NYCU-NLP at SemEval-2026 Task 9: Stacking Small Language Models for Multilingual, Multicultural and Multievent Polarization Detection

Ding-Xiang Lin, Po-Chun Chu, Lung-Hao Lee*

Institute of Artificial Intelligence Innovation
National Yang Ming Chiao Tung University

***Contact:** lhlee@nycu.edu.tw

Abstract

This paper presents the NYCU-NLP system for SemEval-2026 Task 9 on online polarization analysis. Our approach explores the effectiveness of instruction-tuned small language models (SLMs), including Phi-4 (14B), Mistral-small-3.2 (24B), and Gemma-3 (27B), with task-specific prompting strategies and combined them via a stacking ensemble to leverage complementary modeling capacities. Evaluated across 22 languages and three subtasks, our system achieved macro-averaged F1 scores of 0.8071 for Polarization Detection (Subtask 1), 0.6108 for Polarization Type Classification (Subtask 2), and 0.5111 for Polarization Manifestation Identification (Subtask 3). Notably, our approach ranked first in 15, second in 12, and third in 10 of the 62 language-specific leaderboards, demonstrating the robustness and competitiveness of stacking-based SLM ensembles in multilingual settings.

1 Introduction

Online polarization, defined as the sharp division and hostility among social, political, or identity-based groups (Waller and Anderson, 2021), has attracted increasing attention due to its potential to escalate into hate speech, offensive language, harassment, real-world violence, and broader social fragmentation (Piazza, 2023; Martínez-España et al., 2024). Consequently, the identification of polarized opinions has become a critical task for determining whether a given text exhibits explicit attitudinal polarization, thereby contributing to the promotion of a healthier online environment.

SemEval-2026 Task 9 (Naseem et al., 2026a) focuses on analyzing how polarization is manifested in textual discourse across different languages, cultures, and socially significant contexts. The shared task comprises three subtasks: 1) **Subtask 1: Polarization Detection**, which determines whether a given text expresses polarization; 2) **Subtask 2: Polarization Type Classification**, which identifies the underlying social dimension(s) of polarization, including political, racial/ethnic, religious, gender/sexual, and other categories; and 3) **Subtask 3: Polarization Manifestation Identification**, which detects rhetorical manifestation(s) of polarization, such as stereotyping, vilification, dehumanization, extreme language, lack of empathy, and invalidation. While Subtask 1 is formulated as a binary classification problem, Subtasks 2 and 3 are designed as multi-label classification tasks, reflecting the possibility that multiple forms of polarization may co-occur within a single text. The datasets cover 22 languages and are provided by the task organizers. Participating teams may choose to compete in one or more languages and subtasks according to their preference.

This paper presents the **NYCU-NLP** (**N**ational **Y**ang Ming **C**hiao Tung **U**niversity, **N**atural **L**anguage **P**rocessing Lab) system for the SemEval-2026 Task 9. Our approach leverages instruction-tuned Small Language Models (SLMs), including Phi-4 (14B) (Abdin et al., 2024), Mistral-small-3.2 (24B), and Gemma-3 (27B) (Kamath et al., 2025), to identify and classify polarized content. We further integrate these three fine-tuned models using a stacking ensemble (Pavlyshenko, 2018) as our primary system architecture, aiming to achieve stable and robust performance across languages and subtasks. Experimental results on the official test set, averaged across all languages, demonstrate that our stacking-based SLM ensemble achieves macro-averaged F1 scores of 0.8071 for Subtask 1, 0.6108 for Subtask 2, and 0.5111 for Subtask 3. Our system ranked first in 15, second in 12, and third in 10 out of 62 language-specific leaderboards, highlighting the robustness and competitiveness of the stacking-based SLM ensemble in multilingual polarization analysis.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1258–1267

2 Related Work

Online polarization denotes the division of individuals or communities into opposing groups with distinct ideological positions, values, or identities in online discourse. Prior studies have characterized such polarization through differences at both group and corpus levels. Polarized concepts have been modeled via salience and framing in Reddit communities (Hofmann et al., 2022), while topic-level polarization has been quantified by associating opposing poles in aligned distributional representations (Bianchi et al., 2021). More recently, large language model (LLM)-based simulations have been leveraged to study echo chambers and opinion polarization in social networks (Wang et al., 2025).

In polarized discourse, ideological groups often exhibit divergent evaluations of the same issue or entity. Aspect-based sentiment analysis (ABSA) has been employed to capture stance-dependent affective differences (Vorakitphan et al., 2020) and to track sentiment toward multiple geopolitical or ideological entities in large-scale online comments (Miehling et al., 2025). Beyond this, online polarization has been formulated as a supervised text classification task, where models predict whether a text is polarized and, in more fine-grained settings, classify its associated social dimensions or manifestations. Polarization may also emerge at the annotation level, particularly in toxic language detection. For example, polarized annotations have been modeled using normalized distance from unimodality, enabling their treatment as an additional class and yielding improved classification performance (Pavlopoulos and Likas, 2024).

