

List of changes to Manuscript

INFOTEC-NLP at SemEval-2026 Task 9: Comparing Regional
Transformers and Bag-of-Words Approaches for Polarization Detection
in Spanish

May 3, 2026

META-REVIEW

Comment 1.1. *Please note that your result ranks 19th on the leaderboard, but in the final leaderboard it is reported as 22nd in your paper.*

Response 1.1. Thank you for pointing this out. We reviewed the leaderboard entry and corrected the ranking throughout the manuscript so that it consistently matches the official final leaderboard.

Comment 1.2. *Please implement the reviewers' suggestions and run ACL pubcheck to ensure your paper pass all checks before submitting camera-ready version.*

Response 1.2. We have implemented the reviewers' suggestions in the revised manuscript, including clearer result tables, additional analysis, explicit limitations, and an appendix with implementation and reproducibility details. Before submitting the camera-ready version, we will also run ACL pubcheck and address any remaining formatting or compliance issues.

Reviewer #1

Comment 2.1. *The system-description papers should have ensured replicability. Please, either ensure all hyperparameters including random seeds and version of libraries and frameworks used in the experiments, or even better, provide the source code in the online repository.*

Response 2.1. We thank the reviewer for highlighting the importance of replicability. In the revised manuscript, we improved transparency by adding Appendix A, which now reports the main software libraries and framework versions used for the lexical, Transformer-based, and ensemble experiments, together with additional implementation details. We also added a dedicated subsection on reproducibility considerations. In particular, we now state explicitly that some parts of the pipeline used a fixed seed of 42, such as internal train-validation splits and selected Transformer and ensemble configurations. At the same time, we clarify an important limitation of the original experiments: random seeds were not controlled or logged uniformly across all runs, and not all intermediate checkpoints were preserved after the competition. Therefore, the revised version improves transparency and partial reproducibility, while explicitly acknowledging that exact reproduction of every explored configuration is not guaranteed.

Comment 2.2. *The ground truth have been released. Please, use it for further analysis of the compared alternatives and provide the results in a clear table (Macro F1 values located sparsely in the text are really bad for comparison).*

Response 2.2. We addressed this comment in two ways. First, we reorganized the Results section so that development-set performance is now reported in a clear comparative table (Table 1), instead of having Macro-F1 values scattered throughout the text. Second, after the release of the labels, we added a post-hoc analysis using preserved checkpoints and the released test labels. This analysis is reported in Section 5.3 and summarized in Table 2. In particular, we reevaluated two preserved BILMALAT configurations and showed that the official submission with `_es` achieved a Macro-F1 of 0.7701 on the released test labels, while the preserved `[MASK]` configuration reached 0.7628. Table 2 also reports the number of instances per label together with per-class precision, recall, and F1 scores, which provides a more detailed view of the behavior of the preserved checkpoints.

Reviewer #2

Comment 3.1. *The authors do a good job of introducing the task and their methodology is sound, but I had some difficulties when reading the paper: The authors use 'a prefix-based conditioning strategy and evaluate different ensemble schemes through linear meta-classifiers'. However, the background section does not clearly discuss prior work in this area. This made the start of the system review section difficult to read as there was missing context. The results section is not very detailed. A table or graph to show the performance of the different methods and/or regional_token values would be nice. Some more analysis would be interesting as well. The title of the paper follows the required format.*

Response 3.1. We appreciate this detailed feedback and revised the manuscript accordingly. To provide better context before the system description, we expanded the Background section with a short methodological discussion of prefix-based input conditioning and stacking with linear meta-classifiers, so that these design choices are introduced before Section 3. We also reorganized the Results section by adding Table 1 to summarize development-set performance more clearly. In addition, we expanded Section 5.3 with further analysis of the behavior of lexical and Transformer-based models, the effect of REGIONAL_TOKEN, and the limited gains obtained with the ensemble strategies. Finally, we incorporated a post-hoc evaluation on the released test labels for preserved BILMALAT checkpoints (Table 2), which provides additional evidence on generalization and model selection. We also thank the reviewer for noting that the title follows the required format.

