ABARUAH at SemEval-2026 Task 9: Multilingual Polarization Detection across Seven Indic Languages using Qwen3

Arup Baruah
Dept. of CSE, Assam Don Bosco University
arup.baruah@gmail.com, arup.baruah@dbuniversity.ac.in

Abstract

Online polarization creates division within the society. As such, it is important to detect and remove polarized messages from social media. This study presents fine-tuned Qwen3-8B Large Language Model (LLM) based models to identify online polarization, its specific categories, and its manifestation types. This study used Quantized Low-Rank Adaptation (QLoRA) to fine-tune the model in seven Indic languages: Bengali, Hindi, Nepali, Oriya, Punjabi, Telugu, and Urdu. The experimental results demonstrate the efficacy of this approach, achieving macro F1-scores of 0.82, 0.78, 0.90, 0.76, 0.78, 0.87, and 0.79, respectively, for polarization detection. The proposed model surpassed the established baseline systems in several of the subtasks, suggesting that parameter-efficient fine-tuning is a viable and powerful strategy for addressing linguistic diversity in low-resource and high-variability Indic language datasets. To leverage cross-lingual transfer, a unified model was developed by fine-tuning on a concatenated dataset of seven Indic languages. This approach proved superior to standalone language-specific models, yielding substantial improvements in F1-score (most notably a 28.76 point gain in Subtask 2 for Punjabi language). This provides strong evidence for the benefits of cross-lingual knowledge transfer in low-resource settings.

1 Introduction

Although social media enables interconnectivity, community building, and social awareness, they also introduce significant societal challenges. Online polarization is a prominent example of these adverse effects. Ali et al. (2025) defined polarization as "animosity directed at individuals outside one's group, coupled with a sense of unity and support for those within one's own group". Cruz et al. (2025) mentions that the increase in Internet and social media usage coincides with increase in political polarization throughout the world. Social media enables like-minded people to form groups and factors such as ease of sharing views, speed, anonymity, misinformation, echo-chambers, and filter-bubbles contribute to the growth of online polarization (Sunstein, 2018; Brown et al., 2022). Considering the societal challenges arising from polarization, it is important to detect and control the spread of online polarization.

The shared task "POLAR @ SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multi-event Online Polarization" was organized to detect online polarization (Naseem et al., 2026a). The shared task covered twenty-two different languages from different parts of the world. There were three subtasks in this shared task. The first subtask was a binary classification task that required to determine if a given text contains polarized content. The second subtask required determining the type of polarization present in a given text, the possibilities being Political/Ideological, Racial/Ethnic, Religious, Gender, and Other. It is possible for a given text to be in more than one class. Thus, it was a multi-label classification problem. The third subtask required determining if the polarized text exhibited Stereotype, Vilification, Dehumanization, Extreme Language and Absolutism, Lack of Empathy or Understanding, or Invalidation. This subtask was also a multi-label classification problem.

2 Related Work

Various Natural Language Processing (NLP) techniques have been utilized to detect online polarization. For instance, (Badami et al., 2017) focused on determining polarization within recommender systems by extracting features from user ratings to train a Random Forest classifier, which subsequently calculated polarization scores for different items. Similarly, (Mall et al., 2024) employed a Gradient Boosting regressor to analyze YouTube comments, utilizing features such as view counts

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3002–3009

and share statistics to derive polarization scores.

With the emergence of Large Language Models (LLMs), recent research has shifted toward using these architectures for polarization detection. (Rules et al., 2026) utilized the LLaMA-3.1-70B model to extract features such as stance, affective tone, and agreement patterns from social media discussions. These were then processed by a rule-based system to score affective polarization. Furthermore, (Juvino Santos et al., 2025) explored the generative capabilities of GPT-4 to reduce polarization, where the model was provided with formal definitions of polarization and prompted to paraphrase text to mitigate its divisive tone.

3 Dataset

The POLAR dataset was used for SemEval-2026 Task 9 (Naseem et al., 2026b). Tables 1 through 3 (see Appendix A) present the statistics for this dataset. This study specifically focused on seven Indic languages: Bengali, Hindi, Nepali, Oriya, Punjabi, Telugu, and Urdu. As illustrated in Table 1, the distributions for the Nepali, Punjabi, Telugu, and Bengali datasets in Subtask 1 were relatively balanced. In contrast, the Hindi and Urdu datasets exhibited a significant class imbalance, with the *Polarized* category accounting for approximately 80% and 70% of the samples, respectively. Similarly, the Oriya dataset was skewed, with only 30% of the examples belonging to the *Not Polarized* category.

