NLP-CIMAT at SemEval-2026 Task 9: LLM-Based One-Shot and Cross-Lingual Data Augmentation for Polarization Detection

Miriam Calderón-Reyes[1], Fernando Sánchez-Vega[1, 2], Adrián Pastor López-Monroy[1]
[1] Computer Science Department, Mathematics Research Center (CIMAT)
Jalisco S/N, 36023, Guanajuato, Guanajuato, México
Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI)
Av. Insurgentes Sur 1582, 03940, CDMX, México
{miriam.calderon, fernando.sanchez, pastor.lopez}@cimat.mx

Abstract

This paper describes our participation in SemEval 2026 Task 9: Multilingual Text Polarization. The task requires estimating polarization levels across languages, where linguistic variability and limited annotated data pose significant challenges. To address data scarcity, we propose a pipeline that combines cross-lingual translation, synthetic data augmentation via LLMs, and domain-specific pre-trained models. Our approach leverages the hypothesis that polarization signals can transfer across languages without substantial loss of semantic alignment, enabling effective data augmentation through translation. Notably, one-shot synthetic example generation emerges as a viable strategy for enriching training data in topic-specific scenarios. Experimental results demonstrate high stability and competitive performance, achieving a macro F1-score of 0.7869 for Spanish and 0.7939 for English on the test set, ranking 21th on the official English leaderboard, while our Spanish results are competitive with top-performing systems, corresponding to 7th place.

1 Introduction

Digital platforms have changed public discourse into a fragmented space where differences of opinion turn into isolated groups that define themselves by rejecting opposing voices (Sunstein, 2017; Bakshy et al., 2015; Pariser, 2011). This division becomes a social problem when conflict stops being about ideas and starts targeting people: the other side is no longer someone who thinks differently, but an enemy to be excluded through stereotypes (Bramson et al., 2016; Mason, 2015). The result is attitude polarization—a deep hostility that reduces empathy, encourages hate speech, and weakens public debate (Cinelli et al., 2021; Quattrociocchi et al., 2016).

Detecting polarization automatically is particularly challenging because hostile discourse rarely presents itself as such. Polarized language often hides behind opinions, making the line between critical expression and antagonistic intent hard to draw (Davidson et al., 2017). Unlike explicit hate speech, polarization works through framing, insinuation, and in-group signals—subtle patterns that require contextual and cultural knowledge to identify (Bail et al., 2018; Basile et al., 2019).

This makes supervised models hard to train: they need large amounts of labeled data to learn these subtle patterns, but annotated datasets are scarce, especially outside English (Röttger et al., 2021; Magueresse et al., 2020). A direct response is data augmentation. Since polarization is tied to real social context, translating existing datasets from other languages provides authentic examples (Kumar et al., 2020; Conneau et al., 2020), while using LLMs to generate examples across underrepresented topics helps the model cover a wider range of domains (Møller et al., 2024).

This study introduces a binary classification method for polarization, using cross-lingual data augmentation to address limited training data by translating polarized datasets into Spanish[1] and English and further enriching them with synthetic examples covering underrepresented topics. Developed for SemEval 2026 Task 9 - Subtask 1: Multilingual Text Classification Challenge (Naseem et al., 2026a)[2], this framework provides a scalable way to detect antagonistic intent and monitor digital public debate. Our findings suggest that cross-lingual data supports polarization detection, especially for closer languages, while synthetic data expands coverage and improves robustness beyond conventional domains.

[1] Due to the submission policy, only the last submitted run is shown on the leaderboard, corresponding to our English model. However, our Spanish results were also evaluated under the official framework.

[2] Code and resources are available at https://github.com/miricalderonr/nlpcimat-at-POLAR

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2886–2893

2 Related Work

A central challenge in polarization detection is the scarcity of labeled data, particularly in languages other than English (Bender, 2019; Joshi et al., 2020). Data augmentation has been widely studied as a mitigation strategy, with back-translation emerging as the predominant technique: a text is translated into a pivot language and back, yielding a paraphrase of the original (Sennrich et al., 2016; Wei and Zou, 2019; Feng et al., 2021a). However, back-translation operates within the same language, primarily preserving the communicative intent and core semantics of the source (Edunov et al., 2018).

Cross-lingual transfer offers a more diverse alternative. Unlike back-translation, which recycles existing examples within the source language, annotated instances from other languages are translated directly into the target language, introducing texts that were originally produced with different communicative intentions and discourse patterns (Fadaee et al., 2017; Feng et al., 2021b). This distinction is particularly relevant for polarization detection, where diversity in rhetorical strategies and framing may be as important as lexical variation.

