mdok-style at SemEval-2026 Task 9: Finetuning LLMs for Multilingual Polarization Detection

Dominik Macko[1], **Alok Debnath**[2], **Jakub Simko**[1]
[1] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
[2] ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin
{dominik.macko, jakub.simko}@kinit.sk, {alok.debnath}@adaptcentre.ie

Abstract

SemEval-2026 Task 9 is focused on multilingual polarization detection. Specifically, it covers the identification of multilingual, multicultural and multievent polarization along three axes (in subtasks), namely detection, type, and manifestation. Online polarization presents a concern, because it is often followed by hate speech, offensive discourse, and social fragmentation. Therefore, its detection before it escalates is crucial for a safer and more inclusive online space. We have coped with this SemEval task by finetuning mid-size LLMs for the sequence-classification task using the QLoRA parameter-efficient finetuning technique. The training data augmented the multilingual (22 languages) training sets by anonymized, lower-cased, upper-cased, and homoglyphied counterparts, making the detection more robust.

1 Introduction

Automated detection of online polarization is an important problem, as it could help to mitigate polarization before it escalates. This challenge is dealt with by the POLAR shared task (Naseem et al., 2026b), officially SemEval-2026 Task 9. Within its three subtasks, it is aimed at binary polarization detection (subtask 1, determining whether a given text expresses polarization), polarization type classification (subtask 2, identifying the social dimension underlying the polarization), and manifestation identification (subtask 3, detecting how polarization is rhetorically manifested). All of them are handled multilingually, covering 22 languages.

The submitted multilingual system[1] is heavily based on mdok (**m**achine **d**etector **o**f **K**InIT), a robust detector of machine-generated text presented at PAN@CLEF2025 (Bevendorff et al., 2025). Besides a simple transferring of the existing system to a new task, we also explore possibilities of using appraisal-theoretic[2] detection of emotions (including various appraisal dimensions, e.g., pleasantness, self-control, or suddenness) (Hofmann et al., 2020; Debnath et al., 2025), for the tasks of polarization detection.

2 Background

We have utilized our experience in a binary machine-generated text detection task and transferred our best approach to polarization detection task. Our first experience with finetuning LLMs for a binary classification task was at SemEval-2024 Task 8 (Spiegel and Macko, 2024b), where we have identified better detection performance achievable by finetuned a 7B-parameters LLM than the traditional small pre-trained models (BERT-like). We have further explored and increased the robustness of the finetuning process (Macko et al., 2025), and eventually resulted into the mdok (Macko, 2025) finetuning approach, ranking 1st in both subtasks of PAN@CLEF2025 (Bevendorff et al., 2025). In one of the subtasks, mdok was extended to the multiclass scenario (hybrid human-AI collaboration) and further modified to multilingual authorship attribution (Cava et al., 2025). In this SemEval-2026 Task 9 shared task (Naseem et al., 2026a), we have modified the multiclass mdok version to the multiclass multi-label classification required in the subtasks 2 and 3.

3 System Overview

The proposed system contains multiple components that are described below. The appraisal annotation has not been used in the submitted system (due to timing); however, we include it as an alternative we have explored.

[1] https://github.com/kinit-sk/mdok-style-polar2026

[2] appraisals are cognitive signals of emotion

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 314–321

3.1 Data Augmentation

To increase the number of training samples and to make the detection more robust, we augment the original training data. In our approach for binary polarization detection (subtask 1), we use various augmentation techniques described below. Essentially, for each technique, the training texts are duplicated and the duplicated part is modified by the corresponding technique. Afterwards deduplication takes care about the redundant (unmodified) texts.

Anonymization. This text-prepocessing procedure replaces identified email addresses, user mentions, and phone numbers by the tags of [EMAIL], [USER], [PHONE]. Such an anonymization approach reduces biases towards specific personal information, which could appear in the training data and focuses more on the meaning of the text. We use this technique as a data augmentation, duplicating the original texts, preprocessing by this anonymization procedure and then combining with the original texts.

Lower-casing and Upper-casing. To further make the detection more sensitive on actual meaning than on the visual appearance of the texts (especially important on social media using wilder style of communication), we convert the duplicated texts to all lowercase, and also to all uppercase. This effectively increases the training size, making the decision of the detection case insensitive.

