wangkongqiang at SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization

Kongqiang Wang[1], **Qingli Tan**[2]
[1]School of Information Science and Engineering, Yunnan University,
[2]School of College of Ecology and Environment, Yunnan University,
Kunming 650500, Yunnan, China
[1]wangkongqiang60@gmail.com, [2]tanqingli@stu.ynu.edu.cn

Abstract

This paper presents our system developed for the SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization. on Subtask 1: Multilingual Text Classification Challenge - Polarization Detection. on Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification. on Subtask 3: Multilingual Text Classification Challenge - Manifestation Identification. To this end, we focus on English and Spanish language use two different pre-trained languages models: models–google-bert–bert-base-uncased, and models–microsoft–deberta-v3-base. We experiment with (1) the training set data is analyzed visually, (2) use the gemma-3-27b-it generative model to perform data augmentation on the training dataset through prompts, and (3) multiple numbers of single models are trained on the training set data. We further study the influence of different hyper-parameters on the single model and select the best single model for the prediction of the test set. Our submission achieved the good ranking place in the test set. All subtasks evaluated using Macro F1 score across different languages and cultural contexts. For Subtask 1, the English and Spanish language tasks are Macro F1 Score 0.7805 and 0.7155 respectively. For Subtask 2, the English and Spanish language tasks are Macro F1 Score 0.2603 and 0.4647 respectively. For Subtask 3, the English and Spanish language tasks are Macro F1 Score 0.2766 and 0.3322 respectively. For the final ranking, organizers will use the Macro F1 score. Even so, my approach has yielded good results from an overall perspective.

1 Introduction

Online polarization is the sharp division and hostility between social, political, or identity groups. Online polarization has become a growing concern, as it often precedes hate speech, offensive discourse, and social fragmentation. In its extreme form, polarization can create a fragmented society where individuals or groups are unable to engage in constructive dialogue, leading to a breakdown in community cohesion and social unity. Thus, detecting and mitigating polarization before it escalates is crucial to ensure safer and more inclusive online spaces. For the first time, the SemEval-2026 competition (Ghosh et al., 2026) organizers introduce a polarization task, aimed at detection of online polarization (Naseem et al., 2026b). The task focuses on the identification of multilingual, multicultural and multievent polarization, capturing the complexity of online discourse across diverse contexts. Participants may participate in one or more of the following three sub-tasks: Polarization Detection, Polarization Type Classification and Polarization Manifestation Identification. Based on the predictive task background of predictive post text, we propose the data augmentation learning method based on pre-train language model.

We developed for the SemEval-2026 Task 9: Detecting Multilingual, Multicultural And Multievent Online Polarization (Naseem et al., 2026a). on **Subtask 1: Multilingual Text Classification Challenge - Polarization Detection**. Binary classification to determine whether a post contains polarized content (polarized or not polarized). on **Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification**. Given a target text snippet, predict the target types of polarization. Specifically, select whether each of the following categories applies: $political$, $racial/ethnic$, $religious$, $gender/sexual$, or $other$. on **Subtask 3: Multilingual Text Classification Challenge - Manifestation Identification**. Given a target text snippet, predict how polarization is expressed. Specifically, select whether each of the following manifestations applies: $stereotype$, $vilification$, $dehumanization$, $extreme\_language$, $lack\_of\_empathy$, or $invalidation$. The code of this experiment

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 133–139
July 3-4, 2026

method is available on my GitHub website.[1]

2 Related Work

SemEval competition in previous years has introduced tasks focusing on multi-label text classification and text binary classification (Wang et al., 2024; Su and Zhou, 2024; Tran and Tran, 2024; Brekhof et al., 2024) to evaluate internal potential elements and potential content of the text. These tasks provided datasets with human labeled similarity scores, which have been extensively utilized for training sentence embedding models and conducting semantic evaluations.