Despite these advances, multilingual online polarization detection remains challenging, as polarized expressions are shaped by language, culture, and sociopolitical context. The importance of language-specific resources for analyzing political polarization, propaganda techniques, and media bias has been emphasized (Szwoch et al., 2022). Evidence further indicates that low- and mid-resource languages remain underexplored, and that modeling choices, such as monolingual, multilingual, cross-lingual, or zero-shot LLM settings, can substantially affect performance (Davoudi and Goharian, 2026). These challenges motivate the development of robust multilingual approaches that mitigate model-specific weaknesses across languages.

3 The NYCU-NLP System

As SemEval-2026 Task 9 (Naseem et al., 2026a) represents the first shared task on online polarization analysis, we design our SLM-based ensemble architecture by drawing on insights from our previous task participations (Lee et al., 2024; Lin et al., 2024; Xu et al., 2025). Figure 1 illustrates the overall architecture of the NYCU-NLP system. We first fine-tune three small language models (SLMs) using carefully designed task-specific prompts, and subsequently integrate them through a stacking ensemble framework. Specifically, we train a logistic regression meta-model to aggregate the predictions from multiple fine-tuned SLMs and generate the final outputs for all three subtasks across languages.

3.1 Small Language Models

SLMs are lightweight versions of large language models (LLMs), designed to operate efficiently under constrained computational resources while preserving core modeling capabilities. We investigate the effectiveness of the following three SLMs.

(1) **Phi-4 (14B)** (Abdin et al., 2024), developed by Microsoft Research, adopts a training strategy centered on high-quality synthetic data and reasoning-oriented curricula. Despite its relatively moderate scale, Phi-4 achieves strong reasoning performance through optimized training schedules, and post-training techniques such as supervised fine-tuning (SFT)(Wei et al., 2022) and Direct Preference Optimization (DPO) (Rafailov et al., 2023).

(2) **Mistral-small-3.2 (24B)** (MistralAI, 2026) is a compact language model designed to deliver competitive within a lightweight architecture. It surpasses several larger-scale language models on multiple official benchmarks in instruction following, multilingual understanding, and inference efficiency, demonstrating that carefully optimized small language models can rival, or even outperform substantially larger counterparts.

(3) **Gemma-3 (27B)** (Kamath et al., 2025), developed by Google DeepMind, incorporates architectural refinements and advanced post-training techniques to enhance instruction-following and multilingual capabilities. It achieves competitive performance comparable to significantly larger language models across a broad range of language understanding tasks.

