PolDeck at SemEval-2026 Task 9: Multilingual Online Polarization Detection via Hybrid Model Ensembling and Data Augmentation

Ben Grandy and Daniel Khir
University of Augsburg, Germany
{ben.grandy,muhammad.bin.mohd.khir}@uni-a.de

Abstract

In this paper, we address SemEval 2026 Task 9: Multilingual Online Polarization Detection. We present our hybrid ensemble framework, integrating few-shot prompting with Qwen3-30B, a native multilingual XLM-R encoder, and a translation-augmented DeBERTa encoder. To mitigate label imbalance, we implement a multi-stage augmentation pipeline leveraging LLMs for synthetic paraphrasing and cross-lingual translation. Our system ranked in the Top 10 on the English and German leaderboards, proving that integrating both high-capacity monolingual models and flexible multilingual models in a holistic system is a viable method for detecting online polarization. Our code is available on GitHub[1].

1 Introduction

The proliferation of social media has significantly accelerated the spread of online discourse, bringing with it the challenge of monitoring vast quantities of user-generated content for harmful language. While NLP research has traditionally prioritized overt hate speech detection, attitude polarization presents a more nuanced challenge characterized by subtle, context-dependent divisive language. This phenomenon fosters societal fragmentation by reinforcing group mistrust, yet its detection is hampered by the scarcity of high-quality annotated data. Consequently, state-of-the-art transformer-based models must evolve to capture these complex semantic signals that go beyond simple keyword matching. In this paper, we address the challenges of SemEval Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization (Naseem et al., 2026a).

We propose a hybrid ensemble framework for the detection and classification of multilingual online polarization. The framework integrates three parallel inference streams: (1) few-shot prompting with Qwen3-30B, (2) a native multilingual XLM-R encoder, and (3) a translation-augmented DeBERTa encoder. The streams are combined by a weighted ensemble calibrated with per-label temperature scaling.

To overcome data limitations and enhance model robustness, our approach centers on the strategic application of Large Language Models (LLMs) for data enrichment. First, we utilize LLMs to generate additional synthetic training data, exposing our models to diverse polarized contexts. Second, we implement a cross-lingual pipeline specifically for the DeBERTa encoder, using LLMs to translate German texts into English, allowing the encoder model to benefit from the broader semantic knowledge captured in English-centric pre-training.

Our system placed within the Top 10 for the English and German languages in subtasks 1 and 2, with peak Macro-F1 scores of 0.8189 for English and 0.7317 for German in the binary classification task (Subtask 1). Furthermore, the application of synthetic data augmentation proved particularly vital for the multi-label classification task (Subtask 2), where it yielded a performance increase of approximately 3 percentage points compared to non-augmented models.

While individual language models provide a robust foundation for semantic reasoning, our findings additionally show that the integration of calibrated encoder outputs via optimized weighting is central in significantly enhancing final classification reliability in multilingual contexts.

2 Background

Task 9 involves identifying polarization in multilingual online interactions, characterised by language that encourages group-based division and intolerance. We participate in two subtasks for the English and German languages.

[1] https://github.com/danielkhir/poldeck-semeval-task-9

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 879–885

Subtask 1 (Polarization Detection). A binary classification task to determine if a given social media post contains polarized content. For example, a post such as "Fascist oligarchs now control the USA" is labelled as polarized (1), whereas a neutral informational post like "House drafts bill to strike Iran proxies amid Israel-Hamas war" is labelled as non-polarized (0).

Subtask 2 (Polarization Type Classification). A multi-label classification task to identify targets of polarization. Up to five different categories may be found in a single text. For instance, the text "Find yourself a west bank settler gf" indicates both political and gender polarization. Refer to the dataset paper for a complete list of the categories and their definitions (Naseem et al., 2026b).

The task dataset contains texts from a total of 22 languages, extracted from multiple social media platforms such as X, BlueSky, Reddit, and regional news forums. The texts feature discourse regarding global and regional events, such as elections, war, and social movements.

