

Cross-Lingual Transfer Learning for Multilingual Emotion Classification and Generation

Adrián Maldonado

Universidad Politécnica de Madrid
adrian.maldonado@alumnos.upm.es

Oscar Araque 

Universidad Politécnica de Madrid
o.araque@upm.es

Abstract

This paper studies multilingual emotion analysis by exploring classification and conditioned text generation within language models. We examine how mixed-language training and sequential transfer affect cross-lingual generalization in multilabel emotion classification. Focusing on English and Spanish, our experiments show that limited target data yields substantial gains, while pretraining on high-resource or related languages supports robust transfer in low-resource settings. We extend this framework to emotion-conditioned text generation, producing text via prompting and evaluating alignment using our optimized classifier as an automated proxy. Results reveal that while few-shot prompting provides marginal gains in alignment, zero-shot generation produces more diverse outputs. These findings indicate that targeted multilingual training reduces annotation bottlenecks in target languages and that classification models can serve as scalable, robust evaluators for conditioned generation.

1 Introduction

Emotion analysis in text is a central task in natural language processing, with applications in areas such as sentiment analysis, human-computer interaction, and educational technologies. Although recent advances in transformer-based models have led to strong performance in high-resource settings, these improvements are uneven across languages. Many languages still lack sufficient annotated data, making it difficult to train reliable emotion classifiers.

Multilingual language models offer a potential solution by enabling knowledge transfer across languages. However, it remains unclear how different training strategies, such as mixing languages during training or adapting models through sequential fine-tuning, affect cross-lingual generalization, especially when target-language data are limited.

At the same time, large language models have made it possible to generate text conditioned on specific attributes, including emotional tone. This raises a complementary question: whether such models can be reliably steered to express intended emotions, and how this alignment can be evaluated in a scalable way.

In this work, we address these challenges by combining multilingual emotion classification with emotion-conditioned generation. Specifically, we investigate the following research questions:

RQ1: How do multilingual training strategies affect cross-lingual generalization and knowledge transfer in emotion classification?

RQ2: To what extent can language models be steered to generate emotion-conditioned text, and can a classifier be used to evaluate this alignment?

To address these questions, this work investigates two core components: cross-lingual classification dynamics and evaluation-steered text generation.

First, we analyze two complementary training regimes: mixed-language training and sequential transfer. By systematically varying both the proportion of languages in the training corpus and the volume of target-language supervision, we trace how emotional knowledge propagates across languages. This approach allows us to reflect on the empirical trade-offs between joint multilingual training and staged fine-tuning, particularly under low-resource constraints.

Second, we implement a conditional generation pipeline where language models produce text under various emotional prompting strategies. We leverage our optimized classifier as an automated evaluator to measure how closely the generated outputs align with the target emotions. This setup enables us to quantify steering reliability and analyze the trade-off between stylistic control and semantic diversity in emotion-conditioned generation.

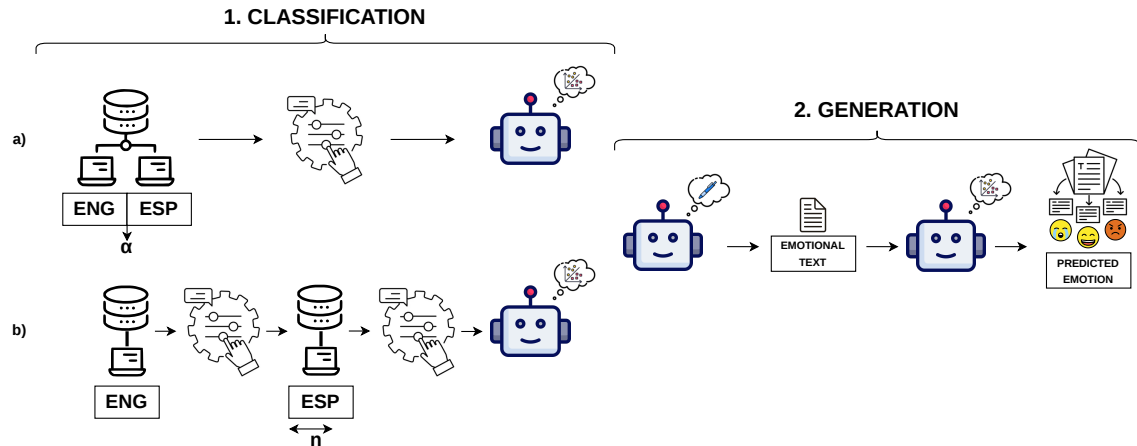


Figure 1: Overview of the proposed pipeline. Left: classification pipeline with (a) mixed-language training, where English and Spanish data are combined using a mixing coefficient α , and (b) sequential transfer, where a model is first trained on a source language and then adapted to Spanish using n samples. Right: generation pipeline, where a language model generates emotion-conditioned text from a prompt, and the trained classifier is reused to predict the expressed emotion and assess alignment with the intended label.

