VCR at Hope-EXP2026: Parameter-Efficient Large Language Models and Multilingual Transformers for Hope Speech Analysis*

Jesus Armenta-Segura[1,2,*], Grigori Sidorov[2]

[1]Monterrey Institute of Technology and Higher Education (ITESM), 2501 Eugenio Garza Sada Sur Av., Monterrey, 64700, Mexico.
[2]Instituto Politécnico Nacional, Centro de Investigación en Computación (CIC IPN), Juan de Dios Bátiz Av., Gustavo A. Madero, 07738 Ciudad de México, México.

Abstract

Hope is a fundamental human trait, frequently expressed on social media. It has the potential to enhance mental health, making it crucial to detect automatically. This paper presents the submission of the VCR team at the Hope-EXP2026 competition, which achieved first place overall. Our approach combined a QLoRA-fine-tuned Qwen3.5-4B model for Tasks 2 (multilabel detection), 3 (hope trigger span identification), 4 (actor identification), and 5 (stance classification), with a small XLM-RoBERTa model for Task 1 (hope speech detection). The proposed system obtained an overall score of 0.785. The code to reproduce our experiments can be find at https://github.com/JesusASmx/HopeEXP-2026

Keywords

Large Language Models, Applied QLoRa, Parameter Efficient LLMs, Hope Speech, Sentiment Analysis

1. Introduction

Hope is a fundamental psychological construct that plays a central role in shaping human behavior, motivation, and future-oriented thinking [1]. As a positive emotional state, it contributes to resilience, goal-directed action, and the capacity to cope with adversity across diverse social and cultural contexts. In recent years, social media platforms have emerged as prominent spaces for the expression of personal aspirations, collective expectations, and visions of societal progress, providing researchers with unprecedented access to large-scale manifestations of hopeful discourse [2].

The automatic identification and analysis of hope-related content, however, remain challenging tasks in Natural Language Processing (NLP) [1, 3]. Expressions of hope are frequently intertwined with figurative language, sarcasm, irony, implicit sentiment, and multilingual phenomena, making their computational modeling considerably more complex than traditional sentiment analysis. Developing robust approaches for detecting and characterizing hope speech is therefore of both theoretical and practical importance.

Beyond the development of datasets and modeling techniques [1, 4, 5], shared evaluation campaigns have played a fundamental role in advancing hope speech detection. Recent community-driven initiatives have established standardized multilingual benchmarks, promoted the exploration of both coarse- and fine-grained hope categories, and highlighted challenges related to linguistic diversity, cultural variation, and the distinction between genuine and sarcastic expressions of hope [3, 6]. These efforts have contributed to the maturation of the field by fostering methodological comparisons and encouraging the development of models capable of capturing increasingly nuanced forms of hopeful discourse. Continuing on this line, the 2026 HOPE-EXP shared task [2, 7] extends the scope of previous research by shifting the focus from the mere identification of hope toward the analysis of outcome-oriented expectations and future reasoning in social media texts, representing a natural next step in the computational study of hope and related affective phenomena.

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
jesusarmenta@tec.mx (J. Armenta-Segura); sidorov@cic.ipn.mx (G. Sidorov)
0009-0003-8729-7096 (J. Armenta-Segura); 0000-0003-3901-3522 (G. Sidorov)

This paper presents the submission of team VCR at the 2026 HOPE-EXP shared task [2, 7]. By fine-tunning a quantized 4-bit Qwen3.5:4b model [8] along with a RoBERTa model [9] to support Task 1, we achieved the global first place on the competition. The structure of this work is as follows: Section 2 presents a deeper insight into related work with respect of both Hope studies and the involved Transformer-based models. Section 3 explains the methodology, including what is quantization and low rank. Section 4 presents the experiments and results. Finally, Section 5 presents the conclusions.

