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

Eduardo C. C. Hernandez-Garcia[†] and **Guillermo Ruiz**[†] and **Mario Graff**[†,‡]
[†] INFOTEC Centro de Investigación e Innovación en
Tecnologías de la Información y Comunicación, Aguascalientes, México
[‡] Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), México
eduardo.cc.hernandez@gmail.com
luis.ruiz@infotec.mx
mario.graff@infotec.mx

Abstract

Polarization detection in short texts is a challenging and relevant problem in Natural Language Processing, particularly in social media environments where regional variations and subtle discursive nuances converge. In this paper, we describe our participation in Subtask 1 (Spanish) of SemEval-2026 Task 9 (Naseem et al., 2026a), which focuses on binary polarization classification. We evaluate two main strategies: lexical models based on Bag-of-Words representations and regionally pre-trained Transformer models for Spanish. In addition, we explore a logistic stacking framework that combines lexical and contextual representations. Our experiments show that regionally adapted Transformers generally outperform purely lexical approaches, with `BILMALAT` achieving the strongest performance in this task. The results highlight the importance of regionally aligned pre-training on social media data for effective polarization detection in Spanish.

1 Introduction

Online polarization represents a growing challenge for the analysis of public discourse and constitutes a relevant problem for Natural Language Processing, particularly on platforms where regional variation, informal language, and subtle discursive cues interact. SemEval-2026 Task 9 (Naseem et al., 2026a), in its Subtask 1, addresses the automatic identification of polarization across multiple languages. In this work, we focus on the Spanish language and adhere to the task definition and evaluation protocol established by the workshop.

Our approach was based on comparing two families of methods. First, we implemented lexical Bag-of-Words models using EvoMSA (Graff et al., 2020), which serve as an efficient baseline. Second, we employed the regionally pre-trained Transformer models for Spanish `mex_large` and `BILMALAT` (Tellez et al., 2023; Ruiz et al., 2025), fine-tuned in a supervised manner. In addition, we further introduced a region-specific prefix token (`regional_token`) to condition the model during fine-tuning. We also evaluated stacking-based ensemble methods by combining model outputs using logistic regression over predicted probabilities.

During our participation in the competition, we observed that regionally pre-trained Transformers generally outperformed lexical approaches during development, with `BILMALAT` emerging as the most robust model among our configurations. In the final evaluation phase, our last submission (see Section 5.1) was made under the username `INFOTEC-NLP` and achieved a Macro-F1 score of 0.7701, ranking 18th on the Spanish leaderboard. The official competition results reported in this paper correspond to the evaluation scores provided by the platform during the competition phase. Although the gold labels of the final test set were later released, they were not used for further model modification. In the revised version of this paper, they are considered only for post-hoc analysis of preserved checkpoints.

2 Background

SemEval-2026 Task 9 addresses the automatic detection of polarization in online discourse within a multilingual and multicultural setting. The benchmark includes multiple languages and events, and proposes several subtasks targeting different levels of analysis of the phenomenon. In this work, we participate in Subtask 1: Multilingual Text Classification Challenge – Polarization Detection, which is formulated as a binary text classification problem.

Formally, given a text x, the model must predict a binary label $y \in \{0, 1\}$, where 1 denotes a polarizing text and 0 a non-polarizing text. The task is evaluated using Macro-F1 as the official performance metric.

Our participation focuses exclusively on the Spanish subset, using the data provided by the

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2946–2953

workshop organizers.

The Spanish training set contains 3305 instances, each with a label. During the competition, the development set consisted of 165 instances whose gold labels were not available to participants and was used for model comparison through Codabench submissions. During the final phase of the workshop (evaluation phase), an additional test set of 1488 unlabeled instances was provided for official evaluation on the Codabench[1] platform. The gold labels of the development and test sets were released only after the competition, and in this paper they are used only for post-hoc analysis of preserved or reproducible configurations.

The texts correspond to fragments of online discourse, consisting of opinions and comments associated with various social and political contexts, as described in the benchmark overview paper (Naseem et al., 2026b).