2.1 Related Work

Detecting and classifying polarization in multilingual online texts exhibits similarities to sentiment and stance analysis tasks. Transformer-based architectures and Large Language Models (LLMs) have been shown to excel in this domain due to their capability in modelling implicit semantic dependencies (Conneau et al., 2020; Ji et al., 2025).

Fine-tuning multilingual pretrained language models, such as XLM-RoBERTa (Conneau et al., 2020), provides competitive baselines but suffers from the "curse of multilinguality". Concretely, learning simultaneous representations for multiple languages results in a performance trade-off when compared to dedicated monolingual architectures.

Attempts to overcome this problem include cross-lingual transfer learning, such as the approach demonstrated by Schuster et al. (2019), where training data in the source language is directly translated to the target language, outperforming cross-lingual embeddings when the target language data is limited. This is especially effective for linguistic neighbours such as English and German, where there is minimal structural divergence that hinders zero-shot transfer (Lauscher et al., 2020).

Besides fine-tuning, direct prompting of LLMs have also been shown be effective for identifying nuanced concepts like polarization, especially when leveraging advanced strategies like Chain-of-Thought prompting (Ji et al., 2025).

To leverage both the discriminative precision of fine-tuned language models and the emergent reasoning of LLMs, we propose a hybrid ensemble framework. We specifically address the multilingual challenge by pairing a multilingual model with a monolingual architecture augmented via LLM-generated translations.

3 System Overview

Our pipeline is organised into three parallel branches, as illustrated in Figure 1. The top stream employs the Qwen3-30B-Instruct LLM to derive classification labels via few-shot prompting. The middle stream uses a native multilingual XLM-RoBERTA model fine-tuned on the original source texts. The bottom stream leverages a monolingual DeBERTa model fine-tuned on English data and augmented with German-to-English translations generated by Qwen3-30B-Instruct.

We compute a final prediction by ensembling the streams through a learned weighted average of temperature-scaled logits from the fine-tuned models and the log-probabilities of the classification tokens from the LLM. Furthermore, we implement a synthetic data augmentation strategy using the same LLM to paraphrase existing texts to mitigate the class label imbalance present in each subtask.

3.1 Fine-tuning

We select two transformer-based encoders for the discriminative branches of the ensemble.

Multilingual Encoder (XLM-R). We use XLM-RoBERTa-base[2] as our multilingual model due to its robust cross-lingual transfer capabilities. As no language-specific architectures are required, it also serves as our baseline for comparison.

Monolingual Encoder (DeBERTa). To account for the curse of multilinguality, we incorporate DeBERTa-v3-base[3]. We chose this model due to its gradient-disentangled embedding sharing and disentangled attention mechanism, which have demonstrated empirical improvements over standard BERT and RoBERTa architectures (He et al., 2021).

For both encoders, we opt for the *base* model variants due to limited GPU compute and to speed

[2] https://huggingface.co/FacebookAI/xlm-roberta-base

[3] https://huggingface.co/microsoft/deberta-v3-base

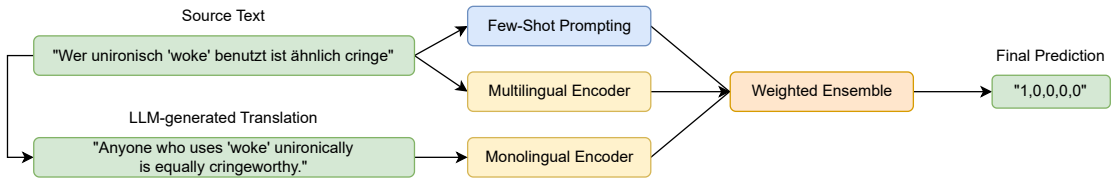


Figure 1: Our system pipeline.

up training cycles, allowing for faster evaluation and iteration during the ensemble tuning process.

3.2 Prompting

For the generative branch of the ensemble, we use Qwen3-30B-A3B-Instruct-2507[4] due to its state-of-the-art performance in multilingual instruction-following and reasoning benchmarks (Yang et al., 2025). We opt for the non-thinking variant of the 30B model to minimise inference latency.

We construct system prompts that instruct the model to role-play as an experienced social media content moderator. The prompt includes the definition of attitude polarization provided in the task documentation, supplemented by few-shot examples to enforce the required output format. We also explicitly remind the LLM to consider the context and overall meaning of the text, ensuring that the model is aware of the necessary nuance.