2.1 Sentence Embeddings

Word embedding models such as BERT, GloVe, RoBeRTa and Word2Vec are frequently employed to assess the semantic distance between words. They are also some of the more commonly used methods in text classification tasks. Sentence embeddings with a fixed length are often generated via mean/max pooling of word embeddings or employing CLS embedding in BERT. The semantic distances are commonly measured using the cosine similarity of embeddings of two expressions. Siamese or triplet network architectures are frequently employed in sentence embedding training. For example, models such as Sentence-BERT utilize a dual-encoder architecture with shared weights for predicting sentence relationships (e.g., semantic contradiction, entailment, or neutral labeling) or for similarity score prediction using regression objectives. e.g., the difference between human annotated similarity score (sim) of two sentences and the cosine of two sentence embeddings.

2.2 Data Augmentation Learning

In previous studies (Muhammad et al., 2025a), data augmentation learning presents several advantages. The data augmentation approach can reduce the errors from insufficient examples by expanding the learning of the text content or can make the system more robust. In our study, using the gemma-3-27b-it generative model to perform data augmentation on the training dataset through prompts to generate large amounts of similar instances while making use of information from the large data samples during pre-training. Previous research (Muhammad et al., 2025b) has demonstrated that data augmentation learning can achieve remarkable success in the field of text classification.

In our study, we aim to use multiple pre-train learning models to assess task-specific of the text classification. When models are trained on diverse datasets with different architectures, they may produce varied predictions on task-specific of text classification, and fine-tune them using different hyperparameters may improve final performance. We use sentence embeddings mainly from the following models: $models--google-bert--bert-base-uncased$, and $models--microsoft--deberta-v3-base$.

3 Methodology

3.1 Overall Architecture

The pursued approach involves using a data augmentation system composed of the gemma-3-27b-it generative model and then training two different pre-train transformers. We trained several state-of-the-art natural language processing (NLP) models on a large dataset of annotated tweets to create two separate of classifiers with different architectures and configurations. We then feed the augmentation data from a data augmentation system into these classifiers to obtain the final predictions. We have used the following transformers for the predictions: $models--google-bert--bert-base-uncased$, and $models--microsoft--deberta-v3-base$.

For each instance, the final classification decision is based on the independent outputs of these models. The novel data augmentation system presented to improved the prediction of the pre-trained model to downstream tasks in most sub-tasks.

3.2 Implementation Step

First, set the API key, get from Gemini large language model (LLM) calling platform. The API key can be applied from the official website of Google AI Studio[2].

Second, since the content of the text columns provided by the three subtasks is the same, they are joined/merged with the id as the key. The text column is retained once, and all other tag columns (polarization, political, stereotype, etc.) are concatenated together. By invoking the API key of Google AI Studio using the $gemma-3-27b-it$ generative model to perform data augmentation on the training dataset through prompts to generate

[1] https://github.com/WangKongQiang//SemEval2026-Task-9

[2] https://aistudio.google.com/app/api-keys

large amounts of English or Spanish twitter texts, which is as similar as original text. Only the content of the text column has been changed. The label of this newly generated instance remains unchanged (Belay et al., 2025).

Third, the transformers python library that will be used below requires the data to be presented in a specific form. The following data samples split to contain only two columns: text and labels, where the text is an array equal in size to the number of multiple labels list.

Fourth, model definition. In this section, the different pre-trained transformer models to be evaluated are defined individually. The implementation primarily relies on the Trainer and TrainingArguments classes from the Hugging Face transformers python library, which enable training and evaluation of transformer models in a few straightforward steps.

Fifth, model training. Each of the aforementioned models is trained separately using either the complete training dataset or an augmented version of the dataset. The training process is conducted based on the previously defined parameters for convenience and consistency.

Sixth, model evaluation. Each pre-trained transformer model is evaluated independently using the validation split. The main evaluation metrics — accuracy, F1 score, precision, and recall — are computed and stored. Predictions for each validation instance are generated accordingly. After calculating these metrics, the model achieving the highest Macro F1 score can be identified. This model will be selected as the final model for testing. The input instances for the models are obtained through a data augmentation system. Once the outputs for the validation instances are computed, the validation results are saved to a csv file for further analysis.