Homoglyphication. In our previous work on machine-generated text detection, we have identified homoglyph attack (replacing some characters to visually similar counterparts of another script) as making one of the highest confusion on text classifiers (disrupting their tokenization). Such modification causes detection evasion in many cases; however, can be effectively dealt with when considered during training. We have applied such homoglyph attack to the duplicated portion of the training texts, making the polarization detection more robust against obfuscation.

3.2 Finetuning Process

For finetuning the multilingual LLM for polarization task, we have combined the train samples of all languages, which effectively increases the train samples count and enhances the focus on the task rather than on the language.

Since the provided development-set samples are imbalanced across languages, we have combined the provided development data with the training data and de-duplicated such a train set. Afterwards, for subtask 1, we have sampled 100 texts per each language and label (i.e., 200 samples per language $\times$ 22 languages) resulting in 4400 samples for validation during the finetuning. For subtasks 2 and 3, we have just pseudorandomly sampled 100 texts per language, reflecting multi-labels distribution of the train set (i.e., not perfectly balanced).

In subtask 1, we have increased the train set by above mentioned four augmentation technique by 20% (i.e., 5% each). It makes the trained classifier more robust (invariant to the visual form of the text), focusing on the meaning rather than on representation.

For all subtasks, we have used QLoRA (Dettmers et al., 2023) parameter-efficient finetuning (PEFT) approach with 4-bit quantization using bitsandbytes library within HuggingFace transformers[3] python framework. Regarding hyperparameters, we have applied constant learning rate of 2e-5, a warmup ratio of 0.03, paged Adamoptimizer, a batch size of 1 sample, and validation each 500 steps. The finetuning process has taken a single epoch with the final checkpoint selection based on the best metric achieved on validation set (AUC[4] for subtask 1, macro avg. F1 for subtasks 2 and 3).

We have published the full source code to retrain the detectors (see the footnote on the first page), making the detectors replicable. To install the dependencies, just clone the IMGTB framework (Spiegel and Macko, 2024a) repository[5], install the corresponding conda environment, and update the transformers library (for support of the newest models).

3.3 Base Model Selection

We have focused on a single model for all the languages (22 in total) combined; therefore, we have limited the base-model selection to multilingual models that can cope with low-resource languages as well. For subtask 1, we have experimented with multiple models, while Qwen3-32B (Yang et al., 2025) (support of >100 languages) has been selected to be the best performant. For subtasks 2 and 3, we had problems with multi-label classification

[3] https://github.com/huggingface/transformers
[4] area under curve of the receiver operating characteristic
[5] https://github.com/kinit-sk/IMGTB

finetuning of Qwen3 models; therefore, we have resulted with Gemma-3-27B-pt (Team et al., 2025) (support of >140 languages) in these subtasks.

3.4 Appraisal Annotation

Appraisal theory is an evolutionary theory of emotion that suggests that emotions are outcomes of how individuals “appraise” or value an event or experience on cognitive dimensions such as its *pleasantness*, *predictability*, and the *control* the experiencer has on the situation.

The ISEAR corpus is a benchmark dataset of categorical emotion annotations for private perspectives on emotional scenarios to understand the causes and reactions of emotions as represented in text (Scherer and Wallbott, 1994). For multilingual appraisal analysis, the **enISEAR** and **deISEAR** datasets were used (Troiano et al., 2019a). Both corpora provide crowd-sourced classification annotations for seven emotion categories (anger, disgust, fear, sadness, shame, joy, and guilt), as well as binary ratings for five appraisal categories (consequences to self, consequences to others, degree of control, degree of responsibility, alignment with social values), and four event descriptions (general, past, future, or prospective). Both deISEAR and enISEAR comprise 1001 event-centered descriptions: deISEAR includes 1084 sentences, while enISEAR contains 1366 sentences.

Appraisal estimation was treated as a multitask regression task, with an architecture similar to AppraisePLM (Debnath et al., 2025), except that for handling multilingual inputs, the model uses an XLM-Roberta (Conneau et al., 2020) encoder representation for text embeddings. The loss function is also adjusted to account for the binary nature of the appraisal and event description categories, using a weighted combination of MSE and BCE loss.