The task falls within a growing body of research on polarization analysis and divisive discourse in digital media. The POLAR project extends previous efforts by incorporating multiple languages and events under a systematic annotation framework (Naseem et al., 2026b). Beyond computational approaches, polarization has been extensively studied from social, communicational, and policy perspectives. Recent work conceptualizes polarization as a multidimensional phenomenon that includes affective polarization, characterized by identity-based antagonism and increasing animosity in online interactions (Ali et al., 2025). Similarly, policy-oriented frameworks have proposed measurable indicators of polarization at the platform level, emphasizing its implications for civic discourse, democratic resilience, and systemic digital risk (Banim, 2025).

From a methodological perspective, polarization detection can be addressed as a standard text classification task. In this context, lexical representation-based approaches such as Bag-of-Words have proven competitive across multiple text classification scenarios (Graff et al., 2025). At the same time, contextualized models based on Transformers (Devlin et al., 2018; Vaswani et al., 2017) have established the state of the art in a wide range of Natural Language Processing tasks.

Beyond the general contrast between lexical and contextual approaches, our system also relies on two practical modeling choices. First, it uses prefix-based input conditioning, where controlled tokens are prepended to the input sequence in order to provide additional contextual cues to the model. Second, it uses stacking with a linear meta-classifier to combine predictions from heterogeneous models, a standard ensemble strategy in supervised learning (Wolpert, 1992). These two elements provide the methodological context for the system configurations described in Section 3.

3 System Overview

Our system combines traditional lexical approaches and regionally pre-trained Transformer models under a supervised binary classification framework. In addition, we incorporate a prefix-based conditioning strategy and evaluate different ensemble schemes through linear meta-classifiers. The following subsections describe each component in detail.

3.1 Lexical Models

As a baseline, we implemented lexical representation models using the EvoMSA library[2] (Graff et al., 2020). In particular, we evaluated three configurations reported in (Graff et al., 2025): Bag-of-Words, DenseBoW, and StackGeneralization (Wolpert, 1992).

3.1.1 Bag-of-Words

The Bag-of-Words model represents each text as a high-dimensional vector associated with a fixed vocabulary, where each component reflects the frequency of a term (word, q-gram, n-gram, etc.). In the approach described by (Graff et al., 2025), the representation may incorporate a TF-IDF scheme, in which term frequencies are weighted by inverse document frequency (IDF), resulting in highly sparse vectors suitable for linear classifiers. In our experiments, we used the EvoMSA implementation in Spanish with a vocabulary size of 2^{15}.

The model directly produces a binary prediction based on the learned lexical representation.

3.1.2 DenseBoW

Unlike sparse term-frequency-based representations, DenseBoW constructs a lower-dimensional space derived from multiple auxiliary classifiers trained on specific textual signals. In particular, (Graff et al., 2025) describes variants based on Emojis and Keywords, where each classifier learns to predict the presence of or compatibility

[1] https://www.codabench.org/competitions/10522/

[2] https://evomsa.readthedocs.io/en/docs/

with a given semantic signal. The outputs of these auxiliary classifiers (e.g., decision functions or predicted probabilities) are concatenated to form a dense document representation.

This strategy enables the incorporation of additional semantic information while maintaining a framework grounded in lexical representations and linear classifiers. In our experiments, we used the DenseBoW implementation available in EvoMSA for Spanish, through the parameters `emoji=True` and `keyword=True`.

3.1.3 StackBoW

StackBoW corresponds to a stacking scheme designed to combine lexical models. Graff et al. (2025) describe a stack generalization strategy that integrates BoW and DenseBoW models, highlighting that an interpretable option consists of combining components through a convex combination, which allows inspection of the relative "importance" of each component.

In our experiments, we implemented this idea using the `StackGeneralization` configuration in EvoMSA, which internally combines BoW and DenseBoW models through a stacking-based approach.

3.2 Regionally Pre-trained Transformer Models

In addition to lexical models, we evaluated classification models based on regionally pre-trained Transformers for Spanish, i.e., models trained to account for linguistic variation across regions of the Spanish-speaking world. This approach is particularly relevant in the social media domain, where lexical, pragmatic, and discourse-level differences exist among Spanish varieties.