SemEval 2026 Task 9 (POLAR) (Naseem et al., 2026b) provides precisely such a setting, offering a multilingual framework covering 22 languages that enables the systematic incorporation of diverse intentional and rhetorical patterns through cross-lingual augmentation.

LLM-based synthetic data generation provides a complementary strategy. Prompted with task definitions and a small number of examples, instruction-tuned models can produce labeled instances with variation in framing, tone, and rhetorical strategy (Møller et al., 2024; Whitehouse et al., 2023), proving particularly effective in low-resource and domain-specific settings (Chung et al., 2023).

In this work, we propose a pipeline that combines direct cross-lingual translation from multiple languages and LLM one-shot synthetic generation as complementary augmentation strategies, evaluated through BERT-based ensemble classifiers for both Spanish and English. We examine whether polarization signals transfer effectively across languages and whether synthetic data can serve as a viable substitute for scarce native annotations.

3 System Overview

The following overview describes the methodology used to develop the system; for a visual summary, refer to the pipeline diagram in Figure 1.

3.1 Data augmentation

Polarization is a complex task where the distinction between polarized and neutral content often relies on subtle intent-related nuances. Since capturing these patterns effectively requires a high volume of training examples (Sun et al., 2017), we expanded our training set through data augmentation.

3.1.1 Cross-lingual Augmentation

As a first data augmentation strategy, we implemented a cross-lingual approach to address the limited number of labeled examples. Since polarization tends to manifest through rhetorical patterns that transcend specific languages and cultures, translated examples are expected to preserve the structural signatures of polarized discourse. As the primary focus of our system is the Spanish language, we used the Google Cloud Translation API to convert polarized texts from all other available sources into Spanish. This process included data from Amharic, Arabic, Bengali, Burmese, English, German, Hausa, Hindi, Italian, Khmer, Nepali, Odia, Persian, Polish, Punjabi, Russian, Swahili, Telugu, Turkish, Urdu, and Chinese. By projecting the original labels onto the translated versions, we significantly increased the training volume, providing the model with a broader range of patterns to identify polarization. A similar procedure was followed for our second proposed model in English, where texts from the other available languages, including Spanish, were translated to expand the training set.

3.1.2 LLM-driven Data Synthesis

The second strategy involved generating synthetic data to expand the training set beyond the standard categories provided in the competition. While the original dataset focused on traditional conflict domains such as politics, race, religion, and gender, two additional areas were introduced: science and technology, and sports. This approach aimed to demonstrate that polarization is not limited to commonly debated social categories, but can also arise in a broader range of contexts. By exposing the model to this variation, the objective was to develop a more robust classifier capable of identifying structural features of polarized discourse, such as in-group and out-group dynamics and antagonistic framing, even when vocabulary and thematic content differ substantially.

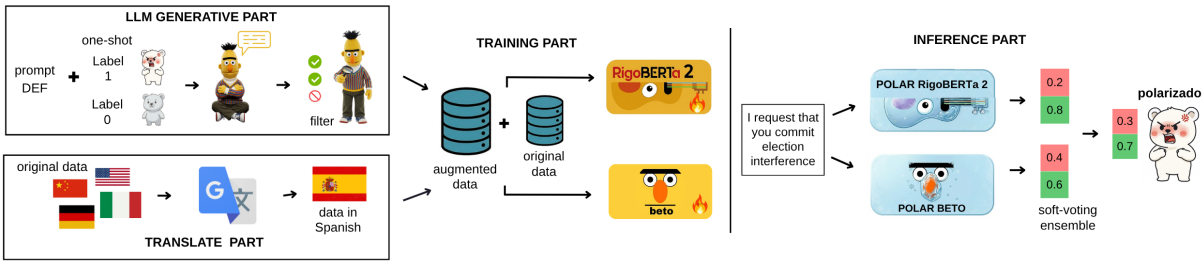

Figure 1: Flowchart of the system architecture, showing the data augmentation process, training phase, and inference stage.

To generate this content, we used RigoChat 2 (Gómez et al., 2025), a 7B parameter model specifically aligned for Spanish tasks, as this augmentation strategy was applied exclusively to Spanish. To ensure the quality of these examples, we applied a validation process in which the model classified each generated text using a one-shot approach, employing a formal definition of attitudinal polarization and a single example per label. This allowed the system to learn that polarization is a structural linguistic pattern that transcends specific thematic keywords. The prompt used for both generation and validation, including the one-shot example, is shown in appendix A.