Seventh, selecting the best model. After obtaining the predicted labels for all validation instances from each pre-trained transformer model, their respective evaluation metrics are calculated. Since the task involves binary or multi-class classification, the best-performing model is determined based on the highest Macro F1 score.

Finally, the selected model is used to generate predictions for the test dataset. These results are further utilized to produce evaluation visualizations, including the confusion matrix and the ROC curve.

Table 1: The English text experiment data situation and the number of Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification labels are described.

<table>
<tr><th>Training Set Text</th><th>Value</th></tr>
<tr><td>|count|</td><td>3222.000000</td></tr>
<tr><td>|mean|</td><td>13.565177</td></tr>
<tr><td>|std|</td><td>9.041417</td></tr>
<tr><td>|min|</td><td>5.000000</td></tr>
<tr><td>|25%|</td><td>8.000000</td></tr>
<tr><td>|50%|</td><td>10.000000</td></tr>
<tr><td>|75%|</td><td>17.000000</td></tr>
<tr><td>|max|</td><td>62.000000</td></tr>
</table>

<table>
<tr><th>Training Set Label</th><th>Value</th></tr>
<tr><td>|political|</td><td>1150</td></tr>
<tr><td>|racial/ethnic|</td><td>281</td></tr>
<tr><td>|religious|</td><td>112</td></tr>
<tr><td>|gender/sexual|</td><td>72</td></tr>
<tr><td>|other|</td><td>126</td></tr>
</table>

Table 2: The Spanish text experiment data situation and the number of Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification labels are described.

<table>
<tr><th>Training Set Text</th><th>Value</th></tr>
<tr><td>|count|</td><td>3305.000000</td></tr>
<tr><td>|mean|</td><td>12.271710</td></tr>
<tr><td>|std|</td><td>6.240053</td></tr>
<tr><td>|min|</td><td>3.000000</td></tr>
<tr><td>|25%|</td><td>7.000000</td></tr>
<tr><td>|50%|</td><td>11.000000</td></tr>
<tr><td>|75%|</td><td>16.000000</td></tr>
<tr><td>|max|</td><td>58.000000</td></tr>
</table>

<table>
<tr><th>Training Set Label</th><th>Value</th></tr>
<tr><td>|political|</td><td>901</td></tr>
<tr><td>|racial/ethnic|</td><td>623</td></tr>
<tr><td>|religious|</td><td>525</td></tr>
<tr><td>|gender/sexual|</td><td>443</td></tr>
<tr><td>|other|</td><td>443</td></tr>
</table>

4 Results and Analysis

4.1 Training Dataset Analysis

The text and label columns of training set is described in Table 1 and Table 2 respectively for English and Spanish. Labels here only analyzes Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification. The length and quantity distribution of training text experiment data samples are analyzed in Figure 1 and Figure 2 respectively for English and Spanish language. Distribution of the size of texts for each class in Figure 3 and Figure 4 respectively for English and Spanish language. It shows the number of percentages relative to each class for various text size cases. Classes here only analyze on Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification.

4.2 Experimentation Configuration

For the sake of completeness and in an attempt to improve the results obtained by the transformer pretrain model. For $models - google - bert - bert - base - uncased$, and $models - microsoft - deberta - v3 - base$, the two different hyperparameters were used respectively: See Table 3.

Table 3: Experimentation configuration hyperparameters for the two different transformer pre-train models.

