OZemi at SemEval-2026 Task 9: A Cross-Lingual Approach to Online Text Polarization Classification Using Multilingual Models and Adaptive Loss Formulation

Hidetsune Takahashi[1], Eleale Nusi Tee[1], Aika Yu[1], Ruri Furukawa[1], Shuta Niinomi[1], Sooeun Kim[1], Dingyu Zhang[1], Emily Ohman[1]

[1]Waseda University

Correspondence: ohman@waseda.jp

Abstract

We propose a unified multilingual approach that addresses multiple languages and subtasks efficiently. Our system combines multilingual models with data-level techniques and a class-weighted cross-entropy loss to mitigate data imbalance across languages, subtasks, and categories. Results show consistent performance across languages, achieving macro F1 scores above 70% in most languages for Subtask 1, with our highest rank for Persian (1st out of 44). These results suggest that the proposed framework provides a flexible foundation for multilingual and multi-task polarization analysis.

1 Introduction

This paper addresses SemEval-2026 Task 9 (Naseem et al., 2026a), which focuses on multilingual online text classification from the perspective of polarization. Subtask 1 involves binary classification to determine whether a given text is polarized, covering 22 languages. Subtasks 2 and 3 are formulated as multi-label classification tasks: Subtask 2 aims to identify polarization types, while Subtask 3 focuses on polarization characteristics. These subtasks span 22 and 18 languages, respectively.

We utilize open-source encoder-based BERT-based (Devlin et al., 2019) transformer models to construct a unified framework for multilingual and multi-subtask classification. A central component of our approach is the use of a class-weighted cross-entropy loss, which aims to enable relatively stable and effective learning under varying degrees of class imbalance across subtasks and languages. In addition to the loss design, we explore complementary strategies at both the model and data levels, including the selection of multilingual encoder models and the application of back-translation for selected languages.

Our approach shows consistent multilingual performance, particularly in Subtask 1, which was the main focus of our team. The proposed system achieves macro F1 scores above 70% in 20 of the 22 languages, with several languages exceeding 85%. Although the remaining subtasks are more challenging due to their label complexity and diverse data distributions, the unified framework with a class-weighted cross-entropy loss still attains competitive performance in several languages. Overall, these results indicate the potential of the proposed strategy as a general framework for multilingual online text polarization tasks.

2 Background

Online polarization, which is defined as the increasing divergence of opinions, attitudes, or affective orientations between individuals or groups within a public or political domain (Arora et al., 2022), has emerged as a critical challenge in digital communication. Polarized speech contributes to the deepening of societal divisions and undermines constructive dialogue (Naseem et al., 2026b).

Subtask 1 is formulated as a binary classification task that aims to determine whether a given post contains polarized content (Polarized or Not Polarized). The binary class distributions are highly skewed in most languages. Polarized samples dominate in Khmer (90.8%), Hindi (85.5%), and Amharic (75.6%), while Hausa (10.7%) and Odia (28.8%) data are heavily biased toward the non-polarized class. This multilingual disparity motivates both our back-translation strategy for selected languages and the class-weighted loss applied across all languages (Naseem et al., 2026b).

Subtask 2 follows a data structure similar to that of Subtask 1, however, it differs in that this subtask requires classification of the polarization type. Accordingly, it comprises the following five labels: political/ideological polarization, racial or ethnic

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 182–192

polarization, religious polarization, gender/sexual orientation polarization, and other.

Subtask 3 focuses on how polarization is expressed, with the following five labels: stereotype, vilification, dehumanization, extreme language and absolutism, lack of empathy or understanding, and invalidation. Subtask 3 covers 18 languages.

Subtasks 2 and 3 are multi-label classification tasks with 5 and 6 categories respectively. Full per-language distributions are provided in Tables 6 and 7.

Both subtasks exhibit severe label sparsity in certain language-category combinations. In Subtask 2, the Italian subset contains zero samples for both "political" and "other" categories. Beyond this extreme case, Hausa has fewer than 5% positive samples in all five Subtask 2 categories, and Bengali shows similar patterns with several categories below 1%. In Subtask 3, Hausa "invalidation" contains only 9 positive samples out of 3,651, and Odia "dehumanization" has 16 out of 2,368.

