POLAR: A Benchmark for Multilingual, Multicultural, and Multi-Event Online Polarization

Usman Naseem[1], Juan Ren[1], Saba Anwar[2], Sarah Kohail[6], Rudy Alexandro Garrido Veliz[2], Robert Geislinger[2], Aisha Jabr[6], Idris Abdulmumin[5], Laiba Qureshi[2], Aarushi Ajay Borkar[2], Maryam Ibrahim Mukhtar[7], Abinew Ali Ayele[2,3], Ibrahim Said Ahmad[7,8], Adem Ali[2,3], Martin Semmann[2], Shamsuddeen Hassan Muhammad[4,7], Seid Muhie Yimam[2]

[1]Macquarie University, [2]University of Hamburg, [3]Bahir Dar University,[4]Imperial College London

[5]University of Pretoria, [6]Zayed University, [7]Bayero University Kano, [8]Northeastern University

Abstract

Online polarization poses a growing challenge for democratic discourse, yet most computational social science research remains monolingual, culturally narrow, or event-specific. We introduce POLAR, a multilingual, multicultural, and multievent dataset with over 23k instances in seven languages from diverse online platforms and real-world events. Polarization is annotated along three axes: presence, type, and manifestation, using a variety of annotation platforms adapted to each cultural context. We conduct two main experiments: (1) we fine-tune six multilingual pretrained language models in both monolingual and cross-lingual setups; and (2) we evaluate a range of open and closed large language models (LLMs) in few-shot and zero-shot scenarios. Results show that while most models perform well on binary polarization detection, they achieve substantially lower scores when predicting polarization types and manifestations. These findings highlight the complex, highly contextual nature of polarization and the need for robust, adaptable approaches in NLP and computational social science. All resources will be released to support further research and effective mitigation of digital polarization globally.

1 Introduction

Online polarization, defined as sharp division and antagonism between social, political, or identity groups, has become a pervasive threat to democratic institutions, civil discourse, and social cohesion worldwide (Waller and Anderson, 2021; Iandoli et al., 2021). It is often fueled by biased or inflammatory content on social media, reinforcing echo chambers and undermining mutual understanding (Garimella, 2018). Polarized discourse not only amplifies ideological divides but can also escalate into hate speech, harassment, and real-world violence. As such, early detection of polarization is critical to designing interventions that promote healthier online ecosystems.

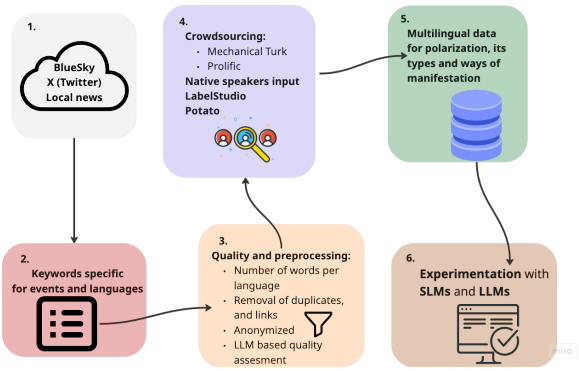


Figure 1: Pipeline for POLAR construction: multi-platform data curation in 7 languages, annotation workflow with quality control, and benchmarking.

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).

Unlike prior work, POLAR supports three complementary tasks:

1. **Binary Polarization Detection:** Is a message polarized or not?
2. **Polarization Type Classification:** What social dimension is targeted (e.g., political, religious, racial)?
3. **Manifestation Identification:** How is polarization rhetorically expressed (e.g., stereotyping, deindividuation, extreme language)?

We design a cross-cultural annotation protocol tailored for each language's sociopolitical context. Extensive qualitative analysis reveals how event salience, linguistic norms, and platform affordances shape polarization dynamics across languages. The complete pipeline can be found in Figure 1. We benchmark a range of multilingual language models (MLMs) and large language models (LLMs) under zero-shot, few-shot, and cross-lingual scenarios. Our experiments highlight the challenges of generalization and the limitations of current models in capturing nuanced rhetorical patterns across languages. Our contributions are as follows:

- We release POLAR, the first large-scale, multilingual, fine-grained dataset for polarization detection across 7 languages and diverse global events.
- We define a taxonomy of polarization types and manifestations, operationalized through a robust cross-lingual annotation protocol.
- We provide comprehensive benchmarks using state-of-the-art MLMs and LLMs across multiple evaluation settings (monolingual, cross-lingual, and few-shot).

