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

Anonymous ACL submission

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 multi-event dataset with over 110K instances in 22 languages drawn from diverse online platforms and real-world events. Polarization is annotated along three axes, namely *detection*, *type*, and *manifestation*, using a variety of annotation platforms adapted to each cultural context. We conduct two main experiments: *(1)* fine-tuning six pretrained small language models; and *(2)* evaluating a range of open and closed large language models in few-shot and zero-shot settings. The results show that, while most models perform well in binary polarization detection, they achieve substantially lower performance when predicting polarization types and manifestations. These findings highlight the complex, highly contextual nature of polarization and demonstrate 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). It is often fueled by biased or inflammatory content in digital media, strengthening echo chambers and undermining mutual understanding (Garimella, 2018). Polarized discourse amplifies ideological divides and can escalate into hate speech, harassment, and real-world violence (Piazza, 2023; Martínez-España et al., 2024). Therefore, early detection of polarization is essential for designing interventions that promote healthier online ecosystems.

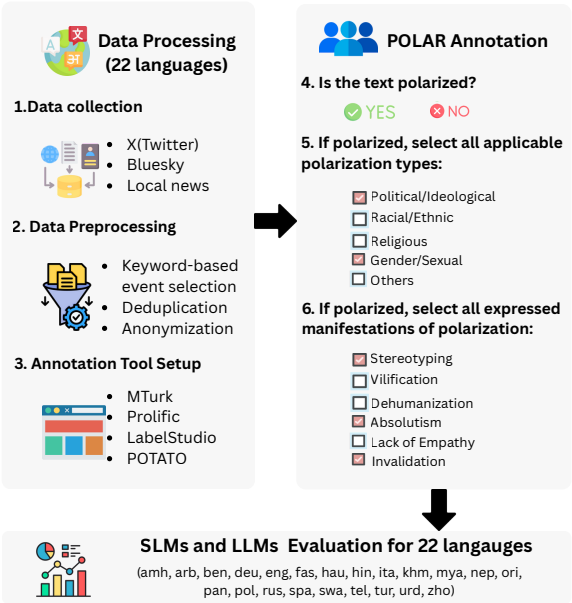

Figure 1: **Pipeline for POLAR construction.** Data curation in 22 languages, annotation workflow, and benchmarking.

Despite increasing attention, computational approaches to polarization face several 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 (Simchon et al., 2022; Piazza, 2023; Rojo Martínez, 2025). *Second*, previous studies are event-specific or monodomain, such as U.S. elections or Western political debates, limiting their generalizability (Demszky et al., 2019; Casal Bértoa and Rama, 2021; Sinno et al., 2022; Piazza, 2023). *Third*, the conceptualization of polarization in NLP has largely been binary or topic-focused (Hofmann et al., 2022), overlooking the multifaceted ways in which polarization is expressed through vilification, dehumanization, stereotyping, or other rhetorical tactics (Donohue

and Hamilton, 2022). These tactics are often employed in political rhetoric, social debates, or campaigns to solidify support within a group and increase hostility to others.

To address these gaps, we introduce POLAR, a large-scale, multilingual, multicultural, and multi-event dataset for fine-grained polarization detection. POLAR supports 22 languages spanning seven language families and balances high-, medium-, and low-resource languages (see Table 5). The wide extent of our efforts can be seen in Figure 2. Unlike prior work, POLAR supports three complementary tasks:

- **Binary Polarization Detection:** Determine whether a given text expresses polarization. We refer to this task as **POLARDETECT**.
- **Polarization Type Classification:** Identify the social dimension underlying the polarization (e.g., political, religious, racial). We refer to this task as **POLARTYPE**.
- **Manifestation Identification:** Detect how polarization is rhetorically manifested, including strategies such as stereotyping, deindividuation, vilification, dehumanization, extreme language, and other rhetorical devices. We refer to this task as **POLARMANIFEST**.

For each task, we develop a cross-cultural annotation protocol tailored for each language's sociopolitical context. The complete data construction pipeline is illustrated in Figure 1. We benchmark a range of Small Language Models (SLMs) and Large Language Models (LLMs) under zero-shot and few-shot settings. 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 analysis across 22 languages and diverse global events, comprising 110K instances.
- 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 SLMs and LLMs across multiple evaluation settings.

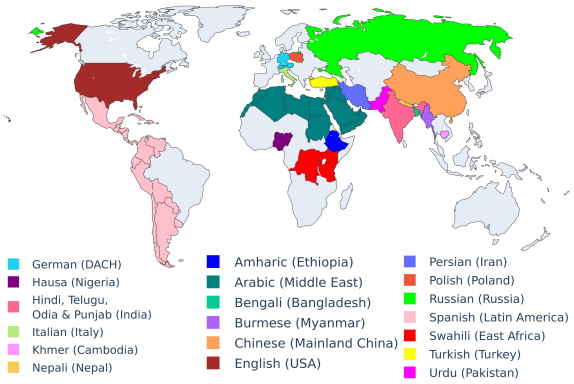

Figure 2: Languages represented in a world map covered by POLAR, covering diverse linguistic and regional contexts. The language and societal context can present itself across varied areas. Language assignments to countries and regions are approximate.

2 Related Work

Online polarization poses a threat to social cohesion, exacerbated by 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. Polarization detection is important for content moderation, peace building, policy development, responsible digital governance and healthy democracy. Foundational research has defined polarization as both intergroup hostility and blind ingroup cohesion (Arora et al., 2022), and has highlighted its relationship with hate speech, fragmentation, and incivility (Mathew et al., 2021).

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 for low-resource languages 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 et al., 2019; Pamungkas et al., 2020) have expanded the lan-

guage and domain coverage, yet remain less fine-grained regarding polarization's diverse types and rhetorical manifestations. Our work addresses this gap by presenting the 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 Operational Definitions

In this work, we define polarization as the increasing extremity of opinions, beliefs, or behaviors, resulting in heightened intergroup divisions and conflict. Polarization types are classified as:

- **Political polarization**: Focuses on division, intolerance, and conflict between political parties and followers.
- **Racial or ethnic polarization**: Focuses on ethnic identity or racial origin and incites division, intolerance, and conflict between ethnic groups or races.
- **Religious polarization:** Focuses on religious identity and incites division, intolerance, and conflict between religious followers
- **Gender/ Sexual polarization**: Refers to the exclusion, discrimination, and marginalization of individuals based on their gender or sexual orientations.
- **Other**: Polarized texts targeting other groups or identities not covered above, such as economic class, technology or media.

In addition to topical categories, we further distinguish polarization by its rhetorical manifestations, defined as follows:

- **Stereotype**: A generalized belief that attributes specific characteristics to all members of a group, often neglecting individual differences, thereby reducing complex personalities to simplistic and uniform representations.
- **Vilification**: The act of defaming or demonizing a particular group, person, or entity by inciting fear, often through exaggeration, misrepresentation, or biased framing that portrays the subject negatively and harmfully.
- **Dehumanization**: The process of depriving a group or individual of their human qualities or personality by comparing them to animals, machines, or objects, or otherwise denying their humanity, dignity, or individuality.
- **Extreme Language and Absolutism**: The use of language that is extreme or makes definitive, all-encompassing statements, often involving words like "always", "never", "worst", or "best", and presenting issues in a dichotomous manner such as "us versus them" or "right versus wrong".
- **Lack of Empathy**: The absence of compassion or recognition for other viewpoints or experiences in the text.
- **Invalidation**: The act of denying or dismissing the identity and existence of individuals or groups, thereby rejecting their sense of self and their presence.

Appendix E contains more details and examples for each manifestation of polarization.[1]

3.2 Data Collection

Data Sources: We collected data from a range of online platforms, including major social media sites (e.g., X, Facebook, Reddit, Bluesky, Threads, YouTube comments, Weibo, and Zhihu) and local news or commentary forums (see Table 6 in the Appendix). For several languages (Chinese, Turkish, Polish, Burmese, and Italian), we sampled and re-annotated instances from existing toxic or hate speech datasets, including ToxiCN (Bai et al., 2025), COLD (Deng et al., 2022), the Turkish Hate Speech Dataset (Çöltekin, 2020), Myanmar Hate Speech (Kyaw et al., 2024), HaSpeede2 (Sanguinetti et al., 2020), HODI (Nozza et al., 2023), and BAN-PL (Kolos et al., 2024).

Event Selection: We curated the dataset to cover diverse real-world events, grounding event selection in the sociopolitical and socioeconomic contexts specific to each language and cultural setting. The data span a broad range of events and issues, including armed conflicts (e.g., the Tigray War in Ethiopia, the Russia–Ukraine conflict, and the Gaza genocide), elections and party politics (e.g., the 2024 U.S. and 2025 German elections), public health crises, large-scale migration, climate change, and broader socioeconomic debates. The dataset also includes discussions related to gender and indigenous rights, religion, and ideology. For some languages, such as Bengali, a broader sampling strategy was adopted due to the lack of sufficiently large event-specific data on the selected platforms.