A more detailed description of appraisal theory, the appraisal dimensions used for analysis, and the AppraisePLM architecture are provided in Appendix B.

4 Experimental Setup

In each subtask, the submitted systems used the official train and development sets for training as mentioned above and the official test set for testing.

The official metric for all subtasks is the macro average of per-class F1 scores (F1 is a harmonic mean of precision and recall), separately for each language. Percentile value is calculated as a number of systems ranked worse than our system our of all systems.

For exploration of usage of appraisals annotations for polarization detection, we have annotated only the train set (due to time constraints), which has been further split for training (80% of data) and testing (20% of data) of the LogisticRegression classifier with default parameters (random state of 42). For each language, a separate classifier has been trained. The results of predictions has been evaluated per subtask also using macro F1 metric.

5 Results

Performance of the submitted system (trained as described above) in a form of macro avg. F1 (Macro F1) is provided in Table 1. Darker color of the background gradient reflects higher performance. Four languages without the performance value (-) have not been included in subtask 3.

Based on the lowest achieved macro F1 value in subtask 1 (0.63 in case of Khmer), we can see that the submitted detection system can distinguish the two classes (very well above chance). It performs much worse in the other two subtasks, where multi-label nature makes the detection challenging. Only

	Macro F1		
Language	**Subtask 1**	**Subtask 2**	**Subtask 3**
amh	0.6619	0.5116	0.4310
arb	0.8348	0.6279	0.5157
ben	0.8415	0.3050	0.1272
deu	0.7398	0.5399	0.4044
eng	0.8058	0.4519	0.3697
fas	0.7690	0.5250	0.3208
hau	0.7401	0.1689	0.0000
hin	0.7974	0.7573	0.7453
ita	0.7303	0.3019	-
khm	0.6293	0.6323	0.2482
mya	0.8788	0.6835	-
nep	0.8915	0.8026	0.5669
ori	0.8013	0.4628	0.1227
pan	0.7736	0.3734	0.4933
pol	0.8158	0.5350	-
rus	0.8077	0.4952	-
spa	0.7788	0.6256	0.3965
swa	0.7658	0.4226	0.4570
tel	0.8818	0.2573	0.2143
tur	0.8008	0.5692	0.4125
urd	0.7743	0.7791	0.8108
zho	0.9237	0.8199	0.4912

Table 1: Per-language performance of the submitted system based on official results.