Specifically, we used two models:

- `BILMALAT`[3], belonging to the `BILMA` (BERT in Latin America) family proposed by (Tellez et al., 2023).
- `mex_large`, publicly available on the Hugging Face Hub[4].

Both models contain approximately 100 million parameters, allowing efficient fine-tuning using standard computational resources.

[3] https://huggingface.co/guillermoruiz/bilmaLAT

[4] https://huggingface.co/guillermoruiz/mex_large

While `BILMALAT` is formally described in (Tellez et al., 2023), the `mex_large` model is currently under peer-reviewed academic publication detailing its pre-training process and training data. Therefore, the available methodological reference corresponds to the public model card hosted on Hugging Face.

3.2.1 Architecture and Regional Pre-training

The `BILMA` models described in (Tellez et al., 2023) follow the standard `BERT` (Devlin et al., 2018) architecture and are pre-trained using the Masked Language Modeling (MLM) objective on region-specific Twitter corpora.

The regionalized approach proposed by (Tellez et al., 2023) consists of pre-training BERT-style models on corpora segmented by geographical region, with the goal of capturing lexical and discourse-level variation specific to different Spanish-speaking countries.

3.2.2 Supervised Fine-tuning and Prefix Conditioning

For the polarization task, we instantiated both models using the `Hugging Face Transformers` library (Wolf et al., 2020), specifically through the class `AutoModelForSequenceClassification` with `num_labels = 2`. This implementation follows the standard `BERT` fine-tuning scheme, in which the representation associated with the initial `[CLS]` token is used as an aggregated sequence representation for classification tasks.

Conceptually, if h_{CLS} denotes the contextualized representation produced by the `Transformer` encoder, the prediction $\hat{y}$ can be expressed as:

$$\hat{y} = \text{softmax}(W h_{\text{CLS}} + b)$$

where W and b correspond to the parameters of the linear classification layer learned during supervised fine-tuning. This formulation reflects the standard behavior implemented in `AutoModelForSequenceClassification`.

The models were fully fine-tuned in a supervised manner, meaning that all parameters of the Transformer encoder were updated jointly with the classification head. We did not employ partial freezing techniques or parameter-efficient fine-tuning methods.

Training was conducted using the training set provided on the Codabench platform, applying an internal split to separate training and validation data to 95% and 5% respectively.

In addition, we incorporated a prefix-conditioning strategy by prepending controlled tokens before tokenization. In our experiments, we used two input templates:

For BILMALAT, the input text was transformed as follows:

<REGIONAL_TOKEN> _2023 _02 <text>

with a maximum of 128 tokens per text and discarding the excess.

For mex_large, the input text was transformed as:

<REGIONAL_TOKEN> _GEO <text>

with a maximum length of 200 tokens and discarding the excess.

In both cases, <REGIONAL_TOKEN> was experimentally varied. For BILMALAT, we evaluated the values {'[MASK]', '_ar', '_es'}. In this context, '[MASK]' was incorporated solely as a fixed token at the beginning of the sequence. For mex_large, <REGIONAL_TOKEN> corresponded to different regional identifiers associated with Mexican states.

Tokenization was performed using the tokenizer corresponding to each model, ensuring consistency between the vocabulary and the pre-training scheme of the selected model.

3.3 Ensemble Strategies

To explore the complementarity between lexical and contextual representations, we evaluated different stacking-based ensemble strategies combining BoW models and regionally pre-trained Transformers.

3.3.1 Lexical Stacking with EvoMSA

We used the stacking implementation provided in EvoMSA through the StackGeneralization class to combine lexical variants of BoW and DenseBoW. This scheme internally trains a meta-model that integrates the outputs of the base models. No additional modifications were made to the standard implementation of the library.

3.3.2 BoW + mex_large Ensemble

We additionally implemented an ensemble combining a BoW model with the mex_large model. For this purpose, we extracted the predicted probabilities for the positive class from each base model and used them as input features to a linear meta-classifier (logistic regression). This approach allows the meta-model to assign different weights to each source of information.