2 Related Work

Online polarization has long been recognized as a threat to democracy and social cohesion, intensifying through social media echo chambers and biased content (Waller and Anderson, 2021; Iandoli et al., 2021; Garimella, 2018). As social media and other online platforms become key arenas for political and cultural discourse, the need for early detection and nuanced understanding of polarization has grown significantly. Such efforts are critical not only for content moderation, but also for peacebuilding, policy development, and responsible digital governance. Foundational research has defined polarization as both intergroup hostility and ingroup cohesion (Arora et al., 2022), and has highlighted its relationship with hate speech, fragmentation, and incivility (Bozdag and van den Hoven, 2020; Mathew et al., 2021).

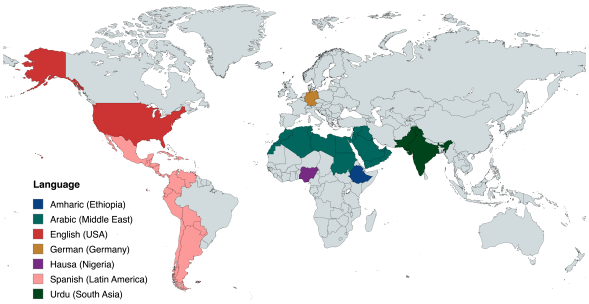


Figure 2: Languages represented in the dataset, covering diverse linguistic and regional contexts.

A growing body of research has documented the role of online spaces in intensifying polarization across different regions (Kubin and von Sikorski, 2021; Barberá, 2020; Gitlin, 2016; Soares and Recuero, 2021). However, most computational work focuses on high-resource languages and event- or region-specific datasets, limiting generalizability (Kubin and von Sikorski, 2021). This leaves a significant gap in our ability to generalize findings across cultures, languages, and events, especially in the Global South or multilingual regions.

The lack of standardized datasets across languages has hindered progress in developing and evaluating polarization detection models with cross-lingual or cross-cultural capabilities. Recent shared tasks on hate speech and toxicity (Basile and others, 2019; Mohammad and others, 2021; Pamungkas et al., 2020; Ousidhoum et al., 2024) have expanded the language and domain coverage, yet remain less fine-grained regarding polarization's diverse types and rhetorical manifestations. Our work addresses this gap by presenting the first comprehensive, fine-grained dataset benchmark for multilingual, multicultural, and multievent online polarization, enabling robust cross-lingual and context-aware modeling.

3 POLAR Dataset Construction

3.1 Data Collection

We collected data from various online platforms, including X (formerly Twitter), Facebook, Reddit, Bluesky, Threads, and news/commentary forums.

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.

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.

The Amharic dataset focuses on the Northern Ethiopia or Tigray War in Ethiopia; Arabic texts cover a broad array of social and economic topics; English data centers on US elections and international conflicts; the German corpus features election discourse and migration debates; the Hausa subset captures religious and ideological discussions in West and Central Africa; Spanish content spans abortion, migration, and indigenous/gender rights; and Urdu texts reflect political and sectarian divides.

3.2 Annotation Process

Given the cultural and linguistic breadth of POLAR , we developed detailed, multilingual annotation guidelines and deployed a hybrid strategy combining crowdsourcing with trained community annotators.

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.

Annotator Recruitment and Annotation Process We used both expert and crowd-sourced annotation strategies. Trained annotators used POTATO and Label Studio, while crowdsourced annotators used Mechanical Turk and Prolific. To evaluate inter-annotator agreement, we report both Cohen's Kappa and Fleiss' Kappa. Since different teams were responsible for dataset creation across various languages, we provide a detailed description of the annotation process for each language:

