which polarization manifests. Second, current benchmarks are often event-specific or monodomain, such as U.S. elections or Western political debates, limiting their generalizability. Third, the conceptualization of polarization in NLP has largely been binary or topic-focused, overlooking the multifaceted ways in which polarization is expressed through vilification, dehumanization, stereotyping, or other rhetorical tactics.

To address these gaps, we introduce POLAR, a novel multilingual, multicultural, and multievent dataset for fine-grained polarization detection. It spans seven languages across diverse regions, including low-resource languages such as Amharic and Hausa. Our data is sourced from various platforms (e.g., Twitter/X, Facebook, BlueSky, Reddit, and local news outlets), reflecting authentic, event-driven discourse ranging from armed conflict (e.g., the Tigray War) to social justice movements (e.g., abortion rights, migration crises).

We collected data from various online platforms, including X (formerly Twitter), Facebook, Reddit, Bluesky, Threads, and news/commentary forums. We use a dynamic keyword-driven strategy tailored for each language. Human experts curated keyword lists to reflect culturally and politically significant discourse across regions and events. Table 1 shows the languages covered, data splits and total number of instances annotated for each language.



Language	Source(s)	Train	Dev	Test	Total
Amharic	**Facebook, X**	3500	500	1000	5000
Arabic	**Facebook, X, Threads, News**	1482	212	424	2118
English	**X, BlueSky, News**	2117	303	605	3025
German	**X, BlueSky, Reddit**	2426	347	694	3467
Hausa	**Facebook , X**	3893	557	1113	5563
Spanish	**X, BlueSky**	1400	201	401	2002
Urdu	**X**	1960	280	560	2800

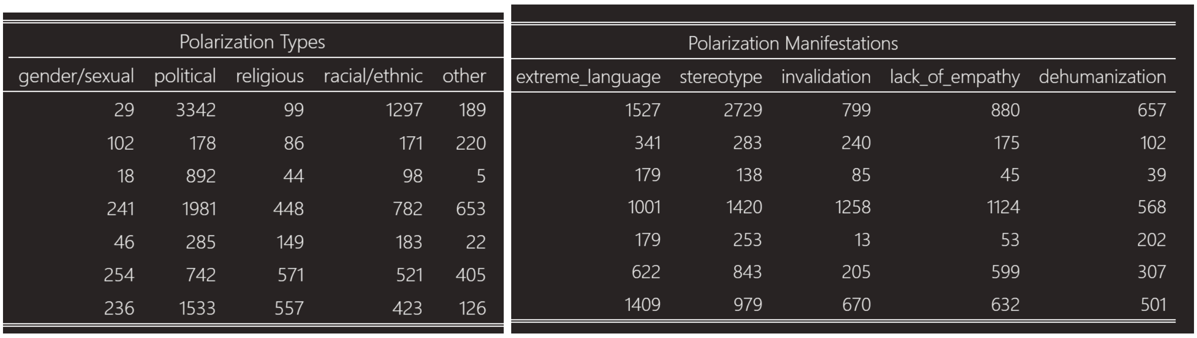
Table 1: Dataset sources and split sizes for all three tasks as per each language.

Annotation Guidelines:

We developed the guidelines in English and Amharic, and then translated and culturally adapted them for each target language. Annotators were instructed to:

- Identify whether a text is polarized;
- If the text is classified as polarized, tag the type of polarization (political, racial/ethnic, religious, gender/sexual identity, other);
- If the text is classified as polarized, tag its manifestations (stereotyping, vilification, dehumanization, deindividuation, extreme language, lack of empathy, invalidation).