	Subtask 1	Subtask 2					Subtask 3					
Lang.	**polarization**	**political**	**racial/ ethnic**	**religious**	**gender/ sexual**	**other**	**stereotype**	**vilification**	**dehuman- ization**	**extreme language**	**lack of empathy**	**invalidation**
amh	0.6619	0.7447	0.7533	0.7180	0.5987	0.6815	0.6805	0.7282	0.6927	0.6003	0.4516	0.5488
arb	0.8348	0.8382	0.7787	0.8542	0.7577	0.6902	0.8096	0.8322	0.7176	0.8084	0.6004	0.5111
ben	0.8415	0.8134	0.4972	0.6513	0.5987	0.6036	0.4847	0.7098	0.4899	0.5738	0.4951	0.4955
deu	0.7398	0.7222	0.7552	0.8059	0.8018	0.5470	0.7138	0.6884	0.6239	0.6780	0.6039	0.4646
eng	0.8058	0.8014	0.7353	0.7535	0.7034	0.5194	0.6868	0.7821	0.5391	0.7560	0.5742	0.4894
fas	0.7690	0.8198	0.4937	0.8482	0.7202	0.7793	0.5919	0.7004	0.5445	0.6492	0.6763	0.4793
hau	0.7401	0.6783	0.6176	0.6031	0.4979	0.4991	0.4891	0.4966	0.4907	0.4920	0.4979	0.4994
hin	0.7974	0.8019	0.8896	0.9214	0.8853	0.6536	0.7137	0.7407	0.7130	0.6674	0.6554	0.7480
ita	0.7303	0.4227	0.7741	0.8374	0.6011	0.4617	-	-	-	-	-	-
khm	0.6293	0.8052	0.7220	0.8673	0.7207	0.7873	0.7275	0.5464	0.5233	0.4942	0.6960	0.5168
mya	0.8788	0.8955	0.7417	0.8071	0.7927	0.8399	-	-	-	-	-	-
nep	0.8915	0.8438	0.9281	0.9369	0.9132	0.8187	0.8650	0.7952	0.7537	0.8244	0.6643	0.5589
ori	0.8013	0.7828	0.7305	0.7847	0.6254	0.6590	0.5720	0.6428	0.4984	0.5680	0.4960	0.4914
pan	0.7736	0.7096	0.5914	0.6745	0.6923	0.5558	0.6384	0.7562	0.7214	0.6834	0.5423	0.6547
pol	0.8158	0.8314	0.7368	0.8523	0.7761	0.5392	-	-	-	-	-	-
rus	0.8077	0.8231	0.7451	0.8207	0.7389	0.5469	-	-	-	-	-	-
spa	0.7788	0.8069	0.6994	0.7785	0.8831	0.7310	0.6754	0.7521	0.4766	0.7227	0.5639	0.6442
swa	0.7658	0.6926	0.8357	0.9040	0.5085	0.5304	0.7573	0.7393	0.4658	0.6425	0.6995	0.6073
tel	0.8818	0.5560	0.6134	0.5720	0.6190	0.5444	0.4702	0.6228	0.4938	0.4853	0.7034	0.5792
tur	0.8008	0.8092	0.7910	0.8515	0.7867	0.5221	0.7673	0.7570	0.5862	0.8122	0.4747	0.4861
urd	0.7743	0.7806	0.7192	0.7379	0.7168	0.7103	0.7527	0.7701	0.7276	0.7634	0.7309	0.7248
zho	0.9237	0.8571	0.9212	0.9651	0.9245	0.8294	0.8393	0.8696	0.7958	0.6997	0.5491	0.5937

Table 2: Per-language per-feature (label) performance of the submitted system based on the released test-set gold labels.

	Macro F1		
Language	**Subtask 1**	**Subtask 2**	**Subtask 3**
amh	-0.0532	0.1400	-0.0123
arb	0.0391	0.1424	0.1255
ben	-0.0113	0.0163	0.0404
deu	0.0684	0.1321	0.0559
eng	0.0256	0.1186	-0.0403
fas	-0.0734	0.0624	0.1204
hau	-0.0352	-0.0349	-0.7456
hin	0.0595	-0.0338	0.5105
ita	0.0530	-0.0740	0.0000
khm	-0.0299	0.0055	-0.3613
mya	0.0578	0.2063	0.0000
nep	0.0117	0.0807	0.4355
ori	0.0248	-0.0972	-0.2614
pan	-0.0162	0.0084	0.0372
pol	0.0917	0.0859	0.0000
rus	0.0620	-0.0952	0.0000
spa	0.0522	0.0321	-0.1123
swa	0.0087	-0.0191	0.2365
tel	0.2378	-0.0572	-0.4595
tur	0.1051	0.0984	-0.3568
urd	-0.0147	0.0664	0.2792
zho	0.0546	0.1502	0.4912
Average	0.0326	0.0425	-0.0008

Table 3: Per-language performance comparison to the official baseline.

	Percentile		
Language	**Subtask 1**	**Subtask 2**	**Subtask 3**
amh	7.1	40.0	26.3
arb	86.7	77.8	15.8
ben	83.7	46.7	23.8
deu	88.9	50.0	21.1
eng	78.3	33.3	8.3
fas	11.4	14.8	10.5
hau	26.7	28.6	0.0
hin	53.2	46.7	71.4
ita	97.7	57.7	-
khm	27.9	48.0	26.3
mya	83.3	52.0	-
nep	27.3	89.3	25.0
ori	82.2	40.7	19.0
pan	56.8	12.5	45.0
pol	86.0	59.3	-
rus	86.0	37.0	-
spa	70.0	58.6	9.5
swa	18.2	33.3	10.0
tel	71.1	18.5	23.8
tur	85.7	53.8	5.3
urd	50.0	83.3	81.0
zho	89.1	80.0	13.6

Table 4: Per-language rank percentile of the submitted system based on unofficial results. Higher percentile is better in comparison to others.

in 6 cases, it was able to achieve above 0.7 macro F1 value.