2 Related Work

2.1 Emotion Classification

Emotion classification has been widely studied in both monolingual and multilingual settings, with pretrained transformer encoders such as BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) enabling cross-lingual transfer through shared representations. These models support both joint multilingual training and sequential fine-tuning, which are commonly used to transfer knowledge from high-resource to low-resource languages. Prior work has shown that multilingual pretraining captures partially aligned semantic spaces, allowing models to generalize across languages even with limited target-language supervision.

Beyond standard supervised setups, cross-lingual emotion classification has been explored in zero-shot and few-shot regimes, where models trained on a source language are directly applied or minimally adapted to a target language (Artetxe et al., 2020). These approaches highlight the importance of language similarity, domain alignment, and label consistency, as well as the limitations of transfer when annotation schemes differ across datasets.

Cross-lingual emotion classification has been specifically investigated by Bostan and Klinger (2018), who analyze transfer effectiveness across multiple languages using multilingual encoders. More recent work studies robustness under varying data conditions, including low-resource settings,

label imbalance, and domain shifts. Shared tasks provide standardized multilingual benchmarks and emphasize the challenges of generalization beyond high-resource languages (Muhammad et al., 2025a).

Recent work has also incorporated large language models (LLMs), including GPT-4 (OpenAI et al., 2024) and LLaMA 3 (Grattafiori et al., 2024), which perform emotion classification through prompting rather than task-specific fine-tuning. While these models exhibit strong cross-lingual generalization, their performance is sensitive to prompt design, label verbalization, and output calibration, making controlled evaluation more difficult. Instruction-tuned models such as FLAN-T5 (Chung et al., 2022) and multilingual instruction-following variants (Muennighoff et al., 2023) improve robustness, particularly in low-resource and cross-lingual scenarios.

Recent surveys suggest that foundation models are reshaping affective computing by enabling unified modeling of classification and generation, as well as multimodal extensions, while emotion recognition remains a key benchmark for controlled and interpretable evaluation (Schuller et al., 2026).

2.2 Emotion Generation

Emotion-conditioned text generation has been addressed using both supervised and prompt-based approaches. Early work focused on explicitly conditioned generation, where models are trained to produce text given emotion labels or latent affective

representations (Rashkin et al., 2019). Subsequent approaches explored data augmentation and style transfer techniques to better control emotional expression (Zheng et al., 2023).

With the emergence of large language models, emotion control is increasingly achieved through natural language prompting. Models such as GPT-4 (OpenAI et al., 2024) and LLaMA 3 (Grattafiori et al., 2024) can generate emotionally grounded text across domains and languages, although consistency depends on prompt structure, instruction clarity, and decoding strategies. Few-shot prompting typically improves adherence to target emotions but can reduce lexical and semantic diversity, leading to more templated outputs.

Evaluation of emotion-conditioned generation remains an open challenge. Human evaluation provides the most reliable signal but is costly and difficult to scale. As a result, prior work has explored automatic metrics, including classifier-based evaluation, embedding similarity, and learned evaluators (Celikyilmaz et al., 2021). More recent approaches use LLMs as judges, though these introduce additional sources of variance and may inherit biases from the underlying models.

In this work, we adopt a classifier-based evaluation strategy, using a multilingual emotion classifier as an automatic signal to assess alignment between intended and generated emotions. Generated texts are passed through the classifier, and the predicted labels are directly compared to the target emotions used during prompting. This setup enables consistent, large-scale evaluation across prompting strategies within the same experimental framework.

3 Architecture

We propose a two-stage pipeline for multilingual emotion processing that combines classification and generation. The first stage performs multi-label emotion classification, and the second stage generates emotion-conditioned text and evaluates it using the classifier. Figure 1 provides an overview of the pipeline.