As illustrated in Table 2, the *Political* category was the dominant class across the majority of the polarized examples. Since the dataset follows a multi-label annotation scheme where a single text may be assigned multiple categories, the percentage totals for some languages exceed 100%. In the Bengali dataset, 34% of the samples were categorized as *Political*, while the *Racial*, *Religion*, and *Gender* categories each represented approximately 1%. A similar trend was observed in the Hindi dataset, where *Political* and *Religion* categories were prominent at 70% and 42%, respectively, while other categories remained negligible at roughly 1%. The Oriya and Punjabi datasets also showed a political bias (21% and 31%, respectively), with other categories ranging between 5% and 10%. The Telugu dataset showed a more distributed spread, with *Political*, *Racial*, and *Other* categories each accounting for approximately 20%. The Urdu dataset displayed high co-occurrence across labels, with *Political* texts at 66% and all other categories maintaining a significant presence at approximately 50% each.

Table 3 illustrates that the manifestations of polarization vary significantly across the seven Indic languages. In the Bengali dataset, *Vilification* was the most frequent method, appearing in 25% of the samples, followed by *Dehumanization* at 11%; all other manifestations occurred in fewer than 5% each. In Hindi, *Vilification* and *Invalidation* each was present in about 65% of the examples, while *Stereotype*, *Extreme Language*, and *Lack of Empathy* each appeared in over 50% of the texts. The Urdu dataset showed the most uniform distribution, with all manifestation categories each appearing in approximately 60% of the samples. In the Nepali dataset, *Vilification*, *Stereotype*, and *Extreme Language* each appeared about 31%, 27% and 25% of the samples. These same categories each appeared in about 10% of the examples in the Oriya dataset. The Punjabi dataset showed a strong presence of *Vilification* (40%), while *Dehumanization*, *Invalidation* and *Extreme Language* each appeared in about 22% of the examples. In the Telugu dataset, *Vilification*, *Invalidation* and *Lack of Empathy* each appeared in about 23% of the examples, while *Stereotype* and *Extreme Language* each appeared in about 12% of the examples.

4 Methodology

This section describes the model that was used in this study, the data preparation phase, and the method used for supervised fine-tuning of the model.

4.1 Model

This study used Qwen3-8B base and instruction-tuned large language model (Yang et al., 2025) to perform classification. The instruction-tuned model supports both thinking and non-thinking mode in the same model. The Qwen3-8B model has a total of 8.2 billion parameters including 6.95 billion non-embedding parameters. The model was trained on 36 trillion tokens covering 119 different languages including the seven Indic languages that this study participated in (Hindi, Bengali, Oriya, Urdu, Telugu, Punjabi, and Nepali). In non-thinking mode, Qwen3-8B demonstrated superior performance on multilingual benchmarks such as MMMLU and INCLUDE compared to LLaMA-3.1-8B-Instruct, while performing only marginally below Gemma-3-12B-IT (Yang et al., 2025).

4.2 Data Preparation

The original dataset consisted of separate files for each of the twenty-two languages. For any given language, the files were further sub-divided into three distinct subtask files. In the data integration phase, the separate files for each subtask were integrated so that for any given text the polarization, type, and manifestation labels were available together. This consolidated structure provided a unified multi-label ground truth for every text, facilitating joint multi-task training. Additionally, to leverage cross-lingual transfer learning, the datasets for the seven Indic languages that this study participated in were concatenated into a single file. This allowed the model to share semantic and cultural representations across linguistically related scripts.

Each integrated example in the dataset was then structured to be in the Chat Markup Language (ChatML) format, consisting of system, user, and assistant roles. The *system message* was used to set the model's persona as an expert in social media analysis with knowledge of cultural nuances for the seven Indic languages. The *user message* was used to specify the target and its language context. The *assistant message* was used to provide the ground truth labels for fine-tuning purposes. Finally, role-specific control tokens (<|*im_start*|>, <|*im_end*|>) and reasoning tokens (<*think*>, </*think*>) were added to the messages to align the data with the processing requirements of the Qwen-3 architecture. An example in Hindi after the data preparation phase is shown in 1. The prompt persona was similarly tailored for each of the seven languages to ensure domain-specific expertise in the target language.