At the opposite extreme, Urdu exhibits dense multi-labeling across both subtasks, with all categories exceeding 50%. In contrast, the Hausa subset of Subtask 2 contains under 5% samples for all categories, indicating differences in data collection or annotation scope for the different languages.

3 System Description

Our system first investigates several open-source encoder-based Transformer models. The system was designed to emphasize broad applicability rather than language-specific tuning, following previous work in which a similar approach produced stable results in semantic relatedness tasks (Takahashi et al., 2024). To address the data imbalance, we apply back-translation for selected languages. In addition, we employ a class-weighted cross-entropy loss to further mitigate class imbalance across different subtasks and categories, enabling more effective learning under imbalanced data distributions. This work does not aim to achieve strong language-specific optimization, but rather, it seeks a solution that can be broadly applied to multiple languages and related tasks. Since we participate in all available subtasks and languages, we adopt multilingual models and begin with simple experimental settings, which are then refined in subsequent experiments.

4 Experimental Setup

Due to the data imbalances and annotation differences, we focused on data augmentation for the less balanced languages. Under our class-weighted loss method, class weights are inversely proportional to sample counts. For extremely sparse categories such as Hausa "invalidation" in Subtask 3, this yields $w_k = 3,651/9 \approx 405.7$, which may cause training instability.

4.1 Multilingual Stylistic Analysis

To understand how polarization manifests linguistically across languages, we analyze 11 surface-level features capturing social media conventions and emphatic markers: emoji-presence and -density, @mention and #hashtag frequencies, repeated punctuation (*!!, ???, !?*), all-caps ratio, overall punctuation density, and text length. This analysis addresses two objectives: identifying cross-linguistic patterns in polarization expression, and explaining model performance variations, particularly for languages exhibiting extreme social media characteristics.

Amharic and Odia show higher @mention rates in polarized texts (+0.193, and +0.170 respectively), suggesting direct user engagement as the primary polarization mechanism. Amharic and Russian further distinguish this pattern by simultaneously reducing hashtag usage (–0.104 and –0.164), prioritizing interpersonal targeting over topical indexing. The Italian case provides evidence for model selection: the polarization-driven rise in @mention count (+0.064) and the increase in repeated punctuation (+0.031) shifts the discourse from topical argumentation to interactive confrontation. This structural change directly favors TwHIN-BERT (Zhang et al., 2023), which encodes social graph dependencies, over generic web-crawled models like XLM-RoBERTa (Conneau et al., 2020). For more information on stylistic features in the data, see Table 8.

4.2 Back-Translation vs Language-specific Models

Previous work suggested that data augmentation can be beneficial in multilingual classification tasks, particularly under imbalanced data conditions (Takahashi et al., 2025). Based on this, back-translation (Sennrich et al., 2016; Edunov et al., 2018) was implemented in the present study to address the significant class imbalances observed in

our datasets. This process aimed to introduce syntactic diversity while preserving semantic meaning by targeting samples from the minority class. For the English data set, we utilized MarianMT models (Junczys-Dowmunt et al., 2018) ($Helsinki-NLP/opus-mt-en-de$ and $de-en$) to pass a fraction of the polarized samples through an English-German-English pipeline. Similarly, for Hindi, we applied an English-intermediate pipeline ($Helsinki-NLP/opus-mt-hi-en$ and $en-hi$) to increase underrepresented samples, thus synthesizing a more balanced training distribution (Tiedemann and Thottingal, 2020).

However, these augmentation efforts varied significantly between the two languages. For the English data set, back-translation of the polarized minority class did not yield the intended performance gains; the RoBERTa-Base model's F1-macro score decreased from 0.85 on the raw data to 0.8058 after augmentation, suggesting that the added syntactic variations may have introduced noise that hindered generalization. In contrast, the strategy proved highly effective for Hindi, where augmenting the non-polarized samples allowed the XLM-RoBERTa-Base model to reach a peak F1-macro score of 0.7749, a clear improvement over the 0.7475 achieved using raw data alone. Similarly to English, the Spanish and Russian data also saw worse F1 scores with data augmentation using back-translation.

In addition to our general experiments, we explored the efficacy of language-specific models for the Italian, Chinese, and Hindi datasets to leverage pre-trained architectures optimized for these linguistic families. For the Hindi dataset, we specifically evaluated MuRIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021), a model pre-trained on a diverse corpus of Indian languages. However, despite these optimizations, MuRIL achieved a test 0.72 F1-macro score, which is lower than our baseline.