The classification component maps input text to a set of emotion labels. Following the annotation schema of the used dataset (see Sect. ref:sec:evaluation), we consider five non-exclusive categories, *anger*, *fear*, *joy*, *sadness*, and *surprise*. Each input text may express multiple emotions simultaneously, so the output is represented as a five-

dimensional binary vector.

3.1 Multilingual training strategies

We consider three training configurations that differ in how cross-lingual information is incorporated.

Monolingual baseline: Models are trained and evaluated on a single language, providing a reference for language-specific performance.

Mixed-language training: Training data from a source language and a target language is combined into a single dataset. The mixture is controlled by a coefficient $\alpha \in [0, 1]$, where α denotes the proportion of target-language samples and, conversely, $1 - \alpha$ represents the proportion of source-language samples. Formally, the training mixture is:

$$\alpha D_{\text{target}} + (1 - \alpha) D_{\text{source}}$$

where D_{source} and D_{target} denote the source and target language training sets, respectively. The evaluation is performed independently on each language.

Sequential transfer: Training is performed in two stages. A model is first fine-tuned on a source language and then adapted to the target language using n labeled examples.

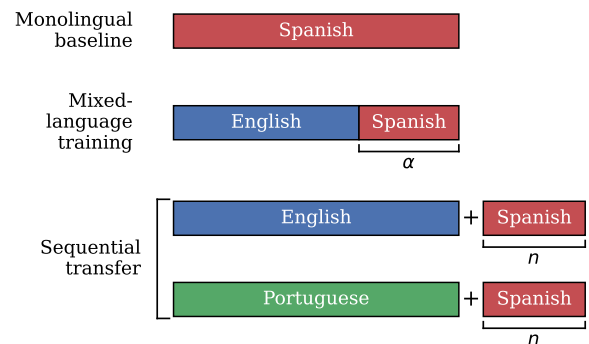


Figure 2: Schematic overview of the training configurations used in this work. Monolingual training uses only Spanish data. Mixed-language training combines English and Spanish with mixing coefficient α , where α controls the proportion of Spanish samples and $1 - \alpha$ the proportion of English samples. Sequential transfer fine-tunes a model first on English or Portuguese and then adapts it to Spanish using n labeled examples.

To ground our analysis and reflect a common real-world scenario, we identify English as our primary source language, capitalizing on its extensive availability of pre-training and fine-tuning data. The target language for our experiments is Spanish, representing a language with comparatively fewer readily available resources. This dichotomy

allows us to specifically probe the efficacy of cross-lingual transfer mechanisms in bridging this data disparity and achieving robust performance in a lower-resource setting. Portuguese is included as a linguistically related Romance language to Spanish, allowing us to examine transfer from a typologically closer source. This setup allows us to analyze the effect of prior cross-lingual training.

3.2 Emotion-conditioned generation

The generation stage extends the pipeline by producing emotion-conditioned text using a generative language model. Given a target emotion and contextual constraints, the model generates short text passages intended to reflect the specified emotional category.

To assess generation quality, we reuse the trained classifier as an automatic evaluator. The generated samples are fed to the classifier, and the predicted labels are compared with the intended emotion. We also explore a few-shot prompting variant in which the generation model is provided with example texts exhibiting clear emotional labels (see Figure 3). This setting improves classification-based metrics, but tends to reduce diversity in generated outputs.

PROMPT TEMPLATE

You are a teaching assistant speaking directly to a student. Write a short first-person paragraph expressing [intensity] [emotion] toward a situation related to [subject] class. Address the student using “you”, keep a supportive and respectful tone, and focus on realistic thoughts and feelings. Limit the response to 100 words.

LABELS

Intensity: mild, moderate, intense

Emotion: anger, fear, joy, sadness, surprise

Subject: math, history, science, literature, art

EXAMPLE

(emotion = sadness, intensity = intense, subject = art)
I’m really sad about what happened in art class today. I know how much you were looking forward to showing your finished project, and it’s heartbreaking that it was damaged during lunchtime. You poured your heart into creating something special, and it’s frustrating when things don’t go as planned. Please know that it’s not your fault - accidents happen, and we can always find a way to fix or recreate it. Let’s talk about what you want to do next, okay?

Figure 3: Prompt template used for emotion-conditioned generation with a single example output. The template varies across intensity, emotion, and subject.

4 Evaluation

4.1 Materials and Methods

We evaluate the classification models under three experimental settings using consistent train/test splits constructed independently for each language. For each language, the original training data is divided into training and test partitions using an 80/20 split.