<|im_start|>system
You are an expert linguist specializing
in Hindi cultural nuances and social media
analysis.<|im_end|>
<|im_start|>user
Language Context: Hindi
Text to analyze: #जुम्मा के ढील द्द #शनिचर के
छिल द्द.....😅😂 #बजरंग_दल वाला फ़ील द्द💪 💪
Provide the Task 1, 2, and 3 labels in
JSON format.<|im_end|>
<|im_start|>assistant
<think>

</think>
{"Task1": 1, "Task2": [1, 1, 1, 0, 1],
"Task3": [1, 1, 1, 1, 1]}<|im_end|>
<|im_end|>

Figure 1: A Hindi Example after Data Preparation Phase

4.3 Supervised Fine-Tuning Phase

The examples in the ChatML format, obtained from the Data Preparation phase, were used to fine-tune the Qwen3-8B base and instruct models. To optimize the fine-tuning process for a 16GB VRAM environment (NVIDIA GeForce RTX 3080), the Quantized Low-Rank Adaptation (QLoRA) technique was employed. QLoRA integrates four primary optimization techniques: (i) 4-bit NormalFloat (NF4) quantization, (ii) Double Quantization, (iii) Paged Optimizers, and (iv) Low-Rank Adapters (LoRA). NF4 quantization maps the model's weights to an information-theoretically optimal distribution. As only 4-bits are used to represent the weights, this step achieves a 4:1 compression of the weights reducing the memory requirement for the model. Double Quantization further reduces memory overhead by quantizing the quantization constants themselves. Following quantization, the base model weights were frozen, and trainable LoRA adapters were integrated. Both the LoRA rank (r) and the scaling hyperparameter (α) were set to 16. The target modules for these adapters included all linear layers, specifically the attention projections (q, k, v, o) and the feed-forward network modules (gate, up, down). To ensure maximum memory efficiency, gradient checkpointing was used to reduce activation memory requirements. The Unsloth-optimized implementation of gradient checkpointing was used.

The 8-bit AdamW optimizer was used for weight updates, with the β_1, β_2, and ϵ parameters maintained at their default values of 0.9, 0.999, and 1e-8, respectively. The learning rate was set to 2e-4, and Cross-Entropy was used as the loss function. An effective batch size of 8 was implemented using a physical batch size of 2 and gradient accumulation over four steps. The model was trained for 6780 steps corresponding to 3 epochs over the consolidated Indic language dataset. Model convergence was monitored via validation loss at 100-step intervals. In the case of both Qwen3-8B Instruct and Qwen3-8B Base, the validation loss achieved its lowest point (0.5366 and 0.7259, respectively) after 4500 steps. In case of the Qwen3-8B instruct model, the training process was stopped after 5500 steps. The weights from the 4500-step checkpoint

were selected for final evaluation.

In addition to the two unified model mentioned above, seven language-specific models were trained to evaluate the impact of concatenating the multi-language datasets. Each model was derived by fine-tuning the Qwen3-8B instruct model on each respective Indic language dataset.

5 Results and Discussion

Table 4 presents the results obtained by the fine-tuned Qwen3-8B intruct model used in this study (due to space constraints, detailed performance metrics for all subtasks are provided in Appendix A). The macro F1-score was used as the evaluation metric for this shared task. This table also compares the performance of the Qwen3-8B instruct model against the baseline and top-performing systems from the shared task. Table 5 compares the scores of the Qwen3-8B instruct and base models. The table also compares the scores of these two unified models against the language specific models. Tables 6 and 7 show the absolute F1-score improvement between the Qwen3-8B instruct and base models, as well as the delta between the unified and language-specific instruct models.