These findings suggest that data augmentation was not uniformly beneficial; therefore, we used it selectively rather than as a general strategy across languages.

4.3 Model Selection

In the early stages of our experiments, we evaluated several Transformer-based models to identify models that could effectively improve performance. We first experimented with relatively simple models, such as RoBERTa-Base (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and subsequently adopted TwHIN-BERT (Zhang et al., 2023) to better align the system with the characteristics of the target data, which consist of relatively short contexts related to social media posts.

The performance of two models fine-tuned on the task dataset during the development phase is compared in Table 1. For RoBERTa-base, the English base model was applied to English data, while XLM-RoBERTa-base was used for the remaining languages. In contrast, TwHIN-BERT-base was employed across all languages reported in Table 1.

Overall, TwHIN-BERT-base achieves better performance than XLM-RoBERTa-base in many languages, although slight performance decreases are observed for several languages. However, this comparison is limited to multilingual models. When compared with a language-specific model, different trends are observed. For English, the RoBERTa-base model clearly outperforms the multilingual alternatives. Unsurprisingly (Pàmies et al., 2020), for some non-English languages, the multilingual model shows clear advantages. Notably, TwHIN-BERT-base achieves a substantial improvement in Italian, with an increase of approximately 25 percentage points. This result is likely attributable to the model's pre-training on Twitter data, as the Italian subset tends to contain relatively extreme and informal expressions characteristic of social media text.

Given our system design strategy of developing a unified solution that efficiently addresses multiple languages and use cases, we select TwHIN-BERT-base as our base model. Its multilingual nature and pre-training on SNS-related text allow subsequent fine-tuning and experimental configurations to generalize more effectively across diverse languages and task settings.

4.4 Class-weighted Cross-entropy Loss

In order to address data imbalance across multiple languages, categories and subtasks, the cross-entropy loss was customized to assign relatively heavier weights to labels with fewer samples. Here, let N_k denote the number of training samples belonging to class k, and let

$$w_k = \frac{N}{N_k}, \tag{1}$$

where $N = \sum_k N_k$ is the total number of samples.

Using these class weights, the modified cross-

Model	Amh	Ara	Deu	Eng	Fas	Hau	Hin	Ita	Nep	Spa	Tur	Urd	Zho
RoBERTa-base	0.669	0.726	0.624	0.828	0.803	0.662	0.707	0.352	0.880	0.694	0.747	0.698	0.892
TwHIN-BERT-base	0.728	0.731	0.673	0.704	0.803	0.722	0.859	0.606	0.850	0.671	0.800	0.732	0.864

Table 1: Performance comparison across languages

Hyperparameter	Value
Batch size (train)	16
Batch size (validation)	32
Training epochs	10
Learning rate	2×10^{-5}
Warm-up steps	40
Weight decay	0.01
Early stopping patience	4
Maximum sequence length	256

Table 2: Hyperparameter settings

entropy loss is defined as

$$L_{\mathrm{WCE}}(\theta) = -\sum_{k} w_k \, t_k \log y_\theta(k \mid x), \quad (2)$$

where $L_{\mathrm{WCE}}(\theta)$ denotes the weighted cross-entropy loss.

From Eq. (1), it follows that classes with fewer samples are assigned larger weights w_k, which increases their contribution to the weighted cross-entropy loss L_{WCE}. Consequently, misclassification errors for underrepresented classes yield larger gradient magnitudes during optimization. This class-weighting scheme mitigates the bias toward majority classes and promotes more balanced parameter updates in the presence of data imbalance. Additionally, early stopping was used, with the hyperparameter setting summarized in Table 2.

The class-weighted cross-entropy loss was employed for 16 languages in Subtasks 1, all 22 languages in Subtask 2, and for all 18 languages in Subtask 3. In particular, Subtask 2 and Subtask 3 exhibited substantial class imbalance in several labels across multiple languages, motivating the use of class weighting. In these scenarios, our modified loss function is designed to effectively address data imbalance in a manner that is consistently applicable across all languages and categories.