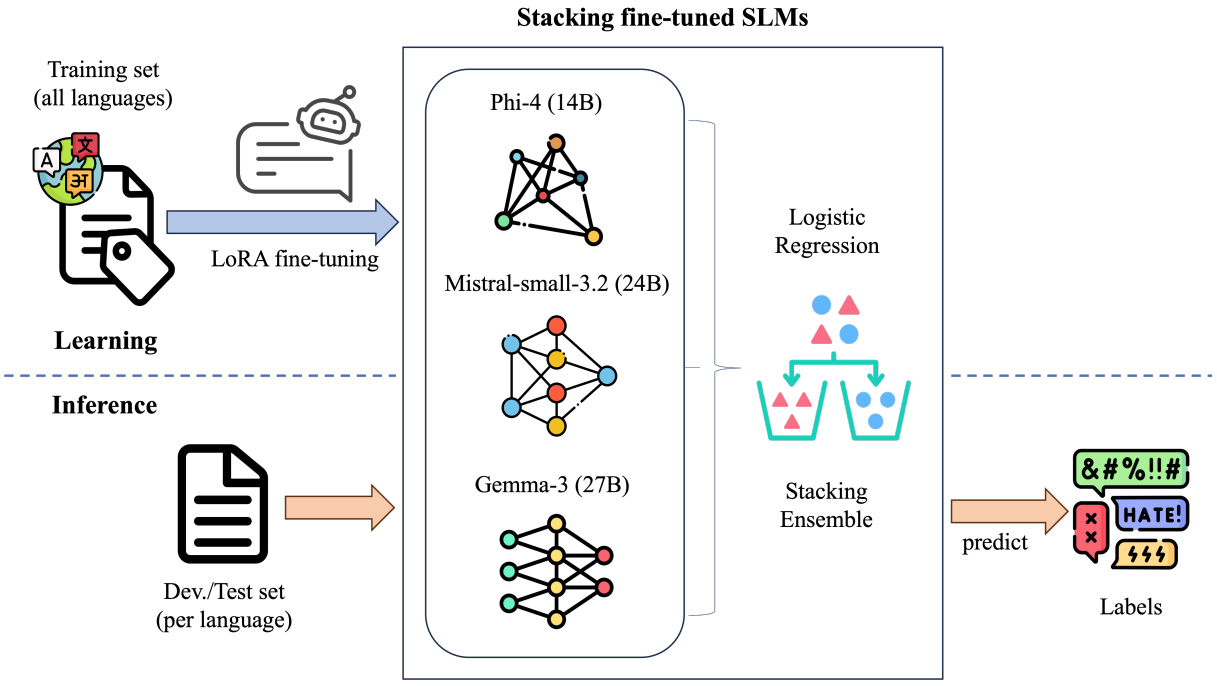

Figure 1: Our NYCU-NLP system architecture for the SemEval-2026 Task 9.

3.2 Instruction Fine-tuning

We use instruction tuning (Wei et al., 2022) in combination with LoRA (Hu et al., 2021) to adapt the three pre-trained SLMs for each subtask. The prompts are designed based on the annotation guidelines provided by the task organizers (see Appendix A). For each subtask, the SLMs are configured to perform the corresponding classification task, including binary polarization detection and multi-label classification for polarization targets and manifestation categories. Each predicted label indicates the presence (1) or absence (0) of the corresponding category.

3.3 Assembly Mechanism

To enhance robustness and reduce the variance of individual classifiers, we adopt an ensemble-based aggregation strategy that integrates predictions from three SLMs. Let $M = \{1, 2, 3\}$ denote the set of base models. Given an input instance x, each SLM produces a binary prediction $\hat{y}_m(x) \in \{0, 1\}$, where $m \in M$.

To further exploit complementarities among the SLMs, we employ a stacking ensemble framework (Pavlyshenko, 2018), in which the meta-model is implemented as a logistic regression classifier. Each SLM outputs a confidence score for the positive class, denoted as $p_m(x) \in [0, 1]$. These three confidence scores serve as meta-level features. Formally, for an instance x, the feature vector is defined as:

$$\mathbf{z}(x) = [p_1(x), p_2(x), p_3(x)]^\top. \quad (1)$$

The meta-model computes the logit value $t(x)$ and applies the sigmoid function to obtain the final predicted probability $\hat{p}(x)$.

$$t(x) = \mathbf{w}^\top \mathbf{z}(x) + b. \quad (2)$$

$$\hat{p}(x) = \sigma\big(t(x)\big). \quad (3)$$

where $\mathbf{w}$ and b denote the weight vector and bias term learned by the logistic regression meta-classifier, and $\sigma(\cdot)$ is the sigmoid activation function. The final binary prediction is obtained by thresholding the probability:

$$\hat{y}(x) = \begin{cases} 1, & \text{if } \hat{p}(x) \geq 0.5, \\ 0, & otherwise. \end{cases} \quad (4)$$

3.4 Task-specific Heuristic Rule

To improve performance on the *other* category in Subtask 2, we incorporate the prediction from Subtask 1 as an auxiliary signal during inference. Our preliminary analysis indicates that the *other* label is comparatively difficult to learn. In some cases, the Subtask 2 classifier predicts all labels as 0, even when Subtask 1 identifies the instance as polarized (i.e., prediction is 1).

Lang. Family	Lang. (ISO-639)	Subtask 1			Subtask 2			Subtask 3		
		training	dev	test	training	dev	test	training	dev	test
Indo-European	English (eng)	3222	160	1452	3222	160	1452	3222	160	1452
	German (deu)	3180	159	1432	3180	159	1432	3180	159	1432
	Urdo (urd)	3563	177	1606	3563	177	1606	3563	177	1606
	Bengali (ben)	3333	166	1501	3333	166	1501	3333	166	1501
	Hindi (hin)	2744	137	1236	2744	137	1236	2744	137	1236
	Odia (ori)	2368	118	1066	2368	118	1066	2368	118	1066
	Nepali (nep)	2005	100	903	2005	100	903	2005	100	903
	Punjabi (pan)	1700	100	809	1700	100	809	1700	100	809
	Spanish (spa)	3305	165	1488	3305	165	1488	3305	165	1488
	Italian (ita)	3334	166	1538	3334	166	1538		–	
	Russian (rus)	3348	167	1508	3348	167	1508		–	
	Polish (pol)	2391	119	1077	2391	119	1077		–	
	Persian (fas)	3295	164	1484	3295	164	1484	3295	164	1484
Afro-Asiatic	Hausa (hau)	3651	182	1644	3651	182	1644	3651	182	1644
	Arabic (arb)	3380	169	1521	3380	169	1521	3380	169	1521
	Amharic (amh)	3332	166	1501	3332	166	1501	3332	166	1501
Sino-Tibetan	Chinese (zho)	4280	214	1927	4280	214	1927	4280	214	1927
	Burmese (mya)	2889	144	1301	2889	144	1301		–	
Austroasiatic	Khmer (khm)	6640	332	2988	6640	332	2988	6640	332	2988
Niger-Congo	Swahili (swa)	6991	349	3147	6991	349	3147	6991	349	3147
Dravidian	Telugu (tel)	2366	118	1066	2366	118	1066	2366	118	1066
Turkic	Turkish (tur)	2364	115	1093	2364	115	1093	2364	115	1093

