UNED at SemEval-2026 Task 9: Sentiment-Aware Transformer Models with Back-Translation Augmentation for Online polarisation Detection

Victor García-Sanabria, Álvaro Rodrigo, Roberto Centeno
NLP & IR group at UNED
Madrid / Spain
vgarcia1859@alumno.uned.es

Abstract

This paper describes our submission to SemEval-2026 Task 9 (Subtask 1) on Spanish online polarisation detection. We investigate whether sentiment-adapted pretrained language models provide an advantage over general-purpose multilingual models for binary polarisation classification. Under a controlled training setup, we compare a base XLM-RoBERTa model, an emotion-adapted model, and a sentiment-adapted XLM-R model trained on Twitter data. To mitigate overfitting in the relatively small training dataset, we additionally apply back-translation as a data augmentation strategy. Experimental results show that the sentiment-adapted checkpoint consistently outperforms the alternative pretrained models under identical conditions. When combined with back-translation augmentation, the final system achieves a macro-averaged F1 score of 0.743 on the preliminary competition leaderboard. These findings suggest that prior adaptation to affective signals in social media can provide beneficial inductive bias for polarisation detection.

1 Introduction

Online polarisation has become a central phenomenon in digital communication, characterised by strong evaluative language, antagonistic framing, and the expression of extreme or divisive viewpoints. Automatically detecting polarised discourse is therefore an important task for understanding online dynamics and moderating potentially harmful content. SemEval-2026 Task 9 introduces several polarisation-related subtasks, among which Subtask 1 formulates polarisation detection as a binary classification problem over short social media texts; in this work, we participate only in Subtask 1 (Naseem et al., 2026a).

Polarisation is not merely a matter of topical disagreement; it is often realised linguistically through intensified evaluation, affective positioning, and emotionally charged framing (Kiesel and Amlani, 2025). Expressions of approval, condemnation, outrage, or moral judgement frequently accompany polarising discourse. This overlap between affective language and polarisation motivates the central research question of this work: Do pretrained models that have been explicitly adapted to sentiment signals encode representations that are particularly suitable for polarisation detection? This hypothesis is motivated by the observation that polarised discourse frequently relies on strong evaluative language and affective intensity, suggesting that models adapted to sentiment may capture relevant signals for this task.

To investigate this question, we conducted a controlled comparison of several pretrained transformer checkpoints, including XLM-RoBERTa-base (XLM-R) [1] (Conneau et al., 2020) as base model, an emotion-adapted model: RoBERTa-base fine-tuned on GoEmotions[2] (Demszky et al., 2020), and a sentiment-adapted XLM-R model trained on Twitter data: Twitter-XLM-RoBERTa-base for Sentiment Analysis (XLM-T)[3] (Barbieri et al., 2022). All models are fine-tuned under identical training conditions. In addition, we explore back-translation as a data augmentation strategy to mitigate overfitting in the relatively small training dataset.

Under this setup, all pretrained models are evaluated under identical training conditions, enabling a direct comparison of how different forms of prior affective adaptation influence polarisation detection performance. Our findings indicate that the sentiment-adapted checkpoint provides more consistent validation performance than the base and

[1] https://huggingface.co/FacebookAI/xlm-roberta-base
[2] https://huggingface.co/SamLowe/roberta-base-go_emotions
[3] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1366–1371

emotion-adapted alternatives under the same optimisation setup. These results suggest that prior exposure to affective signals in social media may provide a useful inductive bias for modelling polarisation, even without incorporating explicit sentiment features at training time. More broadly, understanding whether sentiment-adapted representations capture polarisation may provide insight into how affective signals encode deeper dimensions of stance and personality in language, offering a direction for future investigation.

2 Related Work

2.1 Computational Modelling of Online Polarisation

Online polarisation has been increasingly addressed within computational social science and natural language processing as a measurable linguistic phenomenon (Naseem et al., 2026b; Kiesel and Amlani, 2025). Recent work has moved beyond binary sentiment or toxicity detection toward modelling intergroup hostility, ideological division, and rhetorical antagonism in online discourse.

The POLAR benchmark (Naseem et al., 2026b) provides one of the most comprehensive recent resources in this space, introducing a multilingual, multicultural, and multi-event dataset covering 22 languages and more than 110,000 annotated instances. Crucially, POLAR formalises polarisation along three complementary dimensions: (i) binary polarisation detection, (ii) polarisation type classification (e.g., political, racial, religious), and (iii) identification of rhetorical manifestations such as stereotyping, vilification, and dehumanization.