5 Results

Table 9 in the Appendix reports the macro F1 scores across languages from the task leaderboard, together with language-specific rank information for our submission (*hidetsune*). Overall, our unified multilingual setup performs most strongly on Subtask 1. Across the 22 languages in this subtask, the mean macro F1 is 0.7728, and performance exceeds 0.70 in 20 of 22 languages. The highest scores are obtained for Nepali (0.8870), Telugu (0.8734), and Burmese (0.8631), while German (0.6693) and Italian (0.4981) are the most challenging cases.

Subtasks 2 and 3 are considerably more difficult. In Subtask 2, the mean macro F1 across 22 languages is 0.5195. The strongest results are observed for Hindi (0.7702), Nepali (0.7637), Urdu (0.7563), and Chinese (0.7082), while Italian (0.2080) and Hausa (0.2321) remain challenging. In Subtask 3, which covers 18 languages, the mean macro F1 is 0.4734. Urdu (0.7835) and Hindi (0.7356) perform best, whereas Hausa (0.2058) shows the lowest score.

The lower overall performance and higher variance in Subtasks 2 and 3 are consistent with the data properties discussed earlier. Unlike Subtask 1, these are multi-label settings in which some language-label combinations contain very few positive examples, and in some cases no positive examples at all. Under such conditions, weighted loss remains useful, but it is not sufficient on its own: sparse supervision makes it difficult to learn stable decision boundaries, while fixed decision thresholds may further reduce performance by producing a poor precision-recall balance for rare labels. This helps explain why our approach remains comparatively robust in Subtask 1 but is less stable in the more imbalanced multi-label subtasks.

6 Error Analysis

To better understand where the model fails, we conducted an error analysis for Subtask 1 across all 22 languages in the test set. Italian and German emerged as particularly difficult cases, but for different reasons. In Italian, the model frequently overpredicts polarization, whereas in German it more often fails to detect it. These two languages therefore provide useful case studies for understanding the limitations of our approach beyond overall performance scores. In particular, they suggest that cross-language variation is shaped not only by resource level or dataset size, but also by how polarization is expressed stylistically and rhetorically.

6.1 Italian

Italian is the most difficult language in Subtask 1 for our system. The error profile is dominated by false positives: the model produces 538 false positives compared with 215 false negatives, and the predicted polarization rate (68.3%) is substantially higher than the gold rate (47.3%). This indicates that the model tends to over-predict polarization in Italian.

Representative examples are shown in Table 3. A common pattern is that many false positives contain strong emotional or interactional cues, such as repeated punctuation, elongated spelling, insults, or highly expressive phrasing. However, these posts do not necessarily express the kind of group-directed antagonism or ideological opposition that the task is intended to capture. Instead, many of them reflect personal frustration, casual confrontation, or emotionally charged commentary without a clear polarized target.

We therefore hypothesize that the model has learned an association between expressive intensity and polarization that is too broad in the Italian setting. This interpretation is consistent with the stylistic analysis in Section 4.1, which showed that polarized Italian texts are characterized by increased @mentions and repeated punctuation. These features may be informative overall, but they also make it easier for the model to overgeneralize from emotional tone to polarization, especially in informal discourse where expressive language is common even outside political or intergroup contexts.

#	Text	Characteristics
FP-1	*"Ma cosa chiedete alla Bruzzone, ma perché la invitate a pagamento poi!!!!"* (Why do you invite her, and even pay for it?!)	Frustration toward a public figure; no clear political content
FP-2	*"lui ha il 44 di piede, l'assassino il 42. Fine"* (His shoe size is 44; the killer's is 42. End of story.)	Matter-of-fact comment on a crime case; non-ideological
FP-3	*"Ma che caxxo dici se non Sebastano pagherà sulla Terra pagherà davante a Dio"* (What are you saying—he will face God's judgment.)	Personal attack with religious framing; not directed at a group
FP-4	*"Una marea di ipocrisia! Svegliaaaaaaa!!"* (A flood of hypocrisy! Wake uuuup!!)	Strong emotional tone with exaggerated spelling; no obvious political axis

Table 3: Representative false positive (FP) cases in Italian

6.2 German

German shows a different error pattern. Although false positives are still more frequent than false negatives overall, a large share of the false negatives belongs to the political category. This suggests that the main issue in German is not simple over-triggering, but rather the model's difficulty in identifying certain forms of implicitly expressed polarization.