Table 1: **Task dataset statistics**. For each dataset, we report the number of instances in the train/dev/test splits for each subtask. For Subtask 3, the task does not provide datasets for Italian, Russian, Polish and Burmese languages, denoted as "-".

To enforce cross-task consistency, we introduce a heuristic rule: if Subtask 1 predicts an instance as polarized and all Subtask 2 labels are 0, we assign the *other* label a value of 1. This post-processing strategy recovers missed cases and mitigates false negatives, thereby improving consistency between the two subtasks.

4 Evaluation

4.1 Datasets

The datasets were provided by the task organizers (Naseem et al., 2026b) and partitioned into mutually exclusive training, development, and test sets. Table 1 summarizes the data statistics. Subtasks 1 and 2 shared identical textual instances across 22 languages. In contrast, Subtask 3 covers 18 of these languages, while retaining the same instances for the overlapping languages. Overall, the dataset comprises 73,681 instances for model training, 3,687 for validation, and 33,288 for final evaluation.

4.2 Settings

We obtained pre-trained models from HuggingFace[1] and further fine-tuned them using only the training data for each subtask in a multilingual setting. All experiments were conducted on a server with two Nvidia H100 GPUs (Total 160GB memory). The hyperparameters for the SLMs were configured as follows: training epochs of 3; batch size of 12; Paged AdamW (32 bit) optimizer; learning rate of 1e-4; LoRA rank (r) of 8; LoRA alpha of 32 and LoRA dropout of 0.05.

Task organizers provided baseline results by fine-tuning LaBSE (Feng et al., 2022) using the training set for each language for all three subtasks. Each participating team was permitted to submit up to five runs. For all three subtasks, model performance was evaluated using the macro-averaged F1 score. The unofficial rankings are based on the final submissions of all registered teams. The offi-

[1] https://huggingface.co/microsoft/phi-4
https://huggingface.co/mistralai/Mistral-Small-3.2-24B
https://huggingface.co/google/gemma-3-27b-it

Lang.	Subtask 1				Subtask 2				Subtask 3			
	Phi	**Mistral**	**Gemma**	**Ensemble**	**Phi**	**Mistral**	**Gemma**	**Ensemble**	**Phi**	**Mistral**	**Gemma**	**Ensemble**
eng	**0.8290**	0.8077	0.8046	0.8138	0.3768	0.4230	0.3625	**0.4264**	0.4240	0.4349	0.3908	**0.4978**
deu	0.7734	0.7649	0.7790	**0.7922**	0.5471	0.5665	0.4945	**0.5949**	0.5136	0.4828	0.4618	**0.5487**
urd	0.7427	0.7406	**0.8249**	0.7991	0.7606	0.7657	**0.7985**	0.7937	0.7828	**0.8389**	0.8090	0.8358
ben	0.8411	0.8582	0.8582	**0.8701**	0.3118	**0.3270**	0.3243	0.3196	0.0928	0.1899	0.1469	**0.2134**
hin	0.8587	**0.8633**	0.8385	0.8171	0.7737	0.8324	0.8009	**0.8435**	0.7799	0.7994	0.8147	**0.8240**
ori	**0.8687**	0.6949	0.8376	0.7714	0.5911	0.3722	0.6382	**0.6807**	0.1135	0.1294	0.2094	**0.2374**
nep	0.8900	0.8900	**0.9099**	0.8899	0.7703	0.7802	**0.8421**	0.8001	0.6744	0.7083	0.7063	**0.7398**
pan	0.7895	0.8200	0.8091	**0.8493**	0.2929	0.3980	**0.4934**	0.4786	0.4403	0.4246	0.5236	**0.5350**
spa	0.6843	**0.7331**	0.7030	0.7210	0.5515	0.6158	0.6117	**0.6581**	0.3316	0.4143	0.4165	**0.4758**
ita	0.6528	0.6652	**0.6713**	0.6577	0.3781	0.3906	**0.4218**	0.4049	–	–	–	–
rus	0.8100	0.8123	0.7988	**0.8308**	0.5457	0.5087	0.5619	**0.6957**	–	–	–	–
pol	0.8022	0.8194	0.8429	**0.8620**	0.5712	0.6247	0.5597	**0.6960**	–	–	–	–
fas	0.7851	**0.8443**	0.7794	0.7981	0.5578	0.5997	**0.6376**	0.6045	0.1425	0.2480	0.3129	**0.3164**
hau	0.7259	0.7729	**0.8261**	0.8053	0.2810	0.3625	0.3356	**0.5582**	0.0333	0.0000	0.0972	**0.1378**
arb	0.7962	0.8153	0.8798	**0.8801**	0.6530	0.5951	0.6393	**0.6860**	0.4895	0.5388	0.6027	**0.6852**
amh	0.6742	0.6813	0.7758	**0.8019**	0.3509	0.4278	0.4759	**0.4856**	0.2298	0.3121	0.5457	**0.5705**
zho	0.9159	0.9346	0.9346	**0.9439**	0.7882	0.7982	0.7954	**0.8538**	0.6494	**0.7620**	0.7491	0.7558
mya	0.8294	0.8442	**0.8794**	0.8646	0.4450	0.4851	**0.6068**	0.5709	–	–	–	–
khm	**0.5564**	0.5505	0.5255	0.5534	0.5402	0.6353	**0.7631**	0.7321	0.2102	0.2488	0.3995	**0.4350**
swa	0.8052	**0.8363**	0.8216	0.8166	0.4139	0.4591	**0.5017**	0.4949	0.4313	0.4770	0.4748	**0.5638**
tel	0.8978	0.9237	0.9152	**0.9406**	0.2403	0.2626	0.2883	**0.4539**	0.1809	0.2529	0.2874	**0.4024**
tur	0.8346	0.7998	0.8429	**0.8783**	0.6589	0.6540	0.6265	**0.6666**	0.3514	0.5359	0.5539	**0.5909**
AVG.	0.7892	0.7942	0.8117	**0.8162**	0.5182	0.5402	0.5718	**0.6136**	0.3817	0.4332	0.4723	**0.5203**