Dataset. The dataset used in this work is based on SemEval-2025 Task 11 dataset for text-based emotion detection (Muhammad et al., 2025b). It provides manually annotated emotion labels across multiple text sources and supports multi-label emotion prediction. The corpus is manually annotated and includes diverse sources such as speeches, social media, news, literature, and reviews, with labels for six emotion classes: joy, sadness, anger, fear, surprise, and disgust. However, disgust was excluded from our experiments because it was annotated in the English data but was not available in the Spanish or Portuguese data.

To ensure comparability, we control for dataset size by fixing the maximum number of training examples per language. The Spanish dataset (1,996 examples) defines this upper bound, and we use up to 1,600 training samples after splitting. English and Portuguese data are subsampled accordingly. In mixed-language settings, the total training size remains fixed, with proportions determined by α , while in sequential transfer the number of examples used in the second stage is varied as $n \in [0, 1600]$.

Model implementation. All classification models are implemented using the Hugging Face Transformers library and configured for multi-label prediction, with one output logit per emotion category.

We evaluate a range of encoder-based transformer models, including DistilBERT, BERT, RoBERTa, XLM-R, and DeBERTa, in both English-specific and multilingual variants. Input texts are tokenized with a maximum sequence length of 128 tokens and padded to a fixed length. For large language models, we use LLaMA-3.2-3B with parameter-efficient fine-tuning. We apply 8-bit quantization and Low-Rank Adaptation (LoRA), restricting training to adapter modules. See Appendix A for references of the models.

While we evaluate multiple encoder-based models in the monolingual setting, we focus the mixed-language and sequential transfer experiments on Llama-3.2-3B. This model achieves the strongest

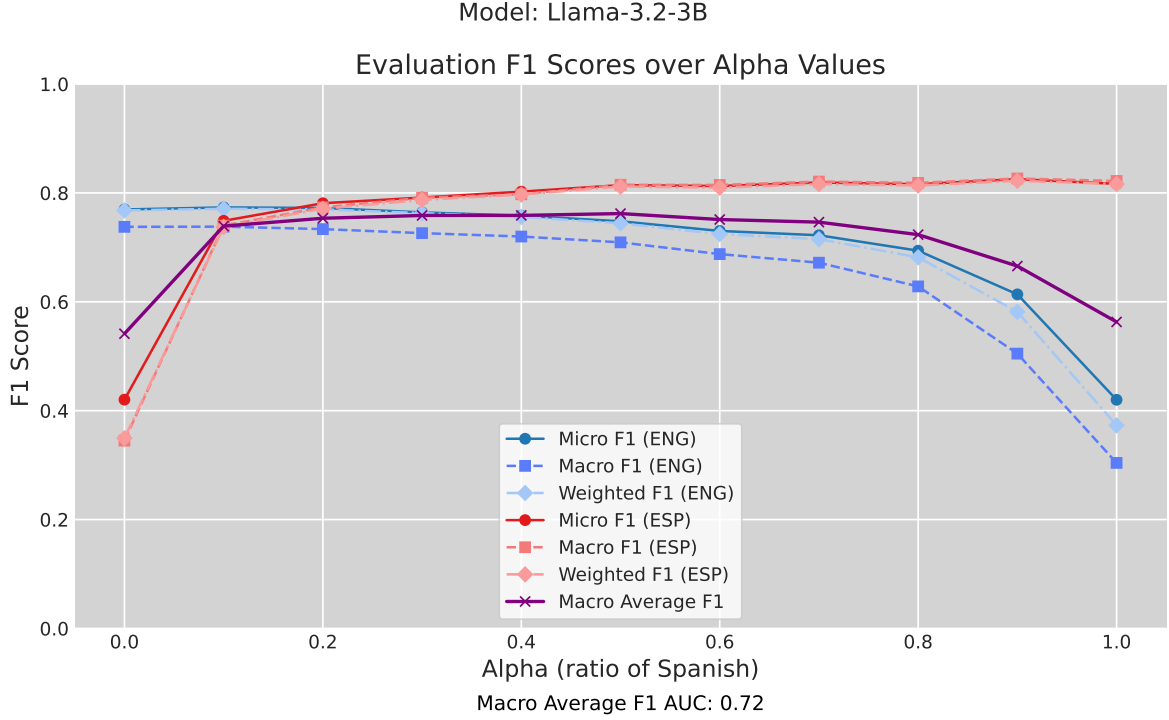


Figure 4: Mixed-language training results for Llama-3.2-3B. Blue and red curves show English and Spanish evaluation, respectively, as the mixture varies with α , the proportion of Spanish samples. The purple curve shows the macro-F1 averaged across both languages. The area under this curve (AUC) as a function of α is reported as a summary measure of stability across mixing ratios.

overall performance in our initial comparisons and provides a consistent basis for analyzing cross-lingual training strategies without introducing additional variability across architectures.