[1] Since we define the above polarization manifestations as rhetorical tactics, we have used the terms "manifestation" and "rhetorical tactics" interchangeably.

#	Lang.	Inner Agr. (κ)	Total	POLARDETECT	POLARTYPE					POLARMANIFEST					
				Polarized (%)	Political	Racial / Ethnic	Religious Polarization	Gender / Sexual	Other	Stereo-type	Vilifi-cation	Dehuman-ization	Extreme Language	Lack of Empathy	Invalid-ation
1	eng	0.39	4,834	1,767 (37%)	1,726	422	168	108	190	730	1,272	586	1,156	536	879
	deu	0.10*	4,771	2,274 (48%)	1,959	883	531	281	658	1,728	1,435	712	1,038	1,272	775
	urd	0.29 / 0.70*	5,346	3,714 (69%)	3,603	2,908	2,954	2,739	2,713	3,328	3,460	2,973	3,324	3,007	3,059
	ben	0.59	5,000	2,127 (43%)	1,701	38	97	26	503	298	1,199	535	236	95	89
	hin	0.49	4,117	3,510 (85%)	3,051	500	2,417	472	540	2,047	2,683	750	2,082	2,336	2,703
	ori	0.46	3,552	1,021 (29%)	744	179	225	119	130	354	385	24	476	56	120
	nep	0.79	3,008	1,510 (50%)	518	422	239	158	354	806	947	198	816	318	450
	pan	0.55*	2,609	1,280 (49%)	803	153	205	291	233	424	1,038	574	624	324	637
	ita	0.39	5,038	2,165 (44%)	412	926	368	461	219				–		
	spa	0.26	4,958	2,479 (50%)	1,351	945	787	665	665	1,355	1,517	443	1,199	1,187	526
	rus	0.39	5,023	1,525 (30%)	696	494	205	284	119				–		
	pol	0.46	3,587	1,504 (42%)	1,313	323	131	165	232				–		
	fas	0.78	4,943	3,656 (74%)	2,170	120	476	296	1,197	649	2,850	213	835	487	394
2	hau	0.48	5,477	587 (11%)	267	173	139	44	21	234	68	193	165	48	13
	arb	0.25	5,070	2,268 (45%)	1,205	874	424	553	847	1,691	1,896	555	1,540	863	411
	amh	0.59	4,999	3,747 (75%)	3,339	1,296	99	29	1,239	2,728	2,398	657	1,527	879	799
3	zho	0.64	6,421	3,208 (50%)	376	1,475	127	1,085	552	1,931	1,188	323	522	506	307
	mya	0.13	4,334	2,508 (58%)	1,095	228	133	459	1,956				–		
4	swa	0.56	10,487	5,257 (50%)	279	3,721	371	234	833	4,160	4,324	1,340	2,509	3,120	2,456
5	khm	0.83	9,960	9,042 (91%)	1,825	147	336	169	6,565	6,799	152	122	225	1,093	651
6	tel	0.70	3,550	1,885 (53%)	766	603	318	471	842	398	781	88	477	933	809
7	tur	0.46	3,566	1,776 (50%)	1,569	579	557	221	193	1,453	1,169	378	1,575	369	159
	Total		**110,650**	**58,810 (53%)**	**30,768**	**17,409**	**11,307**	**9,330**	**20,801**	**31,113**	**28,762**	**10,664**	**20,326**	**17,429**	**15,237**

Table 1: Number of samples labeled as positive for each annotation task across languages. Inner agreement values denote inter-annotator agreement per language (Fleiss's κ unless otherwise noted). * denotes exceptions: German uses Krippendorff's α; Punjabi reports identical Krippendorff's α and Cohen's κ; Urdu reports Fleiss's κ / Cohen's κ. Polarization manifestation annotations are not available for Italian, Russian, Burmese, and Polish. Languages are ordered by language families and sub-branches as defined in Table 5.

To collect data for these events, we adopted a dynamic, keyword-driven strategy tailored to each language and topic. Keyword lists were curated by human experts and native speakers to capture culturally and politically salient discourse and were used to retrieve data from online platforms. Table 7 in the Appendix provides an overview of the dataset composition, size, and event coverage.

Dataset Quality Control: To ensure high-quality annotations across languages, we implemented several steps to ensure data quality throughout the quality control process. Before annotation, native speakers developed language-specific preprocessing pipelines. These pipelines included standard NLP procedures such as tokenization, word count filtering, and duplicate detection. Instances that were either too short or excessively long based on language-specific thresholds were removed. For anonymization, all usernames and URLs were replaced with standardized placeholders. For some languages, LLMs were used during the pre-filtering stage to increase the proportion of polarized content.

3.3 Annotation Process and Guidelines

We used a hybrid annotation strategy, leveraging crowd-sourced annotators and trained community annotators for low-resource languages where crowd-sourced annotation support is limited. For the crowd-sourced setting, annotators were selected based on their prior experience and annotation quality. Specifically, we filtered candidates using historical annotation agreement scores and conducted pilot rounds to identify those with consistent performance. Only annotators achieving a Fleiss' Kappa score of at least 0.8 were retained for the main task. For community annotation, we recruited native speakers with at least a bachelor's degree. Annotators received training, followed by a pilot round to assess their understanding and performance. Subsequent annotation was assigned in batches, with annotator performance monitored continuously to ensure consistency and adherence to guidelines.

The trained annotators used POTATO (Pei et al., 2022) and Label Studio[2] as annotation platforms, while the crowd-sourced annotators used Mechanical Turk[3] and Prolific[4].

Annotation Guidelines: Given the cultural and linguistic breadth of POLAR, we developed detailed, multilingual annotation guidelines (see Appendix E) in English, and then translated and cul-

[2] https://labelstud.io
[3] https://www.mturk.com
[4] https://www.prolific.com

turally 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 or rhetorical tactics (stereotyping/deindividuation, vilification, dehumanization, extreme language, lack of empathy, invalidation).

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

3.4 Annotators' Reliability

To evaluate annotation quality, we report Fleiss' Kappa as the inter-annotator agreement (IAA) metric. As shown in Table 1, the IAA scores vary between languages, with the majority showing moderate agreement and a few, such as "khm" and "tel" achieving good agreement. Although guidelines were standardized, their interpretation was influenced by cultural and political context, especially in languages with lower agreement, where some terms may not have direct equivalents across cultures. Latent content or sarcasm often required annotators to draw on their own socio-political knowledge, highlighting the perspectivist nature of polarization (Cabitza et al., 2023). Thus, low agreement can indicate socio-pragmatic complexity rather than error, signaling that polarization markers may not have universal meanings and that divergences can reveal inherent ambiguity in stimuli or interpretation (Aroyo and Welty, 2015). Examples illustrating such ambiguities are provided below.

Unanimous polarized: *"However I got tanned in Bodrum and I don't wait in line at the hospital."*
This sentence was rated as polarized by all three annotators with high confidence. Here, the sentence refers to perceived privileges granted to the immigrant community.

Non-unanimous polarized: *"Do you agree with the idea that new citizens in our country shouldn't be allowed to vote for ten years?"* This sentence did not receive consensus among annotators, leaning toward being polarized. The topic mentioned here is more complex and sensitive compared to the previous example.

3.5 Dataset Statistics

A comprehensive analysis of the dataset was conducted using the annotated labels to support systematic examination. Table 1 provides a quantitative breakdown of positive labels for each annotation task. For further details on data sources, language-wise composition, and polarization statistics, including distributions of polarized instances, types, and manifestations, see Appendix B (Tables 6–7 and Figure 4). Sample data is shown in Table 9, 10, and 11.

4 Experimentation and Results

4.1 Experimental Setup

To evaluate POLAR, we conducted baseline experiments on three polarization detection tasks: (1) classifying texts as polarized or not, (2) identifying polarization types, and (3) detecting polarization manifestations. For data splitting, we used 70% for training, 10% for validation, and 20% for testing, as summarized in Table 7. All experiments were conducted using the EncouRAGe framework proposed by Strich et al. (2025), which provides a standardized evaluation protocol for language model benchmarking. We benchmarked SLMs and LLMs. The list of evaluated models are listed below and Appendix C:

- **Fine-tuning SLMs:** We fine-tuned six SLMs, including four general-purpose models: mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), RemBERT (Chung et al., 2021), and LaBSE (Feng et al., 2022). In addition, we evaluated two models build on social media and multilingual training corpus: twitter-roberta-hate (Antypas and Camacho-Collados, 2023), a RoBERTa-based encoder specialized for hate-speech detection, and AfroXLMR-large-76L (Adelani et al., 2023), an XLM-R variant optimised for African and other low-resource languages.