Table 2: **Evaluation results of fine-tuned SLMs on the development set**. Bold indicates the best result within each setting for each dataset. "-" denotes not available due to no data.

cial rankings, to be reported in the task description paper, will include only those teams that submit a corresponding system description paper.

4.3 Results

Table 2 presents the macro-averaged F1 scores on the development set for each language. For Subtask 1 (binary classification task), neither the individual SLMs nor the proposed ensemble consistently outperformed the others across all languages. For Subtask 2, the stacking ensemble of the three SLMs slightly outperformed Gemma-3 (27B). For Subtask 3, the proposed stacking ensemble clearly surpassed the individual SLMs, demonstrating the effectiveness of ensemble learning in more challenging multi-label classification scenarios.

Table 3 presents the macro-averaged F1 scores on the test set for each language. While similar trends are observed, the proposed ensemble underperformed the baseline on several low-resource languages, including Odia, Persian, Hausa, Khmer, and Telugu. Overall, the stacking ensemble achieved macro-averaged F1 scores of 0.8071, 0.6108, and 0.5111 across all languages for Subtasks 1, 2, and 3, respectively.

The official rankings released by the task organizers for each language-specific leaderboard are also summarized in Table 3. We submitted the stacking-based SLM ensemble as our final system for all subtasks. Our system ranked first in 15, second in 12, and third in 10 of 62 language-specific evaluations, demonstrating the robustness and competitiveness of stacking-based SLM ensembles in multilingual contexts.

4.4 Discussion

We evaluated several SLMs that have shown strong performance on public leaderboards, including Vicuna-1.5 (13B) (Zheng et al., 2023), Llama-2 (13B) (Touvron et al., 2023), Ling-mini-2.0 (16B) (Tian et al., 2025), Reka-Flash-3.1 (21B) (reka.ai, 2026), OLMo-2 (32B) (Walsh et al., 2025), Yi-1.5 (34B) (Young et al., 2025), and Qwen-3 (30B) (Yang et al., 2025). Based on this comparison, we identified Phi-4 (14B), Mistral-small-3.2 (24B), and Gemma-3 (27B) as the best-performing models and adopted them as our primary SLMs for polarization detection.