As illustrated in table 4, the proposed model outperformed the baseline across several experiments, including: Subtask 3 for Bengali; Subtask 1 for Hindi; Subtasks 1 and 2 for Nepali; Subtask 3 for Oriya and Punjabi; Subtasks 1 and 3 for Telugu; and Subtasks 2 and 3 for Urdu. Detailed analysis of the classification reports revealed that for Bengali Subtask 2, the model struggled to identify *Racial* and *Gender* classes. It can also be seen from table 2, that these two classes were minority classes comprising only 0.75% and 0.5% of the training examples, respectively. Interestingly, the model successfully classified the *Religion* category despite it representing only 2% of the training data. Conversely, in Bengali Subtask 3, the model achieved low macro F1-scores (0.02 to 0.06) for *Stereotype*, *Extreme Language*, *Lack of Empathy*, and *Invalidation*, which were all minority classes.

Regarding the Hindi and Nepali datasets, the model demonstrated well, performing consistently across all categories in Subtasks 2 and 3 despite the inherent imbalances. For Oriya Subtask 3, however, the model exhibited a bias toward the *Extreme Language* category, resulting in poor performance for other manifestations. In the Punjabi dataset, the model remained unbiased and achieved average performance across all categories for Subtasks 2 and 3. Similarly, performance on Telugu Subtasks 2 and 3 was moderate. Finally, the model performed optimally on the Urdu Subtasks 2 and 3; as these datasets were well-balanced, the model successfully learned to classify all target classes with high accuracy.

As illustrated in Table 5, the Qwen3-8B base model outperformed the instruct version in the majority of tested scenarios, achieving higher F1 scores in 14 out of the 21 total subtasks. While the instruct model performed better in Bengali and Telugu, the base model performed better across all three subtasks for Nepali, Oriya, Punjabi, and Urdu. Furthermore, Table 6 reveals significant performance gains for the base model, exceeding 3 percentage points in F1 for Nepali (Subtask 3) and Oriya (Subtasks 1 and 3), and 2 percentage points for Punjabi (Subtasks 1 and 2) and Urdu (Subtasks 2 and 3). These findings suggest that instruction tuning may degrade model performance in multi-label classification tasks for certain Indic languages.

As illustrated in Table 5, the Qwen3-8B instruct-based unified model outperformed the language-specific baselines in 20 out of the 21 tested subtasks. Table 7 further quantifies these improvements, highlighting that the most significant gains occurred in Subtask 2. Specifically, the F1-score for Punjabi rose from 0.2069 to 0.4945 (a gain of 28.76 points), while Oriya and Hindi saw substantial increases of 21.35 and 15.02 points, respectively. In Subtask 1, the unified model achieved gains ranging from 2 to 7.5 percentage points. A gain of 7.5 points was obtained in Subtask 1 of the Punjabi language. Subtask 3 also showed positive trends, with improvements of 9 points in Punjabi and 2 to 3 points in Hindi and Telugu. These results provide strong empirical evidence for cross-lingual transfer. By leveraging a multi-language training corpus, the unified model developed a superior semantic understanding compared to its language-specific counterparts.

This study experimented with data augmentation as a solution for the observed data imbalance. The Qwen3-8B instruct model was used to generate synthetic minority class examples based on seeds from the training set. To vary the tone of the generated text, personas were randomly assigned as either frustrated users using slang language or aggressive commentators. The target volume for each minority class was set to 30% of the majority

class. Despite this expansion of the training set, model performance did not improve significantly, highlighting the need for more robust strategies to handle the data imbalanced problem.

6 Conclusion

This study presented a parameter-efficient approach to detecting online polarization across seven Indic languages using a fine-tuned Qwen3-8B-Instruct model. This study used Quantized Low-Rank Adaptation (QLoRA) and NF4 quantization to fine-tune the model. The experimental results demonstrate that the model is highly effective in identifying polarization, achieving a macro F1-score peak of 0.90 for Nepali and surpassing established baselines in many of the subtasks. The findings of this study also confirm that the unified model is significantly more effective than language-specific models. For example, the unified approach obtained an improvement of 28.76 percentage points in F1-score for Subtask 2 of the Punjabi language. This and the other results obtained provide evidence of cross-lingual transfer, demonstrating that joint training on related Indic languages enhances the model's performance across diverse linguistic contexts. Analysis performed also revealed challenges in multi-label classification for minority classes, particularly in the Bengali dataset, where low F1-socres were obtained for the minority classes such as *Stereotype* and *Lack of Empathy*. In contrast, the model showed remarkable robustness in Urdu, where balanced data allowed for the successful learning of overlapping labels. This research confirms that Supervised Fine-Tuning via QLoRA is a viable and powerful strategy for addressing linguistic diversity in low-resource Indic language datasets. Synthetic data generation for minority classes using Qwen3-8B instruct model did not improve the performance of the model significantly. Future work will focus on addressing these challenges through more robust techniques, such as cost-sensitive learning and advanced resampling methods.