The examples in Table 4 illustrate this point. Several of the false negatives convey stance through contrast, irony, provocation, or rhetorical framing rather than through overtly emotional or abusive wording. On the surface, such texts may appear neutral, analytical, or ambiguous, even when they clearly position the speaker in relation to a polarized issue. In these cases, lexical cues alone appear insufficient.

Taken together, the German results suggest that our model struggles when polarization is communicated indirectly rather than through emotionally salient surface features. Whereas the Italian errors point to over-reliance on expressive style, the German errors indicate difficulty with discourse-level meaning, including irony and argumentative framing. Addressing this limitation would likely require stronger modeling of context and rhetoric, rather than relying primarily on local lexical or stylistic cues.

#	Text	Characteristics
FN-1	*"Klimaschutz != Umweltschutz. Umweltschutz ist 'hier', da geht es um Bienen und Blumen … Klimaschutz dagegen ist ein weltweites Problem"* (Climate protection $\neq$ environmental protection. The latter is local; the former requires fundamental behavioral change.)	Analytical tone; stance conveyed through framing rather than emotion
FN-2	*"4 Superreiche sind gestorben, ist gut fürs Klima"* (4 ultra-rich people died—good for the climate.)	Short and ambiguous; could be read as satire or provocation
FN-3	*"Putin ist unser Sponsor"* (Putin is our sponsor.)	Likely ironic; meaning depends on context
FN-4	*"Sich für Klimaschutz einsetzen schadet dem Klima."* (Advocating for climate protection harms the climate.)	Paradoxical statement; rhetorical rather than literal

Table 4: Representative false negative (FN) cases in German

7 Limitations and Future Work

One limitation of our work is the limited exploration of language-specific models. Since our primary goal is broad applicability across multiple languages and subtasks, we did not conduct sufficient experiments to investigate how, and under what conditions, language-specific models might outperform our unified approach based on fine-tuned multilingual models. Further experimental analyses and direct comparisons would help clarify the trade-offs and applicability of these approaches.

Additionally, while early stopping was employed to mitigate potential overfitting, we did not perform detailed hyperparameter tuning for individual languages or categories. This simplified setup may be suboptimal in certain cases, for example when a language has a limited number of training samples, where fixed hyperparameters, such as a large batch size or a high learning rate, can lead to instability during fine-tuning. Moreover, the number of warm-up steps could have been more carefully designed based on data-specific conditions, including the dataset size and the degree of class imbalance.

Furthermore, adjusting the decision thresholds based on softmax outputs may have been beneficial for certain languages, particularly in cases where straightforward fine-tuning and inference did not yield satisfactory performance. For example, as discussed earlier, the Italian data in Subtask 1 exhibit relatively extreme expressions, suggesting that increasing or decreasing the decision threshold could potentially influence performance. Such post-hoc analyses, conducted outside the fine-tuning process, might have provided a more efficient performance improvement compared to other strategies.

8 Conclusion

Throughout this study, we employed multilingual models to develop an efficient and unified solution for handling multiple languages and subtasks. We began by fine-tuning a baseline model and subsequently compared it with a model more closely aligned with the task across several selected languages. In addition to model-side exploration, we also investigated approaches focusing on the data, including data augmentation through back-translation and analyses of emoji usage. Furthermore, our system is built upon a modified cross-entropy loss designed to consistently mitigate various forms of data imbalance across languages, subtasks, and categories. Together, these design choices establish a flexible foundation for a solution that can effectively generalize to diverse languages and use cases.

Our approach demonstrates robust performance across multiple languages, particularly in Subtask 1, which was the primary focus of this study. Specifically, our system achieves a macro F1 score exceeding 70% in 20 out of the 22 languages provided for Subtask 1, with several languages reaching values above 85%. While performance on the remaining subtasks tends to be suboptimal due to their inherent difficulty, such as the presence of five to six label categories and various data distributions across categories, our unified system incorporating a class-weighted cross entropy loss still attains macro F1 scores above 70% for some languages. These results indicate that our overall strategy is broadly applicable across a wide range of languages and use cases related to online text polarization, although there remains room for further improvement. Our highest ranks were 1 out of 44 for Persian in Subtask 1, 3 out of 45 for Hausa in Subtask 1, 5 out of 27 for Swahili in Subtask 2, and 2 out of 19 for both Persian and Hausa in Subtask 3.