Lang.	Subtask 1			Subtask 2			Subtask 3		
	Baseline	**Individual** (P/M/G)	**Ensemble** (rank/total)	**Baseline**	**Individual** (P/M/G)	**Ensemble** (rank/total)	**Baseline**	**Individual** (P/M/G)	**Ensemble** (rank/total)
eng	0.7802	0.8173 / 0.8161 / **0.8186**	0.8172 (4/44)	0.3333	0.4614 / 0.4606 / 0.4899	**0.5135** (3/28)	0.4100	0.3288 / 0.3999 / 0.4187	**0.5010** (4/17)
deu	0.6714	0.7369 / 0.7528 / 0.7503	**0.7608** (1/32)	0.4078	0.4864 / 0.5501 / 0.5734	**0.6157** (2/23)	0.3485	0.3951 / 0.4599 / 0.4645	**0.5176** (1/14)
urd	0.7890	0.7774 / 0.8024 / **0.8216**	0.8169 (2/34)	0.7127	0.7716 / 0.7861 / 0.7882	**0.7889** (3/24)	0.5316	0.8000 / 0.8128 / 0.8182	**0.8213** (1/16)
ben	0.8528	0.8409 / 0.8493 / **0.8611**	0.8538 (3/36)	0.2887	0.2969 / 0.3329 / 0.3396	**0.4007** (2/25)	0.0868	0.0921 / 0.1395 / 0.1704	**0.2139** (10/16)
hin	0.7379	0.7852 / **0.8331** / 0.8192	0.8157 (6/34)	0.7911	0.7016 / 0.7832 / 0.7840	**0.8013** (2/24)	0.2348	0.6971 / 0.7428 / 0.7399	**0.7704** (2/16)
ori	0.7765	0.7835 / 0.7268 / **0.8044**	0.7786 (19/32)	0.5600	0.4278 / 0.3962 / 0.5258	**0.5779** (3/22)	**0.3841**	0.1408 / 0.1910 / 0.2519	0.2973 (3/16)
nep	0.8798	0.8948 / 0.9003 / 0.9125	**0.9236** (1/32)	0.7219	0.7857 / 0.7913 / **0.8134**	0.8104 (1/22)	0.1314	0.6356 / 0.6749 / 0.7050	**0.7127** (1/15)
pan	0.7898	0.7688 / 0.7849 / **0.8107**	**0.8107** (3/32)	0.3650	0.3401 / 0.4350 / 0.4608	**0.4835** (5/19)	0.4561	0.4115 / 0.4697 / 0.5134	**0.5441** (1/15)
spa	0.7266	0.7782 / 0.7902 / 0.7854	**0.7996** (2/36)	0.5935	0.6239 / 0.6232 / 0.6609	**0.6806** (1/24)	0.5088	0.3497 / 0.4343 / 0.4174	**0.5198** (2/16)
ita	0.6773	**0.7049** / 0.5792 / 0.5689	0.5527 (26/31)	**0.3759**	0.2851 / 0.2431 / 0.2826	0.3673 (6/21)	–	–	–
rus	0.7457	0.8150 / 0.8180 / 0.8062	**0.8232** (2/30)	0.5904	0.5314 / 0.5884 / 0.5690	**0.6295** (2/22)	–	–	–
pol	0.7241	0.8039 / 0.8245 / 0.8332	**0.8350** (2/31)	0.4491	0.5155 / 0.5439 / 0.6061	**0.6400** (2/22)	–	–	–
fas	**0.8424**	0.7150 / 0.7750 / 0.7501	0.7855 (25/31)	0.4626	0.5069 / 0.5409 / 0.5657	**0.5980** (6/21)	0.2004	0.1879 / 0.2826 / 0.3237	**0.3996** (8/14)
hau	0.7753	0.7500 / 0.7658 / **0.7777**	0.7679 (19/30)	0.2038	0.1496 / 0.2939 / 0.2935	**0.4796** (1/21)	**0.7456**	0.0116 / 0.0268 / 0.1196	0.1596 (7/15)
arb	0.7957	0.8254 / 0.8319 / **0.8468**	0.8427 (3/32)	0.4855	0.5992 / 0.6328 / 0.6550	**0.6698** (1/22)	0.3902	0.5220 / 0.5538 / 0.6037	**0.6456** (1/14)
amh	0.7151	0.6217 / 0.7002 / **0.8133**	0.7894 (4/29)	0.3716	0.3445 / 0.4465 / 0.5457	**0.5487** (12/20)	0.4433	0.2398 / 0.3236 / 0.4885	**0.5587** (2/14)
zho	0.8691	0.9092 / 0.9232 / 0.9170	**0.9273** (3/32)	0.6697	0.8066 / 0.8180 / 0.8284	**0.8436** (1/22)	0.0000	0.6719 / 0.7004 / 0.6858	**0.7191** (1/16)
mya	0.8210	0.8456 / 0.8286 / 0.8837	**0.8868** (3/29)	0.4772	0.5655 / 0.6243 / **0.6978**	0.6938 (8/20)	–	–	–
khm	**0.6592**	0.5677 / 0.6023 / 0.6532	0.6462 (23/30)	0.6268	0.5926 / 0.5390 / 0.6948	**0.6965** (4/20)	**0.6095**	0.2496 / 0.2704 / 0.3109	0.3357 (8/14)
swa	0.7571	0.7795 / 0.7925 / 0.7920	**0.7975** (4/30)	0.4417	0.4218 / 0.4769 / 0.5097	**0.5224** (4/21)	0.2205	0.3754 / 0.4431 / 0.4485	**0.5487** (5/15)
tel	0.6440	0.8674 / **0.8967** / 0.8940	0.8921 (5/32)	0.3145	0.2567 / 0.2974 / 0.3330	**0.4296** (8/22)	**0.6738**	0.2248 / 0.2516 / 0.2777	0.3964 (6/16)
tur	0.6957	0.8079 / 0.7958 / 0.8149	**0.8329** (1/30)	0.4708	0.5610 / 0.6039 / 0.6396	**0.6462** (3/21)	**0.7693**	0.4319 / 0.4532 / 0.4803	0.5381 (1/14)
AVG.	0.7603	0.7816 / 0.7904 / 0.8061	**0.8071**	0.4870	0.5014 / 0.5367 / 0.5753	**0.6108**	0.4203	0.3759 / 0.4239 / 0.4577	**0.5111**