2. Related Work

The automatic identification and analysis of hope-related content, however, remain challenging tasks in Natural Language Processing (NLP) [1, 4]. Expressions of hope are frequently intertwined with figurative language, sarcasm, irony, implicit sentiment, and multilingual phenomena, making their computational modeling considerably more complex than traditional sentiment analysis. Developing robust approaches for detecting and characterizing hope speech is therefore of both theoretical and practical importance.

Early efforts in this direction focused on establishing reliable annotation frameworks and taxonomies capable of capturing the multifaceted nature of hope. In particular, previous work introduced a hierarchical formulation that distinguished between the presence of hope and several fine-grained manifestations of hopeful expectations, highlighting the importance of carefully designed annotation guidelines and well-defined conceptual boundaries for the development of high-quality corpora [1]. These foundations contributed to framing hope as a complex affective phenomenon rather than a simple positive sentiment category.

Subsequent research explored the suitability of transformer-based architectures for modeling hope-related discourse, emphasizing the role of contextual representations in capturing subtle linguistic cues that characterize hopeful expressions [5]. These studies demonstrated that the interpretation of hope often depends on broader contextual information and semantic relationships that cannot be adequately represented through surface-level lexical features alone. Such findings reinforced the relevance of contextual language models for affective computing tasks involving nuanced emotional states.

Recent investigations have expanded the scope of hope speech analysis beyond English and beyond hope itself. Notably, the introduction of multilingual resources and the incorporation of hopelessness as a complementary category have provided a more comprehensive perspective on future-oriented emotional expression [4]. By considering both positive expectations and their absence, this line of work aligns computational approaches more closely with psychological theories and enables a richer understanding of how individuals communicate aspirations, uncertainty, and despair in online environments.

Another important direction has involved the refinement of hope taxonomies through the explicit modeling of more subtle communicative phenomena. Research has shown that hopeful discourse may manifest as realistic expectations, generalized optimism, exaggerated wishfulness, or even sarcastic expressions that superficially resemble hope while conveying opposing intentions [10]. The introduction of such distinctions has highlighted the limitations of coarse-grained classification schemes and underscored the need for models capable of performing fine-grained semantic and pragmatic reasoning.

The growing interest in multilingual hope speech detection has further motivated the creation of datasets spanning multiple languages and cultural contexts [10, 11]. These resources have facilitated comparative analyses of how hope is expressed across linguistic communities and have provided benchmarks for evaluating the generalization capabilities of modern NLP systems. Collectively, these efforts contribute to a broader understanding of hope as a culturally situated yet universally relevant psychological construct, opening new avenues for cross-cultural affective computing research [11].

3. Methodology

This section presents the methodology employed by our team to address the task. On a nutshell, it consists in a multi-model system who addresses all five tasks at once. For the single-label classification task of hope detection, the small model of RoBERTa was employed [9], since it has shown robust results on previous editions of this contest [12, 6]. As an interesting note, it was compared with the prompt-oriented few-shot learning focus, outperforming it.

3.1. Tasks Description

As hinted on the Introduction, this benchmarking competition comprises the follow five shared tasks:

1. **Hope Speech Primary Label Classification (Single-label)**:
   This task, similar to previous editions of this competition, consists on classify text on the follow categories: **Realistic Hope** (if expectations are plausible), **Unrealistic Hope** (otherwise), **General Hope** (when expectations are vague or broad), **Hopelessness** (expectations without hope), **Sarcastic Hope** (adding a mocking expression) and **Not Hope** (no expectations at all). The evaluation metric was Macro F1 score.
2. **Trigger Emotion Detection (Multi-label)**:
   Beyond detecting types oh hope speech on text, the authors also pushed for emotion identification. Since there is not a single emotion attached to a unique kind of hope, this is a multilabel task. The range considered on this dataset was: sadness, joy, love, anger, fear, surprise, Neutral/unclear[1]. The evaluation metric was Macro F1 score.
3. **Span Extraction (Outcome Identification)**:
   Pushing further, this year the authors decided to introduce a nuance to the task, by including the task of identifying which exact part(s) of the text corresponded with hope speech. The evaluation metric was ROUGE1.
4. **Outcome Role Classification (Conditional)**:
   Finally, given an extracted span, the final task consisted on identifying the actor triggering hope on it, considering the follow classes: **Self** (the author triggers it) **Other** (2nd or 3rd person/s), **World/System** or **Unclear**. Then, to determine the stance of such actor with respect of the triggered hope: **Desired** (in favour of) or **Avoided** (against it). The evaluation metric for both actor and stance was Macro F1 score.