#	Lang.	POLARDETECT						POLARTYPE						POLARMANIFEST					
		mBERT	XLM-R	RemBERT	LaBSE	twitter-roberta-hate	afro-xlmr-large	mBERT	XLM-R	RemBERT	LaBSE	twitter-roberta-hate	afro-xlmr-large	mBERT	XLM-R	RemBERT	LaBSE	twitter-roberta-hate	afro-xlmr-large
1	eng	74.83	76.34	77.51	77.27	79.59	75.54	31.29	27.33	43.17	34.68	36.99	19.09	41.83	44.50	47.90	46.58	49.04	47.82
	deu	65.69	69.03	69.81	68.57	65.97	59.14	49.34	48.89	54.23	53.26	49.39	51.34	43.74	45.24	47.46	47.13	44.28	45.18
	urd	71.60	68.74	75.62	74.20	63.45	73.51	68.76	72.16	72.96	73.91	71.05	67.16	72.13	72.83	74.75	77.09	72.49	43.00
	ben	79.03	83.68	82.74	82.50	36.81	83.04	24.61	16.59	10.14	26.82	10.14	13.62	21.08	24.12	5.05	25.83	9.59	5.64
	hin	76.37	75.06	45.98	78.23	61.80	65.49	61.40	63.58	73.30	69.97	38.83	57.87	53.71	69.12	72.69	70.13	63.58	62.07
	ori	41.72	41.72	71.14	77.60	42.42	68.73	19.44	28.14	32.92	42.27	19.23	29.47	8.37	4.84	2.78	24.00	5.77	9.47
	nep	85.04	84.04	87.48	88.26	69.86	76.09	45.75	48.78	70.53	66.42	32.69	49.16	55.15	53.92	58.11	60.19	39.19	35.55
	pan	69.98	66.50	73.57	74.91	61.03	62.28	29.59	28.22	40.88	38.73	27.32	29.08	42.07	40.07	45.80	48.35	38.57	43.31
	spa	70.36	71.27	75.24	74.99	71.96	74.24	58.22	54.54	65.01	59.34	57.78	58.64	41.17	43.34	50.63	48.04	43.76	46.45
	ita	54.35	34.50	60.93	56.40	53.00	47.78	21.48	21.83	25.99	26.13	21.08	23.47	–	–	–	–	–	–
	rus	70.20	71.61	77.39	74.77	42.33	74.80	31.97	21.93	47.33	40.94	15.81	36.44	–	–	–	–	–	–
	pol	74.45	74.54	77.96	77.26	68.45	67.62	31.63	31.29	46.84	41.55	24.65	24.48	–	–	–	–	–	–
	fas	80.09	78.34	77.50	83.50	54.64	80.89	46.69	47.21	54.18	52.46	27.72	27.75	35.53	34.42	41.35	39.51	24.72	34.78
2	hau	82.97	47.19	75.91	82.09	78.32	81.82	21.44	15.75	15.53	21.58	16.68	2.76	19.97	16.43	21.44	20.55	18.29	22.16
	arb	78.24	78.43	81.12	81.23	68.37	82.20	49.49	50.43	58.89	55.89	33.40	56.16	49.94	53.05	55.53	56.76	43.21	56.30
	amh	42.45	68.83	72.60	76.43	49.46	58.28	29.32	39.48	24.27	47.09	28.19	28.43	41.21	26.94	42.47	51.16	41.28	45.75
3	zho	83.96	86.45	87.08	86.44	72.49	85.83	57.39	47.83	69.35	63.12	34.05	45.17	40.50	37.91	39.09	46.10	27.72	38.87
	mya	84.10	82.44	84.33	86.07	70.35	82.31	42.62	34.54	25.55	55.06	26.05	42.52	–	–	–	–	–	–
4	khm	47.58	47.58	76.41	73.70	47.58	68.38	23.06	50.12	65.79	58.60	15.89	17.20	14.54	29.77	35.60	34.28	14.57	14.26
5	swa	76.54	76.02	78.17	79.04	75.93	78.77	36.39	33.20	43.35	40.18	35.63	38.46	55.69	54.32	56.21	56.51	53.79	54.82
6	tel	84.86	79.63	88.93	88.93	65.16	78.62	41.16	32.36	43.68	42.56	30.77	36.73	37.87	21.55	39.79	39.24	29.65	35.83
7	tur	69.80	74.47	70.94	75.02	65.42	73.06	36.55	43.92	51.03	48.41	30.76	31.33	36.92	44.83	44.48	44.92	36.83	44.36

Table 2: Average macro-F1(%) scores for **POLARDETECT**, **POLARTYPE**, and **POLARDETECT** across languages and multilingual encoders. The best and second performance scores are highlighted in blue and orange respectively.

- **Evaluating LLMs in Zero- and Few-shot Settings:** We evaluated large language models in zero- and few-shot settings, including Qwen2.5-7B-Instruct, Qwen-3-8B, LLaMA-3.1-8B-Instruct, Ministral-3-14B-Instruct-2512, Gemma-3-27B-IT, GPT-4.1-Nano, and GPT-OSS-120B. For brevity, we refer to them as Qwen2.5, Qwen3, LLaMA3.1, Mistral3, Gemma3, GPT4.1 and GPT-OSS respectively. The exact models used are stated in the Appendix section C. The prompts used for LLM zero-shot and few-shot settings are shown in Appendix F.

4.2 Results and Analysis

SLMs Models: Table 2 presents the results of six small language models. Overall, RemBERT and LaBSE show comparable performance across most languages, achieving the best or the second best macro-F1 scores. RemBERT is designed to balance representation across high-, mid-, and low-resource languages, while LaBSE relies on bilingual sentence-level alignment between English and low-resource languages. These training strategies enhance their ability to mid- and low-resource language understanding, which is reflected in their consistent improvements over mBERT and XLM-R, particularly for languages such as Amharic, Odia, Italian.

The twitter-roberta-hate model is finetuned on English Twitter dataset covering emoji, stance, hate speech, and emotion. This training boosts its performance on English polarization detection, polarization types classification and polarization manifestation recognition. AfroXLMR-large-76L is finetuned upon XLM-R on African languages datasets. As a result, it performs well on African language in our dataset, including Arabic, Hausa, and Swahili.

Polarization detection is comparatively easier for multilingual BERT-based models, which achieve relatively high macro-F1 scores. In contrast, recognizing polarization types (politics, gender, racial, religious, and others) is substantially more challenging. Performance drops even further for polarization manifestation recognition (stereotype, vilification, dehumanization, extreme language, lack of empathy, invalidation), where macro-F1 scores decrease markedly. This gap highlights the limitations of current models in capturing fine-grained and implicit polarization manifestations and the importance of POLAR for advancing research on nuanced polarization understanding.

LLMs Performance: The Tables 2, 3, and 4 present an overall picture of model performance in polarization detection, types, and manifestations across the languages, using the Macro-F1 metric as a measure of accuracy. Across all experiment setups, models consistently achieve their highest scores in Polarization Detection, followed by the classification of Polarization Types, while identifying Polarization Manifestations remains the most challenging task for both encoder models and LLMs likely due to the latent nature of how polarization is expressed, requiring semantic understanding that state-of-the-art models find difficult to generalize across diverse cultural contexts.

Table 2 presents the the performance of encoders (mBERT, XLM-R, RemBERT, LaBSE, twitter-roberta-hate, and afro-xlmr-large), demonstrating