Comment 3.2. *How did your model perform on each label (polarized/not polarized)? Was the model biased towards predicting either label?*

Response 3.2. We examined this point using the preserved official BILMALAT checkpoint with `_es` on the released test labels. The model achieved similar performance across both classes. For label 0 (non-polarized), precision was 0.7858, recall was 0.7503, and F1 was 0.7677. For label 1 (polarized), precision was 0.7555, recall was 0.7905, and F1 was 0.7726. Since the class supports were also relatively balanced (753 for label 0 and 735 for label 1), these results suggest that the model was not strongly biased toward a single label. There is a mild asymmetry, with slightly higher recall for the polarized class and slightly higher precision for the non-polarized class, but overall the behavior is fairly balanced. This information is now reflected in the revised post-hoc analysis Table 2.

Comment 3.3. *How did your results vary based on the regional_token? Do you have a theory for why some regional_token values were more effective than others?*

Response 3.3. Our results varied noticeably depending on the selected REGIONAL_TOKEN. For `mex_large`, different token choices produced visible differences in development performance, suggesting that the model is sensitive to how regional information is encoded in the input. For BILMALAT, the effect was also clear, although less straightforward: the preserved `[MASK]` configuration achieved the best development-set result (0.7333), while the preserved official configuration with `_es` performed better on the released test labels (0.7701 vs. 0.7628 for `[MASK]`). A plausible explanation is that the prefix acts as an additional cue for the model, and its usefulness depends

on both the model’s pretraining and the characteristics of the evaluation data. Since BILMALAT was pretrained on regional Twitter corpora, `_es` may have offered a more stable signal for the final test set, whereas `[MASK]` may have happened to fit the smaller development set better. This suggests that the observed differences are likely due both to real conditioning effects and to some variability introduced by the limited size of the development set.

Reviewer #3

Comment 4.1. *Strengths:*

- 1. The paper is written well - it's easy to understand and the required concepts have been described adequately.*
- 2. The paper is structured as per requirements - Relevant citations, all sections present, page limit adhered to and official submission clearly mentioned.*
- 3. Clear baselines (BOW models), approach used clearly outperforms baselines.*

Response 4.1. We appreciate your positive assessment of the manuscript. We are glad that the paper was clear, well structured, and that the comparison against lexical baselines was useful and easy to follow.

Comment 4.2. *Weaknesses:*

- 1. While the approach used is described clearly, it is quite simple and not very novel.*
- 2. The results have been detailed, but no further analysis has been undertaken. Some analysis to see where the current approach failed/comparison to prior work would be appreciated.*
- 3. The paper would benefit from the mention of limitations.*

Response 4.2.

1. We agree that the approach is intentionally simple. As a system-description paper, our main goal is to provide a clear empirical comparison between efficient lexical baselines, regionally pre-trained Transformer models, and lightweight ensemble strategies for Spanish polarization detection. In the revised manuscript, we make this practical evaluation perspective more explicit.
2. We addressed this by expanding the Results section. The revised manuscript now includes Table 1 for clearer comparison across configurations and a new Further Analysis subsection discussing the behavior of lexical versus Transformer-based models, the effect of `REGIONAL_TOKEN`, the limited gains of logistic stacking, and a post-hoc evaluation on the released test labels (Table 2), including per-class performance for preserved BILMALAT checkpoints.
3. We added an explicit discussion of limitations in Section 5.3. These include the relatively small size of the development set, the fact that the Transformer models were pre-trained on Twitter data, and the partial reproducibility of some experiments because not all runs were preserved and random seeds were not uniformly controlled across all configurations.

Comment 4.3. *Has the code been released? If so, the paper would benefit from a link to the repository.*

Response 4.3. At this stage, we have not released a public repository. To improve transparency, however, the revised manuscript now includes Appendix A, which reports the main software libraries and versions used for the lexical, Transformer-based, and ensemble experiments, together with additional implementation and reproducibility details. We acknowledge that a public repository would further strengthen reproducibility, but it is not available for the current version.