As mentioned previously, our approach for the first task was with XLM-RoBERTa, who outperformed Qwen3.5 in this single case. All other tasks were addresed by Qwen3.5 trained with QLoRa (Figure 2). Follow sections explains further details.

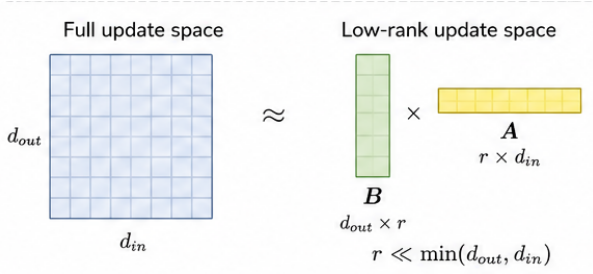
3.2. Qwen 3.5

Qwen3.5 is a new generation of large language models developed by the Qwen Team [8], designed to address the emerging paradigm of agentic artificial intelligence. Unlike conventional language models, Qwen3.5 adopts a native multimodal architecture that integrates textual and visual information from the outset, enabling unified understanding and reasoning across different modalities. The model family spans a broad range of scales and incorporates a hybrid architecture combining linear attention mechanisms with sparse Mixture-of-Experts (MoE) structures, allowing efficient long-context processing while maintaining high computational efficiency. In addition, Qwen3.5 supports more than 200 languages and provides open-weight variants released under the Apache 2.0 license, facilitating both research and industrial deployment.

A major objective of Qwen3.5 is to serve as a foundation for autonomous AI agents capable of performing complex, multi-step tasks involving reasoning, coding, tool usage, and multimodal interaction. The flagship models exhibit strong performance across benchmarks related to mathematical reasoning,

[1]Written as Nuetral/unclear on the actual dataset.

Figure 1: A visual example of Low-rank adaptation. A sample hidden layer L (purple) gets factorized by matrices A (yellow) and B (green). If one of the factors is freezed this reduces computational costs of training.



software development, visual understanding, and agentic workflows, while offering significantly improved inference efficiency compared with previous generations. Furthermore, extended context capabilities, reaching up to one million tokens in premium variants, enable large-scale document analysis and long-horizon reasoning. These characteristics position Qwen3.5 as a versatile foundation model for the development of next-generation intelligent systems and multimodal applications.

This work uses the 4b-parameters version and, although no visual inputs are involved, we utilize the full model.

3.3. Low-Rank and Quantization

Quantized Low-Rank adaptation fine-tuning was introduced on [13] as a solution to manipulate very big LLMs with relatively small resource settings. It consists on two methods: Quantization and Low Rank. The first consisting on *rounding* the neural weights into a smaller format (in this case, 4 bits), while the second consists on factorizing a hidden layer L into a matrix product $L = AB$ such that one of the factors (e.g. A) shall be freezed while the other is trained, in a cheaper fashion, due its reduced size. Figure 1 presents an example.