3.1.3 Data Selection Strategy

To identify the most effective data configurations, we adopted an incremental greedy selection strategy, starting with the original dataset and adding new sources — cross-lingual translations and LLM-generated examples — only when they improved the F1-score, ensuring a controlled and empirically grounded expansion of the training data.

3.2 Classification models

This subsection details the architectures selected for the classification task. We adopted a transformer-based ensemble strategy to capture the linguistic nuances of polarization by merging the strengths of two specialized models. For the primary system in Spanish, we developed a soft-voting ensemble using the following architectures:

- *BETO*: We used BETO as our foundational model. It remains the most widely validated BERT-based architecture for Spanish, providing a stable understanding of the language's formal and syntactic structures through its Whole Word Masking (WWM) pre-training (Cañete et al., 2023).
- *RigoBERTa 2.0*: To complement BETO, we incorporated RigoBERTa 2.0. This model is a recent RoBERTa-based architecture specifically optimized for modern Spanish tasks. It is particularly effective at processing the informal and linguistic variations found in social media contexts (IIC, 2025).

We chose this ensemble to achieve broader linguistic coverage. By averaging the class probabilities of both models, the system balances the formal syntactic stability of BETO with the modern linguistic sensitivity of RigoBERTa 2.0, effectively reducing the specific biases that a single model might inherit from its pre-training data.

For the English experiments, we followed a similar logic by combining *BERT-large-uncased* and *RoBERTa-large*. These models serve as a methodological mirror, allowing us to evaluate the consistency of our approach across different languages using high-capacity, standard benchmarks.

4 Experimental Setup

This section describes the tools and parameters used in our experiments. Our pipeline integrates a self-validated generation stage followed by a robust ensemble classification strategy.

4.1 One-Shot Data Generation

To expand the training set, we used *RigoChat-7b-v2* with a one-shot prompting strategy. Specifically, we provided a single representative example for each domain (sports and science).

We applied a Self-Consistency filter to ensure label reliability while preserving data diversity in the generated data. For each generated text, the model performed five independent classification trials using a temperature of $\tau = 0.6$. We retained only those samples for which the predicted label achieved confidence $c \geq 0.65$. By setting

Table 1: Performance evolution for Spanish and English using a greedy strategy. Bold indicates the best configuration. Δ F1 is the change from the previous best F1, $\Delta F1_T$ according to the original data. N denotes the size.

Language	Data Configuration	N	Accuracy	F1-score	Δ F1	Δ $F1_T$
Spanish	O: Original data	3305	0.7212	0.7386	–	–
	A: O + English	6527	0.7467	0.7628	+0.0242	+0.0242
	B: A + Arabic	9907	0.7333	0.7471	-0.0157	+0.0085
	C: A + Telugu	8893	0.7333	0.7442	-0.0186	+0.0056
	D: A + Odia	8895	0.7333	0.7442	-0.0186	+0.0056
	E: A + Persian	9822	0.7091	0.7143	-0.0485	-0.0243
	F: A + Khmer	13167	0.6970	0.7222	-0.0406	-0.0164
	G: A + Italian	9861	0.6970	0.7222	-0.0406	-0.0164
	H: A + German	9707	0.7515	0.7684	+0.0056	-0.0298
	I: H + Russian	13055	0.7273	0.7458	-0.0226	-0.0072
	J: H + Chinese	13987	0.7333	0.7500	-0.0184	+0.0114
	K: H + Burmese	12596	0.7394	0.7624	-0.0060	+0.0238
	L: H + LLM sports	11256	**0.7576**	**0.7701**	**+0.0315**	**-0.1795**
	M: L + LLM science	11799	0.7212	0.7416	-0.0268	+0.0030
English	O: Original data	3223	0.8063	0.7680	–	–
	A: O + Russian	6570	0.8187	0.8000	+0.0320	+0.0320
	B: A + Nepali	8575	0.8125	0.7922	-0.0078	+0.0242
	C: A + Spanish	9875	0.8063	0.7824	-0.0176	+0.0144
	D: A + Swahili	13561	0.8063	0.7862	-0.0138	+0.0182
	E: A + German	9750	**0.8250**	**0.8077**	**+0.0077**	**+0.0397**
	F: E + Turkish	12114	0.8125	0.7904	-0.0173	+0.0224

the threshold below a perfect confidence score, we allow a controlled degree of noise in the dataset. This design choice encourages the final models to learn from ambiguous or soft instances of polarization, ultimately improving robustness in real-world scenarios.