<table>
<tr><th>model</th><th>Hyperparameter</th><th>Values</th></tr>
<tr><td>models-google-bert-bert-base-uncased</td><td>|num_train_epochs|</td><td>3</td></tr>
<tr><td></td><td>|learning_rate|</td><td>2e-5</td></tr>
<tr><td></td><td>|per_device_train_batch_size|</td><td>64</td></tr>
<tr><td></td><td>|per_device_eval_batch_size|</td><td>8</td></tr>
<tr><td></td><td>|logging_steps|</td><td>100</td></tr>
<tr><td>models-microsoft-deberta-v3-base</td><td>|num_train_epochs|</td><td>10</td></tr>
<tr><td></td><td>|learning_rate|</td><td>2e-5</td></tr>
<tr><td></td><td>|per_device_train_batch_size|</td><td>32</td></tr>
<tr><td></td><td>|warmup_steps|</td><td>100</td></tr>
<tr><td></td><td>|weight_decay|</td><td>0.01</td></tr>
<tr><td></td><td>|logging_steps|</td><td>100</td></tr>
<tr><td></td><td>|save_total_limit|</td><td>1</td></tr>
</table>

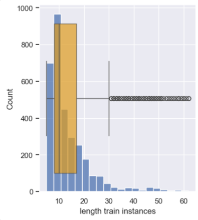

Figure 1: The length and quantity distribution of training text experiment data samples in English are analyzed.

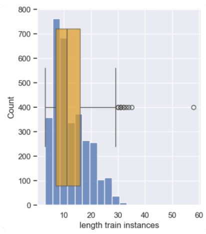

Figure 2: The length and quantity distribution of training text experiment data samples in Spanish are analyzed.

4.3 Development Dataset Result

The following Table 4 records the official results of SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization, on Subtask 1: Multilingual Text Classification Challenge - Polarization Detection, Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification and Subtask 3: Multilingual Text Classification Challenge - Manifestation Identification. The metrics recorded by black bold text is the best (winning) approach in the evaluation task of the development dataset for English and Spanish language respectively.

4.4 Test Dataset Result

The following Table 5 records the official results of SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization, on Subtask 1: Multilingual Text Classification Challenge - Polarization Detection, Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification and Subtask 3: Multilingual Text Classification Challenge - Manifestation Identification. The metrics recorded by black bold text is the best (winning) approach in the evaluation task of the test dataset for English and Spanish language respectively.

Table 4: The development dataset experiment situation detailed results are described.