- Amharic: We used train annotators to annotated 5,000 samples in five batches using the POTATO annotation tool. Annotators completed three rounds of guideline training and received continuous feedback, with ongoing supervision to ensure quality control. The annotation quality achieved a Fleiss' Kappa of 0.49, with full agreement in 69.62% of cases and a peak pairwise Cohen's Kappa of 0.65.
- Arabic: Three topic-aware annotators labeled 2,118 samples using POTATO. Final polarization distribution was 16% positive, with Cohen's Kappa at 0.296. Annotations were revised to ensure thematic consistency.
- German: Annotation was conducted through Prolific. Annotators were screened via surveys and underwent pilot testing. Final annotation used a locally hosted POTATO interface adapted for multilabel inputs. Inter-annotator agreement was moderate, with notable consistency on polarized cases.
- Hausa: We used trained annotators and Label Studio with iterative feedback and internal consistency checks. Data focused on religion and ideology on social media.
- Spanish and English: Amazon Mechanical Turk (MTurk) was used with strict quality filters, such as task completion history and approval ratings. Annotators received feedback and were contacted throughout for quality assurance. English annotation reached a Kappa of 0.52 (up from 0.31 after worker filtering); Spanish showed lower agreement with an average Kappa of 0.24.
- Urdu: Two approaches were used: *(1) Manual Annotation:* 1,792 samples labeled in batches with average Cohen's Kappa of 0.55 (batch 1: 0.79; batch 2: 0.56; batch 3: 0.39). Disagreements were resolved by a fifth annotator. *(2) Prolific:* 1,460 samples were annotated in three batches. Initial agreement was low (Fleiss' Kappa = 0.05). Additional filters (education, region) raised agreement in later

Language	Total	Polarization	Polarization Types					Polarization Manifestations					
			gender/sexual	political	religious	racial/ethnic	other	vilification	extreme_language	stereotype	invalidation	lack_of_empathy	dehumanization
Amharic	5000	3753	29	3342	99	1297	189	2398	1527	2729	799	880	657
Arabic	2118	338	102	178	86	171	220	413	341	283	240	175	102
English	3025	1021	18	892	44	98	5	342	179	138	85	45	39
German	3467	1596	241	1981	448	782	653	1371	1001	1420	1258	1124	568
Hausa	5563	625	46	285	149	183	22	74	179	253	13	53	202
Spanish	2002	1057	254	742	571	521	405	798	622	843	205	599	307
Urdu	2800	1860	236	1533	557	423	126	1548	1409	979	670	632	501

Table 2: Number of samples labeled positive for each annotation task across languages. Labels are grouped by task: Polarization (Task 1), Polarization Types (Task 2), and Polarization Manifestations (Task 3).

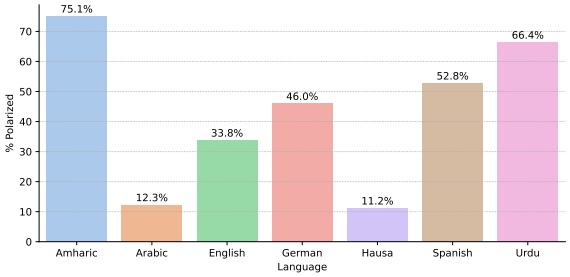


Figure 3: Percentage of polarization per language

batches to 0.28 and 0.30. The final dataset contains 2,800 high-quality Urdu samples.

3.3 Dataset Characteristics

Polarization Prevalence: As shown in Figure 3, Amharic (75%) and Urdu (66%) has the highest polarization rates, tied to their sources - Northern Ethiopia's Tigray conflict and Pakistan's political and sectarian climate. Spanish (53%), German (46%), and English (34%) fall in the mid-range. Arabic (12%) and Hausa (11%) reflect broader topical focus and possibly platform moderation effects. Hausa shows the lowest polarization rate despite covering sensitive religious and ideological topics, possibly due to moderation or cultural norms in expression. Datasets with acute sociopolitical events (e.g., Amharic, Urdu) show higher polarization. Topic scope, moderation, and platform dynamics significantly influence polarization prevalence.
Types of Polarization: Figure 4 show types of polarization. Political polarization dominates in Amharic (67%), German (57%), Urdu (55%), and Spanish (37%). Racial/ethnic and religious types also appear prominently in Spanish, Amharic, and German. Gender/sexual identity polarization is generally low but visible in Spanish (13%) and Urdu (8%). Thematic emphasis (e.g., migration, elections) shapes type distribution. Spanish and Urdu show diverse forms of identity-based polarization, while political conflict dominates Amharic, Urdu, and German.
Polarization Manifestations: Figure 4 show types of Manifestation. Stereotyping and vilification are most frequent in Urdu, Amharic, German, and Spanish. Urdu's vilification rate is highest (55%). Extreme language is strong in Urdu (50%), Spanish (31%), and Amharic (31%). Invalidation and dehumanization peak in German and Urdu. Conflict-heavy datasets (Urdu, Amharic) correlate with higher rates of hostile manifestations. Platform moderation and topic scope likely suppress such features in Arabic and Hausa.
Together, these analyses highlight how linguistic, cultural, and contextual variables - alongside event salience and platform dynamics - shape the structure and tone of online polarization. The POLAR dataset (see Table 2 for detailed statistics of POLAR dataset) offers a robust benchmark for examining these effects at scale.