In comparison to the baseline included in the official results, the submitted system performs better in about two thirds of cases (see Table 3). On average, it achieved by 3% in subtask 1 and by 4% in subtask 2 higher Macro F1 score. The performance in subtask 3 was about the same on average, highly influenced by F1 score of 0.0 in Hausa. The highest increase is in Hindi (by 51%).

Lang.	Subtask 1	Subtask 2					Subtask 3					
	polarization	political	racial/ ethnic	religious	gender/ sexual	other	stereotype	vilification	dehuman- ization	extreme language	lack of empathy	invalidation
amh	0.6804	0.6730	0.6736	0.6119	0.4525	0.6442	0.6278	0.7156	0.5981	0.5955	0.5175	0.4666
arb	0.6908	0.7041	0.6336	0.6365	0.7463	0.7232	0.7105	0.7306	0.6755	0.6903	0.6845	0.6336
ben	0.7616	0.7437	0.3429	0.6989	0.6501	0.7159	0.6252	0.7445	0.6716	0.7356	0.6544	0.7139
deu	0.6757	0.6610	0.6127	0.6877	0.7300	0.6135	0.6608	0.6520	0.6586	0.6593	0.5729	0.5564
eng	0.7411	0.7127	0.6892	0.7476	0.6712	0.6345	0.7446	0.7062	0.6559	0.6829	0.6463	0.6667
fas	0.6797	0.7348	0.6425	0.7294	0.6557	0.7453	0.5391	0.6181	0.6809	0.6473	0.6963	0.6137
hau	0.6741	0.6493	0.6188	0.5637	0.4809	0.3897	0.6100	0.6724	0.7488	0.5643	0.7623	0.4691
hin	0.7382	0.6586	0.6610	0.6830	0.6762	0.6118	0.7053	0.6842	0.6687	0.6009	0.6798	0.6795
ita	0.5902	-	0.6859	0.7392	0.7882	-	-	-	-	-	-	-
khm	0.6655	0.8094	0.6734	0.6940	0.6212	0.6924	0.6988	0.8790	0.8948	0.7842	0.7422	0.7383
mya	0.7579	0.6894	0.6638	0.5986	0.6831	0.7539	-	-	-	-	-	-
nep	0.7890	0.6932	0.7250	0.6352	0.8812	0.7895	0.7844	0.8110	0.8801	0.6829	0.6829	0.7784
ori	0.6632	0.7023	0.7530	0.7685	0.7954	0.8405	0.6083	0.6624	0.8273	0.7572	0.6170	0.6822
pan	0.7115	0.6731	0.6152	0.6902	0.7031	0.5822	0.6467	0.7561	0.6129	0.5963	0.6628	0.6422
pol	0.7179	0.7253	0.6160	0.7058	0.6362	0.6929	-	-	-	-	-	-
rus	0.7043	0.7363	0.7156	0.6133	0.6507	0.6583	-	-	-	-	-	-
spa	0.6996	0.7635	0.6197	0.5660	0.7219	0.7051	0.6639	0.6741	0.5960	0.6978	0.6511	0.7765
swa	0.6668	0.5789	0.6913	0.6182	0.5206	0.5358	0.6472	0.6260	0.5488	0.5789	0.5975	0.6174
tel	0.7573	0.6484	0.6851	0.6470	0.6514	0.6746	0.6419	0.7517	0.7323	0.6188	0.6460	0.6500
tur	0.7660	0.7450	0.7020	0.7119	0.6125	0.6019	0.7627	0.7543	0.7033	0.7438	0.5513	0.6579
urd	0.6750	0.6480	0.6067	0.5986	0.6342	0.5915	0.6148	0.6162	0.6056	0.6341	0.6431	0.6381
zho	0.8543	0.7551	0.7795	0.6801	0.8347	0.6882	0.7730	0.7896	0.7893	0.7724	0.8162	0.7412

Table 5: Per-language per-feature AUC performance of the appraisals-classifier system based on hold-out part of the train set.