#	Lang.	PolarDetect							PolarType							PolarManifest						
		Qwen2.5	Qwen3	LLaMA3.1	Mistral3	Gemma3	GPT4.1	GPT-OSS	Qwen2.5	Qwen3	LLaMA3.1	Mistral3	Gemma3	GPT4.1	GPT-OSS	Qwen2.5	Qwen3	LLaMA3.1	Mistral3	Gemma3	GPT4.1	GPT-OSS
1	eng	71.19	69.42	69.17	70.67	79.84	77.19	77.43	72.45	66.76	69.08	73.79	74.37	69.59	77.73	63.33	62.38	60.62	63.91	64.18	63.39	68.33
	deu	60.02	58.76	61.53	61.34	73.86	70.89	69.72	66.61	61.26	64.17	67.31	71.71	63.63	70.04	57.74	60.10	56.82	58.71	57.93	59.41	65.18
	urd	40.97	48.46	58.55	51.26	75.88	72.02	70.42	42.03	44.24	49.27	45.11	50.65	46.92	49.32	42.20	50.32	52.14	54.74	53.22	48.44	52.46
	ben	65.62	59.75	68.90	69.83	80.85	76.15	80.28	68.96	66.10	65.71	71.05	74.04	72.64	77.13	58.47	54.54	51.23	50.47	51.89	56.99	60.34
	hin	43.94	36.53	48.52	45.67	67.13	59.04	58.96	65.32	54.48	59.96	68.80	75.12	69.73	71.68	48.09	53.62	55.74	60.50	58.43	53.04	58.53
	ori	53.35	50.14	49.30	42.38	65.03	58.39	67.20	65.87	64.74	65.11	48.61	75.81	66.68	71.29	55.64	55.22	54.27	49.76	59.46	57.88	61.66
	nep	49.03	53.48	65.01	61.84	85.28	74.16	82.68	69.31	67.29	65.09	73.78	80.95	72.30	82.29	62.67	67.08	61.88	70.01	73.34	69.59	76.30
	pan	40.69	40.42	49.80	41.64	71.95	71.67	65.92	56.51	54.20	56.98	54.21	68.20	61.73	66.19	51.10	54.90	53.20	55.22	56.73	58.48	61.94
	spa	66.64	68.25	70.42	70.04	71.43	73.64	76.91	67.69	66.07	61.82	69.05	69.81	65.93	72.18	60.00	59.86	57.46	57.30	59.84	60.34	66.36
	ita	66.24	63.20	67.52	67.36	73.72	72.08	74.60	68.48	64.68	63.59	70.19	73.27	68.47	74.70	–	–	–	–	–	–	–
	rus	73.77	71.01	71.12	72.88	70.57	72.85	74.36	70.26	66.33	57.78	73.43	69.88	68.47	77.94	–	–	–	–	–	–	–
	pol	57.15	58.33	63.46	57.25	68.88	65.42	76.94	70.56	62.91	65.26	68.87	75.88	70.42	79.46	–	–	–	–	–	–	–
	fas	32.17	28.44	40.10	29.71	54.79	48.71	51.61	59.53	52.96	63.23	54.93	71.17	60.74	66.33	58.62	55.68	56.52	55.40	56.63	58.98	62.93
2	hau	51.40	49.25	55.37	50.44	46.36	52.18	55.02	55.41	53.65	50.69	54.59	54.04	56.84	59.03	51.01	47.28	43.09	43.56	43.66	48.43	48.33
	arb	65.22	58.82	70.02	65.47	80.70	77.72	78.87	63.22	62.77	62.62	66.17	71.22	64.39	72.25	62.74	64.57	62.64	67.04	67.97	65.02	73.93
	amh	25.16	26.97	39.14	23.50	74.47	50.58	57.00	52.85	51.55	61.67	47.26	77.01	61.20	67.31	50.78	53.99	56.51	53.43	59.77	55.98	61.52
3	zho	58.20	59.70	70.57	62.62	82.48	78.51	80.52	77.38	73.32	68.54	75.32	80.63	74.76	81.15	69.99	67.72	58.16	64.35	65.82	70.44	77.90
	mya	40.63	48.31	51.68	37.72	78.40	58.41	59.96	52.65	55.76	58.48	49.23	67.78	58.09	59.48	–	–	–	–	–	–	–
4	khm	10.56	8.94	11.04	8.65	13.32	11.25	12.53	50.51	46.30	57.32	45.32	55.21	48.71	49.43	47.96	50.04	50.14	49.28	49.30	48.66	48.50
5	swa	40.50	46.76	57.42	52.29	65.64	62.00	66.15	61.31	57.52	61.63	62.87	69.49	66.34	68.81	52.94	54.04	54.63	55.11	56.89	56.22	60.41
6	tel	42.99	44.23	43.02	36.82	42.97	45.23	43.51	52.81	53.88	55.82	52.24	57.88	52.70	50.95	50.78	54.96	54.08	55.68	55.29	56.59	52.13
7	tur	64.11	58.86	66.68	66.23	76.69	74.54	78.97	67.43	63.41	62.19	70.38	75.41	70.27	77.49	60.38	61.54	60.84	62.94	65.08	62.73	71.20

Table 3: F1-Macro resulting from the zero-shot LLM experiments with the POLAR dataset. The highest value per language is highlighted in blue.

significant stability across languages, especially in the **PolarDetect** task. On the other hand, Table 3 presents a more mixed picture through zero-shot LLM experiments. While high-resource languages like English (eng) and Chinese (zho) maintain high detection scores (reaching 79.84% and 82.48% respectively), the models exhibit a significant drop in performance for languages with unique scripts or less resources. For example, Khmer (khm) detection scores drop to a range between 8.65% and 13.32%, indicating that without prior examples, these generative models may not generalize. Table 4 demonstrates the transformative impact of few-shot prompting, where providing the LLMs with a few examples significantly improves performance for many languages. For instance, Urdu (urd) has the detection score rise from 40.97% in the zero-shot setting to 73.81% with GPT4.1 in the few-shot setting. In this few-shot setting, LLMs also outperform BERT-family encoders in the complex **PolarManifest** task, for example, with Arabic (arb) reaching a peak of 75.49%. In summary, while BERT-family models remain the most efficient for binary detection, LLMs with few-shot prompting show potential for handling more complex, fine-grained classification tasks when sufficient examples are provided.

Finally, some of the linguistic families exhibit clustered behaviors. For instance, a language cluster emerges among South Asian languages like Bengali (ben), Hindi (hin), Nepali (nep), and Oriya (ori), as these languages exhibit similar performance improvements in the few-shot LLM experiments, compared to zero-shot settings (Table 4). Nevertheless, the most dominant predictor of model performance seems to be resource-tier of the language, which often aligns with geographic and economic regions.

5 Discussion

Performance within the same language family: It is not necessary that models uniformly achieve a high performance across all languages within the same language family. For example, while a model may perform strongly on Chinese, its performance on Burmese still lags behind, demonstrating that high-resource language performance does not guarantee transferability to related lower-resource languages such as Burmese.

Few-shot vs. Zero-shot: Few-shot performance is not always superior to zero-shot results. For larger models, such as Gemma3 and GPT-OSS, the few-shot setting underperforms zero-shot, indicating that large LLMs are inherently capable of recognizing polarized sentences without requiring in-context exemplars. In contrast, for smaller models, including Qwen2.5, Qwen3, LLaMA3.1, and Mistral3, in-context examples (3-shot prompting) can improve performance, showing that smaller LLMs benefit more from in-context learning.

LLMs vs. SLMs: SLMs demonstrate improved performance in detecting polarization after fine-tuning, but still struggle to distinguish polarization types and manifestations, reflecting limited semantic knowledge of domain-specific concepts. Conversely, LLMs generally exhibit a stronger understanding of social-science polarization constructs, enabling better recognition of polarization types and manifestations, yet they remain weaker in direct polarization classification tasks. This suggests that LLMs possess richer implicit knowledge of social-science constructs, whereas SLMs currently lack comparable semantic grounding for fine-