References

Adem Chanie Ali, Seid Muhie Yimam, Abinew Ali Ayele, Chris Biemann, and Martin Semmann. 2025. Silenced voices: social media polarization and women's marginalization in peacebuilding during the northern ethiopia war. *i-com*, 24(2):407–432.

Mahsa Badami, Olfa Nasraoui, Welong Sun, and Patrick Shafto. 2017. Detecting polarization in ratings: An automated pipeline and a preliminary quantification on several benchmark data sets. In *2017 IEEE international conference on big data (big data)*, pages 2682–2690. IEEE.

Megan A Brown, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2022. Echo chambers, rabbit holes, and algorithmic bias: How youtube recommends content to real users. *Available at SSRN 4114905*.

Cesi Cruz, Horacio Larreguy, and Ernesto Tiburcio. 2025. Political polarisation.

Lucas Raniére Juvino Santos, Leandro Balby Marinho, Claudio Elizio Calazans Campelo, Filippo Menczer, and Alessandro Flammini. 2025. Can large language models effectively mitigate polarization in social media text? In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 348–357.

Raghvendra Mall, Mridul Nagpal, Joni Salminen, Hind Almerekhi, Soon-gyo Jung, and Bernard J Jansen. 2024. Politics on youtube: Detecting online group polarization based on news videos' comments. *Sage Open*, 14(2):21582440241256438.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"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 (SemEval-2026)*. ACL.

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.

Heuristic Rules, Jawad Chowdhury, Rezaur Rashid, and Gabriel Terejanu. 2026. Measuring social media polarization using large language models. In *Social Networks Analysis and Mining: 17th International Conference, ASONAM 2025, Niagara Falls, ON, Canada, August 25–28, 2025, Proceedings, Part III*, page 429. Springer Nature.

Cass R Sunstein. 2018. Republic: Divided democracy in the age of social media.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Appendix A: Dataset Details and Experimental Results

Lang	Type	Polar	Not Polar	Total
Bengali	Train	1424 (43%)	1909 (57%)	3333
	Dev	70 (42%)	96 (58%)	166
	Test	868 (58%)	633 (42%)	1501
Hindi	Train	2346 (85%)	398 (15%)	2744
	Dev	112 (82%)	25 (18%)	137
	Test	1114 (69%)	492 (31%)	1606
Nepali	Train	1008 (50%)	997 (50%)	2005
	Dev	51 (51%)	49 (49%)	100
	Test	451 (50%)	452 (50%)	903
Oriya	Train	683 (29%)	1685 (71%)	2368
	Dev	35 (30%)	83 (70%)	118
	Test	303 (28%)	763 (72%)	1066
Punjabi	Train	840 (49%)	860 (51%)	1700
	Dev	47 (47%)	53 (53%)	100
	Test	393 (49%)	416 (51%)	809
Telugu	Train	1274 (54%)	1092 (46%)	2366
	Dev	59 (50%)	59 (50%)	118
	Test	552 (52%)	514 (48%)	1066
Urdu	Train	2476 (69%)	1087 (31%)	3563
	Dev	124 (70%)	53 (30%)	177
	Test	1114 (69%)	492 (31%)	1606
Total	Train	10051 (56%)	8028 (44%)	18079
	Dev	498 (54%)	418 (46%)	916
	Test	4795 (56%)	3792 (44%)	8587