4.2 Model Training and Ensemble

We fine-tuned the RigoBERTa (IIC/RigoBERTa-2.0) and BETO (dccuchile/bert-base-spanish-wwm-cased) models using the following parameters:

- **Pre-processing:** We normalized the text by replacing URLs and user mentions. We also reduce character repetitions.
- **Training Settings:** Both models were trained for 3 epochs with a learning rate of 2×10^{-5} and a batch size of 16. We used a fixed seed (93330) to ensure all experiments are reproducible.
- **Soft-Voting Ensemble:** The final predictions are calculated by averaging the probability distributions from both models. This ensemble strategy improves system stability and reduces errors that a single model might introduce.

5 Results

This section presents the evolution of our model's performance under the greedy selection strategy.

5.1 Development Analysis

Results for the Spanish subtask are detailed in Table 1. From a baseline F1-score of 0.7386 on the development set, the inclusion of English and German translations, alongside the LLM-generated sports category, reached a peak F1-score of 0.7701. In contrast, translations from sociolinguistically distant languages, such as Persian or Khmer, decreased performance by up to 0.0400. A similar trend was observed in the English subtask in Table 1, where selecting specific linguistic sources led to a consistent improvement of +0.0397, confirming the strategy's generalizability.

Discussion

These findings suggest that polarization is deeply embedded in cultural context. In their work, Naseem et al. (2026b) assigned the original gold labels through native annotators, thereby capturing locally grounded perceptions of conflict. When these labels are preserved in translations into lin-

Table 2: Official competition results on test set. UTokyo Tsuruoka Lab is the first place on the leaderborad.

Data Configuration	F1-macro
Spanish	
POLAR baseline	0.7266
UTokyo Tsuruoka Lab	0.8030
Original data	0.7613
+ ENG + DEU	0.7746
+ ENG + DEU + LLM sports	0.7869
English	
POLAR baseline	0.7802
UTokyo Tsuruoka Lab	0.8252
Original data + RUSS + DEU	0.7939

guistically and culturally distant languages, they carry over culturally specific judgments that may not fully resonate in the target context, potentially introducing variation. In this light, data augmentation appears to be most effective when the cultural understanding of *opinionated conflict* is closely aligned between the source and target languages.

On the other hand, the success of the sports domain versus the failure of science reflects a fundamental difference in how conflict is structured. The sports category likely improved performance because it shares an inherent *us-versus-them* logic with attitude polarization, where clear, opposing groups are identifiable. In contrast, while scientific discourse can contain strong or even extremist opinions, it lacks the well-defined group-based divisions found in other domains. Conflict in science is often epistemological or focused on individual sentiment toward a theory, rather than a confrontation between established social groups. Consequently, science data may introduce noise by shifting the model's focus toward general sentiment analysis rather than the group-based adversarial markers required for polarization detection.

5.2 Official Evaluation

For the official competition phase, Table 2 presents our proposed models for both languages, alongside the POLAR baseline and the first-place system. Our best configuration achieved an F1-macro of 0.7869 in Spanish, only 0.016 points below the first-place system, demonstrating competitive performance. Furthermore, our result in English of 0.7939 confirms that our approach generalizes effectively across languages. These results validate that the configurations derived from the development phase remain robust on unseen competition data.

6 Conclusion

This study presented an approach for the SemEval-2026 Task 9 on multilingual text polarization, prioritizing Spanish as the primary language of analysis through BETO and Rigoberta. Our results demonstrate that data augmentation, specifically through translation for this task, is an effective strategy for capturing polarized discourse. This method proved functional even when applied to English, showing that the technique scales successfully to the language typically regarded as the field's standard. However, the success of these techniques is not necessarily inherent, as it remains subject to local and cultural contexts. Ultimately, our work highlights that while translation-based augmentation is a practical and robust tool, its effectiveness is mediated by the subjective nature of social media and the specific cultural nuances of the target language.

7 Ethical Considerations

Following responsible data practices, this study builds on the ethical guidelines of the POLAR benchmark (Naseem et al., 2026b), whose data was collected from public sources, anonymized, and annotated by fairly compensated native speakers. Our cross-lingual approach introduces additional ethical complexity, as content that is polarized in one language may be legally or culturally harmful in another due to divergent laws and norms around topics such as gender, religion, or ethnicity. Researchers extending this work should exercise caution when translating sensitive content across language pairs, and remain mindful of the dual-use risks of polarization detection models, which could be repurposed for censorship or political surveillance.