Regarding the comparison to other systems submitted to the shared task (see Table 4), our system ranked in the top 20% in 14 cases (out of 62 combinations of languages and subtasks). In 28 cases, the system ranked in the top 50%. Only in a single case, it has performed the **1st** of all submitted systems (**Italian** in subtask 1). The best rank achieved in subtask 2 was **3rd in Nepali** and **4th in Urdu** in subtask 3. On the other hand, the system ranked in the bottom 20% in 15 cases.

Based on the results, the manifestation identification seems to be the most challenging subtask for our submitted system. After the analysis of the individual monitored features (labels) of each subtask based on the released gold labels (after the system-submission deadline), we can see (Table 2) that "other" class in subtask 2 was the most challenging. In subtask 3, "dehumanization", "lack of empathy", and "invalidation" are particularly misclassified.

5.1 Appraisals for Polarization Detection

As the results in Table 6 indicate, the appraisals can be used also for polarization detection. Such cognitive signals capture features that are important for polarization as well. Since the performance for subtasks 2 and 3 is rather random, we have looked at the AUC values per each label, showing a threshold-independent classification capability. The results are shown in Table 5. Such results clearly indicate a high potential of using (multilingual) appraisals for polarization-related indicators, which might be worthy exploring further in future work.

Language	Macro F1		
	Subtask 1	Subtask 2	Subtask 3
amh	0.5566	0.4897	0.5062
arb	0.6338	0.4633	0.4924
ben	0.6841	0.5168	0.4919
deu	0.6188	0.4869	0.4511
eng	0.6464	0.5174	0.4744
fas	0.5040	0.5217	0.4801
hau	0.4718	0.4939	0.4945
hin	0.4844	0.5113	0.5775
ita	0.4335	0.4615	-
khm	0.4759	0.5196	0.5018
mya	0.6944	0.5189	-
nep	0.7107	0.4770	0.5525
ori	0.4563	0.4788	0.4820
pan	0.6673	0.4903	0.4854
pol	0.6581	0.5086	-
rus	0.5566	0.4811	-
spa	0.6440	0.4907	0.4786
swa	0.6268	0.5074	0.4760
tel	0.7005	0.4534	0.4640
tur	0.6997	0.5217	0.5667
urd	0.4625	0.5611	0.5396
zho	0.7803	0.5382	0.5243

Table 6: Per-language performance of the appraisals-classifier system based on hold-out part of the train set.

6 Conclusion

The submitted multilingual system has shown that a single system can handle all languages combined, competitively in comparison to other systems. However, the performance differs among languages, and for some the approach is not suitable. The multi-label classification subtasks have been more challenging for the system than the binary classification, for which the original system has been tuned. In future, the system might be modified to be finetuned individually for each language or some group or similar languages (not all of them). Furthermore, the appraisal-based alternative and its combination with the pure finetuned detector seem worthy to explore.

Limitations

We have explored only small set of base language models. Others could be better performing, at least for some languages. We have tested only the languages included in the shared task, thus the generalization to other languages is unevaluated. We have limited the training set of samples only to the official train and dev splits, available in the shared task. Other publicly available datasets could be used for training as well.

Acknowledgments

This work was partially supported by *LorAI – Low Resource Artificial Intelligence*, a project funded by Horizon Europe under GA No.101136646 and partially by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00068.

Computational resources. This work was supported by the use of computational resources of the supercomputer PERUN, operated by the Supercomputing Centre at the Technical University of Košice (TUKE), Slovakia with the support of the European Union from the funds of the Recovery and Resilience Plan of the Slovak Republic within the framework of project No. 17I03-04-P03-00001, Development and design of a supercomputer for the National Supercomputing Center. We also acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy.

References

Janek Bevendorff, Yuxia Wang, Jussi Karlgren, Matti Wiegmann, Maik Fröbe, Akim Tsivgun, Jinyan Su, Zhuohan Xie, Mervat T. Abassy, Jonibek Mansurov, Rui Xing, Minh Ngoc Ta, Kareem Ashraf Elozeiri, Tianle Gu, Raj Vardhan Tomar, Jiahui Geng, Ekaterina Artemova, Artem Shelmanov, Nizar Habash, and 5 others. 2025. Overview of the "voight-kampff" generative AI authorship verification task at PAN and ELOQUENT 2025. In *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR-WS.org*.