References

Swapan Deep Arora, Guninder Pal Singh, Anirban Chakraborty, and Moutusy Maity. 2022. Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change*, 183:121942.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, 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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 489–500.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield,

Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, and 1 others. 2018. Marian: Fast neural machine translation in c++. In *Proceedings of ACL 2018, system demonstrations*, pages 116–121.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, 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)*, San Diego, CA, USA. 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.

Marc Pàmies, Emily Öhman, Kaisla Kajava, and Jörg Tiedemann. 2020. Lt@ helsinki at semeval-2020 task 12: Multilingual or language-specific bert? In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1569–1575.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96.

Hidetsune Takahashi, Xingru Lu, Sean Ishijima, Deokgyu Seo, Yongju Kim, Sehoon Park, Min Song, Kathylene Marante, Keitaro-luke Iso, Hirotaka Tokura, and Emily Ohman. 2024. OZemi at SemEval-2024 task 1: A simplistic approach to textual relatedness evaluation using transformers and machine translation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 7–12, Mexico City, Mexico. Association for Computational Linguistics.

Hidetsune Takahashi, Sumiko Teng, Jina Lee, Wenxiao Hu, Rio Obe, Chuen Shin Yong, and Emily Ohman. 2025. OZemi at SemEval-2025 task 11: Multilingual emotion detection and intensity. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 109–115, Vienna, Austria. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 321–322, Lisboa, Portugal.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 5597–5607. Association for Computing Machinery.

Appendix

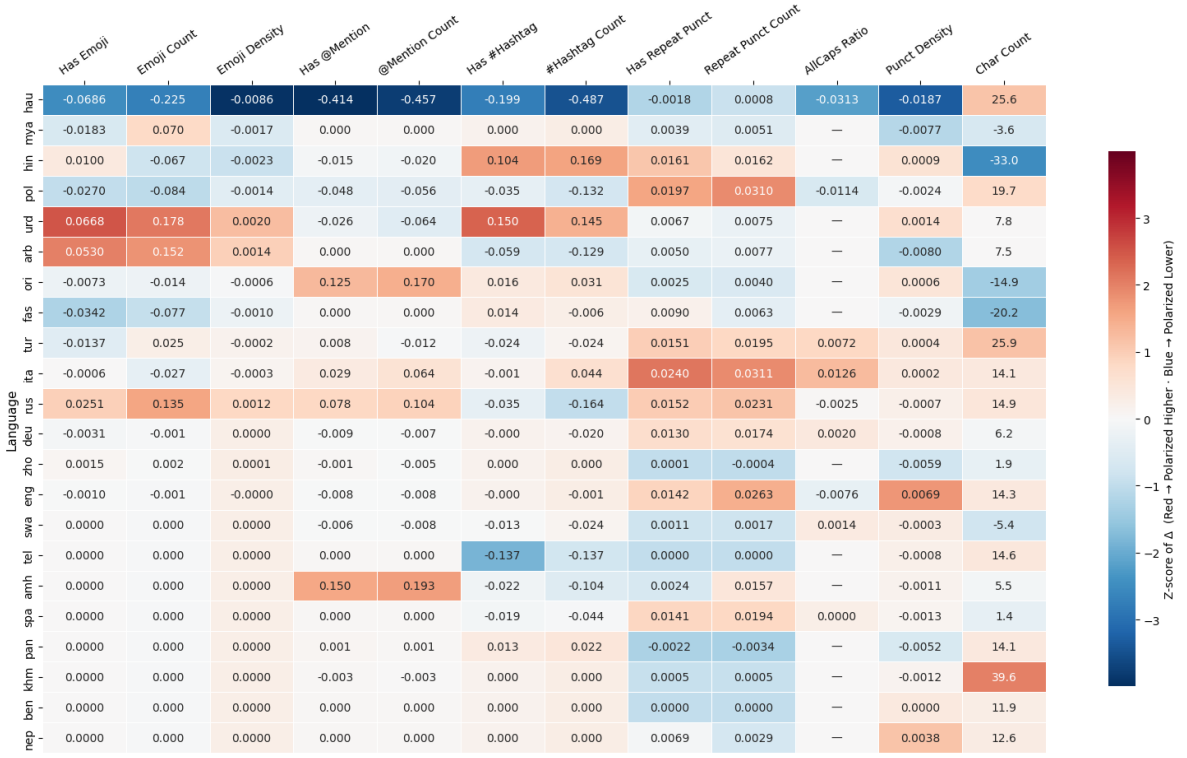

Figure 1: Z-score normalized difference between polarized and non-polarized texts across stylistic features. Red indicates higher values in polarized texts and blue indicates lower.