Table 1: Subtask1 Dataset Statistics

Language	Type	Political	Racial	Religion	Gender	Other	Total
Bengali	Train	1134 (34%)	25 (0.75%)	65 (2%)	17 (0.5%)	335 (10%)	3333
	Dev	57 (34%)	1 (0.6%)	3 (2%)	1 (0.6%)	17 (10%)	166
	Test	510 (34%)	12 (0.8%)	29 (2%)	8 (0.5%)	151 (10%)	1501
Hindi	Train	2023 (74%)	333 (12%)	1611 (59%)	315 (11%)	360 (13%)	2744
	Dev	94 (69%)	17 (12%)	57 (42%)	16 (12%)	18 (13%)	137
	Test	1085 (68%)	874 (54%)	885 (55%)	823 (51%)	815 (51%)	1606
Nepali	Train	345 (17%)	281 (14%)	159 (8%)	105 (5%)	236 (12%)	2005
	Dev	17 (17%)	14 (14%)	8 (8%)	5 (5%)	12 (12%)	100
	Test	156 (17%)	127 (14%)	72 (8%)	48 (5%)	106 (12%)	903
Oriya	Train	496 (21%)	119 (5%)	150 (6%)	79 (3%)	87 (4%)	2368
	Dev	25 (21%)	6 (7%)	7 (6%)	4 (3%)	4 (3%)	118
	Test	223 (21%)	54 (5%)	68 (6%)	36 (3%)	39 (4%)	1066
Punjabi	Train	523 (31%)	100 (6%)	134 (8%)	190 (11%)	152 (9%)	1700
	Dev	31 (31%)	6 (6%)	8 (8%)	11 (11%)	9 (9%)	100
	Test	249 (31%)	47 (6%)	63 (8%)	90 (11%)	72 (9%)	809
Telugu	Train	511 (22%)	402 (17%)	212 (9%)	314 (13%)	561 (24%)	2366
	Dev	25 (21%)	20 (17%)	11 (9%)	16 (14%)	29 (25%)	118
	Test	230 (22%)	181 (17%)	95 (9%)	141 (13%)	252 (24%)	1066
Urdu	Train	2395 (66%)	1938 (54%)	1969 (55%)	1825 (51%)	1808 (51%)	3563
	Dev	123 (70%)	96 (54%)	100 (57%)	91 (51%)	90 (51%)	177
	Test	1085 (68%)	874 (54%)	885 (55%)	823 (51%)	815 (51%)	1606
Total	Train	7427 (41%)	3198 (18%)	4300 (24%)	2845 (16%)	3539 (20%)	18079
	Dev	372 (41%)	160 (18%)	194 (21%)	144 (16%)	179 (20%)	916
	Test	3538 (41%)	2169 (25%)	2097 (24%)	1969 (23%)	2250 (26%)	8587

Table 2: Subtask 2 Dataset Statistics

Language	Subtask	Stereotype	Vilification	Dehumanize	Extreme Lang	No Empathy	Invalidation	Total
Bengali	Train	199 (6%)	802 (24%)	357 (11%)	157 (5%)	63 (2%)	59 (2%)	3333
	Dev	10 (6%)	42 (25%)	18 (11%)	8 (5%)	3 (2%)	3 (2%)	166
	Test	89 (6%)	355 (24%)	160 (11%)	71 (5%)	29 (2%)	27 (2%)	1501
Hindi	Train	1364 (50%)	1788 (65%)	500 (18%)	1388 (51%)	1557 (57%)	1802 (66%)	2744
	Dev	68 (50%)	95 (69%)	25 (18%)	69 (50%)	78 (57%)	95 (69%)	137
	Test	998 (62%)	1038 (65%)	892 (56%)	999 (62%)	903 (56%)	430 (27%)	1606
Nepali	Train	537 (27%)	630 (31%)	132 (7%)	544 (27%)	212 (11%)	300 (15%)	2005
	Dev	27 (27%)	31 (31%)	7 (7%)	27 (27%)	11 (11%)	15 (15%)	100
	Test	242 (27%)	286 (31%)	59 (7%)	245 (27%)	95 (11%)	135 (15%)	903
Oriya	Train	236 (10%)	277 (12%)	16 (0.7%)	317 (13%)	37 (2%)	80 (3%)	2368
	Dev	12 (10%)	11 (9%)	1 (0.9%)	16 (14%)	2 (2%)	4 (3%)	118
	Test	106 (10%)	97 (9%)	7 (0.7%)	143 (13%)	17 (2%)	36 (3%)	1066
Punjabi	Train	276 (16%)	687 (40%)	374 (22%)	407 (24%)	211 (12%)	415 (24%)	1700
	Dev	19 (19%)	39 (39%)	22 (22%)	24 (24%)	12 (12%)	24 (25%)	100
	Test	129 (16%)	312 (39%)	178 (22%)	193 (24%)	101 (13%)	198 (25%)	809
Telugu	Train	265 (11%)	536 (23%)	59 (3%)	318 (13%)	622 (26%)	589 (25%)	2366
	Dev	13 (11%)	24 (20%)	3 (3%)	16 14%)	31 (26%)	27 (23%)	118
	Test	120 (11%)	221 (21%)	26 (2%)	143 (13%)	280 (26%)	243 (23%)	1066
Urdu	Train	2218 (62%)	2307 (65%)	1982 (56%)	2215 (62%)	2004 (56%)	2039 (57%)	3563
	Dev	112 (63%)	115 (65%)	99 (56%)	110 (62%)	100 (57%)	101 (57%)	177
	Test	998 (62%)	1038 (65%)	892 (56%)	999 (62%)	903 (56%)	919 (57%)	1606
Total	Train	5095 (28%)	7027 (39%)	3420 (19%)	5346 (30%)	4706 (26%)	5284 (29%)	18079
	Dev	261 (29%)	357 (39%)	175 (19%)	270 (30%)	237 (26%)	269 (29%)	916
	Test	2682 (31%)	3347 (39%)	2214 (26%)	2793 (33%)	2328 (27%)	1988 (23%)	8587