This fine-grained formulation highlights that polarisation is not just a negative affect, but a structured social phenomenon expressed through identifiable rhetorical strategies. The present work operates within this framework, focusing on improving binary polarisation detection in shared task setting while analysing the role of sentiment-adapted models and data augmentation.

2.2 Multilingual Transformer Architectures

Modern approaches to multilingual text classification are grounded in transformer-based architectures. Multilingual models such as XLM-R (Conneau et al., 2020) provide cross-lingual contextual representations suitable for downstream classification tasks.

The sentiment-adapted model employed in this study builds upon this transformer paradigm. Rather than training from scratch, the approach leverages a pre-trained multilingual encoder that has already been fine-tuned on sentiment-oriented data. This prior adaptation is relevant given the conceptual proximity between sentiment polarity and polarised discourse, while still allowing empirical comparison against alternative modelling strategies.

2.3 Back-Translation for Data Augmentation

Data augmentation is a common strategy to improve robustness and generalisation in low- and medium-resource settings. Back-translation, originally proposed in the context of neural machine translation by Sennrich et al. (2016), generates synthetic training examples by translating text into an intermediate language and translating it back into the source language. This process introduces lexical and syntactic variation while preserving semantic content.

In multilingual NLP, back-translation has been widely adopted as a means of improving model robustness to paraphrastic variation and distributional shift. In this work, back-translation was implemented using OPUS-MT neural machine translation models (Tiedemann and Thottingal, 2020), which provide publicly available transformer-based translation systems trained on large-scale parallel corpora. The Helsinki-NLP models used in the augmentation pipeline belong to this framework.

By combining original and back-translated data in a 1:1 ratio, the training set was expanded while maintaining label consistency. This augmentation strategy was empirically compared against alternative approaches, including cross-lingual transfer from related languages, and yielded superior performance in the final system configuration.

Having outlined the task context and related methodological approaches, we now describe the dataset used for training and evaluation.

3 Methodology

3.1 Base Model

The final system fine-tunes the pretrained checkpoint[4]. This checkpoint is part of the XLM-T framework (Barbieri et al., 2022), where XLM-R (Conneau et al., 2020) is adapted to Twitter-

[4]Twitter-XLM-RoBERTa-base for Sentiment Analysis (XLM-T): https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

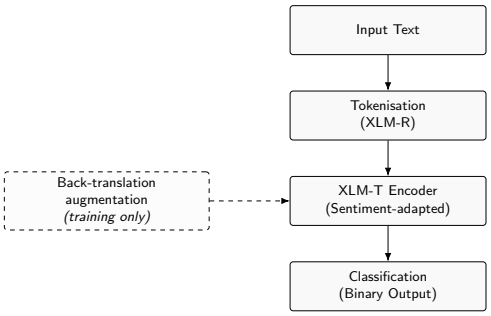

Figure 1: Overview of the proposed system. Input texts are tokenised using the XLM-R tokenizer and processed by a sentiment-adapted XLM-R encoder (XLM-T), followed by a binary classification head. Back-translation is applied during training to augment the dataset.

domain data and subsequently fine-tuned for multilingual sentiment analysis. Prior work shows that domain-specific adaptation improves robustness over general-domain XLM-R models when applied to noisy and multilingual Twitter data (Gururangan et al., 2020; Barbieri et al., 2022). Twitter-domain data is particularly relevant for this task, as it reflects the short, informal, and highly affective nature of social media discourse where polarisation is most prominently expressed. We therefore leverage this sentiment-adapted checkpoint under the hypothesis that its prior exposure to affective and social media signals may provide a beneficial inductive bias for polarisation detection. Figure 1 provides an overview of the system architecture and training-time augmentation.

Inputs are tokenised using the XLM-R tokenizer associated with the selected checkpoint. The classification model is instantiated with a binary output layer (`num_labels=2`) and configured for single-label classification.

All model parameters are jointly fine-tuned on the task-specific training data, including the augmented instances described in Section 3.2. Fine-tuning is performed for 2 epochs with a learning rate of 1×10^{-5}, a batch size of 16, weight decay of 0.01, no warm-up steps, and a maximum sequence length of 256 tokens, using the AdamW optimiser.

3.2 Back-Translation Augmentation

To increase the robustness of the model to lexical and syntactic variation, we apply back-translation as a data augmentation strategy. Back-translation was originally proposed as a method for leveraging monolingual data in neural machine translation (Sennrich et al., 2016) and has since been widely adopted as a general-purpose augmentation technique in NLP.