Subtask	Language	Main Technologies	Accuracy	Precision	Recall	F1 Binary	F1 Macro	F1 Micro	Exact Match
1	|English|	models–google-bert–bert-base-uncased	0.7594	0.7436	0.5686	0.6444	0.7313	0.7594	-
1	|Spanish|	models–google-bert–bert-base-uncased	0.6606	0.6707	0.6548	0.6627	0.6606	0.6606	-
1	|English|	models–microsoft–deberta-v3-base	0.7895	0.7674	0.6471	0.7021	0.7697	0.7895	-
1	|Spanish|	models–microsoft–deberta-v3-base	0.703	0.6923	0.75	0.72	0.7019	0.703	-
1	|English|	models–google-bert–bert-base-uncased + gemma-3-27b-it	0.812	0.7955	0.6863	0.7368	0.7953	0.812	-
1	|Spanish|	models–google-bert–bert-base-uncased + gemma-3-27b-it	0.6667	0.6593	0.7143	0.6857	0.6654	0.6667	-
1	|English|	**models–microsoft–deberta-v3-base + gemma-3-27b-it**	**0.812**	**0.7955**	**0.6863**	**0.7368**	**0.7953**	**0.812**	**-**
1	|Spanish|	**models–microsoft–deberta-v3-base + gemma-3-27b-it**	**0.703**	**0.6923**	**0.75**	**0.72**	**0.7019**	**0.703**	**-**
Subtask	**Language**	**Main Technologies**	**F1 Micro**	**F1 Macro**	**Precision Micro**	**Precision Macro**	**Recall Micro**	**Recall Macro**	**Exact Match**
2	|English|	**models–google-bert–bert-base-uncased**	**0.5846**	**0.3451**	**0.717**	**0.4538**	**0.4935**	**0.2923**	**-**
2	|Spanish|	models–google-bert–bert-base-uncased	0.5652	0.5284	0.6094	0.5624	0.527	0.5182	-
2	|English|	models–microsoft–deberta-v3-base	0.5556	0.2089	0.7143	0.239	0.4545	0.1855	-
2	|Spanish|	**models–microsoft–deberta-v3-base**	**0.6022**	**0.5698**	**0.6694**	**0.6517**	**0.5473**	**0.5396**	**-**
2	|English|	models–google-bert–bert-base-uncased + gemma-3-27b-it	0.5802	0.3213	0.7037	0.3579	0.4935	0.2923	-
2	|Spanish|	models–google-bert–bert-base-uncased + gemma-3-27b-it	0.5842	0.567	0.5944	0.5767	0.5743	0.5608	-
2	|English|	models–microsoft–deberta-v3-base + gemma-3-27b-it	0.5538	0.2147	0.6792	0.2319	0.4675	0.2009	-
2	|Spanish|	models–microsoft–deberta-v3-base + gemma-3-27b-it	0.5862	0.5543	0.5986	0.5607	0.5743	0.5564	-
Subtask	**Language**	**Main Technologies**	**F1 Micro**	**F1 Macro**	**Precision Micro**	**Precision Macro**	**Recall Micro**	**Recall Macro**	**Exact Match**
3	|English|	models–google-bert–bert-base-uncased	0.4646	0.351	0.6211	0.475	0.3711	0.304	0.594
3	|Spanish|	models–google-bert–bert-base-uncased	0.4068	0.353	0.5217	0.4441	0.3333	0.3098	0.4061
3	|English|	**models–microsoft–deberta-v3-base**	**0.5109**	**0.4335**	**0.6087**	**0.4958**	**0.4403**	**0.3891**	**0.5789**
3	|Spanish|	models–microsoft–deberta-v3-base	0.3952	0.3053	0.5593	0.3822	0.3056	0.27	0.4545
3	|English|	models–google-bert–bert-base-uncased + gemma-3-27b-it	0.4591	0.3858	0.602	0.4849	0.3711	0.3277	0.609
3	|Spanish|	models–google-bert–bert-base-uncased + gemma-3-27b-it	0.3989	0.3423	0.5185	0.4274	0.3241	0.2934	0.4242
3	|English|	models–microsoft–deberta-v3-base + gemma-3-27b-it	0.5076	0.4253	0.6381	0.5101	0.4214	0.3714	0.6165
3	|Spanish|	**models–microsoft–deberta-v3-base + gemma-3-27b-it**	**0.4345**	**0.3661**	**0.5455**	**0.457**	**0.3611**	**0.314**	**0.4545**

Table 5: The test dataset experiment situation detailed results are described.

Subtask	Language	Main Technologies	Accuracy	Precision	Recall	F1 Binary	F1 Macro	F1 Micro	Exact Match
1	|English|	**models–microsoft–deberta-v3-base + gemma-3-27b-it**	**0.7948**	**0.714**	**0.7355**	**0.7246**	**0.7805**	**0.7948**	**-**
1	|Spanish|	**models–microsoft–deberta-v3-base + gemma-3-27b-it**	**0.7157**	**0.7191**	**0.6966**	**0.7077**	**0.7155**	**0.7157**	**-**
Subtask	**Language**	**Main Technologies**	**F1 Micro**	**F1 Macro**	**Precision Micro**	**Precision Macro**	**Recall Micro**	**Recall Macro**	**Exact Match**
2	|English|	**models–google-bert–bert-base-uncased**	**0.5687**	**0.2603**	**0.689**	**0.5698**	**0.4841**	**0.2156**	**-**
2	|Spanish|	**models–microsoft–deberta-v3-base**	**0.5117**	**0.4647**	**0.6812**	**0.6437**	**0.4098**	**0.3949**	**-**
Subtask	**Language**	**Main Technologies**	**F1 Micro**	**F1 Macro**	**Precision Micro**	**Precision Macro**	**Recall Micro**	**Recall Macro**	**Exact Match**
3	|English|	**models–microsoft–deberta-v3-base**	**0.4169**	**0.2766**	**0.5481**	**0.4238**	**0.3364**	**0.2645**	**0.5944**
3	|Spanish|	**models–microsoft–deberta-v3-base + gemma-3-27b-it**	**0.399**	**0.3322**	**0.539**	**0.4432**	**0.3168**	**0.2756**	**0.4819**