3.3.3 BILMALAT + mex_large Ensemble

Finally, we evaluated an ensemble combining the two regionally pre-trained Transformer models, BILMALAT and mex_large. Analogously to the previous setting, we used the predicted probabilities for the positive class from each model as input to a logistic regression meta-classifier. This scheme enables the meta-model to capture potential differences in the regional and contextual representations learned by both models.

4 Experimental Setup

4.1 Data and Evaluation

We participated in Subtask 1 for Spanish using the labeled training set provided by the organizers. During the competition, model development was carried out using the official development set (165 instances), whose gold labels were not available to participants at that stage. Predictions generated on this set were submitted to the Codabench platform, which returned the official Macro-F1 score used for model comparison during development.

After the competition, the gold labels of the development set were released by the organizers. In the revised version of this paper, these labels are acknowledged as available for further analysis whenever reproducible configurations can be re-evaluated. However, the official competition results reported in Section 5.1 remain those returned by Codabench during the shared task.

The final test set (1488 instances) was reserved for the official evaluation phase. Our final submission was evaluated on Codabench, and the reported test performance corresponds to the official Macro-F1 score provided by the platform.

4.2 Preprocessing

No advanced text cleaning was applied beyond the preprocessing required by each model. Lexical models were implemented with EvoMSA using its Spanish configuration. For the Transformer-based models, we used the corresponding tokenizer and instantiated the classifiers through the Hugging Face Transformers library.

In both Transformer-based settings, a REGIONAL_TOKEN was prepended before tokenization, following the input templates described in Section 3.2. For BILMALAT, the maximum sequence length was set to 128 tokens, while for mex_large it was set to 200 tokens. During development, Transformer-based models were

trained using an internal split of the provided training data into 95% for training and 5% for validation.

Additional implementation details, including software libraries, input formatting, execution settings, and reproducibility considerations, are provided in Appendix A.

5 Results

5.1 Official Submission and Leaderboard Ranking

Our final official submission for the test phase of Subtask 1 (Spanish) consisted of a `BILMALAT` model using `'_es'` as regional token, fine-tuned for 3 epochs with a learning rate of 10^{-5}.

This model was trained on the provided training set and used to generate predictions on the final test set (1488 instances). The official result obtained on Codabench was a Macro-F1 score of 0.7701, ranking our system 18th in Subtask 1 for Spanish. The top-ranked system achieved a Macro-F1 score of 0.8030, placing our submission within 0.033 points of the best-performing model.

5.2 Results on the Development Set

Table 1 summarizes the best Macro-F1 scores obtained by the evaluated systems on the Spanish development set across the explored configurations. Among all compared models, `BILMALAT` achieved the highest performance (0.7333), outperforming both lexical baselines and the evaluated ensemble strategies. DenseBoW and the best `mex_large` configuration obtained similar results around 0.69, while the stacking-based combinations did not surpass the best individual Transformer model.

In addition, the impact of the `REGIONAL_TOKEN` strategy showed some variability depending on the model. For `mex_large`, different token choices such as `[MASK]` or location-specific tokens (e.g., `Veracruz`) resulted in noticeable differences in performance, suggesting a sensitivity to the conditioning mechanism. In contrast, `BILMALAT` exhibited more stable behavior and achieved the strongest results among the reported configurations, suggesting greater robustness under comparable settings. Finally, the evaluated ensemble strategies based on logistic stacking did not provide a clear performance improvement over the best individual Transformer model, indicating that combining models with similar representations may offer limited gains in this task.

5.3 Further Analysis

The results presented in Table 1 reveal several relevant patterns regarding the behavior of the evaluated approaches. First, Transformer-based models generally outperform lexical baselines, highlighting the importance of contextual representations for capturing subtle linguistic cues associated with polarization. While DenseBoW achieves competitive performance among lexical methods, it remains below the best-performing Transformer configurations.