#	Lang.	PolarDetect							PolarType							PolarManifest						
		Qwen2.5	Qwen3	LLaMA3.1	Mistral3	Gemma3	GPT4.1	GPT-OSS	Qwen2.5	Qwen3	LLaMA3.1	Mistral3	Gemma3	GPT4.1	GPT-OSS	Qwen2.5	Qwen3	LLaMA3.1	Mistral3	Gemma3	GPT4.1	GPT-OSS
1	eng	77.98	77.79	74.90	75.82	79.82	76.81	79.74	72.17	73.52	70.48	77.33	76.96	74.01	79.24	62.37	63.40	61.56	63.41	64.35	63.77	69.40
	deu	65.49	65.79	65.93	64.34	71.26	71.66	69.61	65.03	67.75	68.37	69.58	72.44	68.31	69.61	58.36	60.76	55.97	60.59	53.72	57.59	65.03
	urd	57.89	55.42	55.48	43.96	62.62	73.81	66.26	40.80	49.04	47.02	45.40	51.85	48.14	49.39	43.73	57.34	54.55	59.44	58.42	49.75	56.67
	ben	66.84	63.00	59.91	65.61	79.64	78.89	78.32	70.14	72.51	71.14	73.37	75.11	71.12	76.71	57.01	53.29	51.41	53.12	53.72	58.90	60.36
	hin	52.92	43.45	41.42	52.13	58.54	62.61	56.26	59.85	54.86	55.54	69.68	71.31	66.72	70.99	47.77	51.70	54.76	59.41	58.65	54.29	57.49
	ori	60.85	57.23	60.81	43.60	63.62	65.71	66.09	63.98	66.26	67.26	50.38	75.68	67.56	73.38	54.71	56.56	53.24	48.69	59.73	57.57	62.57
	nep	64.38	74.27	64.25	69.50	84.64	82.44	81.20	68.49	73.98	71.72	77.16	81.57	74.45	82.38	59.81	64.29	60.89	68.35	70.57	72.03	76.92
	pan	51.19	60.47	56.38	52.95	69.77	72.07	69.76	61.04	64.38	62.86	64.38	69.43	66.76	70.57	50.94	53.07	53.95	55.51	57.22	58.66	62.82
	spa	68.01	70.43	67.35	66.91	72.15	74.05	74.61	68.46	68.96	64.16	70.36	71.67	68.65	73.39	59.49	58.69	58.10	59.80	60.00	62.54	66.29
	ita	59.37	55.83	48.57	51.98	65.92	71.82	72.02	60.67	63.55	63.15	64.49	70.46	68.51	72.98	–	–	–	–	–	–	–
	rus	74.96	73.70	62.74	70.59	74.32	74.04	77.45	69.78	70.74	63.57	74.10	73.97	71.19	78.71	–	–	–	–	–	–	–
	pol	58.07	62.86	53.14	59.10	68.38	71.54	79.17	70.56	66.92	69.71	71.19	76.91	72.50	79.79	–	–	–	–	–	–	–
	fas	40.77	41.32	40.45	39.32	54.58	54.22	50.53	57.14	59.51	66.76	59.26	69.72	62.67	65.54	58.08	55.41	54.25	56.35	57.53	59.24	62.47
2	hau	55.01	53.92	55.60	50.47	56.24	55.08	59.28	57.89	55.73	55.49	54.61	57.79	56.69	60.64	50.80	49.01	47.69	46.19	45.98	50.87	49.94
	arb	75.54	72.88	73.42	72.21	79.58	78.37	78.40	64.36	65.53	66.16	70.07	71.44	70.50	72.89	64.89	66.55	61.96	67.07	67.6	67.58	75.49
	amh	38.51	35.70	37.66	28.59	74.58	51.81	54.67	50.41	59.07	64.44	55.74	76.65	59.64	66.07	50.90	55.80	51.50	53.36	58.72	52.91	61.37
3	zho	56.69	68.17	56.79	66.71	69.81	78.22	74.57	71.13	78.00	72.03	76.32	76.87	79.12	77.00	69.97	67.39	61.25	62.78	63.98	72.01	76.47
	mya	62.17	71.12	60.51	50.96	73.90	68.46	63.07	56.23	63.35	68.11	58.24	71.63	65.54	63.89	–	–	–	–	–	–	–
4	khm	24.70	13.66	15.75	12.98	18.90	16.52	15.69	53.99	51.21	62.92	51.81	57.24	51.99	51.41	49.45	50.55	49.55	52.19	50.47	48.27	50.11
5	swa	53.67	58.43	58.75	59.20	67.70	64.56	66.61	61.14	65.17	65.74	66.71	70.80	67.59	69.81	56.27	54.88	53.13	53.58	56.04	57.62	61.13
6	tel	58.52	56.59	53.40	54.56	60.02	59.74	52.15	52.15	55.13	56.16	56.64	57.24	56.24	53.92	53.79	55.79	55.41	54.61	59.03	56.71	52.53
7	tur	72.29	68.12	66.23	68.70	79.62	76.45	78.28	64.02	66.43	62.74	71.08	76.03	69.09	77.67	60.74	58.18	60.60	64.70	64.83	64.14	72.87

Table 4: F1-Macro scores(in %) from the few-shot LLM experiments with the POLAR dataset. The highest value per language is highlighted in blue.

grained polarization distinctions.

6 Error Analysis: Misclassification Cases

Based on our error analysis (See detail in Appendix G Table 8), we identified a misalignment between the model's classification logic and human judgment of polarization. The model relies on a deterministic, surface-level heuristic: it classifies text as polarized only when it detects explicitly named and opposed groups within the sentence (e.g., "Israeli forces vs. Palestinians"). Conversely, if the text expresses hostility but names only one group or relies on implied opposition, the model defaults to labeling it as non-polarized.This error stems from models' reliance on textual pattern alone, while human annotators draw upon cultural and contextual knowledge to interpret hostility and implicit group conflict.

7 Implications

Theoretical Implications: Our dataset highlights the complexity of online polarization, emphasizing its deep cultural and contextual nature. It reveals the current limits of NLP models in detecting implicit rhetorical tactics, underscoring the need for culturally-aware frameworks.

Practical Implications: The dataset provides a valuable benchmark for developing and evaluating models capable of detecting nuanced forms of online polarization across multiple languages and contexts. It supports the creation of more culturally sensitive and robust tools for the monitoring and mitigation efforts of online discourse.

Methodological Implications: Our multi-label, multi-platform annotation approach underscores the importance of culturally sensitive and detailed labeling strategies. The variability of model performance across languages and contexts indicates a pressing need for methods that integrate cultural signals, multimodal data, and contextual embeddings to improve robustness and reduce performance disparities in social NLP applications.

8 Conclusion

In this study, we introduced POLAR, a comprehensive, multilingual, and multi-event dataset designed to advance the understanding and detection of online polarization across diverse linguistic and cultural contexts. By annotating over 110,000 instances along three critical dimensions, we created a nuanced, fine-grained resource. This dataset captures the complex rhetorical tactics and social dimensions that underpin polarized discourse. Our extensive benchmarking of SOTA SLMs and LLMs reveals that while current models are reasonably effective at binary polarization detection, they face significant challenges in accurately identifying polarization types and rhetorical manifestations, especially in low-resource and culturally nuanced settings. These findings emphasize the deep contextual and implicit nature of online polarization and highlight the limitations of existing NLP approaches. Importantly, our work underscores the critical need for culturally aware, adaptable, and context-sensitive models to effectively monitor and mitigate digital polarization globally. The resources and benchmarks provided herein aim to catalyze future research, fostering the development of more inclusive and robust tools for analyzing social phenomena in multilingual and multicultural online environments.

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

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. *Preprint*, arXiv:2309.07445.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

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.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Zewen Bai, Liang Yang, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, and Hongfei Lin. 2025. STATE ToxiCN: A Benchmark for Span-level Target-Aware Toxicity Extraction in Chinese Hate Speech Detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10206–10219, Vienna, Austria. Association for Computational Linguistics.

Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia*, 2(3):315–324.

Pablo Barberá. 2020. *Social Media and Democracy: The State of the Field, Prospects for Reform*, chapter Social Media, Echo Chambers, and Political Polarization. Cambridge University Press Cambridge.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37-6, pages 6860–6868, Washington, DC, USA.

Fernando Casal Bértoa and José Rama. 2021. Polarization: What Do We Know and What Can We Do About It? *Frontiers in Political Science*, 3:1–11.

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.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

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.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. 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 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A Benchmark for Chinese Offensive Language Detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

William Donohue and Mark Hamilton. 2022. A Framework for Understanding Polarizing Language. In *The Routledge Handbook of Language and Persuasion*, pages 207–223. Routledge.

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.

Valentin Hofmann, Xiaowen Dong, Janet Pierrehumbert, and Hinrich Schuetze. 2022. Modeling Ideological Salience and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 536–550, Seattle, United States. Association for Computational Linguistics.

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.

Anna Kolos, Inez Okulska, Kinga Głąbińska, Agnieszka Karlinska, Emilia Wisnios, Paweł Ellerik, and Andrzej Prałat. 2024. BAN-PL: A Polish Dataset of Banned Harmful and Offensive Content from Wykop.pl Web Service. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2107–2118, Torino, Italia. ELRA and ICCL.

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.

Nang Aeindray Kyaw, Ye Kyaw Thu, Thazin Myint Oo, Hutchatai Chanlekha, Manabu Okumura, and Thepchai Supnithi. 2024. Enhancing Hate Speech Classification in Myanmar Language through Lexicon-Based Filtering. In *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 316–323, Phuket, Thailand.

Raquel Martínez-España, Julio Fernández-Pedauye, José Giner-Pérez de Lucía, Jose Miguel Rojo-Martínez, Kaoutar Bakdid-Albane, and Juan José García-Escribano. 2024. Methodology for Measuring Individual Affective Polarization Using Sentiment Analysis in Social Networks. *IEEE Access*.

Binny Mathew, Punyajoy Saha, Seid Muhie 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 2021*, volume 35, pages 14867–14875, online.

Debora Nozza, Andrea Teresa Cignarella, Greta Damo, Tommaso Caselli, and Viviana Patti. 2023. HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian. In *Proceedings of the EVALITA 2023 Evaluation Campaign*, pages 1–8, Parma, Italy.

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.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the*

2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

James A. Piazza. 2023. Political polarization and political violence. *Security Studies*, 32(3):476–504.

José Miguel Rojo Martínez. 2025. *Unravelling Radicalisation: Exploring Concepts, Contexts, and Perspectives*, chapter Affective polarisation, Dehumanisation of the Adversary and Political Violence. Springer.

Manuela Sanguinetti, Gloria Comandini, Elisa di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tomaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*, pages 1–9, online.

Almog Simchon, William J Brady, and Jay J Van Bavel. 2022. Troll and divide: the language of online polarization. *PNAS nexus*, 1(1):pgac019.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022. Political Ideology and Polarization: A Multi-dimensional Approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 231–243, Seattle, United States. Association for Computational Linguistics.

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.

Jan Strich, Adeline Scharfenberg, Chris Biemann, and Martin Semmann. 2025. EncouRAGe: Evaluating RAG Local, Fast, and Reliable. *Preprint*, arXiv:2511.04696.