Language	Samples	Pol. %
Khmer	6,640	90.8
Hindi	2,744	85.5
Amharic	3,332	75.6
Persian	3,295	74.1
Urdu	3,563	69.5
Burmese	2,889	58.2
Telugu	2,366	53.8
Nepali	2,005	50.3
Spanish	3,305	50.2
Swahili	6,991	50.1
Chinese	4,280	49.6
Punjabi	1,700	49.4
Turkish	2,364	48.9
German	3,180	47.5
Arabic	3,380	44.7
Bengali	3,333	42.7
Polish	2,391	41.9
Italian	3,334	41.0
English	3,222	36.5
Russian	3,348	30.6
Odia	2,368	28.8
Hausa	3,651	10.7

Table 5: Training samples and Subtask 1 polarized proportion (%) per language, sorted by polarized proportion. Subtask 3 excludes Italian, Burmese, Polish, and Russian.

Language	Political %	Racial %	Religious %	Gender %	Other %
Amharic	66.8	25.9	2.0	0.6	24.8
Arabic	23.1	17.3	8.4	10.9	16.7
Bengali	34.0	0.8	2.0	0.5	10.1
Chinese	5.9	22.6	2.0	16.9	8.6
English	35.7	8.7	3.5	2.2	3.9
German	40.7	18.5	11.1	5.9	13.8
Hausa	4.9	3.2	2.6	0.8	0.4
Hindi	73.7	12.1	58.7	11.5	13.1
Italian	0.0	22.4	8.6	11.4	0.0
Khmer	18.3	1.5	3.4	1.7	65.9
Burmese	25.3	5.3	3.1	10.6	45.1
Nepali	17.2	14.0	7.9	5.2	11.8
Odia	21.0	5.0	6.3	3.3	3.7
Punjabi	30.8	5.9	7.9	11.2	8.9
Polish	36.6	9.0	3.6	4.6	6.5
Russian	13.9	9.8	4.1	5.7	2.4
Spanish	27.3	18.9	15.9	13.4	13.4
Swahili	2.7	35.5	3.5	2.2	7.9
Telugu	21.6	17.0	9.0	13.3	23.7
Turkish	44.7	16.9	15.2	4.8	4.8
Urdu	67.2	54.4	55.3	51.2	50.7
Persian	43.9	2.4	9.6	6.0	24.2

Table 6: Subtask 2 positive label proportions (%) per language. Multi-label; rows do not sum to 100%.

Language	Stereo. %	Vilif. %	Dehum. %	Extreme %	Lack emp. %	Invalid. %
Amharic	54.6	48.5	13.2	30.6	17.6	16.0
Arabic	33.3	37.2	11.0	30.4	17.0	8.1
Bengali	6.0	24.1	10.7	4.7	1.9	1.8
Chinese	30.1	18.5	5.0	8.1	7.9	4.8
English	15.1	26.6	12.1	23.9	11.1	18.2
German	35.9	30.1	14.9	21.8	26.7	16.3
Hausa	4.3	1.2	3.5	3.0	0.9	0.2
Hindi	49.7	65.2	18.2	50.6	56.7	65.7
Khmer	68.3	1.5	1.2	2.3	11.0	6.5
Nepali	26.8	31.4	6.6	27.1	10.6	15.0
Odia	10.0	11.7	0.7	13.4	1.6	3.4
Punjabi	16.2	40.4	22.0	23.9	12.4	24.4
Spanish	27.5	30.6	8.9	24.2	23.9	10.6
Swahili	39.7	41.2	12.8	23.9	29.8	23.4
Telugu	11.2	22.7	2.5	13.4	26.3	22.8
Turkish	40.8	32.4	10.9	43.1	9.6	4.0
Urdu	62.3	64.8	55.6	62.2	56.2	57.2
Persian	13.1	57.5	4.3	16.9	9.9	8.0

Table 7: Subtask 3 positive label proportions (%) per language. Multi-label; rows do not sum to 100%.