8 Acknowledgments

We thank SECIHTI for the computational resources provided through the CIMAT Bajío Supercomputing Laboratory (Grant #300832). Calderón-Reyes (CVU 2014269) acknowledges SECIHTI's support through the Master's degree scholarship at CIMAT. Sánchez-Vega also acknowledges SECIHTI's support through the program Investigadoras e Investigadores por México (Project ID 11989, No. 1311).

A Prompts and One-Shot Examples

This appendix presents the prompts and one-shot examples in their original language (Spanish) alongside their English translations.

Generation Prompt

Spanish (original)

```
Genera un texto breve en español tipo tweet.
Características:
- Dominio: opiniones sobre ciencia, tecnología,
    IA, pseudociencia /
  Dominio: opiniones sobre deportes (futbol,
      equipos, jugadores, partidos)
- Muy informal y espontáneo
- Escritura descuidada a propósito
- Puede tener:
  - faltas ortográficas
  - palabras pegadas
  - letras cambiadas
  - puntuación incorrecta o ausente
  - mayúsculas inconsistentes
- Suena impulsivo, no redactado
- Polaridad clara
- NO corrijas el texto
Devuelve SOLO el texto.
```

English (translation)

```
Generate a short tweet-like text in Spanish.
Characteristics:
- Domain: opinions on science, technology, AI,
    pseudoscience /
  Domain: opinions on sports (soccer, teams,
      players, matches)
- Very informal and spontaneous
- Deliberately careless writing
- May include:
  - spelling mistakes
  - merged words
  - swapped letters
  - incorrect or absent punctuation
  - inconsistent capitalization
- Sounds impulsive, not drafted
- Clear polarity
- Do NOT correct the text
Return ONLY the text.
```

Classification Prompt

Spanish (original)

```
Clasifica si el texto contiene polarizacion
    social o actitudinal.
La polarizacion implica discurso que promueve
    division entre grupos,
groupismo (ellos vs nosotros), estereotipos,
    vilificacion,
deshumanizacion, intolerancia u odio hacia otros
     grupos.
(1=polarizado, 0=no polarizado).
```

English (translation)

```
Classify whether the text contains social or
    attitudinal polarization.
Polarization implies discourse that promotes
    division between groups,
in-group dynamics (them vs. us), stereotypes,
    vilification,
dehumanization, intolerance or hatred toward
    other groups.
(1=polarized, 0=not polarized).
```

One-shot Examples

Science (Spanish):

```
"yo creo mas en la astrologia q en la ciencia
    moderna". Label: 0
"los antivacunas son un peligro, no deberian ni
    opinar". Label: 1
```

Science (English):

```
"i believe more in astrology than in modern
    science". Label: 0
"anti-vaxxers are dangerous, they shouldnt even
    have an opinion". Label: 1
```

Sports (Spanish):

```
"el partido termino 1 a 0 con gol de cabeza".
    Label: 0
"los hinchas de ese club son todos unos simios
    nunca cambian". Label: 1
```

Sports (English):

```
"the match ended 1-0 with a header goal". Label:
     0
"the fans of that club are all a bunch of apes
    they never change". Label: 1
```

References

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Emily M. Bender. 2019. The technical is political: Ethical considerations in NLP research. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Aaron Bramson, Patrick Grim, Daniel Singer, Steven Fisher, William Berger, Graham Sack, and Carissa Flocken. 2016. Disambiguation of social polarization concepts and measures. *The Journal of Mathematical Sociology*, 40:80–111.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data. *Preprint*, arXiv:2308.02976.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Matteo Cinelli, Gianmarco Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118:e2023301118.

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, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

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, Brussels, Belgium. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021a. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021b. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Gonzalo Santamaría Gómez, Guillem García Subies, Pablo Gutiérrez Ruiz, Mario González Valero, Natàlia Fuertes, Helena Montoro Zamorano, Carmen Muñoz Sanz, Leire Rosado Plaza, Nuria Aldama García, David Betancur Sánchez, Kateryna Sushkova, Marta Guerrero Nieto, and Álvaro Barbero Jiménez. 2025. Rigochat 2: an adapted language model to spanish using a bounded dataset and reduced hardware. *Preprint*, arXiv:2503.08188.

IIC. 2025. Rigoberta-2.0.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Lilliana Mason. 2015. "i disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, 59(1):128–145.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2024. The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. *Preprint*, arXiv:2304.13861.

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.

Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press, New York, NY.

Walter Quattrociocchi, Antonio Scala, and C. Sunstein. 2016. Echo chambers on facebook. *SSRN Electronic Journal*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. *Preprint*, arXiv:1707.02968.

Cass R. Sunstein. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press, Princeton, NJ.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.