Regarding the use of the `REGIONAL_TOKEN` strategy, the observed results indicate that its effectiveness depends on the underlying model. In the case of `mex_large`, performance varied noticeably across different token choices, which suggests that the model is sensitive to how regional information is encoded in the input. On the other hand, `BILMALAT` exhibited more stable behavior, consistently achieving strong performance under similar hyperparameter settings. This may indicate that regional pretraining provides a more robust mechanism for capturing geographically influenced language patterns compared to explicit prefix-based conditioning.

Another relevant observation is that the ensemble strategies based on logistic stacking did not consistently improve over the best individual models. This behavior may be explained by the similarity in the information captured by the combined models, which limits the potential benefits of aggregation. In this context, combining models that produce highly similar output probabilities may not provide sufficient diversity to achieve meaningful performance gains.

After the release of the test labels, we reevaluated two preserved `BILMALAT` checkpoints on the official test set. Table 2 reports the number of instances per label together with per-class precision, recall, and F1 scores. The official submission using `_es` achieved the strongest overall performance, with a Macro-F1 of 0.7701, compared with 0.7628 for the preserved `[MASK]` configuration. The table also shows that the released test set is relatively balanced across both labels, and that the `_es` configuration achieved slightly better per-class F1 values overall. This post-hoc comparison suggests that the best-performing configuration on the development set did not generalize best to the final test set.

Finally, it is important to note several limitations of this study. The development set is relatively

Family	Model	Token	LR / Epochs	Macro-F1
Lexical	BoW	–	–	0.6060
Lexical	DenseBoW	–	–	0.6905
Lexical	StackBoW	–	–	0.6416
Transformer	`mex_large`	`Veracruz`	10^{-4} / 4	0.6920
Transformer	`mex_large`	`[MASK]`	10^{-5} / 3	0.6848
Transformer	`BILMALAT`	`[MASK]`	10^{-5} / 3	**0.7333**
Ensemble	BoW + `mex_large`	–	–	0.6839
Ensemble	`BILMALAT` + `mex_large`	–	–	0.7148

Table 1: Summary of the best Macro-F1 scores on the development set for the evaluated systems across the explored configurations.

Token	**Label**	**n**	**Prec.**	**Rec.**	**F1**
`[MASK]`	0	753	0.7688	0.7596	0.7642
	1	735	0.7567	0.7660	0.7613
	Macro	1488	–	–	0.7628
`_es`	0	753	0.7858	0.7503	0.7677
	1	735	0.7555	0.7905	0.7726
	Macro	1488	–	–	**0.7701**

Table 2: Post-hoc evaluation on the released test labels for two preserved `BILMALAT` configurations, including the number of instances per label and per-class precision, recall, and F1 scores (0 = non-polarized, 1 = polarized).

small, which may introduce variability in the observed performance across configurations. Additionally, the Transformer-based models evaluated in this work were pre-trained on Twitter data, which may limit their generalization to texts from other social media platforms or textual sources with different linguistic characteristics. Some parts of the experimental pipeline used fixed seeds, but random seeds were not systematically controlled or logged across all experiments. Furthermore, not all experimental runs were preserved, which restricts the ability to perform a fully reproducible and exhaustive comparison of all explored configurations. Future work could address these limitations by incorporating more controlled experimental settings, reevaluating additional reproducible configurations on the released labels, and exploring alternative strategies to better exploit complementary information across models.

6 Conclusion

In this paper, we presented our participation in SemEval-2026 Task 9, Subtask 1 for Spanish, comparing lexical approaches based on EvoMSA with regionally pre-trained Transformer models and simple ensemble strategies. Our results show that Transformer-based models, particularly `BILMALAT`, provide the strongest performance for polarization detection in Spanish, while lexical approaches remain useful as efficient baselines.

Our final official submission, based on `BILMALAT`, achieved a Macro-F1 score of 0.7701 and ranked 18th on the Spanish leaderboard. Post-hoc evaluation on the released test labels further showed that the official configuration with `_es` generalized better than the preserved `[MASK]` configuration that had obtained the best development-set result. This suggests that model selection based on a relatively small development set may not always identify the best generalizing configuration. These results indicate that regionally pre-trained models offer a practical and effective alternative for this task.