4 Experimentation and Results

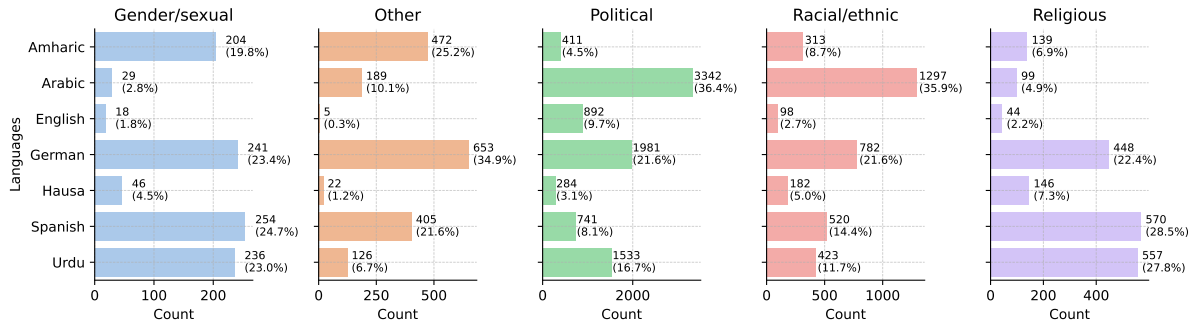
4.1 Experimental Setup

To evaluate POLAR , we conducted baseline experiments on three polarization detection tasks:

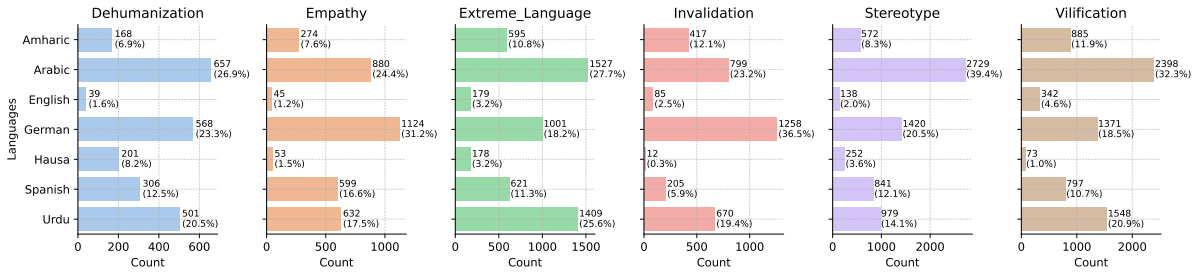
- **Task 1:** Binary classification - determining whether a text is polarized.
- **Task 2:** Multi-label classification - identifying polarization types (e.g., political, religious, ethnic).
- **Task 3:** Multi-label classification - detecting polarization manifestations (e.g., incivility, stereotyping, dehumanization).

For data split, we used 70% for training, 10% for validation, and 20% for testing, as summarized in Table 1. 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),



(a) Types



(b) Manifestation

Figure 4: Annotation statistics for polarization classification, types, and manifestations

mBERT (Devlin et al., 2019), and mDeBERTa (He et al., 2023) for monolingual and crosslingual experiments.

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.

4.2 Results and Analysis

4.2.1 Monolingual Setup Results

To benchmark the performance of state-of-the-art multilingual language models (MLMs) for polarization detection, we fine-tuned six prominent pretrained models: 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). Each model was evaluated in a monolingual (within-language) setting across three tasks: (i) binary polarization classification, (ii) multi-label polarization types classification, and (iii) multi-label polarization manifestations detection.