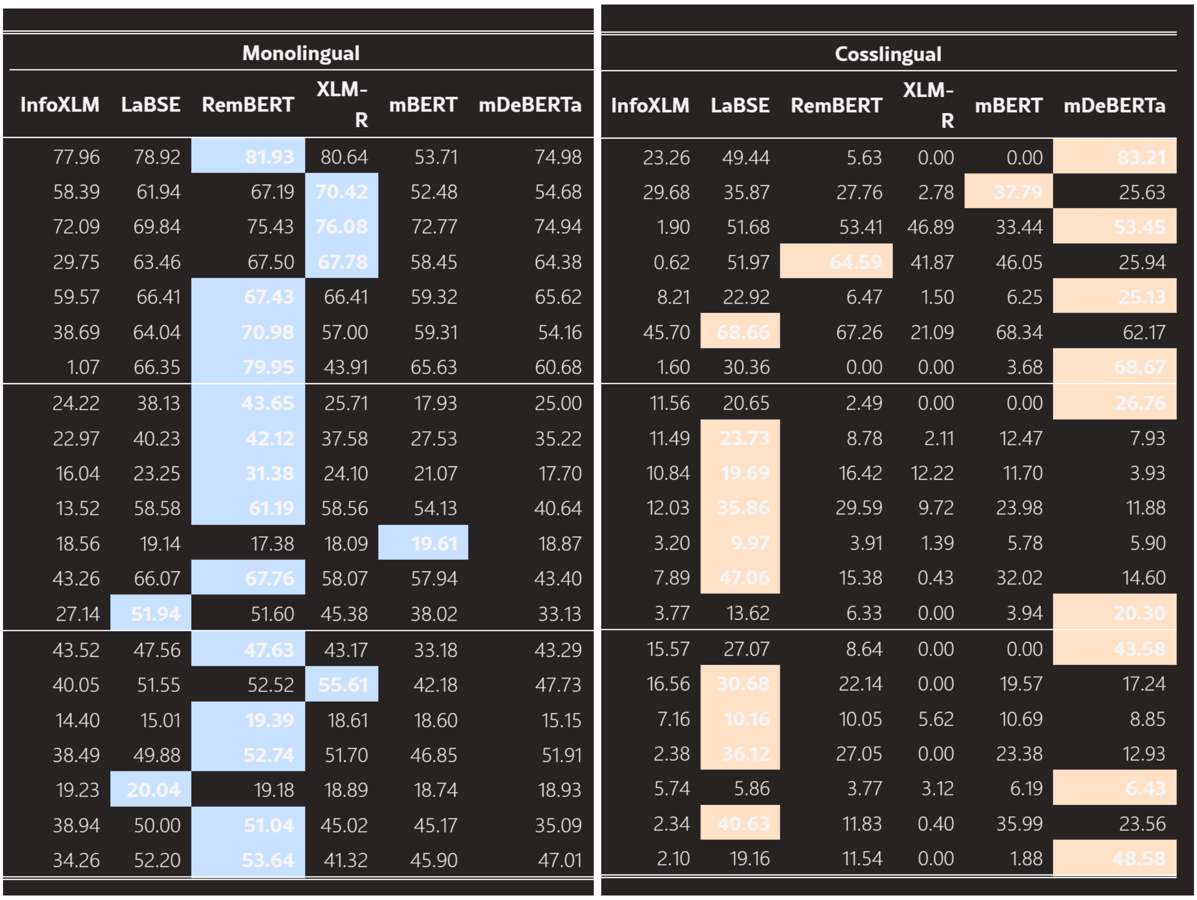
Multiple labels were allowed due to the conceptual and contextual overlap often observed in polarized content.



Polarization Types					Polarization Manifestations				
gender/sexual	political	religious	racial/ethnic	other	extreme_language	stereotype	invalidation	lack_of_empathy	dehumanization
29	3342	99	1297	189	1527	2729	799	880	657
102	178	86	171	220	341	283	240	175	102
18	892	44	98	5	179	138	85	45	39
241	1981	448	782	653	1001	1420	1258	1124	568
46	285	149	183	22	179	253	13	53	202
254	742	571	521	405	622	843	205	599	307
236	1533	557	423	126	1409	979	670	632	501

We pursued two main experimental paradigms:

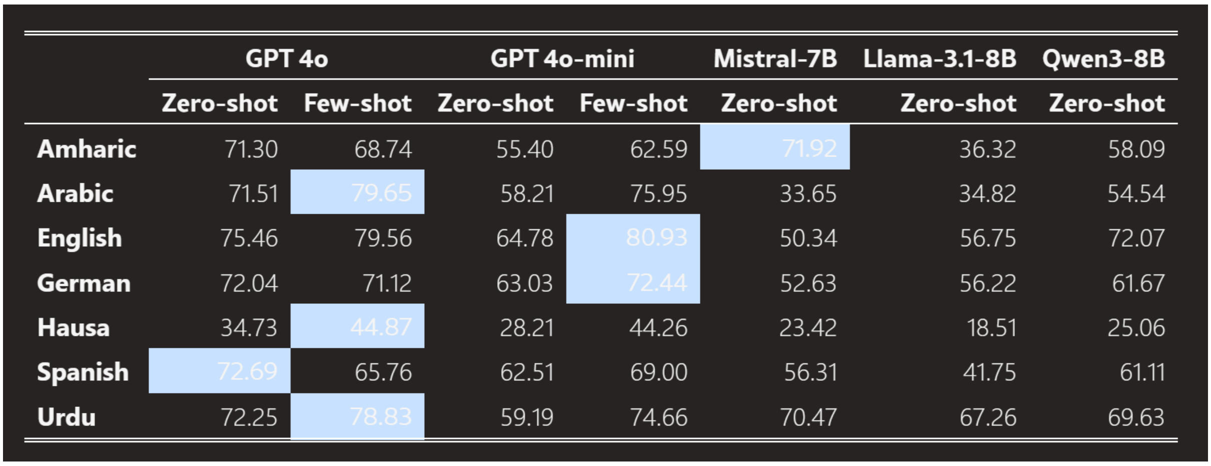
1. **Fine-tuning Multilingual Language Models (MLMs)**: We fine-tuned six multilingual models including InfoXLM (Chi et al., 2021), LaBSE (Feng et al., 2022), RemBERT (Chung et al., 2021), XLM-R (Conneau et al., 2020), mBERT (Devlin et al., 2019), and mDeBERTa (He et al., 2023) for monolingual and crosslingual experiments.



Monolingual						Cosslingual					
InfoXLM	LaBSE	RemBERT	XLM-R	mBERT	mDeBERTa	InfoXLM	LaBSE	RemBERT	XLM-R	mBERT	mDeBERTa
77.96	78.92	81.93	80.64	53.71	74.98	23.26	49.44	5.63	0.00	0.00	83.21
58.39	61.94	67.19	70.42	52.48	54.68	29.68	35.87	27.76	2.78	37.79	25.63
72.09	69.84	75.43	76.08	72.77	74.94	1.90	51.68	53.41	46.89	33.44	53.45
29.75	63.46	67.50	67.78	58.45	64.38	0.62	51.97	64.59	41.87	46.05	25.94
59.57	66.41	67.43	66.41	59.32	65.62	8.21	22.92	6.47	1.50	6.25	25.13
38.69	64.04	70.98	57.00	59.31	54.16	45.70	68.66	67.26	21.09	68.34	62.17
1.07	66.35	79.95	43.91	65.63	60.68	1.60	30.36	0.00	0.00	3.68	68.67
24.22	38.13	43.65	25.71	17.93	25.00	11.56	20.65	2.49	0.00	0.00	26.76
22.97	40.23	42.12	37.58	27.53	35.22	11.49	23.73	8.78	2.11	12.47	7.93
16.04	23.25	31.38	24.10	21.07	17.70	10.84	19.69	16.42	12.22	11.70	3.93
13.52	58.58	61.19	58.56	54.13	40.64	12.03	35.86	29.59	9.72	23.98	11.88
18.56	19.14	17.38	18.09	19.61	18.87	3.20	9.97	3.91	1.39	5.78	5.90
43.26	66.07	67.76	58.07	57.94	43.40	7.89	47.06	15.38	0.43	32.02	14.60
27.14	51.94	51.60	45.38	38.02	33.13	3.77	13.62	6.33	0.00	3.94	20.30
43.52	47.56	47.63	43.17	33.18	43.29	15.57	27.07	8.64	0.00	0.00	43.58
40.05	51.55	52.52	55.61	42.18	47.73	16.56	30.68	22.14	0.00	19.57	17.24
14.40	15.01	19.39	18.61	18.60	15.15	7.16	10.16	10.05	5.62	10.69	8.85
38.49	49.88	52.74	51.70	46.85	51.91	2.38	36.12	27.05	0.00	23.38	12.93
19.23	20.04	19.18	18.89	18.74	18.93	5.74	5.86	3.77	3.12	6.19	6.43
38.94	50.00	51.04	45.02	45.17	35.09	2.34	40.63	11.83	0.40	35.99	23.56
34.26	52.20	53.64	41.32	45.90	47.01	2.10	19.16	11.54	0.00	1.88	48.58