Training setup. Models are trained using the Hugging Face training framework with AdamW optimization (Loshchilov and Hutter, 2017), weight decay, and mixed-precision training. The same pipeline is used across all configurations, varying only the data composition and initialization. We use binary cross-entropy with logits for multi-label classification.

Generation setup. For the generation stage, we use a Llama 3.1-based generative model to produce short emotion-conditioned texts across different emotions, intensities, and subject domains. We select Llama 3.1 for generation due to its broader availability and stable inference behavior in our setup, while Llama 3.2 is used for the classification experiments.

The generated outputs are evaluated using the trained classifier, which serves as an automatic proxy for emotional alignment within the pipeline. Specifically, predicted labels are compared with the intended emotions used during prompting, allow-

ing us to quantify how reliably the model follows emotional constraints.

Performance is measured using standard classification metrics (micro-F1, macro-F1, and weighted-F1) computed over the generated dataset. In addition, a confusion matrix (Figure 6) provides a more detailed view of prediction patterns and common confusions between emotion categories.

We also evaluate a few-shot prompting variant within this framework. While this setting yields modest improvements in quantitative metrics, it tends to produce less diverse and more templated outputs, highlighting a trade-off between alignment and generation diversity.

Evaluation metrics. Performance is measured using micro-F1, macro-F1, and weighted-F1. These metrics are computed from sigmoid outputs using a fixed global threshold of 0.5 for all labels, without additional tuning.

4.2 Results

Classification results. Table 1 reports monolingual performance across encoder models and Llama-3.2-3B using micro-, macro-, and weighted-F1 for English and Spanish. English-specific en-

Model	English			Spanish			Macro Δ
	Micro	Macro	Weighted	Micro	Macro	Weighted	
DistilBERT (en)	0.666	0.596	0.654	0.540	0.468	0.476	-0.128
DistilBERT (multi)	0.586	0.456	0.557	0.667	0.658	0.652	+0.202
BERT (en)	0.698	0.630	0.687	0.598	0.551	0.555	-0.079
BERT (multi)	0.626	0.519	0.604	0.691	0.684	0.675	+0.165
RoBERTa (en)	0.716	0.661	0.709	0.691	0.687	0.682	+0.026
XLNet (en)	0.659	0.533	0.630	0.746	0.734	0.733	+0.201
DeBERTa (en)	0.689	0.591	0.668	0.746	0.734	0.734	+0.143
DeBERTa (multi)	0.661	0.550	0.638	0.759	0.755	0.751	+0.205
Llama-3.2-3B (multi)	0.770	0.738	0.768	0.817	0.822	0.816	+0.084

Table 1: Performance of encoder-based models and Llama-3.2-3B on English and Spanish emotion classification. Macro Δ is computed as Spanish macro-F1 minus English macro-F1.

coders achieve higher scores on English, while multilingual encoders perform more consistently across languages and obtain stronger results on Spanish. Llama-3.2-3B consistently outperforms encoder baselines in both languages, with particularly high macro-F1 on Spanish and a positive macro Δ indicating better performance in the target language.

Figure 4 presents mixed-language training results for Llama-3.2-3B. The horizontal axis shows α , the proportion of Spanish samples in the training data. Blue curves correspond to English evaluation, red curves to Spanish evaluation, and the purple curve shows macro-F1 averaged across both languages. The results show that incorporating even a relatively small proportion of Spanish (target-language) data leads to strong performance.

Figure 5 compares sequential transfer strategies, all evaluated on Spanish. The horizontal axis shows the number of Spanish examples used during the second fine-tuning stage (n). The baseline without pretraining (blue) is compared to English (red) and Portuguese (green) pretraining. Pretrained models outperform the baseline in the low-data regime, with English pretraining providing the largest gains.

Overall, both mixed-language training and sequential transfer improve performance when target-language data is limited.