We generate augmented instances using the `BackTranslationAug` component from the `nlpaug` library, relying on MarianMT translation models from the Helsinki-NLP OPUS-MT project (Tiedemann and Thottingal, 2020). Specifically, for Spanish data we employ the publicly available models `Helsinki-NLP/opus-mt-es-en` (Spanish→English) and `Helsinki-NLP/opus-mt-en-es` (English→Spanish). These models are trained on large-scale parallel corpora within the OPUS-MT framework and are distributed for research use.

For each training instance, the text is translated from Spanish to English and then back to Spanish, using English as a pivot language. This back-translation procedure generates semantically equivalent but lexically varied paraphrases that increase training diversity while preserving the original labels.

The augmented data are combined with the original training instances under the configuration described in Section 3.3.

3.3 Training Configuration

The model is fine-tuned using the Hugging Face `Trainer` API. The original training data are split into training and validation subsets using an 80/20 stratified split based on the binary polarisation label, with `random_state=12` to ensure reproducibility of the split.

The augmented dataset described in Section 3.2 is concatenated with the original training data in a 1:1 ratio, resulting in a combined training set containing 100% of the original instances and their corresponding back-translated variants.

All input texts are tokenised with truncation enabled and padding set to `max_length`, using a maximum sequence length of 256 tokens.

Fine-tuning is performed for 2 epochs with a learning rate of 1×10^{-5}. The per-device batch size is set to 16 for both training and evaluation. No warm-up steps are used, and weight decay is set to 0.01. Logging occurs every 50 steps, and external reporting is disabled. Optimisation uses the AdamW optimiser as implemented in the Transformers library.

Model selection is based exclusively on validation macro-averaged F1 score. No early stopping is applied; training proceeds for the full number of epochs. Apart from the fixed data split

(`random_state=12`), no additional global random seed is set.

Training is conducted on a single NVIDIA Tesla T4 GPU using Google Colab. The model achieving the best validation macro-F1 is retained, and both the fine-tuned model and the corresponding tokenizer are saved.

4 Experiments

4.1 Experimental Setup

The official evaluation metric is macro-averaged F1. The training dataset contains 3,305 Spanish instances with a near-balanced label distribution (50.2% polarised, 49.8% non-polarised). The texts are short social media posts (mean length ≈11 tokens), well within the maximum sequence length of 256 tokens used during fine-tuning.

For local evaluation, we use an 80/20 stratified train–validation split as described in Section 3.3. Each configuration is trained once under identical hyperparameter settings. For augmentation experiments, the original training set is expanded in a 1:1 ratio by pairing each instance with a back-translated variant.

During inference on the official test set, predictions are obtained by selecting the class with the highest logit (argmax). No threshold tuning, ensembling, calibration, or post-processing techniques are applied.

4.2 Development Experiments

This section presents the experiments conducted during model development using the validation split described in Section 3.3. All configurations are trained under identical hyperparameter settings to isolate the effects of pretrained checkpoint selection and data augmentation. Model selection for submission is based exclusively on validation macro-F1.

4.2.1 Pretrained Checkpoint Comparison

We compare three pretrained checkpoints under identical training conditions: XLM-R (base), RoBERTa-GoEmotions, and Twitter-XLM-R-Sentiment.

Without task-specific fine-tuning, the base XLM-R model achieves 0.49 macro-F1 on the validation set, which we use as a baseline for comparison with fine-tuned configurations. After fine-tuning, the emotion-adapted model reaches 0.64 macro-F1, while the sentiment-adapted model achieves 0.69 macro-F1. The fine-tuned base model exhibits perfect training performance (macro-F1 = 1.0), indicating severe overfitting and poor generalisation.

Model	Macro-F1
XLM-R (no fine-tuning)	0.49
RoBERTa-GoEmotions	0.64
XLM-T (sentiment-adapted)	0.69
XLM-T + Back-translation	0.69

Table 1: Validation macro-F1 scores on the 80/20 train–validation split across model configurations.

As shown in Table 1, the sentiment-adapted checkpoint obtains the highest validation score among the compared pretrained models. Based on these results, it is selected for subsequent experiments and final submission.

4.2.2 Effect of Back-Translation

Using the selected sentiment-adapted checkpoint, we evaluate the impact of back-translation as a data augmentation strategy. We compare training on the original dataset (3,305 instances) with training on the combined original and augmented dataset, where each instance is paired with a back-translated variant in a 1:1 ratio.