Rather than parsing free-text responses, we constrain the model to output specific binary tokens for Subtask 1 (e.g. "0" or "1") and a sequence of labels for Subtask 2 (e.g. "0,1,1,0,0"). We then extract the relative log-probabilities of these tokens at the relevant positions as features for the ensemble. As demonstrated by Kauf et al. (2024), soft probabilistic input provides a more nuanced representation of the model's confidences compared to discrete labels.

3.3 Data Augmentation

We implement a multi-stage data augmentation pipeline using Qwen3-30B-A3B-Instruct-2507 to improve the robustness of the discriminators by mitigating the label imbalance in the datasets.

Synthetic Paraphrasing. We employ few-shot prompting with similar instructions in Subsection 3.2 to generate synthetic texts for each subtask. We append 4,000 paraphrased texts for Subtask 1, split equally across both classes. For Subtask 2, we append 2,000 paraphrased texts with at least one polarization category per text. This prevents the encoders from overfitting to a majority class or specific keywords.

Cross-lingual Translation. We translate the German source and paraphrased texts into English for fine-tuning the monolingual DeBERTa encoder. By using an LLM for translation rather than machine translation tools, we aim to maintain the pragmatic meaning and social media tone of the original German posts, which are often lost in literal translations. We perform the translation step after the paraphrasing step to ensure that linguistic variety is sufficiently captured before being mapped into English.

3.4 Hybrid Ensemble

We implement a calibrated weighted ensemble to combine the predictions from our parallel inference branches. To ensure the outputs from the different architectures are comparable, we first follow Guo et al. (2017) and perform temperature scaling for each label to reduce overconfidence in the encoder models. In parallel, we extract the log-probabilities of the target classification tokens from the LLM. Both the scaled logits and log-probabilities are projected into the same probability space to create a consistent feature set for learning per-label weights.

We treat the identification of per-label temperature and ensemble weights as a constrained optimization problem. Concretely, we minimize the cross-entropy loss on the development set using the iterative Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm.

For the multi-label classification in Subtask 2, we extend this optimization by performing a grid search to identify optimal decision thresholds per label. During inference, the system generates a final score for each label by multiplying the model-specific weight with its corresponding probability. A label is only then assigned if the aggregated score exceeds the learned threshold.

[4] https://huggingface.co/unsloth/Qwen3-30B-A3B-Instruct-2507-GGUF

4 Experimental Setup

We fine-tune the encoder models on the training split with and without augmentation, and use the LLM for generative tasks, including few-shot prompting, synthetic data augmentation, and translation. The final ensembles are calibrated using the development set.

Fine-tuning Hyperparameters. To ensure comparability across encoder architectures, we employ a standardized training configuration. We use a base learning rate of 3×10^{-5} for the original dataset. When training on augmented data, we reduce the learning rate to 2×10^{-5} to mitigate overfitting and account for the increased dataset size. We fine-tune both the DeBERTa and XLM-R encoders for 3 epochs. We select the best model for integration into the final hybrid ensemble by retaining the checkpoint that achieves the highest Macro-F1 score on the development set.

Generative Configuration. We perform all generative tasks using Qwen3-30B-A3B-Instruct-2507 deployed via llama.cpp[5]. We maintain the default temperature value of 0.7 to ensure diversity in the paraphrasing and translation.

Evaluation Metrics. We report the per-language Macro-F1 score for English and German on the test splits of subtasks 1 and 2. The Macro-F1 computes the unweighted mean of the per-label F1 scores.

5 Results

The results of our experiments on the test set for both Subtask 1 (Polarization Detection) and Subtask 2 (Polarization Type Classification) are summarized in Table 1.

We compare the individual performance of both encoder approaches (DeBERTa and XLM-R) and the few-shot prompting approach (Qwen3-30B) against the baselines reported by Naseem et al. (2026b), followed by an error analysis and discussion of the impact of our hybrid ensemble and data augmentation strategies.