Generation results. We evaluate the generation pipeline by comparing the intended emotion used for prompting with the emotion predicted by the classifier on the generated text. Emotion detection rate measures the proportion of samples where the intended emotion is assigned a probability above 0.5. For ranking-based metrics (Top-1, Top-2, and mean rank), emotion labels are ordered by predicted probability per sample.

The confusion matrices (Figure 6) show a diagonal structure in both settings, indicating general

agreement between intended and predicted emotions, with a larger diagonal in the few-shot setting. Most categories are well distinguished, although *surprise* is correctly classified only a small fraction of the time, with a substantial share of *surprise* predictions being mistaken for *joy*. Table 2 shows that few-shot prompting leads to marginal improvements across all metrics.

To assess diversity, we compute lexical and embedding-based metrics, including distinct-2 (Li et al., 2016) and average pairwise embedding similarity. We further define a creativity score as a composite measure that combines distinct-2 with three penalties: Self-BLEU (Guo et al., 2018), average embedding similarity, and duplicate rate, such that higher values indicate greater lexical variety, lower self-overlap, less redundancy, and broader semantic spread. The zero-shot setting shows higher diversity, as reflected by distinct-2 and the creativity score, while embedding similarity remains comparable across settings, indicating similar semantic spread.

Metric	Zero-shot	Few-shot
Emotion detection rate	0.82	0.83
Top-1 accuracy	0.70	0.73
Top-2 accuracy	0.92	0.93
Mean emotion probability	0.80	0.81
Mean emotion rank	1.42	1.40
distinct-2	0.46	0.41
Creativity score	0.14	0.10

Table 2: Generation performance evaluated via the classifier.

These results indicate a trade-off between alignment and diversity: few-shot prompting slightly improves emotion control, while zero-shot generation produces more varied outputs. They also support the use of the classifier as an automatic signal for evaluating emotion-conditioned generation.

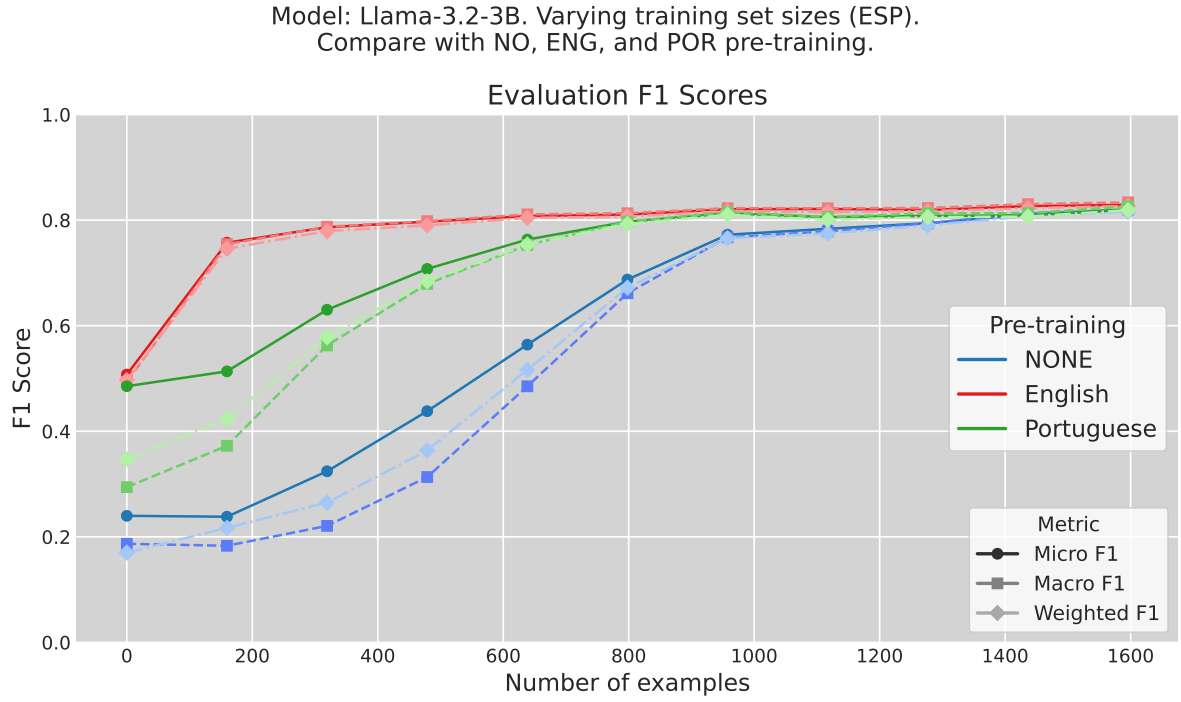


Figure 5: Sequential transfer results on Spanish evaluation. The x-axis indicates the number of Spanish examples used during the second fine-tuning stage. Blue denotes the baseline without pretraining, red denotes English pretraining, and green denotes Portuguese pretraining.

