ServSocIA at Semeval-2026 Task 9: Evaluating Prompt Strategies for Polarization Detection

Jacob Altamirano, Mario León-Pérez
Facultad de Ciencias
UNAM

Bruno Ruiz Juarez
Facultad de Estudios Superiores
de Acatlán
UNAM

Luis Chiruzzo
Facultad de Ingeniería
Universidad de la República
Uruguay

Helena Gómez, Fazlourrahman Balouchzahi
Instituto de Investigaciones en Matemáticas
Aplicadas y en Sistemas - UNAM

Abstract

This paper presents our approach to Subtask 1 of SemEval-2026 Task 9 on multilingual polarization detection in social media texts in English and Spanish. We model the task as a prompt-based binary classification problem and systematically compare zero-shot, one-shot, and few-shot strategies across multiple large language models accessed via commercial APIs, without task-specific fine-tuning. Our controlled experimental setup enforces strict data separation and consistent decoding conditions to analyze the impact of in-context supervision across architectures and languages. Results indicate that well-structured prompting enables competitive performance, though implicit and culturally nuanced polarization remains challenging.

1 Introduction

Online discourse has become increasingly polarized across linguistic and cultural boundaries, motivating the development of computational methods for detecting polarization in multilingual social media data (Cinelli et al., 2021; Naseem et al., 2026b). This paper describes our participation in Subtask 1 of the SemEval 2026 shared task on multilingual polarization detection (Naseem et al., 2026a), which focuses on identifying polarized content in social media texts in both English and Spanish. The task emphasizes the cross-cultural and cross-lingual dimensions of polarization, making robust and adaptable modeling approaches particularly important. Our work is grounded in the task's operational definition of polarization as "stereotyping, vilification, dehumanization, deindividuation, or intolerance of other people's views, beliefs, and identities" (Naseem et al., 2026b). By explicitly aligning our modeling framework with this definition, we ensure conceptual consistency between the sociolinguistic understanding of polarization and its computational implementation.

Our primary strategy relies on prompt-based inference using large language models (LLMs). We systematically explored Zero-shot, One-shot, and Few-shot prompting configurations to evaluate how varying levels of in-context supervision influence performance. To enable a controlled and systematic comparison, we leveraged APIs from Cohere[1], Groq[2], and Gemini[3], allowing us to assess multiple model architectures under consistent experimental conditions. This comparative framework enabled us to identify the most effective model–prompting combinations for multilingual polarization detection and to analyze how different LLMs respond to structured contextual guidance. Participation in the shared task yielded both quantitative and qualitative insights. Our best-performing configuration achieved a Macro-F1 score of 0.7652 for English and 0.7073 for Spanish. We observed that improvements from Few-shot prompting were model-dependent: while certain models benefited substantially from additional in-context examples, others demonstrated relatively stable performance across prompting strategies. Qualitative error analysis further revealed persistent challenges in detecting implicit or context-dependent polarization, particularly in culturally nuanced Spanish instances. These findings highlight both the strengths and the limitations of prompt-based LLM approaches in cross-lingual polarization detection settings.

All code, prompts, and experimental configurations described in this paper are publicly available in our GitHub repository[4].

[1] https://cohere.com/
[2] https://groq.com/
[3] https://gemini.google.com/
[4] https://github.com/PLN-disca-iimas/

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2754–2759

2 Background

Polarization refers to discourse that constructs or reinforces antagonistic divisions between social, political, or identity-based groups, often through stereotyping, delegitimization, or exclusionary framing (Naseem et al., 2026b). In online environments, such polarized narratives can intensify group conflict and contribute to the fragmentation of public discourse.

To facilitate the systematic study of this phenomenon, Subtask 1 of SemEval-2026 Task 9 (Naseem et al., 2026a) formulates polarization detection as a binary classification problem. Given a short user-generated text and its identifier, systems must predict either *Polarized* (1) or *Not Polarized* (0). The POLAR dataset (Naseem et al., 2026a) comprises multilingual and multicultural online texts, released in training, development, and test splits. In this work, we focus on the English and Spanish tracks, modeling each language independently to account for linguistic and cultural variation in the expression of polarization.