Table 3: Subtask 3 Dataset Statistics

Language	Subtask	Precision	Recall	F1	Baseline F1	Best System F1	Rank
Bengali	Subtask 1	0.8236	0.8247	0.8241	0.8528	0.8625	34/49
	Subtask 2	0.3243	0.2296	0.2437	0.2887	0.4216	24/30
	Subtask 3	0.1712	0.1400	0.1369	0.0868	0.2805	15/21
Hindi	Subtask 1	0.8081	0.7652	0.7841	0.7379	0.8281	33/48
	Subtask 2	0.7759	0.7284	0.7488	0.7911	0.8073	18/30
	Subtask 3	0.6966	0.6851	0.6788	0.7456	0.7709	20/21
Nepali	Subtask 1	0.8992	0.8992	0.8992	0.8798	0.9236	24/45
	Subtask 2	0.7912	0.7441	0.7662	0.7219	0.8104	14/28
	Subtask 3	0.6531	0.5690	0.6013	0.6095	0.7127	13/20
Oriya	Subtask 1	0.7834	0.7459	0.7602	0.7765	0.8255	30/45
	Subtask 2	0.5479	0.3939	0.4526	0.5600	0.6027	18/27
	Subtask 3	0.1820	0.1436	0.1522	0.1314	0.3296	16/21
Punjabi	Subtask 1	0.7782	0.7782	0.7775	0.7898	0.8257	11/45
	Subtask 2	0.5120	0.4811	0.4945	0.3650	0.5526	N/A
	Subtask 3	0.4403	0.4128	0.4166	0.3841	0.5441	18/20
Telugu	Subtask 1	0.8717	0.8704	0.8686	0.644	0.9053	28/45
	Subtask 2	0.3690	0.2919	0.3134	0.3145	0.4647	16/27
	Subtask 3	0.3356	0.2471	0.2780	0.2205	0.4445	13/21
Urdu	Subtask 1	0.8046	0.7742	0.7863	0.7890	0.8196	14/47
	Subtask 2	0.6844	0.8636	0.7624	0.7127	0.7978	15/30
	Subtask 3	0.7321	0.8792	0.7987	0.7693	0.8213	10/21