For cross-lingual transfer experiments, each MLM was trained on all other languages within the same language family (Afro-Asian or Germanic), excluding the target language, and evaluated on the target language's test set for the same three tasks.

2. **Evaluating Large Language Models (LLMs) in Zero- and Few-shot Settings**: We tested models including QWEN3-8B, LLAMA-3.1-8B, MIXTRAL-8X7B, and GPT4o/mini.



	GPT 4o		GPT 4o-mini		Mistral-7B	Llama-3.1-8B	Qwen3-8B
	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Zero-shot	Zero-shot
Amharic	71.30	68.74	55.40	62.59	71.92	36.32	58.09
Arabic	71.51	79.65	58.21	75.95	33.65	34.82	54.54
English	75.46	79.56	64.78	80.93	50.34	56.75	72.07
German	72.04	71.12	63.03	72.44	52.63	56.22	61.67
Hausa	34.73	44.87	28.21	44.26	23.42	18.51	25.06
Spanish	72.69	65.76	62.51	69.00	56.31	41.75	61.11
Urdu	72.25	78.83	59.19	74.66	70.47	67.26	69.63

Our choice of models is not exhaustive. Although we included several leading multilingual models and both open and closed LLMs. Adding more language-specific models in the future could improve results, especially for monolingual scenarios.

Finally, for some of the languages in our benchmark, the available data size is still limited, which may constrain the generalizability of model training and evaluation for those cases. Future work should expand dataset size and diversity, and explore language- or region-specific model development to better support underrepresented contexts.

5. Other works

A. “NLP-Driven Approaches to Measuring Online Polarization and Radicalization (Ghafouri, 2024)”

- Focus of the Work:
  This doctoral thesis investigates the limitations of standard NLP models (like BERT) when processing polarized text and proposes new architectures to better detect ideological divides and radicalization in online communities (specifically the "Manosphere").

- Key Contributions:
  - *Stance-Aware Sentence Transformers:*
    Ghafouri identifies a critical flaw in traditional text embeddings: they tend to cluster sentences based on topic rather than viewpoint. (For example, "I love the policy" and "I hate the policy" are often mathematically close because they share the same keywords). The thesis introduces a "stance-aware" fine-tuning approach that forces the model to separate opposing views in the vector space.

  - *Quantifying Echo Chambers:*
    The work moves beyond individual sentence classification to measure the "Echo Chamber Effect." By analyzing the semantic distance between users' posts, the author creates a metric to quantify how insulated a community is from opposing views.

  - *Radicalization Detection:*
    The thesis applies these methods to gender-based polarization (e.g., Incel and Red Pill communities), effectively demonstrating how polarization acts as a gateway to radicalization.

- Relevance to Our Project:
  This work is directly relevant to Subtask 1 (Detection) and Subtask 3 (Manifestation). It provides a blueprint for handling the "semantic overlap" problem, where models confuse healthy debate with polarization. Furthermore, the focus on gender-based radicalization aligns with the "Gender/Sexual Orientation" category in the POLAR dataset.