Subtask 1 (Polarization Detection). Our ensemble approach achieved the highest performance across both languages, reaching a peak Macro-F1 of 0.8189 for English and 0.7317 for German. The ensemble consistently outperformed the best encoder models by approximately 3 to 4 percentage points in both language sets. This validates our hypothesis that a hybrid ensemble approach is capable of unifying the high-precision discriminative features of encoders with the semantic reasoning capabilities of generative models.

For English texts, the DeBERTa model provide the best scores across the board, achieving a maximum Macro-F1 of 0.7955. Notably, few-shot prompting with Qwen3-30B model delivered highly competitive results, trailing DeBERTa by only 4 percentage points without requiring any task-specific fine-tuning. However, only the DeBERTa model and the ensemble performed better than the baseline.

Despite observing a general performance decrease on German texts compared to English, our methods all performed better than the baseline for German. Interestingly, our cross-lingual translation strategy for DeBERTA outperformed the native multilingual XLM-R model by 2 percentage points. This suggests that the curse of multilinguality can be mitigated by leveraging LLMs for text translation.

Subtask 2 (Polarization Type Classification). This subtask proved significantly more challenging due to its multi-label nature. However, our approach using learned decision thresholds resulted in models that achieve a large performance increase over the baseline.

Surprisingly, while translation-based pipelines led in Subtask 1, XLM-R demonstrated superior performance among individual encoder models in Subtask 2. For both languages, the XLM-R encoder consistently outperformed the other individual approaches, achieving a Macro-F1 of 0.3928 on the non-augmented English dataset and 0.4797 on the augmented dataset.

This suggests that while translation is highly effective for binary classification, the multi-label categorization of specific polarization targets may benefit from the broad multilingual pre-training of XLM-R, which captures diverse linguistic and cultural markers of identity more effectively than a translate-then-classify approach.

However, the ensemble approach remained the overall top performer, outperforming the individual XLM-R baseline by a significant margin. Specifically, the ensemble reached a Macro-F1 of 0.4589 in English and 0.5251 in German, proving that integrating multilingual representations into a holistic ensemble is an effective strategy for handling multi-label polarization detection.

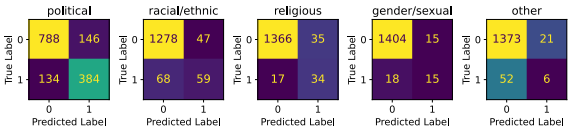
Impact of Data Augmentation. Data augmentation negatively impacted the results for Subtask 1 across all approaches. In contrast, a significant up-

[5] https://github.com/ggml-org/llama.cpp

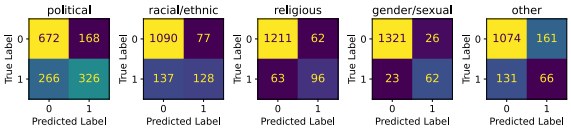
Subtask	Model	ENG_{test}		DEU_{test}	
		Regular	Augmented	Regular	Augmented
1	Baseline	0.7802	–	0.6714	–
	XLM-R	0.7745	0.7560	0.6980	0.6941
	DeBERTa	0.7955	0.7847	0.7123	0.7006
	Qwen3-30B	0.7593	–	0.7063	–
	Ensemble	**0.8189**	0.7970	**0.7317**	0.7271
2	Baseline	0.3333	–	0.4078	–
	XLM-R	0.3928	0.4797	0.4200	0.5719
	DeBERTa	0.3700	0.4964	0.4198	0.5575
	Qwen3-30B	0.4429	–	0.5057	–
	Ensemble	0.4589	**0.4847**	0.5251	**0.5558**

Table 1: Per-language Macro-F1 scores for Subtasks 1 and 2 on the test set. Encoder models are fine-tuned on either the regular training split or the augmented dataset. Bold numbers indicate the leaderboard submissions for each subtask-language pair.

lift in the score is observed for all approaches in Subtask 2. This indicates that while synthetic data is highly beneficial for addressing class sparsity in the multi-label setting of Subtask 2, the original training distribution was already sufficient for the binary classification task.



(a) ENG_{test}.



(b) DEU_{test}.

Figure 2: Per-label performance of the augmented ensemble on the Subtask 2 test split.