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

A Language and its language family

The POLAR dataset covers 22 languages from seven linguistic families. The coverage language and their corresponding branches is presented below:

#	Lang.	ISO-639	Language Family	Sub-branch
1	English	eng	Indo-European	Germanic
	German	deu	Indo-European	Germanic
	Urdu	urd	Indo-European	Indo-Aryan
	Bengali	ben	Indo-European	Indo-Aryan
	Hindi	hin	Indo-European	Indo-Aryan
	Odia	ori	Indo-European	Indo-Aryan
	Nepali	nep	Indo-European	Indo-Aryan
	Punjabi	pan	Indo-European	Indo-Aryan
	Spanish	spa	Indo-European	Romance
	Italian	ita	Indo-European	Romance
	Russian	rus	Indo-European	Slavic
	Polish	pol	Indo-European	Slavic
	Persian	fas	Indo-European	Iranian
2	Hausa	hau	Afro-Asiatic	Chadic
	Arabic	arb	Afro-Asiatic	Semitic
	Amharic	amh	Afro-Asiatic	Semitic
3	Chinese	zho	Sino-Tibetan	Sinitic
	Burmese	mya	Sino-Tibetan	Tibeto-Burman
4	Khmer	khm	Austroasiatic	Mon-Khmer
5	Swahili	swa	Niger–Congo	Bantu
6	Telugu	tel	Dravidian	Dravidian
7	Turkish	tur	Turkic	Turkic

Table 5: Language covered and its language families

B Data Statistics

Table 6 summarizes the distribution of instances across data sources in the POLAR dataset. The data primarily originate from major social media platforms, with Twitter contributing over half of the instances. Additional data are drawn from news websites, online forums, and other social platforms to ensure coverage of diverse discourse contexts. A smaller portion of the data comes from existing datasets that were re-annotated to align with our polarization framework.

Source	No. of Instances	%
X (Twitter)	58,984	53.3%
News/websites	15,208	13.7%
Bluesky	8,099	7.3%
YouTube	7,159	6.5%
Reddit	5,705	5.2%
Existing Dataset	4,794	4.3%
Weibo	4,550	4.1%
Facebook	2,928	2.6%
Zhihu	1,088	1.0%
Threads	852	0.8%
Tieba	783	0.7%
Wikipedia	500	0.5%
Total	**110,650**	**100%**

Table 6: Data sources of the POLAR dataset.

B.1 Dataset Composition by Language

Table 7 presents a language-wise overview of the POLAR dataset, including data sources, targeted events or topics, and inter-annotator agreement. For each language, data collection was tailored to platform availability and sociopolitical relevance, resulting in variation in event focus and discourse type. Inter-annotator agreement is reported primarily using Fleiss's kappa, with alternative reliability measures noted where applicable.

B.2 Polarization Statistics

Figure 3 shows the distribution of polarization types, and Figure 4 illustrates the distribution of polarization manifestations across languages.

C SLMs and LLMs Used

C.1 Multilingual Encoders

- https://huggingface.co/google-bert/bert-base-multilingual-cased
- https://huggingface.co/FacebookAI/xlm-roberta-base
- https://huggingface.co/google/rembert
- https://huggingface.co/sentence-transformers/LaBSE
- https://huggingface.co/cardiffnlp/twitter-roberta-base-hate
- https://huggingface.co/Davlan/afro-xlmr-large-76L

C.2 LLMs

- https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
- https://huggingface.co/Qwen/Qwen3-8B
- https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
- https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512
- https://huggingface.co/google/gemma-3-27b-it
- https://huggingface.co/openai/gpt-oss-120b

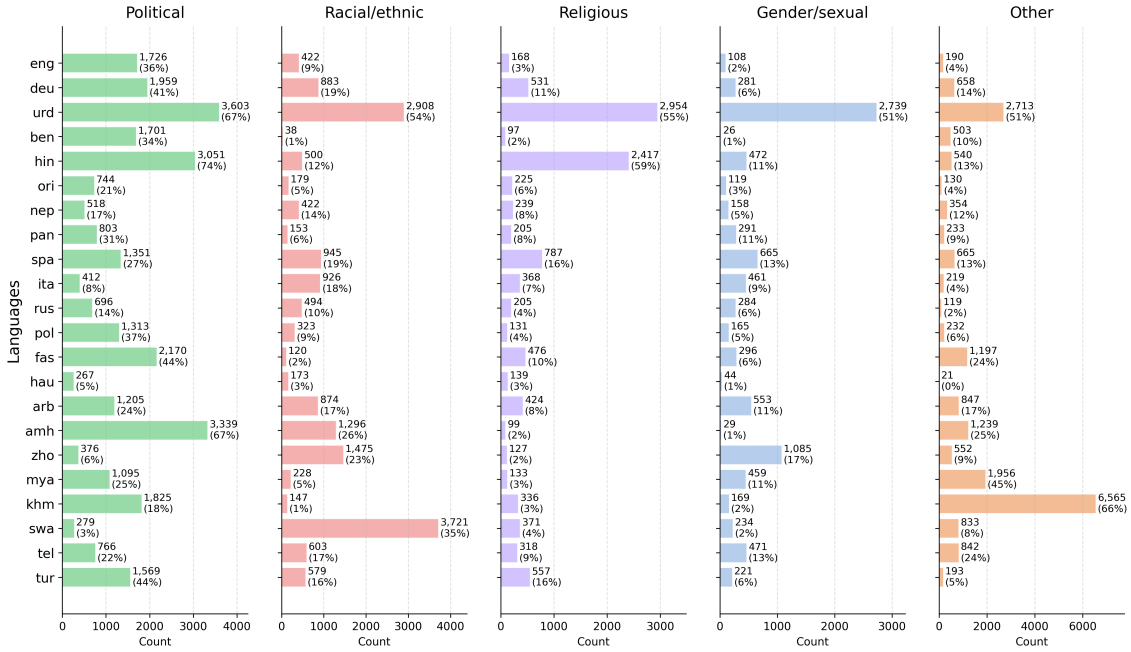

Figure 3: **POLARTYPE** by Languages, for each language, the numeric value shows the count of instances assigned to a given polarization type, while the percentage reflects its share of the total annotated instances for that language.

Language	Data Source(s)	Events/Topics Focused
Amharic (amh)	Facebook, X (Twitter)	The Tigray War
Arabic (arb)	Existing datasets, Facebook, News, Reddit, Threads,X (Twitter)	Social issues regading politics and religion
Bengali (ben)	YouTube comments	Social discourse, generic contemporary topics
Burmese (mya)	Existing datasets (Kyaw et al., 2024), Wikipedia	Social issues regading politics, ethnicity and popular culture
Chinese (zho)	Tieba, Weibo, Zhihu	Social issues regarding racism, sexuality/gender and religious discrimination
English (eng)	Bluesky, Local news, X (Twitter)	US elections and international conflicts
German (deu)	Bluesky, Reddit, X (Twitter)	COVID-19, and contemporary social issues
Hausa (hau)	Facebook, X (Twitter)	Social issues regading politics, ethnicity and religion
Hindi (hin)	Bluesky, Reddit, X (Twitter)	Social issues regarding politics, religion, and caste
Italian (ita)	YouTube, X (Twitter)	Pride parade, immigration crisis , crime news, Italian justice reform
Khmer (khm)	Facebook, Specialised websites, Wikipedia, YouTube, Local news	COVID-19, and contemporary social issues
Nepali (nep)	Facebook, Local news, X (Twitter)	Social discourse, generic contemporary topics
Odia (ori)	Bluesky, Local news, X (Twitter)	Social discourse, generic contemporary topics
Persian (fas)	Bluesky, X (Twitter)	Social discourse, generic contemporary topics
Polish (pol)	Bluesky, Existing dataset (Kolos et al., 2024)	COVID-19, and contemporary social issues
Punjabi (pan)	Existing dataset (pnbTenTen) https://www.sketchengine.eu/pnbtenten-western-punjabi-corpus , Youtube Comments, X (Twitter)	Social issues regarding politics, religion, and caste
Russian (rus)	Bluesky, X (Twitter), COVID-19 chatter dataset (Banda et al., 2021)	Social discourse, generic contemporary topics
Spanish (spa)	Bluesky, X (Twitter)	2010's inmigration movement, "Salvemos las dos vidas" movement, social issues regarding politics and gender inequality
Swahili (swa)	X (Twitter)	Kenyan elections
Telugu (tel)	Facebook, Reddit, X (Twitter)	Social discourse, generic contemporary topics
Turkish (tur)	X (Twitter), Existing dataset (Çöltekin, 2020)	Social discourse, generic contemporary topics
Urdu (urd)	X (Twitter)	Social discourse, generic contemporary topics

Table 7: Summary of dataset composition regarding data sources and events or topics focused.