Lucio La Cava, Dominik Macko, Róbert Móro, Ivan Srba, and Andrea Tagarelli. 2025. Authorship attribution in multilingual machine-generated texts. *Preprint*, arXiv:2508.01656.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alok Debnath, Yvette Graham, and Owen Conlan. 2025. An appraisal theoretic approach to modelling affect flow in conversation corpora. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 233–250, Vienna, Austria. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dominik Macko. 2025. mdok of KInIT: Robustly finetuned LLM for binary and multiclass AI-generated text detection. In *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR-WS.org*.

Dominik Macko, Robert Moro, and Ivan Srba. 2025. Increasing the robustness of the fine-tuned multilingual machine-generated text detectors. *Preprint*, arXiv:2503.15128.

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. *Preprint*, arXiv:2505.20624.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Michal Spiegel and Dominik Macko. 2024a. IMGTB: A framework for machine-generated text detection benchmarking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 172–179, Bangkok, Thailand. Association for Computational Linguistics.

Michal Spiegel and Dominik Macko. 2024b. KInIT at SemEval-2024 task 8: Fine-tuned LLMs for multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 558–564, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, 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, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019a. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy. Association for Computational Linguistics.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019b. Crowdsourcing and validating event-focused emotion corpora for german and english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011.

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.

A Computational Resources

For experiments regarding model finetuning and inference processes, we have used multiple GPU-accelerated systems. Experiments using 1× NVIDIA A100 64GB GPU taken around 200 GPU hours, and using 1× NVIDIA H200 140GB GPU taken around 100 GPU hours. Analysis has been done without the GPU acceleration.

B Appraisal Annotations and Polarization

This appendix summarizes the datasets and architecture of the appraisal estimation framework underlying the empirical results reported in the main paper.

Theoretical basis. Appraisal theory is a cognitive and evolutionary account of emotion, positing that emotional experiences arise from evaluations of surrounding stimuli against context-relevant criteria such as *pleasantness*, *urgency*, and *alignment with social norms*. These appraisal dimensions offer a richer substrate for understanding affective-cognitive links than coarse categorical emotion labels alone.

Corpus. The *International Survey on Emotion Antecedents and Reactions* (ISEAR; Scherer and Wallbott 1994) corpus comprises English-language first-person descriptions of emotional experiences across seven categories: anger, disgust, fear, guilt, joy, shame, and sadness. Its emphasis on subjective experiential accounts distinguishes it from other NLP emotion resources. Building on this, Troiano et al. (2019b) produced *deISEAR* and *enISEAR* — crowdsourced, emotion-annotated datasets that preserve ISEAR's design principles while surfacing divergences between experiencer and reader emotion judgements.

Annotation scheme. From these corpora, appraisal objectives, including *self-consequences*, *consequences for others*, and *situational control*, were annotated to capture whether the experiencer attributed their emotional state to each factor. These appraisal-level annotations, rather than emotion category labels, provide grounded cues to the link between event cognition and affective language. This monolingual, dimensionally richer appraisal mechanism has also proven useful in adjacent tasks such as emotion recognition in conversation (Debnath et al., 2025).

Architecture. Appraisal estimation is implemented as a multitask binary classifier trained with weighted binary cross-entropy loss, designed to minimize appraisal mischaracterization. The resulting appraisal estimates are used as supplemental embeddings for the polarization detection and SemEval subtask objectives, augmenting the primary representation with affective-cognitive signals.

Findings and outlook. Incorporating deISEAR and enISEAR appraisals as additional embeddings yields promising but mixed results for polarization detection. Key limitations include the specificity of the selected appraisal dimensions, the method of appraisal representation learning and integration, and limited multilingual coverage. Nevertheless, performance improvements on subtasks with stronger affective-cognitive content (such as dehumanization, lack of empathy, and related emotional expression markers) suggest meaningful potential for future work exploiting these signals in text and conversation.