4.5 Text and Label Biased Performance

From Figure 1 and Figure 2 of the visual analysis, we can observe that 75% of tweets in training dataset, either in the chart or in the previous model input column content, have no more than 20 words. This information could be useful in determining the size of a network of neurons, or when a sentence length limit needs to be set.

SemEval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization, on Subtask 2: Multilingual Text Classification Challenge - Polarization Type Classification. This task is multi-label classification. Each instance can have 0 to n (n=5) categories, and you need to predict which category each instance belongs to. In the specific cases (English and Spanish language) we focus on, there may be up to five different categories: $political$, $racial/ethnic$, $religious$, $gender/sexual$, and $other$. For this subtask, we will look at the quantity distribution followed by each category, as shown in Table 1 and Table 2 respectively for English and Spanish language. In this case, independent category percentages for each samples cannot be assessed because of the labels intersection.

5 Conclusion

Our multiple model systems employ a data augmentation learning approach to estimate task-specific of text classification (Agirre et al., 2014), integrating results from multiple systems: $models-google-bert-bert-base-uncased$ and $models-microsoft-deberta-v3-base$. The hyperparameter is following: per_device_eval_batch_size is 8, num_train_epochs is 3, learning_rate is 2e-5, logging_stpes is 100, per_device_train_batch_size is 64. The dataset usage is shown in Table 6. Our findings suggest that task-specific of text classification can be deduced from a variety of sources. Although some features (e.g., lexical overlap ratio)

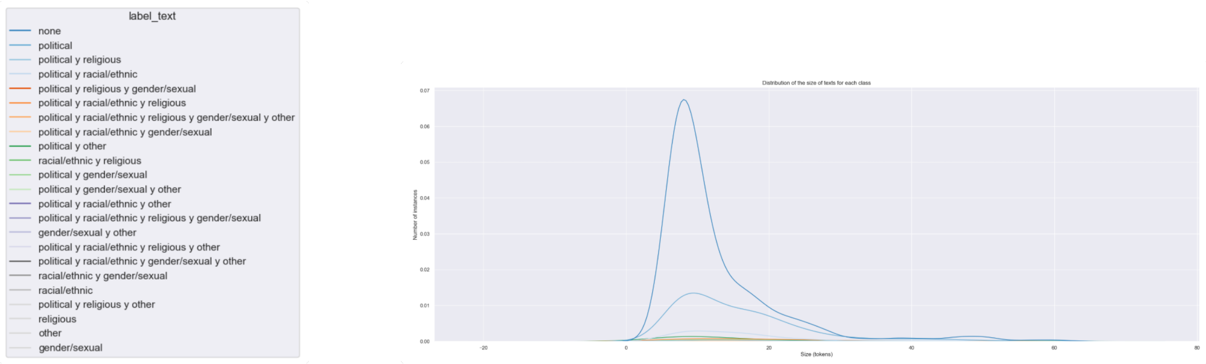

Figure 3: Distribution of the size of English texts for each class.The horizontal axis indicates the size of the English text, while the vertical axis shows the percentage proportion of the corresponding text size.