D Experiment Settings

D.1 Experiments settings

For SLMs, we performed language-specific fine-tuning for 3 epochs using a learning rate of 2e-4.

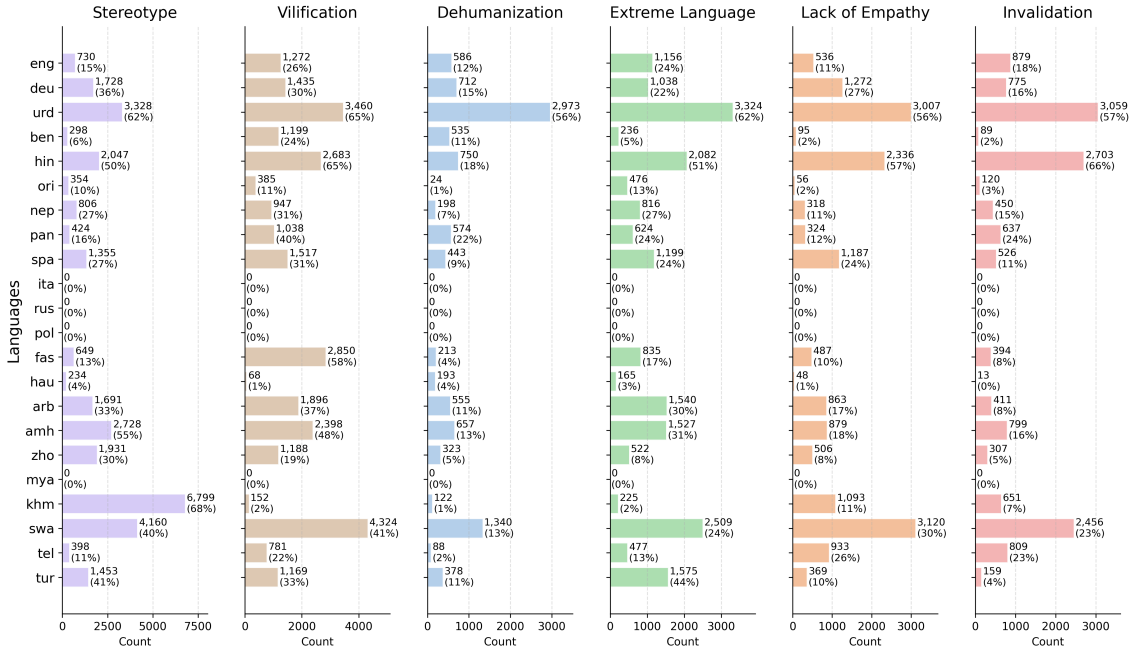

Figure 4: **PolarManifest**-by Languages. For each language, the numeric value shows the count of instances assigned to a given polarization manifestation, while the percentage reflects its share of the total annotated instances for that language.

For LLMs in the few-shot setting , we use three shots. ALL prompts are written in English, while the in-context examples are provided in the target language. The full prompts is reported in Appendix F.

All experiments were conducted on local GPU and implemented with EncouRAGe (Strich et al., 2025) framework to execute experiments and measure results.

E Annotation Guidelines

Annotation Guidelines

To minimize potential misunderstandings, we offer annotation guidelines customized for each event, accompanied by event-specific examples. These guidelines are in the languages of the target datsets. The English version is provided below. This guideline aims to assess whether social media messages

reflect **attitude polarization** and to categorize the various types and manifestations of this polarization. The dataset consists of content sourced from platforms such as Facebook and Twitter.

Warning to annotators: You may encounter polarized and hateful content during this task. Please take breaks as needed and prioritize your mental well-being.

If you choose to participate, please read the following sections carefully. To exit the study at any time, click "**Cancel**" or "**Quit**" and provide feedback if possible.

Task 1: Polarization Classification (Yes/No)

Polarization refers to increasingly extreme, divided beliefs or behaviors between opposing groups. **Attitude polarization** includes:

- Negative attitudes toward out-groups
- Blind support for in-groups
- Stereotyping, vilification, dehumanization, or intolerance

Texts should be labeled:

- **Yes** – if the message clearly reflects attitude polarization
- **No** – if it does not show any polarization indicators

Note: Always consider the overall context and meaning, not just individual words.

Task 2: Type of Polarization (if "Yes" in Task 1)

If the message is polarized, select all applicable types:

2.1 **Political/Ideological Polarization** – Conflict between political parties or ideologies
2.2 **Racial/Ethnic Polarization** – Division based on race or ethnicity
2.3 **Religious Polarization** – Conflict based on religious beliefs
2.4 **Gender/Sexual Polarization** – Discrimination based on gender or sexual orientation
2.5 **Other** – e.g., based on economic class, technology, media, etc.

Note: Select all relevant options if the text contains more than one of the above.

Task 3: Manifestations of Polarization (if "Yes" in Task 1)

Classify the tactics used to express polarization in the message. A text is considered polarizing if it includes one or more of the following:

3.1 **Stereotyping** – Generalizing traits to all group members *Example: "Women are weak."*
3.2 **Vilification** – Defaming or demonizing a group *Example: "Migrants are traitors."*
3.3 **Dehumanization** – Stripping away human qualities *Example: "They are cockroaches."*
3.4 **Extreme Language or Absolutism** – Using words like "always", "never", or "worst". *Example: "We can never trust them."*
3.5 **Lack of Empathy or Understanding** – Dismissing others' perspectives *Example: "Wearing the hijab is extremist."*
3.6 **Invalidation** – Denying a group's identity or existence *Example: "There is no nation called Palestine."*

Note: Select all relevant options if the text contains more than one of the above.

F Prompts for Text Classification

F.1 Prompts for polarization detection

TASK: Determine whether the following {{lang_full}} text is polarized or not polarized.

DEFINITION: Polarization refers to the process or phenomenon in which opinions, beliefs, or behaviors become more extreme or divided, leading to a greater distance or conflict between differing groups. Attitude polarization is the negative attitude that individuals or groups display towards individuals and groups outside their group while also showing blind support and solidarity towards people within their group.

Polarization denotes stereotyping, vilification, dehumanization, deindividuation, or intolerance of other people's views, beliefs, and identities. Speeches and articles that are shared on social media that incite division, groupism, hatred, conflict, and intolerance are classified as containing polarization. Only texts that clearly reflect attitude polarization should be classified as such.

If a text includes one or more of the specified characteristics, it is classified as polarized. Conversely, social media texts that do not display any of these characteristics are classified as non-polarized. Always consider the context and the overall meaning of the text, not just individual words or phrases.

INSTRUCTIONS:
- Analyze the text carefully.
- Classify the polarization of the given text. Use the given definition of polarization.
- Provide a short reason explaining your decision. The reason should be English.
- Assign a binary label: If the text is polarized, polarization is set to 1. If the text is not polarized, polarization is set to 0.
- Do not include any text or formatting outside the JSON.

OUTPUT FORMAT:
```
{"reason": "the reason of the classification decision", "polarization": 0 or 1}
```

Examples (if applicable):
```
{% if examples %}
{{examples}}
{% endif %}
```

Input Text:
```
{{user_prompt}}
```

F.2 Prompt for polarization type

TASK: Determine whether the following {{lang_full}} text is polarized or not polarized and which type of polarization.
DEFINITION: Polarization refers to the process or phenomenon in which opinions, beliefs, or behaviors become more extreme or divided, leading to a greater distance or conflict between differing groups. Polarization denotes stereotyping, vilification, dehumanization, deindividuation, or intolerance of other people's views, beliefs, and identities. Only texts that clearly reflect attitude polarization should be classified as such. If a text includes one or more of the specified characteristics, it is classified as polarized. Conversely, social media texts that do not display any of these characteristics are classified as non-polarized. Always consider the context and the overall meaning of the text, not just individual words or phrases.
TYPES OF POLARIZATION:
- Political/ideological polarization: Political polarization refers to political beliefs and affiliations becoming more extreme.
- Racial or ethnic polarization: This type of polarization focuses on ethnic identity or racial origin and incites division, intolerance, and conflict between ethnic groups or races.
- Religious polarization: This type of polarization focuses on religious identity and incites division, intolerance, and conflict between religious followers.
- Gender/Sexual polarization: This type of polarization refers to the exclusion, discrimination, and marginalization of individuals based on their gender and sexual orientations within society, often leading to heightened tensions, misunderstandings, or conflicts.
- Other: polarization texts targeting other groups/identities such as economy, technology, media, polarization, etc.

INSTRUCTIONS:
- Analyze the text carefully, and classify the polarization of the given text and if the text is polarized the types of polarization.
- Use the given definition of polarization and their types. Provide a short reason explaining your decision. The reason should be English.
- Assign a binary label for polarization: If the text is polarized, polarization is set to 1. If the text is not polarized, polarization is set to 0.
- Classify the text for the different types of polarization.
- The answers need to be in the following order: [political/ideological, racial/ethnic, religious, gender/sexual, other].
- The text can contain more than one type of polarization. If the text is not polarized set all polarization types to 0. Do not include any text or formatting outside the JSON.