Training on the augmented dataset consistently improves validation performance, yielding macro-F1 scores in the range 0.65–0.69, with the best configuration reaching 0.69 prior to submission. In contrast, training on the original data alone results in lower and less stable validation scores.

These findings suggest that back-translation introduces useful lexical and syntactic variation that acts as a regulariser in the relatively small training setting.

4.2.3 Alternative Expansion Strategies

We also experiment with a single translation-based expansion from a neighbouring language (Italian) as a form of cross-lingual transfer. This approach yields substantially lower validation performance (approximately 0.50 macro-F1) and does not improve generalisation.

The negative result suggests that simple cross-lingual data expansion introduces noise rather than useful variation in this setting. Consequently, this strategy is not pursued further.

4.3 Official Submission Results

Based on the development experiments described above, the final submitted system consists of the

sentiment-adapted XLM-R model trained on the combined original and back-translated dataset under the hyperparameter configuration specified in Section 3.3.

On the official SemEval-2026 Task 9 leaderboard, the system achieves a macro-F1 score of 0.743, ranking 29th out of 37 participating systems, and outperforming the provided baseline (0.7266).

These results indicate that while affective adaptation and augmentation provide measurable gains over baseline configurations, additional modelling of the structural dimensions of polarised discourse is necessary to close the performance gap to leading approaches.

5 Conclusions and Future Work

This work investigated whether polarisation detection can be understood, in part, as a function of affective sensitivity encoded in pretrained language models. Empirically, sentiment-adapted representations yielded the most stable performance under controlled training conditions, and back-translation acted as an effective regulariser in a low-resource setting. Beyond their immediate performance implications, however, these findings raise a more fundamental representational question: what aspects of polarised discourse are captured when a model is adapted to sentiment?

The results suggest that a substantial portion of the signal required for binary polarisation detection overlaps with affective polarity. Yet polarisation cannot be reduced to sentiment alone. Polarised discourse is structurally organised around intergroup positioning, identity signalling, antagonistic framing, and rhetorical intensification (Naseem et al., 2026b; Kiesel and Amlani, 2025). Sentiment-adapted encoders appear to capture surface-level evaluative intensity, but they only partially model the deeper social-relational structure that defines polarisation as a phenomenon. The persistent performance gap to the top-ranked system indicates that these additional dimensions remain underrepresented.

These observations motivate a shift from treating polarisation as a standalone classification task toward modelling it as an emergent interaction between multiple linguistic sensitivities: affect (sentiment and emotion), stance, and social alignment. Future work should therefore pursue controlled representational analyses that isolate these dimensions within large language models. Rather than merely improving predictive performance, such investigations would clarify how pretrained encoders internally organise affective and intergroup information.

More broadly, advancing the computational study of polarisation requires moving beyond binary detection toward structured representations of rhetorical manifestation and group-directed antagonism. This includes modelling how language encodes identity boundaries, moral evaluation, and social positioning. Understanding the interaction between affective encoding and socially situated meaning remains an open theoretical challenge with implications for computational personality modelling, discourse analysis, and the formal representation of emotionally grounded social behaviour in language.

These conclusions are drawn from a relatively small training dataset (3,305 instances) and a single-run evaluation protocol, which may amplify variance effects and limit statistical robustness. The submitted system achieved 0.743 macro-F1 on the official SemEval-2026 Task 9 leaderboard, placing 29th out of 37 participating systems. While this result demonstrates measurable gains from affective adaptation and augmentation, it also indicates that sentiment-aware representations alone are insufficient to match approaches that more explicitly capture the structural dimensions of polarised discourse. Cross-lingual expansion strategies did not yield improvements, further suggesting that simple data scaling does not substitute for modelling the social-relational properties underlying polarisation.

Acknowledgments

This publication has been supported in part by the I+D+i project DeepSocial (PID2024-159202OB-C22), funded by the MICIU/AEI/10.13039/501100011033 and the FEDER/UE

References

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 258–266.

Alexis Conneau, Kartik 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 (ACL 2020)*, pages 8440–8451.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4040–4054.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Spencer Kiesel and Sharif Amlani. 2025. Affective polarization in a word: Open-ended and self-coded evaluations of partisan affect. *PLOS ONE*, 20(1):e0310772.

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, 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)*, 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, and 24 others. 2026b. Polar: A benchmark for multilingual, multicultural, and multi-event online polarization. *Preprint*, arXiv:2505.20624.

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 (ACL 2016)*, pages 86–96.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt — building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020)*, pages 479–480.