Table 4: Official Results

Language	Subtask	Qwen3 Instruct (Unified Model)			Qwen3 Base (Unified Model)			Qwen3 Instruct (Language Specific)		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Bengali	Subtask 1	**0.8236**	**0.8247**	**0.8241**	0.8185	0.8160	0.8171	0.8029	0.7604	0.7673
	Subtask 2	**0.3243**	**0.2296**	**0.2437**	0.3092	0.1898	0.1975	0.1849	0.1255	0.1486
	Subtask 3	0.1712	**0.1400**	0.1369	0.1188	0.1318	0.1079	**0.1794**	0.1180	**0.1412**
Hindi	Subtask 1	0.8081	0.7652	0.7841	0.8443	**0.7746**	**0.8034**	**0.8496**	0.7200	0.7631
	Subtask 2	0.7759	**0.7284**	**0.7488**	**0.7914**	0.6490	0.6935	0.5963	0.6095	0.5986
	Subtask 3	0.6966	0.6851	0.6788	**0.7456**	**0.7269**	**0.6964**	0.6415	0.6916	0.6525
Nepali	Subtask 1	0.8992	0.8992	0.8992	**0.9004**	**0.9003**	**0.9003**	0.8806	0.8781	0.8780
	Subtask 2	0.7912	0.7441	0.7662	**0.7919**	**0.7708**	**0.7800**	0.6831	0.6310	0.6532
	Subtask 3	0.6531	0.5690	0.6013	**0.7157**	**0.5991**	**0.6401**	0.6042	0.5972	0.6005
Oriya	Subtask 1	0.7834	0.7459	0.7602	**0.8029**	**0.7876**	**0.7945**	0.7840	0.7340	0.7514
	Subtask 2	0.5479	0.3939	0.4526	**0.5801**	**0.4093**	**0.4581**	0.3677	0.1972	0.2391
	Subtask 3	0.1820	0.1436	0.1522	**0.2313**	**0.2180**	**0.1853**	0.1610	0.1408	0.1463
Punjabi	Subtask 1	0.7782	0.7782	0.7775	**0.8032**	**0.8002**	**0.7982**	0.7276	0.7066	0.7025
	Subtask 2	0.5120	0.4811	0.4945	**0.5611**	**0.4928**	**0.5169**	0.3129	0.1693	0.2069
	Subtask 3	0.4403	0.4128	0.4166	**0.4652**	**0.4243**	**0.4238**	0.4209	0.2640	0.3181
Telugu	Subtask 1	0.8717	**0.8704**	**0.8686**	**0.8749**	0.8703	0.8675	0.8480	0.8484	0.8480
	Subtask 2	0.3690	**0.2919**	**0.3134**	**0.4184**	0.2536	0.2972	0.3064	0.2481	0,.2656
	Subtask 3	0.3356	**0.2471**	**0.2780**	**0.4105**	0.1954	0.2168	0.2660	0.2277	0.2421
Urdu	Subtask 1	0.8046	0.7742	0.7863	**0.8319**	**0.7822**	**0.8001**	0.7924	0.7335	0.7517
	Subtask 2	0.6844	0.8636	0.7624	**0.6850**	**0.9370**	**0.7892**	0.6616	0.8727	0.7516
	Subtask 3	0.7321	0.8792	0.7987	**0.7323**	**0.9371**	**0.8215**	0.7055	0.8918	0.7874

Table 5: Comparison of Qwen3 Instruct (Unified Model), Qwen3 Base (Unified Model) and Qwen3 Instruct (Language Specific Models)

Language	Subtask 1 (Δ)	Subtask 2 (Δ)	Subtask 3 (Δ)
Bengali	-0.0070	-0.0462	-0.0290
Hindi	+0.0193	-0.0553	+0.0176
Nepali	+0.0011	+0.0138	+0.0388
Oriya	+0.0343	+0.0055	+0.0331
Punjabi	+0.0207	+0.0224	+0.0072
Telugu	-0.0011	-0.0162	-0.0612
Urdu	+0.0138	+0.0268	+0.0228
Average	+0.0116	-0.0070	+0.0042

Table 6: Absolute F1-Score Improvement (Δ) of Qwen3-Base over Qwen3-Instruct Models across Indic Languages

Language	Subtask 1 (Δ)	Subtask 2 (Δ)	Subtask 3 (Δ)
Bengali	+0.0568	+0.0951	-0.0043
Hindi	+0.0210	+0.1502	+0.0263
Nepali	+0.0212	+0.1130	+0.0008
Oriya	+0.0088	+0.2135	+0.0059
Punjabi	+0.0750	+0.2876	+0.0985
Telugu	+0.0206	+0.0478	+0.0359
Urdu	+0.0346	+0.0108	+0.0113
Average	+0.0340	+0.1311	+0.0249

Table 7: Cross-lingual Transfer Gains: Absolute F1-Score Improvement (Δ) of Unified over Language-Specific Models