Polarization detection intersects with several established research areas in natural language processing. It is related to sentiment analysis (Pang and Lee, 2008), stance detection (Mohammad et al., 2016), and hate speech detection (Schmidt and Wiegand, 2017). However, it differs conceptually from these tasks. Unlike sentiment analysis, which captures affective polarity, polarization detection emphasizes structural opposition and divisive framing between groups. Similarly, while hate speech detection targets explicitly abusive language, polarized discourse may manifest through more subtle rhetorical mechanisms, including implicit in-group/out-group distinctions or framing conflicts as fundamentally irreconcilable.

Recent advances in transformer-based architectures such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have demonstrated strong performance in social media text classification tasks. More recently, LLMs have shown competitive zero-shot and few-shot capabilities through prompt-based inference (Brown et al., 2020). These models leverage extensive pretraining to generalize to downstream tasks without parameter updates. Our work builds on this paradigm by systematically evaluating prompt-based inference strategies for multilingual polarization detection without task-specific fine-tuning, contributing empirical insights into how prompting configurations influence performance across languages.

3 System Overview

Our system formulates polarization detection as a prompt-based text classification task using LLMs accessed through the Gemini, Cohere and Groq APIs. Rather than training a task-specific classifier, we construct structured prompts that combine task instructions with a small set of labeled examples sampled from the training data. Multiple models were explored during development, including both reasoning-oriented and instruction-tuned variants, allowing us to compare model behavior under identical prompting conditions. We implemented a reproducible pipeline that strictly separates demonstration examples from evaluation inputs.

3.1 Data Handling and Leakage Prevention

We utilize the official dataset partitioning provided by the shared task, ensuring strict separation between training and test splits throughout prompt construction, model selection, and evaluation. At no stage are the splits merged. For prompt-based inference, binary labels are mapped to their textual representations (*Polarized* and *Not Polarized*), aligning the classification objective with the natural language output space of large language models.

For few-shot prompting, we sample a balanced subset of training instances to construct a dedicated demonstration pool. These examples are removed from the evaluation subset, guaranteeing that no instance appears both as a prediction target and as an in-context example. This protocol mitigates information leakage and preserves the validity of performance estimates.

3.2 Prompt-Based Inference

For each input instance x, we construct a structured prompt designed to guide the model toward a constrained binary decision. The prompt consists of (i) an explicit definition of polarization grounded in the task description, (ii) optionally, a set of balanced in-context examples, (iii) the target input text, and (iv) explicit instructions requiring the model to output exactly one of the two predefined labels.

The complete prompt template is provided in Appendix B.

Model responses are post-processed through deterministic label normalization. Outputs matching *Polarized* are mapped to 1, and outputs matching

`polarization-Semeval26`

Language	Split	Class 0	Class 1	Total
English	Train	2047 (63.53%)	1175 (36.47%)	3222
English	Dev	101 (63.12%)	59 (36.88%)	160
English	Test	919 (63.29%)	533 (36.71%)	1452
Spanish	Train	1645 (49.77%)	1660 (50.23%)	3305
Spanish	Dev	81 (49.09%)	84 (50.91%)	165
Spanish	Test	753 (50.60%)	735 (49.40%)	1488

Table 1: Class distribution across training, development, and test splits.

Not Polarized are mapped to 0:

$$\text{Polarized} \rightarrow 1, \qquad \text{Not Polarized} \rightarrow 0.$$

To ensure robustness, responses that deviate from the expected label set are mapped to the majority-safe fallback class *Not Polarized*, preventing malformed generations from affecting evaluation consistency.

3.3 Inference Pipeline

Predictions are produced by iterating over dataset instances, sampling few-shot examples, prompts, and querying the model in batches. The process is summarized below.

```
Algorithm 1: Prompt-based Prediction
Input: dataset D, example pool E, model M
Output: predicted labels Y

for each text x in D do
    sample k examples from E
    construct prompt P
    response ← API_call(M, P)
   label ← map_response_to_label(response)
    append label to Y
return Y
```

Batching reduces API calls and token usage, while temperature is fixed to zero to ensure deterministic outputs.