Table 3: **The test set results and official rankings of our NYCU-NLP system for each language**. Parentheses (P/M/G) respectively denote the SLM: Phi-4 (14B), Mistral-small-3.2 (24B), and Gemma-3 (27B). Parentheses (rank/total) show the official ranking released by the task organizers among all submissions. Yellow indicates the first rank, blue indicates the second rank, and green indicates the third rank.

We further compared several ensemble strategies. Ultimately, we adopted a stacking ensemble as our assembly approach, as it consistently outperformed both the Simple Majority Voting Ensemble (SMVE) and the Weighted Majority Voting Ensemble (WMVE) (Dogan and Birant, 2019) baselines on the development set.

Existing SLMs are typically pretrained on multilingual corpora. By leveraging all available training instances across languages to fine-tune a unified SLM, we consistently achieved better performance than training separate language-specific models. These results indicate that unified fine-tuning more effectively harnesses the benefits of multilingual pretraining and promotes cross-lingual knowledge transfer.

5 Conclusions

This paper presents the NYCU-NLP system for SemEval-2026 Task 9 on multilingual polarization analysis, detailing both the system architecture and its empirical evaluation. Our approach is built on instruction-tuned small language models (SLMs), including Phi-4, Mistral-small-3.2, and Gemma-3, optimized with LoRA and task-specific prompting strategies, and further strengthened through ensemble-based aggregation mechanisms.

Experimental results show that the stacking ensemble consistently achieved the best performance on both the development and test sets across languages. Specifically, our system obtained macro-averaged F1 scores of 0.8071 for polarization detection (Subtask 1), 0.6108 for polarization type classification (Subtask 2), and 0.5111 for polarization manifestation identification (Subtask 3). Overall, the system ranked first in 15, second in 12, and third in 10 out of 62 language-specific evaluations, demonstrating the robustness and competitiveness of stacking-based SLM ensembles in multilingual settings.

Limitations

This work does not propose a new model to address this task for online polarization analysis. Experi-

ments were conducted with basic settings without other advanced explorations due to computational resource limitations.

We do not apply additional prompt engineering techniques to further optimize the prompts. Consequently, the instruction fine-tuning prompts may benefit from further refinement to achieve improved performance. Moreover, our instruction tuning relies solely on the original training data; incorporating data augmentation strategies could potentially yield additional gains in model performance.

Use of AI Assistants

We used ChatGPT and Grammarly to correct grammatical errors, enhancing the coherence of the final manuscript. While these tools have augmented our capabilities and contributed to our findings, it's pertinent to note that they have inherent limitations. We have made every effort to use AI in a transparent and responsible manner. Any conclusions drawn are a result of combined human and machine insights.

Acknowledgments

This study was partially supported by the National Science and Technology Council, Taiwan, under the grant NSTC 114-2221-E-A49-059-MY3. This work was also financially supported by the Co-creation Platform of the Industry Academia Innovation School, NYCU.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Federico Bianchi, Marco Marelli, Paolo Nicoli, and Matteo Palmonari. 2021. SWEAT: Scoring polarization of topics across different corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10065–10072.

Saeedeh Davoudi and Nazli Goharian. 2026. Online polarization detection in Persian (Farsi) social media. In *Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family*, pages 50–59.

Alican Dogan and Derya Birant. 2019. A weighted majority voting ensemble approach for classification. In *Proceedings of the 4th International Conference on Comouter Science and Engineering*, pages 1–6.

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.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Lung-Hao Lee, Chen-Ya Chiou, and Tzu-Mi Lin. 2024. NYCU-NLP at SemEval-2024 Task 2: Aggregating large language models in biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, pages 1129–1135.

Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, and Lung-Hao Lee. 2024. NYCU-NLP at EXALT 2024: Assembling large language models for cross-lingual emotion and trigger detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*, pages 505–510.

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*, 12:102035–102049.

Daniel Miehling, Daniel Dakota, and Sandra Kübler. 2025. Analyzing polarization in online discourse on the 2023-2024 Israel–hamas war. In *Proceedings of the 21st Conference on Natural Language Processing: Workshops*, pages 7–16.

MistralAI. 2026. Mistral Small 3. *mistral.ai*.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acart"urk, Aisha Jabr, Saba Anwar,

Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. POLAR: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint arXiv:2505.20624*.

John Pavlopoulos and Aristidis Likas. 2024. Polarized opinion detection improves the detection of toxic language. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Bohdan Pavlyshenko. 2018. Using stacking approaches for machine learning models. In *Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing*, pages 255–258.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing System*, pages 53728–53741.

reka.ai. 2026. Reniforcement learning for reka flash 3.1. *reka.ai*.

Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. 2022. Creation of Polish online news corpus for political polarization studies. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 86–90.

Changxin Tian, Kunlong Chen, Jia Liu, Ziqi Liu, Zhiqiang Zhang, and Jun Zhou. 2025. Towards greater leverage: Scaling laws for efficient mixture-of-experts language models. *Preprint*, arXiv:2507.17702.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Vorakit Vorakitphan, Marco Guerini, Elena Cabrio, and Serena Villata. 2020. Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 219–224.

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

Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025. Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3913–3923.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Zhe-Yu Xu, Yu-Hsin Wu, and Lung-Hao Lee. 2025. NYCU-NLP at SemEval-2025 Task 11: Assembling small language models for multilabel emotion detection and intensity prediction. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, pages 1129–1135.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, Wenhao Huang, and 12 others. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Appendix: Instruction-tuning Prompts

- Subtask 1

Field	Content
System Prompt	You are a binary classifier for polarization. For one input text, decide if it clearly contains attitude polarization. Respond ONLY with JSON: `{"polarization":1}` for polarized, `{"polarization":0}` otherwise. No other text, no spaces, no line breaks. Definition of polarization (any one clearly present → 1): – Negative attitude toward an out-group and/or blind in-group support. – Stereotyping, vilification, dehumanization, or other individuation of a group. – Intolerance of others' views/beliefs/identities. – Calls to exclude, silence, attack, or refuse coexistence with another group; urges division/hatred/conflict. – Blanket generalizations about a group ("they are all X"). Attacking groups still → 1. Label 0 when: – Neutral reporting/quoting without endorsement or with clear condemnation. – Criticism of policies/ideas without targeting a group identity. – Calls for dialogue/unity/tolerance. – Profanity not directed at a group; personal disputes not framed as groups. – Ambiguous, sarcastic/ironic but unclear, contextless snippets. – Empty, gibberish, or off-topic input. Guidelines: – Consider overall meaning and context, not keywords. – When uncertain, choose 0.
User Prompt	Text: `{text}`
Assistant Prompt	`{Polarization}`

- Subtask 2

Field	Content
System Prompt	You are a multi-label classifier for polarization target. For each input text, decide which groups are polarized (attacked, excluded, degraded, or placed in conflict): – political: political parties, ideologies, or their supporters – racial_ethnic: races, ethnicities, or national/ethnic groups – religious: religions or believers/non-believers – gender_sexual: genders or sexual orientations – other: other groups/identities (e.g., social class, economy, media, institutions, etc.) Rules: 1. Multi-label: any number of categories can be 1. 2. Set a category to 1 only if it is clearly targeted; otherwise 0. 3. Respond ONLY with a JSON object using exactly these keys and integer values 0 or 1.
User Prompt	Text: `{text}`
Assistant Prompt	`{Polarization type}`

- Subtask 3

Field	Content
System Prompt	You are a multi-label classifier for polarization manifestations. For each input text, decide which manifestations are used: – stereotype: generalizing traits of some individuals to an entire group, ignoring individual differences – vilification: defaming or demonizing a group/person/entity, often presenting them as dangerous or harmful – dehumanization: describing people as animals, objects, machines, or otherwise less than human – extreme_language: using very extreme or absolute language (e.g., “always”, “never”, “us vs. them”) – lack_of_empathy: dismissing or showing no willingness to understand others’ experiences or perspectives – invalidation: denying the identity, rights, or existence of a group (e.g., saying a nation or people do not exist) Rules: 1. Multi-label: any number of manifestations can be 1. 2. Set a manifestation to 1 only if it is clearly present; otherwise 0. 3. Respond ONLY with a JSON object using exactly these keys and integer values 0 or 1.
User Prompt	Text: `{text}`
Assistant Prompt	`{Polarization manifestation}`