OUTPUT FORMAT:
```
{   "reason": "the reason of the classification decision",
    "polarization": 0 or 1,
    "polarization Types": [0/1, 0/1, 0/1, 0/1, 0/1]
}
{% if examples %}
**EXAMPLES**:
{{examples}}
{% endif %}
**TEXT**:
{{user_prompt}}
```

F.3 Prompt for polarization manifestation

TASK: Determine the manifestation type(s) of polarization expressed in the following {{lang_full}} text.

LABEL SET: ["stereotype", "vilification", "dehumanization", "extreme_language", "lack_of_empathy", "invalidation"]

DEFINITIONS:
- stereotype: Generalizes traits to all members of a group, ignoring individual differences.
- vilification: Defames or demonizes a group or person through exaggerated or biased framing.
- dehumanization: Denies humanity by comparing people to animals, machines, or objects.
- extreme_language: Uses absolute or hyperbolic language (e.g., “always”, “never”, “us vs. them”).
- lack_of_empathy: Shows indifference or refusal to understand others’ perspectives.
- invalidation: Denies or rejects the identity, legitimacy, or existence of a group or individual.

INSTRUCTIONS:
- Select all labels from the list that apply (one or more).
- If the text is not polarized, set labels = [0].
- Provide a short reasoning sentence explaining your decision.
- Do not include any text or formatting outside the JSON.

OUTPUT FORMAT:

```
{
    "reason": "the reason of the classification decision",
    "polarization": 0 or 1,
    "polarization Types": [0/1, 0/1, 0/1, 0/1, 0/1, 0/1]
}

{% if examples %}
**EXAMPLES**:
{{examples}}
{% end if %}

**TEXT**:
{{user_prompt}}
```

G Error examples

Text	Gold	Type	Reason given by model
The Cybersecurity Infrastructure Security Agency has frozen its election security work is reviewing everything it has done to help state local officials secure elections for the past 8 years. ...The move accommodates Ts false claim of election fraud. @URL	0	FP	The text criticizes a political figure's claim as false and portrays a government agency as accommodating that claim, expressing a negative attitude toward the out-group and contributing to partisan division."
Four key facts Hezbollah is not an Islamic sect or denomination. It is a terrorist org founded in 1985. Hezbollah is an Iranian Shia Islamist proxy based in Lebanon. Not a Lebanese resistance movement.	0	FN	The text vilifies Hezbollah by labeling it a terrorist organization and dismissing its claimed legitimacy, showing hostile stereotyping toward the group, which meets the criteria for attitude polarization.
this new optic the President use that term lightly posted on X of him as a Kingis he serious? Are we no longer a Democracy? We have an election wearly voting starting tomorrow. Im running. Did King Donny stop elections? I appreciate your courage.	1	FN	The text merely questions the President's rhetoric and expresses concern about democratic processes without demeaning, vilifying, or stereotyping any group, so it does not show attitude polarization.
Black Officers Awarded $16M After White Colleague's Racial Comments	1	FN	The headline reports an incident involving race but does not contain demeaning language, vilification, or calls for hostility toward any group, so it is not polarized.
And thats not ethnic cleansing how ?	1	FN	The brief question does not contain explicit negative or vilifying language toward any group, nor does it promote division or intolerance, so it lacks clear attitude polarization.

Table 8: Example texts, the human annotated gold labels, the error type (FN=False Negative, FP=False Positive) and the Reason given by the model

H Dataset Samples

The dataset samples are available at the following link: https://anonymous.4open.science/r/trial-data-05C6/README.md.

Text	Types	Manifestation
"Denazification" as an excuse to explain Russia's deployment of troops to Ukraine, alleging that Hitler was of Jewish ancestry	P	V, EL
"Heck, I voted for Biden three times." Signed, Liberal Asshats in Swing States	P	EL, S, D
Satellite images confirm civilian massacres in towns during Russian military occupation	P	V
"Sufficient evidence of voter fraud?" Where? marjorieisaninsurrectionist	P	EL, I
"They" are now the deep state.	P	S, I, D
15000 dead children claimed by Hamas who uses those children as human shields. Yes yes we covered that already. Now you're just talking in circles	P,Ra	V, EL, I, LE, D
5 Black Officers Awarded $16M After White Colleague's Racial Comments	R	V
A Federal Investigation needs to be launched against TWITTER Fraud and Election Interference	P	EL, I
A Jewish state was created because no one else wanted them. This is why. Zionism and Naziism are the same disease.	P+Re+Ra	V, EL, S, I, LE, D
A lot of lying by the radical left.	P	V, S, I, D
A rational and just society probably wouldn't allow school shootings to be part of their National identity	P	EL, S, I, D
A relationship between the Apartheid state and a tiny tyrant state. Free Palestine	P	V, I, LE, D
A small price to pay for Ukrainian sovereignty and our green future. Stop whining.	P	I, LE
A traitor today, a traitor tomorrow, a traitor always!	P	V
A valid protest of a rigged election, that ended up having some sort of kerfuffle.	P	V, I
A very moral army. IDF Gaza Israel	P	V, LE

Table 9: Examples of polarized statements with annotated types and manifestations. **Types**: **P**: Political, **Re**: Religious, **Ra**:Racial; **Manifestations**: **V**: Vilification, **EL**: Extreme Language, **S**: Stereotype, **I**: Invalidation, **LE**: Lack of Empathy, **D**: Dehumanization.

Text	Lang	Polarized
Gibt sogar recht viele Ex-Muslime. Alles sehr vernünftige Leute. There are actually quite a few ex-Muslims. All very reasonable people.	DE EN	TRUE
Kriegstreiber! Auch ihr, die Propagandapresse!!! Warmongers! You too, the propaganda press!!!	DE EN	TRUE
Steuerzahler und Juden Taxpayers and Jews	DE EN	TRUE
Ja die Türken leiden sehr stark unter ihrem links grün versifften Erdogan. Muss man wissen. :D Yes, the Turks are suffering greatly under their left-green filthy Erdogan. Good to know. :D	DE EN	TRUE
Die Klima Armee Fraktion wird kommen. The Climate Army faction will come.	DE EN	TRUE
Denk ich nicht, nein. Migration ist deren Kernthema. I don't think so, no. Migration is their core issue.	DE EN	TRUE
"Es gibt bereits vereinzelte Einsätze von Bodentruppen in Gaza. @URL" "There are already isolated deployments of ground troops in Gaza. @URL"	DE EN	FALSE
Lass dich nicht aufhalten. Es gibt eine ukrainische Fremdenlegion. Da kannst du dich melden. Don't let yourself be stopped. There is a Ukrainian foreign legion. You can sign up there.	DE EN	FALSE
Klingt für mich nach Long COVID Sounds like long COVID to me.	DE EN	FALSE
Wenn ich schon wieder dieses Woke lese. Der AfD ist sogar unser Grundgesetz zu Woke. Hört auf das Wort so inflationär zu gebrauchen. When I read this "woke" again. Even our constitution is too woke for the AfD. Stop using the word so inflationarily.	DE EN	FALSE
CO2 ist gut für Pflanzen CO2 is good for plants.	DE EN	FALSE

Table 10: German samples with labels and English translations

Text	Lang	Polarized
y vos un random judio que comenta pelotudeces para poder figurar en algun lado And you, a random Jew who comments nonsense just to get noticed somewhere.	SPA EN	TRUE
quiero comer indio viejo I want to eat old Indian.	SPA EN	TRUE
es que a dia de hoy parece que la diferencia es nazis con aborto y nazis sin aborto Nowadays it seems that the difference is Nazis with abortion and Nazis without abortion.	SPA EN	TRUE
orgulloso de sacarle la conchetumare a un pastor evangelico por meterse en weas que no debia. Proud to beat the hell out of an evangelical pastor for meddling in things he shouldn't.	SPA EN	TRUE
el tribunal superior del partido corrupto The supreme court of the corrupt party.	SPA EN	TRUE
claudia es judia, que no manche Claudia is Jewish, don't stain.	SPA EN	FALSE
la escuela feminista de pintura The feminist school of painting.	SPA EN	FALSE
la cancion que me representa ahora que estoy deportado de los estados unidos. The song that represents me now that I am deported from the United States.	SPA EN	FALSE
aunque, siendo justos, esa foto es de una deportacion de 2018, cuando trump estaba gobernando. Although, to be fair, that photo is from a 2018 deportation, when Trump was in power.	SPA EN	FALSE
la propuesta de k!nadim me parece muy interesante de cara a eurovision, es facilona y divertida #benidormfestsemi1 The proposal by k!nadim seems very interesting to me for Eurovision; it's easy and fun. #benidormfestsemi1	SPA EN	FALSE

Table 11: Spanish sentences with labels and English translations.