Language	Emoji %	Emoji Density	@Mention %	#Hashtag %	Repeat Punct. %	AllCaps %	Punct. Density
Italian	6.03	0.13	71.63	25.70	6.48	10.56	4.39
Hausa	30.10	1.23	49.69	30.51	2.71	5.62	4.12
Amharic	0.00	0.00	38.36	14.41	16.03	—	4.07
Turkish	8.93	0.15	29.65	24.70	3.30	7.18	3.42
Urdu	10.44	0.28	22.96	42.41	0.93	—	2.88
Russian	6.27	0.11	22.76	7.53	3.73	4.86	2.75
Odia	8.57	0.25	22.47	2.41	1.14	—	2.30
Polish	10.54	0.29	9.74	10.33	2.05	3.40	3.55
German	3.93	0.09	9.78	1.48	1.70	2.05	3.16
Swahili	0.00	0.00	2.46	2.86	1.06	0.87	0.37
English	0.06	0.00	1.27	0.03	2.67	3.96	1.85
Hindi	13.67	0.48	0.47	24.20	3.39	—	2.85
Chinese	0.54	0.02	0.14	0.00	0.33	—	8.83
Punjabi	0.00	0.00	0.06	3.29	0.59	—	1.99
Khmer	0.00	0.00	0.06	0.03	0.05	—	1.02
Burmese	24.37	0.71	0.00	0.00	0.31	—	0.56
Persian	9.86	0.23	0.00	4.43	1.37	—	1.94
Arabic	8.58	0.28	0.00	5.62	1.78	—	1.63
Spanish	0.00	0.00	0.00	3.66	5.02	0.00	2.53
Telugu	0.00	0.00	0.00	8.45	0.00	—	1.32
Nepali	0.00	0.00	0.00	0.00	0.95	—	1.46
Bengali	0.00	0.00	0.00	0.00	0.00	—	0.00

Table 8: Cross-language stylistic features. Languages sorted by @mention frequency. "—" indicates scripts without case distinction (Devanagari, Arabic, Ethiopic, Khmer, Burmese, Telugu).

Language	Subtask 1	Subtask 2	Subtask 3
Amharic	0.7182 (35/42)	0.4995 (16/25)	0.5022 (10/19)
Arabic	0.8160 (33/45)	0.5770 (18/27)	0.5736 (11/19)
Bengali	0.8053 (43/49)	0.2597 (22/30)	0.2355 (6/21)
Chinese	0.8544 (43/46)	0.7082 (24/30)	0.5315 (16/22)
German	0.6693 (42/45)	0.5142 (19/28)	0.4548 (11/19)
English	0.7636 (55/60)	0.4545 (22/36)	0.4838 (9/24)
Persian	0.8348 (1/44)	0.5966 (6/27)	0.4764 (2/19)
Hausa	0.8313 (3/45)	0.2321 (18/28)	0.2058 (2/19)
Hindi	0.7854 (30/47)	0.7702 (14/30)	0.7356 (9/21)
Italian	0.4981 (42/44)	0.2080 (23/26)	—
Khmer	0.7320 (12/43)	0.5010 (—)	0.2635 (13/19)
Burmese	0.8631 (27/42)	0.5747 (19/25)	—
Nepali	0.8870 (33/44)	0.7637 (16/28)	0.6483 (8/20)
Odia	0.7685 (28/45)	0.4503 (19/27)	0.2183 (13/21)
Punjabi	0.7338 (37/44)	0.5061 (—)	0.5195 (5/20)
Polish	0.7823 (32/43)	0.4838 (20/27)	—
Russian	0.7459 (39/43)	0.4608 (19/27)	—
Spanish	0.7415 (41/50)	0.6121 (19/29)	0.4874 (7/21)
Swahili	0.7829 (19/44)	0.5099 (5/27)	0.5623 (4/20)
Telugu	0.8734 (19/45)	0.4270 (10/27)	0.3734 (9/21)
Turkish	0.7712 (31/42)	0.5636 (16/26)	0.4662 (11/19)
Urdu	0.7442 (39/46)	0.7563 (17/30)	0.7835 (12/21)

Table 9: Macro F1 scores with leaderboard rank (r/N) per language and subtask for our submission. "—" indicates the language was not evaluated for that subtask or the submission is not listed in the language-specific leaderboard.