B. “Quantifying polarization in online political discourse (Muñoz et al., 2024)”

- Focus of the Work:
  This study challenges the US-centric view of polarization (which assumes a binary, two-party system) and aims to develop robust metrics for measuring polarization in multi-party political environments, using the Spanish elections as a case study.

- Key Contributions:
  - *Multi-Party Metrics:*
    The authors demonstrate that existing algorithms (like Random Walk Controversy) fail when applied to complex political landscapes with more than two opposing sides. They introduce a new methodological pipeline that combines network analysis (user interactions) with content analysis to map polarization across multiple political clusters.

  - *Ideological vs. Affective Polarization:*
    The study distinguishes between disagreeing on ideas (ideological) and disliking the other group (affective). They find that while ideological distances might remain stable, the structural division (who talks to whom) drives the perception of polarization.

  - *Contextual Dependency:*
    The paper proves that polarization metrics are highly sensitive to the specific political event (e.g., an election campaign vs. routine governance), suggesting that models need to be "event-aware."

- Relevance to Our Project:
  This paper is crucial for addressing the Multicultural/Multilingual aspect of SemEval Task 9. Since the POLAR dataset covers 20+ languages (many of which, like Italian, German, and Hindi, operate in multi-party systems), Muñoz et al.’s findings warn us that binary classification methods might oversimplify the problem. It highlights the need for our models to be sensitive to the specific political context of each language.

C. Social media polarization during conflict: ideological stance on Israel–Palestine Reddit comments (Ali et al., 2025)

- A classic *stance detection* setup in a highly polarized conflict context. Labels (Pro-Israel / Pro-Palestine / Neutral) are a direct proxy for polarization.

- Setup (Data):

  - 1.8M Reddit comments → filtered down to ~9,969 conflict-relevant comments (keywords like “Hamas terrorist”, “FreePalestine”, etc.).
  - Manually annotated into Pro-Israel, Pro-Palestine, Neutral with high agreement (Fleiss’ κ ≈ 0.93).

- Setup (Models Compared):

  - Classical ML: SVM, logistic regression, etc.
  - Neural nets: RNN, LSTM, GRU, BiLSTM.
  - PLMs: BERT cased/uncased, XLM-RoBERTa, DistilBERT, ELECTRA-small/base.
  - **LLMs:** Mixtral 8x7B, Mistral 7B, Gemma 7B, Falcon 7B with various prompting strategies.

- Setup(Implementation details):

  - Standard train/dev/test split, preprocessing includes lowercasing, cleaning URLs/usernames, etc.
  - PLMs: fine-tuned classification heads with cross-entropy.
  - LLMs: prompt-based classification; they try multiple prompting strategies.

- Setup(Results):

  - PLMs outperform classical ML & bare neural nets.

  - **Best overall:** Mixtral 8x7B with a “Scoring and Reflective Re-read” prompting scheme (basically self-critique / chain-of-thought style), achieving highest accuracy, precision, recall, and F1.

D. HASOC / OffensEval / multilingual hate/offensive detection

- Example: “Multilingual Hate speech and Offensive language detection in English, Hindi, and Marathi” (HASOC 2021 team paper)
- What they did:

  - **Data:** English, Hindi, Marathi (including code-mixed) tweets, with binary offensive vs not and multi-class fine-grained categories (OFFN/HATE/PRFN/NONE)
  - **Models:**
    - Classical models (Naive Bayes, SVM, RF with TF-IDF, doc2vec).

    - PLMs: BERT, RoBERTa, ALBERT, mBERT.

- **Implementation details:** Fine tuning BERT/RoBERTa with:
    - LR 2e-5–3e-5.
    - 7–10 epochs.
    - Macro-F1 as main metric (exactly like POLAR)

- **Results:**
    - Best macro-F1: ~0.79–0.82 for some subtasks (RoBERTa for English, mBERT for Hindi/Marathi).
    - Transformer models clearly beat classical ML across languages.