For Task 1, binary polarization classification, RemBERT consistently achieved the highest macro F1-scores for most languages, while XLM-R performed best in Arabic, English, and German. In the multi-label polarization types classification (Task 2), RemBERT again provided the top scores for Spanish, except for Hausa and Urdu, where mBERT and LaBSE performed best, respectively. For Task 3, multi-label polarization manifestations classification, the results were generally lower across all models, with RemBERT delivering the highest scores for most languages, except for Arabic and Hausa.

4.2.2 Cross-Lingual MLM Results

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.

In the cross-lingual setup for polarization classification (Task 1), mDeBERTa achieved the highest F1-macro for Amharic, English, Hausa, and Urdu. For Arabic, German, and Spanish, mBERT, RemBERT, and LaBSE performed slightly better than the other models, respectively.

For Task 2, polarization types classification, performance was generally lower as expected in the cross-lingual scenario. LaBSE yielded the best results for four languages, while mDeBERTa performed better for Amharic and Urdu.

Similarly, for Task 3, polarization manifestation classification, the results were quite low overall, but LaBSE and mDeBERTa showed relatively better

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

Table 3: Average F1-Macro for all three tasks. In the *monolingual* settings, we train and evaluate the model for each language separately. In the *crosslingual* setting, we train on all languages within a language family (**AfroAsian:** Amharic, Arabic, Hausa, and Urdu), **Germanic:** English, German, and Spanish) except the target language, and evaluate on the test set of the target language. The best performance scores are highlighted in blue and orange, respectively.

performance compared to the other models.

This suggests that general polarization detection transfers more easily across languages than the more nuanced identification of polarization types and manifestations, which are culturally and linguistically specific. Effective cross-lingual transfer likely requires aligned annotation schemes and fine-tuning strategies sensitive to these differences.

4.2.3 Zero and Few-shot LLM Performance

We evaluated five state-of-the-art LLMs (QWEN3-8B, LLAMA-3.1-8B, MIXTRAL-8X7B, GPT4o/mini) in a zero- and few-shot setting across languages and for the first ask. Larger models such as GPT4o and GPT4o-mini outperformed smaller ones, especially in capturing subtle polarization cues.

5 Discussion

Our multilingual and cross-cultural investigation into online polarization reveals that polarization is a deeply contextual and event-driven phenomenon. Its prevalence and intensity are closely tied to the sociopolitical climate, public discourse, and digital environment of each region. In societies experiencing major crises, such as the war in Ethiopia, religious tensions in Pakistan, and international conflicts like the wars in Ukraine and Gaza that influence the US election, we find polarization to be especially amplified on online platforms. Conversely, broader topical selection, moderation, or more diffuse societal tensions are associated with noticeably lower polarization.

The types of polarization evident in online discourse are shaped by the most salient local issues: political (Ethiopia, US, German), religious (Pakistan), ethnic (Ethiopia), or gender-based (Spain) divides tend to dominate where these subjects are at the forefront of public debate or conflict. This underscores the importance of accounting for each region's unique sociopolitical history and cultural landscape when analyzing or intervening in digital polarization.

Polarization also manifests in varied rhetorical forms, including stereotyping, vilification, and exclusionary or hostile language. These patterns are particularly pronounced amid contentious issues and persistently divided communities, such as civil war, election, migration issues, religious divide, and so on.

From a computational perspective, our results demonstrate both the feasibility and the challenges of detecting polarization in multilingual settings. Binary polarization classification is relatively tractable for well-resourced languages and clear-cut contexts, especially when leveraging state-of-the-art multilingual models (Sec. 4. However, classifying the specific types and manifestations of polarization introduces significantly greater ambiguity and requires a much deeper, context-sensitive

	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

Table 4: F1-Macro resulting from the zero- and few-shot LLM experiments with the POLAR dataset. The highest value per language is highlighted in blue.

understanding. Performance on these more nuanced tasks remains limited. This highlights important directions for future work, including the integration of cultural signals and context or event understanding, the development of more robust multilingual embeddings, and the pursuit of consistent annotation (with a great deal of annotator training and follow up, specially for the crowdsourcing annotations). The wide variability across languages, contexts, and events further affirms the necessity of designing language, culture, and event specific approaches, particularly for underrepresented regions in global NLP research.