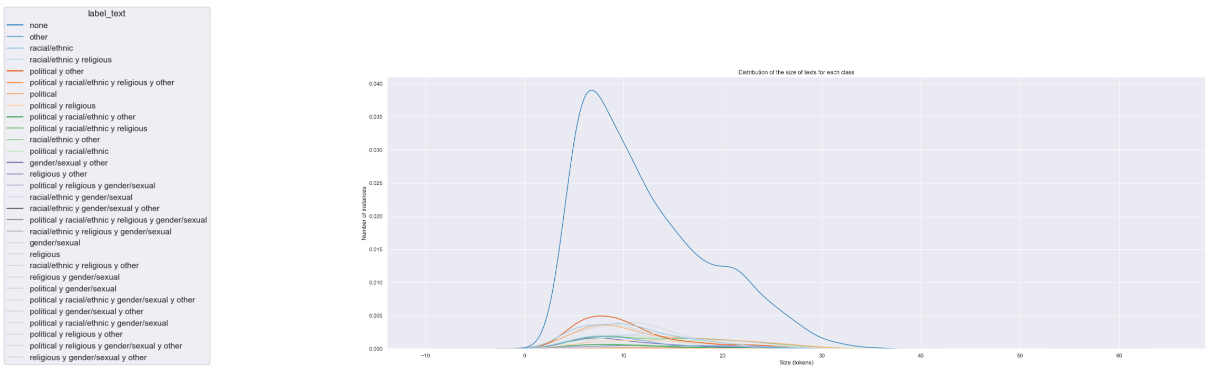

Figure 4: Distribution of the size of Spanish texts for each class. The horizontal axis indicates the size of the Spanish text, while the vertical axis shows the percentage proportion of the corresponding text size.

Table 6: Use dataset supported by Semeval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization. The style is based on raw data without any manual alteration of the dataset.

Dataset Input	Description	Use or Not
POLAR (Naseem et al., 2026b)	|it is a multilingual, multicultural, and multi-event dataset| |with over 110K instances in 22 languages.|	yes
other dataset	|use external or additional corpora.|	no

may not perform as strongly as models specifically designed to obtain sentence representations, the results demonstrate that these features, when used in a combined manner, can outperform many individual systems and collaboratively achieve a better correlation with human judgment on task-specific of text classification (Siino, 2024). The combined use of these features will also be part of our future research efforts.

6 Limitation and Future Work

Our experiments are based on English and Spanish language datasets only. Constrained by the size of the training dataset and the availability of pre-train language models, it is regrettable that we did not offer insights into other Asian and African languages (Vaidya et al., 2024). In future research, studies on low-resource languages will be valuable. Future works including tasks such as data collection, annotation, and training models tailored to these languages.

Acknowledgments

We are very grateful for the assistance and discussions provided by Semeval-2026 Task 9: Detecting Multilingual, Multicultural and Multievent Online Polarization leader and organizer.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, page 81–91, Dublin, Ireland. Association for Computational Linguistics.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In

Proceedings of the 31st International Conference on Computational Linguistics, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Thijs Brekhof, Xuanyi Liu, Joris Ruitenbeek, Niels Top, and Yuwen Zhou. 2024. Groningen team D at SemEval-2024 task 8: Exploring data generation and a combined model for fine-tuning LLMs for multidomain machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. 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, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, Dheeraj Kodati, Sahar Moradizeyveh, Firoj Alam, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Nelson Odhiambo Onyango, Clemencia Siro, Ibrahim Said Ahmad, Lilian Wanzare, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 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, Xintong Wang, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

Marco Siino. 2024. All-mpnet at SemEval-2024 task 1: Application of mpnet for evaluating semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 379–384, Mexico City, Mexico. Association for Computational Linguistics.

Lianshuang Su and Xiaobing Zhou. 2024. NLP_STR_teamS at SemEval-2024 task1: Semantic textual relatedness based on MASK prediction and BERT model. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 337–341, Mexico City, Mexico. Association for Computational Linguistics.

Bao Tran and Nhi Tran. 2024. NewbieML at SemEval-2024 task 8: Ensemble approach for multidomain machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 354–360, Mexico City, Mexico. Association for Computational Linguistics.

Ankit Vaidya, Aditya Gokhale, Arnav Desai, Ishaan Shukla, and Sheetal Sonawane. 2024. CLTeam1 at SemEval-2024 task 10: Large language model based ensemble for emotion detection in Hinglish. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 365–369, Mexico City, Mexico. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.