4 Experimental Setup

4.1 Dataset

All experiments follow the official SemEval data organization. Table 1 shows a label distribution of Subtask 1 for the three splits in English and Spanish.

Training data is used exclusively to construct the few-shot example pool and to compare model configurations, while test data is used only for generating submission predictions.

4.2 Few-shot Example Pool

To avoid leakage, we sample $n = 50$ examples per class from the training split to form a demonstration pool. Since each prompt uses at most five examples, this pool size ensures sufficient diversity across prompts while preventing overlap with evaluation instances.

4.3 Preprocessing

Preprocessing is intentionally minimal. Numeric labels are converted into their textual representations (*Polarized* and *Not Polarized*) to align with the prompt-based inference setup.

No additional normalization, tokenization, lowercasing, or language-specific transformations are applied. The original text is preserved to maintain stylistic and discourse-level cues that may inform polarization detection.

4.4 Models and Hyperparameters

Inference is performed through the Cohere chat API, and direct API call for Gemini. During development, multiple model variants were evaluated under identical prompting conditions. The only decoding parameter explicitly controlled is temperature, fixed at zero to encourage deterministic responses. Prompting mode (zero-, one-, or few-shot) is treated as an experimental factor rather than a learned hyperparameter, and inference is executed in batches of 16 inputs.

4.5 Output Parsing

Predictions are parsed line-by-line and mapped to binary labels. Invalid outputs are logged for analysis and defaulted to the negative class.

5 Results

Table 2 presents the results of the experiments using the zero-, one-, and few-shot strategies in both languages, English and Spanish, on the development set. We can observe that the performance

Model	Prompt Mode	MacroF1-Eng	Acc.-Eng	MacroF1-Spa	Acc.-Spa
command-a-reasoning-08-2025	one	**0.781**	**0.788**	0.742	0.750
command-a-reasoning-08-2025	zero	0.758	0.769	**0.765**	**0.769**
command-a-reasoning-08-2025	few	0.751	0.769	0.731	0.744
command-r-plus-08-2024	one	0.656	0.694	0.648	0.681
command-r-plus-08-2024	zero	0.630	0.644	0.603	0.631
command-r-plus-08-2024	few	0.681	0.719	0.695	0.725
command-a-03-2025	one	0.538	0.569	0.587	0.588
command-a-03-2025	zero	0.632	0.663	0.619	0.619
command-a-03-2025	few	0.539	0.563	0.648	0.650
command-r-08-2024	one	0.587	0.588	0.618	0.619
command-r-08-2024	zero	0.530	0.531	0.540	0.544
command-r-08-2024	few	0.599	0.606	0.633	0.638
c4ai-aya-expanse-32b	one	0.529	0.531	0.573	0.575
c4ai-aya-expanse-32b	zero	0.574	0.575	0.549	0.550
c4ai-aya-expanse-32b	few	0.525	0.525	0.587	0.588
gemini-3-pro-preview	one	0.737	0.750	0.687	0.691
gemini-3-pro-preview	zero	0.760	0.775	0.690	0.673
gemini-3-pro-preview	few	0.737	0.781	0.692	0.709
Kimi-K2	one	0.740	0.744	0.715	0.672
Kimi-K2	zero	0.632	0.638	0.685	0.667
Kimi-K2	few	0.695	0.706	0.715	0.690
Allam-7B	one	0.593	0.594	0.642	0.515
Allam-7B	zero	0.662	0.663	0.613	0.527
Allam-7B	few	0.604	0.606	0.678	0.515
LlaMa-70B	one	0.586	0.587	0.718	0.648
LlaMa-70B	zero	0.563	0.568	0.724	0.654
LlaMa-70B	few	0.606	0.606	0.735	0.672

Table 2: Performance comparison of evaluated models and prompting modes on the development set.