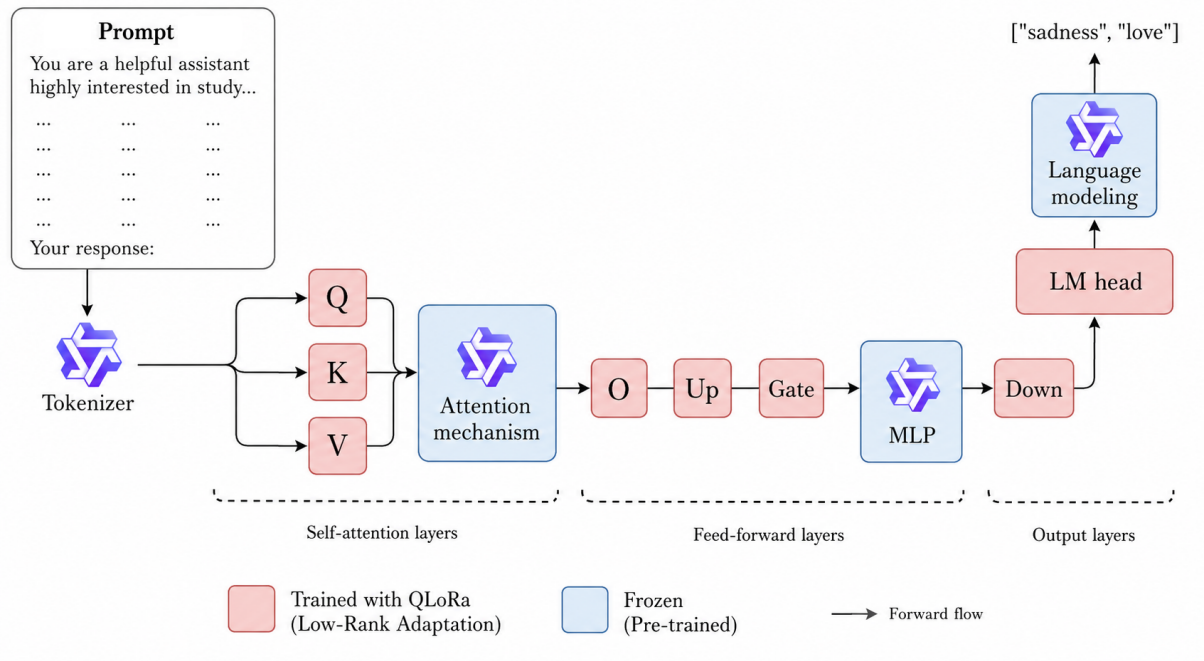
3.4. Prompt

To ensure consistency across experiments, a fixed set of prompts was employed for each task. These prompts were designed to constrain the output format while providing sufficient contextual information for the model to perform the corresponding inference. Table 1 summarizes the prompts used throughout the study. While Task 2 relies on an emotion classification prompt, Tasks 3,4 and 5 were performed at the same time, sharing a common prompt aimed at extracting hope-related spans and their associated semantic properties, including the responsible actor and the desired or avoided nature of the outcome.

Table 1
Prompts used for each task. Full versions can be found on Appendix A and, in yml format, at the github repo.

Task	Prompt
2	`You are a professional psychologist specializing in the underlying emotions expressed in ... Output a list written in the format <LIST>emotion_1, emotion_2, ..., emotion_last </LIST>and nothing else. Your list of emotions:`
3,4 and 5	`You are a helpful assistant highly interested in spans triggering hope in text written in ... DO NOT return "...". Which list of spans fits this text written in {english / spanish}?.`

Figure 2: Feedforward scheme for the Qwen 3.5 model, with a clear highlight between frozen and QLoRa-trained hidden layers. This Figure presents, as example, an arbitrary instance for Task 2 with outputs [“*sadness*”,“ *love*”].



4. Experiments

This section presents the performed experiments along with its specifications.

4.1. Experimental setup

With respect of Task 1, the working pipeline was the Huggingface standard for toy LLMs (e.g. [12]): each batch was fed into the XLM-RoBERTa tokenizer, and then the input ids with the attention masks were passed to the model, which was then fine-tuned via backpropagation using an optimizer (in this case, and for all experiments, Adam with weight decay). Table 2 presents all hyperparameters for RoBERTa.

Table 2
Fine-tuning hyperparameters for the XLM-RoBERTa model.