As future work, we plan to perform a more systematic exploration of hyperparameters, evaluate reproducible subsets of configurations using the released labels when possible, and conduct a finer-grained error analysis to better understand the linguistic phenomena that remain challenging for current models.

References

Adem Chanie Ali, Seid Muhie Yimam, Abinew Ali Ayele, Chris Biemann, and Martin Semmann. 2025. Silenced voices: social media polarization and women's marginalization in peacebuilding during the northern ethiopia war. *i-com*, 24(2):407–432.

G. Banim. 2025. "very large online platforms—how big is your polarization footprint?" towards a metric to give eu citizens transparency around an online systemic risk driving conflict in our societies. *Build Up*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of

deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mario Graff, Sabino Miranda-Jiménez, Eric Sadit Tellez, and Daniela Moctezuma. 2020. Evomsa: A multilingual evolutionary approach for sentiment analysis. *Computational Intelligence Magazine*, 15:76 – 88.

Mario Graff, Daniela Moctezuma, and Eric S. Téllez. 2025. Bag-of-word approach is not dead: A performance analysis on a myriad of text classification challenges. *Natural Language Processing Journal*, 11:100154.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acarürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Elena Tutubalina, Aung Kyaw Htet, Xintong Wang, Surendrabikram Thapa, Tanmoy Chakraborty, 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)*, San Diego, CA, USA. Association for Computational Linguistics.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Garrido Veliz, P Sam Sahil, Yiran Zhang, Marco Antonio Stranisci, Idris Abdulmumin, Özge Alacam, Cengiz Acartürk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, 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.

Guillermo Ruiz, Rogelio Campos, Tania Ramirez-delreal, Daniela Moctezuma, Mario Graff, and Eric Sadit Tellez. 2025. Infotec-nlp at homo-lat 2025: Testing a novel multi-region spanish model to monitor opinion in latin american lgbtqi+ social media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2025)*, volume 4098 of *CEUR Workshop Proceedings*, page paper 4. CEUR-WS.org.

Eric S. Tellez, Daniela Moctezuma, Sabino Miranda, Mario Graff, and Guillermo Ruiz. 2023. Regionalized models for spanish language variations based on twitter. *Language Resources and Evaluation*, 57(4):1697–1727.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

A Implementation Details

This appendix provides additional implementation details to improve the transparency of our experimental setup.

A.1 Software and Libraries

The lexical experiments were conducted in Python 3.12.13 using EvoMSA 2.0.14, microtc 2.4.13, scikit-learn 1.6.1, NumPy 2.0.2, and pandas 2.2.2. The evaluated lexical configurations correspond to BoW, DenseBoW, and `StackGeneralization` under EvoMSA's Spanish setting. For each configuration, the trained model was used to generate predictions for the official development set, which were exported as CSV files and submitted to Codabench to obtain the official Macro-F1 score during the competition.

The Transformer-based experiments were conducted in Python 3.12.13 using Transformers 5.0.0, PyTorch 2.10.0+cu128, datasets 4.0.0, scikit-learn 1.6.1, NumPy 2.0.2, and pandas 2.2.2.

The ensemble experiments were conducted in the same Colab environment. The BoW + `mex_large` configuration used EvoMSA 2.0.14, microtc 2.4.13, Transformers 5.0.0, PyTorch 2.10.0+cu128, datasets 4.0.0, scikit-learn 1.6.1, NumPy 2.0.2, and pandas 2.2.2. The `BILMALAT` + `mex_large` configuration used Transformers 5.0.0, PyTorch 2.10.0+cu128, datasets 4.0.0, scikit-learn 1.6.1, NumPy 2.0.2, and pandas 2.2.2.

A.2 Reproducibility Considerations

Some parts of the experimental pipeline used fixed seeds. In particular, internal train-validation splits

were generated with a fixed seed of 42, and this value was also used in selected Transformer and ensemble configurations. However, random seeds were not systematically controlled or logged across all experiments, and not all intermediate runs were preserved after the competition. Therefore, the revised version improves transparency and partial reproducibility, but exact reproduction of every explored configuration is not guaranteed.