Metric	**English**	**Spanish**
Model	gemini-3-pro-preview	command-a-reasoning
Prompt Mode	zero	zero
Accuracy	0.7679	0.7097
Precision	0.6283	0.7447
Recall	0.9006	0.6272
F1 Binary	0.7402	0.6809
F1 Macro	0.7652	0.7073
F1 Micro	0.7679	0.7097

Table 3: Final results on the Test sets.

varies substantially across architectures. The strongest overall results in English are obtained by `command-a-reasoning-08-2025` (one-shot, Macro-F1 = 0.781) and `gemini-3-pro-preview` (zero-shot, Macro-F1 = 0.760). In Spanish, the highest Macro-F1 is achieved by `command-a-reasoning-08-2025` (zero-shot, 0.765), followed closely by its one-shot configuration (0.742). This indicates that reasoning-oriented instruction-tuned models are particularly effective for this task. With respect to the prompting strategies, the impact is clearly model-dependent. For some models (e.g., `command-r-plus-08-2024` and `LLaMa-70B`), few-shot prompting yields noticeable gains over zero-shot. In contrast, other models (e.g., `gemini-3-pro-preview` in English) achieve their best performance in zero-shot settings, suggesting that additional in-context examples do not universally improve results.

The best-performing models were evaluated on the test dataset. The optimal performance for English was achieved utilizing a zero-shot approach with Gemini. Table 3 shows the performance on the test set of the best of the models that we submitted to the shared task.

6 Conclusions

In this paper, we presented ServSocIA's participation in Subtask 1 of SemEval-2026 Task 9 on multilingual polarization detection. Our approach framed polarization detection as a prompt-based binary classification task, systematically evaluating zero-shot, one-shot, and few-shot prompting strategies across multiple large language models under controlled experimental conditions.

Our findings demonstrate that prompt-based inference can achieve competitive performance without task-specific fine-tuning, reaching a Macro-F1 of 0.7652 in English and 0.7073 in Spanish on the official test sets. Notably, performance gains from in-context examples were model-dependent: while certain models benefited from few-shot prompting, others achieved strong results in zero-shot settings. This variability suggests that prompting effectiveness is closely tied to model architecture and

instruction-following capabilities rather than the mere presence of demonstrations.

We also observed that development performance did not always transfer consistently to the test set, highlighting the sensitivity of prompt-based systems to data distribution and example sampling. In particular, implicit or culturally nuanced forms of polarization remain challenging, especially in Spanish, where contextual framing and discourse-level cues play a central role.

Future work may explore more principled demonstration selection strategies, such as diversity-aware or scenario-based example curation, instead of random sampling. Additionally, structured output constraints or lightweight calibration methods could further improve robustness. Overall, our study provides empirical insights into how prompting configurations influence multilingual polarization detection and contributes to understanding the practical trade-offs of LLM-based classification in socially sensitive tasks.

Acknowledgments

This work was supported by compute credits from a Cohere Labs Research Grant, these grants are designed to support academic partners conducting research with the goal of releasing scientific artifacts and data for good projects. Fazlourrahman Balouchzahi acknowledges the support from the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico, through the Postdoctoral Fellowship Program (EPM 2025). This research is with support from Google.org and the Google Cloud Research Credits program for the Gemini Academic Program.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the national academy of sciences*, 118(9):e2023301118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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. In *Computing Research Repository*, volume arXiv:1907.11692.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, "Ozge Alacam, Cengiz Acar"urk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, and 15 others. 2026a. SemEval-2026 task 9: Detecting multilingual, multicultural and multievent online polarization. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *arXiv preprint arXiv:2505.20624*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Appendix

A Packages used

Experiments were conducted using Python 3.13 with pandas 2.3.3, numpy 2.3.5, scikit-learn 1.7.2, python-dotenv 1.1.1, and tqdm 4.67.1, with model inference performed via the Cohere and Groq APIs.

B Prompt Template (ENG)

The following prompt template was used in our experiments. Few-shot examples were inserted be-

tween the instruction block and the input text when applicable.

```
You are a linguistic analyst tasked with
    detecting polarization in short texts.

Polarization occurs when a statement presents
    issues in a divisive manner,
frames opposing sides as incompatible, uses
    emotionally charged language
to emphasize conflict, or portrays one group or
    viewpoint negatively
in contrast to another.

Your task is to classify the following text as
    either:

Polarized
Not Polarized

Respond with exactly one of these two labels and
     no additional text.

Text:
{INPUT_TEXT}

Label:
```