Taken together, these findings highlight that effective detection and mitigation of online polarization must move beyond generic or monocultural solutions. There is a pressing need for context-aware, adaptable methodologies that acknowledge both the universal and the local characteristics of polarized discourse, ensuring relevance and effectiveness across the world's diverse digital landscapes.

6 Conclusion

In this study, we introduced POLAR, the first multilingual, multicultural, and multievent dataset for benchmarking online polarization. Our findings showed that polarization was deeply context-dependent, manifesting in diverse types and rhetorical forms shaped by local sociopolitical dynamics. Theoretically, we provide new evidence that polarization is a multifaceted, cross-lingual phenomenon strongly influenced by language, culture, and current events, rather than being uniform across settings. On a practical level, our benchmark revealed that while binary detection of polarization was feasible with current models, accurately identifying specific types and manifestations remained a persistent challenge, especially for multilingual and low-resource contexts.

This work established a foundation for further computational studies of digital polarization and for developing more culturally robust moderation and intervention tools. Looking ahead, future research should integrate richer contextual and cultural signals into model architectures, refine annotation guidelines for cross-regional consistency, and expand evaluations to include additional languages, social domains, and further events. By releasing our data and benchmarks, we aimed to catalyze further innovation toward nuanced detection and effective mitigation of online polarization worldwide.

Limitations

While POLAR represents an important step toward multilingual, multicultural, and multievent polarization analysis, several limitations remain. First, annotator understanding - particularly in crowdsourced setups - was sometimes limited, potentially impacting label quality. We mitigated this through strict quality assurance methods, including control questions, pre-study surveys, and ongoing annotator assessment, but some variability in interpretation may persist.

Second, in-house annotation, while yielding higher consistency, sometimes introduced psychological challenges for annotators given the sensitive or hostile nature of polarized content. To address this, we provided detailed instructions and support resources to reduce stress and clarify expectations, but some emotional burden may have remained.

Third, 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.

Ethics Statement

This research uses only publicly available, anonymized data and addresses sensitive topics around polarization in diverse cultures. All annotation was conducted by native speakers using culturally appropriate guidelines; annotators were informed of the project's social good aims, possible distress, and could opt out anytime. Annotators received prompt and fair compensation above local wage standards or per Prolific's requirements. Despite rigorous protocols, labeling polarization remains subjective; we encourage responsible, ethically grounded use of this resource and discourage misuse.

References

Swapan Deep Arora, Guninder Pal Singh, Anirban Chakraborty, and Moutusy Maity. 2022. Polarization and Social Media: A Systematic Review and Research Agenda. *Technological Forecasting and Social Change*, 183:121942.

Pablo Barberá. 2020. Social media, echo chambers, and political polarization. *Social media and democracy: The state of the field, prospects for reform*, pages 34–55.

Valerio Basile and others. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63.

Engin Bozdag and Jeroen van den Hoven. 2020. Managing polarization and fake news: A new responsibility for digital platforms? *Ethics and Information Technology*, 22(1):77–80.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*, pages 1–17, Online.

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, Online.

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, Minneapolis, Minnesota.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland.

Kiran Garimella. 2018. *Polarization on Social Media*. Ph.D. thesis, Aalto University, Finland.

Todd Gitlin. 2016. The Outrage Industry: Political Opinion Media and the New Incivility By Jeffrey M. Berry and Sarah Sobieraj Oxford University Press. *Social Forces*, 95(1):e26–e26.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda.

Luca Iandoli, Simonetta Primario, and Giuseppe Zollo. 2021. The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change*, 170:1–12.

Emily Kubin and Christian von Sikorski. 2021. The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, 45(3):188–206.

Binny Mathew, Punyajoy Saha, Seyed Mahbub Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Saif M. Mohammad and others. 2021. Conan: The contest of online offensive language identification. In *Proceedings of the 2021 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 4712–4727.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024. Semeval task 1: Semantic textual relatedness for african and asian languages. *Preprint*, arXiv:2403.18933.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: A multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Felipe Bonow Soares and Raquel Recuero. 2021. Hashtag wars: Political disinformation and discursive struggles on twitter conversations during the 2018 brazilian presidential campaign. *Social Media+ Society*, 7(2):1–13.

Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7887):264–268.