Error Analysis. The confusion matrices in Figure 2 reveals that the ensemble consistently struggles with recall, frequently failing to identify secondary polarization categories when multiple types are present. For example, the system fails to detect racial/ethnic polarization in primarily political texts.

Furthermore, we observe precision issues where polarization types are falsely flagged due to the presence of specific keywords. Selected examples are provided in Table 2. For example, neutral factual texts are often incorrectly labelled as containing political polarization due to the usage of political terminology. Similarly, texts focused on political polarization are misclassified as racial/ethnic polarization when words such as *racism, country,* or *apartheid* appear, suggesting that the models rely on surface-level associations rather than societal or cultural context.

In the German subset, we notice a higher rate of false positives for the *other* category. The ensemble appears to use this label as a general label for German political topics. This behaviour may reflect a conceptual misalignment between the annotators and the model. For instance, annotators may categorize environmental discourse as a political topic, whereas the individual models of the ensemble may not have learned this specific connection.

6 Conclusion

In this paper, we presented our system for SemEval-2026 Task 9, focusing on the detection and categorization of online polarization in English and German. Our methodology centered on a hybrid ensemble that combined the strengths of specialized encoder architectures like DeBERTa and XLM-R with the generative reasoning capabilities of LLMs like Qwen3-30B.

Our experimental results highlight two key takeaways for the community. First, we demonstrated that **ensembling is critical** for nuanced tasks. By combining the probabilistic outputs of diverse models, we consistently outperformed any single-model architecture. Second, we found that **augmentation strongly benefits complex, data-scarce tasks**, with LLM-generated synthetic data proving essential for Subtask 2 to mitigate the challenges of label imbalance.

To conclude, our findings suggest that integrating both high-capacity monolingual models and

Language	Text	Misclassification	Logic
ENG	*"Thats like liberals suddenly embracing the conservative values."*	Political	Keyword (Political terminology)
	"It should be deportation to his ancestral country of origin!"	Racial/ethnic	Keyword (Surface-level association)
DEU	*"Rückwärts fahren fürs Klima!"*	Other	Misalignment (Annotator vs. Model)
	"Dieser ganze Protest ist halt lächerlich und nutzlos"	Other	Catch-all (Confidence failure)

Table 2: Selected false positives in Subtask 2, grouped by language.

flexible multilingual models in a holistic system is the way forward for capturing the nuanced markers inherent in online polarization. In future work, we aim to explore more nuanced augmentation strategies and the possibilities of scaling our approach to lower-resource languages.

Limitations

Our experiments were restricted to English and German texts, and similar performance may not be observed in low-resource or non-Germanic languages. Methodically, using an ensemble results in larger computational overhead during inference compared to individual models, while also decreasing system explainability.

Furthermore, our data augmentation strategy lacks a comprehensive automated filtering and validation step, which may have impacted the consistency of the results in Subtask 1. Our analysis was also limited to three specific model families, therefore a broader investigation into alternative architectures may yield better results.

Finally, our error analysis revealed that the models frequently struggle with recall, particularly in the multi-label setting of Subtask 2. Future work could benefit from a multi-step pipeline that first performs binary detection to detect if text is polarized before attempting type classification. Alternatively, the multi-label task could be reformulated as a one-vs-all approach to improve the recall of the individual classifiers.

Acknowledgements

We would like to thank the Chair of Computational Linguistics at the University of Augsburg for the opportunity to participate in this shared task as part of the Search Engines and Neural Information Retrieval course. We also express our gratitude to our coursemates for their insightful discussions and supportive exchange of ideas.

References

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. *Preprint*, arXiv:1911.02116.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Yanxu Ji, Jinzhong Ning, Yijia Zhang, Zhi Liu, and Hongfei Lin. 2025. LLM-driven implicit target augmentation and fine-grained contextual modeling for zero-shot and few-shot stance detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5872–5884, Suzhou, China. Association for Computational Linguistics.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A. Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. *Preprint*, arXiv:2403.14859.

Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *CoRR*, abs/2005.00633.

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)*. 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. *arXiv preprint arXiv:2505.20624*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.