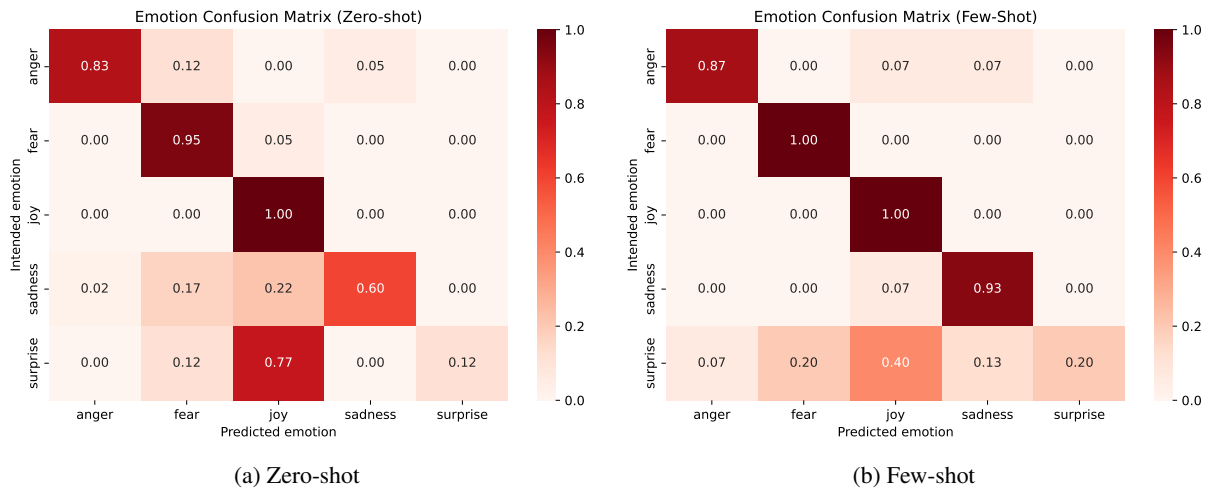


Figure 6: Confusion matrices for emotion prediction on generated texts. Rows correspond to intended emotions used in prompting, and columns correspond to emotions predicted by the classifier. Both settings show a diagonal tendency, with notable confusion between *surprise* and *joy*.

4.3 Discussion

The results highlight three main findings about multilingual emotion recognition. First, Table 1 shows that pretraining language strongly determines monolingual performance, with English models excelling on English and multilingual models performing better on Spanish.

Second, Figure 4 shows that combining languages during training enables effective cross-lingual transfer. Incorporating even a limited proportion of target-language data leads to strong performance, while training on a high-resource language supports generalization across both languages.

Third, and most importantly, Figure 5 shows that sequential transfer, especially from English, improves adaptation to Spanish in the low-data regime. Both English and Portuguese pretraining outperform the baseline without pretraining, indicating that cross-lingual features reduce the amount of target-language data required.

These findings have implications for low-resource emotion recognition. Rather than relying solely on large annotated datasets in the target language, models can leverage high-resource languages through mixed training or staged fine-tuning. The Llama-3.2-3B configuration with LoRA is effective in this setting.

The generation experiments provide a complementary perspective. The classifier serves as an automatic signal to evaluate whether generated texts reflect the intended emotion, and the results show consistent alignment between intended and predicted labels. Few-shot prompting slightly improves this alignment, while zero-shot generation produces more diverse outputs, indicating a trade-off between control and variability.

5 Conclusions

This paper addresses multilingual emotion processing by combining multi-label classification and emotion-conditioned text generation. By systematically evaluating diverse training regimes and prompting paradigms, we investigate our two core research questions:

Regarding **RQ1**, our findings demonstrate that multilingual training strategies substantially enhance cross-lingual generalization and knowledge transfer in emotion classification, particularly in low-resource target settings. Specifically, mixed-language training shows that incorporating even a

minor fraction of target-language data (α) alongside source-language samples yields performance comparable to fully supervised monolingual settings. Furthermore, our sequential transfer experiments indicate that pretraining on a high-resource language (such as English) or a typologically related language (Portuguese) accelerates target-language adaptation when the volume of target supervision is strictly limited ($n < 1,600$).

Regarding **RQ2**, our results suggest that generative language models can be steered to produce emotion-conditioned text across varying domains and intensities through target-oriented prompting. Within our classifier-based evaluation setup, the trained multilingual classifier provides a scalable automated signal for assessing emotional alignment. However, because the classifier uses the same task and label space as the generation evaluation, these results should be interpreted as exploratory evidence rather than independent validation. The confusion matrices indicate generally strong agreement between intended emotions and classifier predictions, although some categories exhibit substantial confusion. Finally, the comparison between prompting techniques highlights an operational trade-off: while few-shot prompting yields marginal improvements in classifier-based alignment measures, zero-shot generation preserves greater lexical diversity and creative semantic spread.

In summary, this work outlines a viable pathway for resource-limited affective computing, showing how cross-lingual transfer can reduce annotation dependencies and how classifier-based frameworks can support scalable exploratory evaluations of generation utility. Future work will focus on evaluating these findings across a broader range of architectures and a more typologically diverse group of target languages, corroborating the automatic classifier-based metrics with human judgments, and investigating how to sustain emotional and semantic control over longer, open-ended textual outputs.

Acknowledgements

This project has been funded by the project UNICO I+D Cloud AMOR (TSI-063100-2022-0002), financed by Ministry of Economic Affairs and Digital Transformation and the European Union through Next Generation EU; as well as the funding of the project CPP2023-010437 financed by

the MCIN/AEI/10.13039/501100011033/FEDER, UE. This work has been partially supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement 2023–2026 with Universidad Politécnica de Madrid in the Line A, Emerging PhD researchers (project MORA, DOCTORES-EMERGENTES-24-9UMLXZ-37-IIGW); Spanish version of this acknowledgement: “Financiado por la Comunidad de Madrid a través del convenio-subservención para el fomento y la promoción de la investigación y la transferencia de tecnología en la Universidad Politécnica de Madrid, en la Línea A, Doctores Emergentes”.

References

- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of monolingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Laura-Ana-Maria Bostan and Roman Klinger. 2018. [An analysis of annotated corpora for emotion classification in text](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. [Evaluation of text generation: A survey](#). *Preprint*, arXiv:2006.14799.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416.
- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 110–119.
- Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. [Crosslingual generalization through multitask finetuning](#). *Preprint*, arXiv:2211.01786.
- Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermimo Dario Mario Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025a. [SemEval-2025 task 11: Bridging the gap in text-based emotion detection](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2558–2569, Vienna, Austria. Association for Computational Linguistics.
- Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Sa’id Ahmad, and 1 others. 2025b. Semeval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th international workshop on semantic evaluation (SemEval-2025)*, pages 2558–2569.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. [Towards empathetic open-domain conversation models: A new benchmark and dataset](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Björn Schuller, Adria Mallol-Ragolta, Alejandro Peña Almansa, Iosif Tsangko, Mostafa M Amin, Anastasia Semertzidou, Lukas Christ, and Shahin Amiriparian. 2026. Affective computing has changed: The foundation model disruption. *npj Artificial Intelligence*, 2(1):16.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. [AugESC: Dialogue augmentation with large language models for emotional support conversation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

A Models used in the experiments

Table 3 shows the models used in the experimentation, along with their corresponding HuggingFace URLs.

Model Name	Hugging Face URL
DistilBERT	https://huggingface.co/distilbert-base-uncased
DistilBERT	https://huggingface.co/distilbert-base-multilingual-cased
BERT	https://huggingface.co/bert-base-uncased
BERT	https://huggingface.co/bert-base-multilingual-cased
RoBERTa	https://huggingface.co/roberta-base
XLM-R	https://huggingface.co/xlm-roberta-base
DeBERTa V3	https://huggingface.co/microsoft/deberta-v3-base
mDeBERTa V3	https://huggingface.co/microsoft/mdeberta-v3-base
Llama-3.2-3B	https://huggingface.co/meta-llama/Llama-3.2-3B

Table 3: Models and Corresponding Hugging Face URLs