Parameter	Value
Model name	XLM-RoBERTa-base
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	16
Epochs	5
Weight decay	0.01
Evaluation frequency	Every epoch
Checkpoint frequency	Every epoch
Model selection metric	Macro F1-score
Padding	Dynamic
Truncation	Yes
Input type	Sentence pair classification

The other tasks involved a different experimental setup. The Unsloth python library [14] via its FastLanguagModel class was employed to load and speed-up the QLoRa model and tokenizer. Unfrozen

modules were selected with the aim to *simulate dropout mechanisms*, but also inspired from similar proposals such as [15]. Then, the Transformers library was employed to load a training pipeline (similar to the RoBERTa case). Table 3 presents the QLoRa specifications and Table 4 presents the hyperparameters.

Table 3
Configuration of the Qwen3.5-4B model and QLoRA adaptation.

Parameter	Value
Base model	Qwen3.5-4B
Quantization	4-bit (QLoRA)
Maximum sequence length	2048
LoRA rank (r)	16
LoRA scaling factor (α)	16
LoRA dropout	0
Bias	None
Target modules	$q_{\text{proj}}, k_{\text{proj}}, v_{\text{proj}}$, $o_{\text{proj}}, gate_{\text{proj}}, up_{\text{proj}}$, $down_{\text{proj}}, lm\_head$
Use RS-LoRA	False
Gradient checkpointing	Unsloth

Table 4
Training hyperparameters used for QLoRA fine-tuning.

Parameter	Value
Optimizer	AdamW 8-bit
Learning rate	2×10^{-4}
Learning rate scheduler	Linear
Batch size	2
Weight decay	0.001
Warmup steps	5
Training batch size	2
Gradient accumulation steps	4
Effective batch size	8
Number of epochs	5
Random seed	67
Dataset processes	1
Logging steps	1
Reporting	None

4.2. Results and Discussion

Table 5 presents the competition leaderboard. Figure 3 presents the loss curves for all three performed experiments (XLM-RoBERTa, Qwen3.5 for emotions and Qwen3.5 for spans, actor and outcomes).

With respect of task 1, XLM-RoBERTa had a good average performance, standing on the top submissions (at 0.02 points from the best task 1 model). Previous submissions with Qwen3.5 showed a best performance of 0.725 F1 score, dramatically inferior than the toy RoBERTa model. It remains as an open question to unveil the actual reason behind this particular behavior, considering that the same RoBERTa model underperformed the very similar Task 2, by obtaining an F1 score of 0.64 (and using the same hyperparameters).

On task 2, our Qwen3.5 model was the clear winner, with a gap of 0.039 points with respect of its closest competitor. On task 3, clearly strongly related with generative answers, showed an average performance. However, for the correctly extracted spans, actor and stance detection highly performed.

Table 5
All results of the competition. Best values are shown in bold. Our submission is highlighted on light blue.

Participant	Overall	Task 1 (F1)	Task 2 (F1)	Task 3 (ROUGE1)	Task 4 (F1)	Task 5 (F1)
jesus_as (us)	**0.785**	0.915	**0.789**	0.599	0.833	0.788
davidorm	0.783	**0.938**	0.750	**0.753**	0.829	0.644
UC-UCO-Plenitas	0.780	0.861	0.664	0.375	**1.000**	**1.000**
franrodrigo	0.755	0.924	0.710	0.664	0.819	0.658
adrianfernandez	0.749	0.933	0.701	0.736	0.772	0.600
wangkongqiang	0.619	0.829	0.655	0.374	–	–
abdollah	0.584	0.798	0.551	0.515	0.588	0.470
fazlfrs	0.533	0.861	0.633	0.411	0.614	0.148

Considering that Tasks 4 and 5 were, from a technical point of view, simpler versions of Task 1, it highlights a lot how Qwen3.5 outperformed RoBERTa on them, but not on Task 1. This may motivates research towards the field of Mechanistic Interpretability, concretely on the study of how emergent properties [16] via scalability [17] may influence the ability of LLMs to deal with small datasets.

5. Conclusion

Hope, a fundamental psychological construct expressed extensively on social media, has motivated the development of NLP approaches and multilingual shared tasks, culminating in the 2026 HOPE-EXP challenge.

This paper presented the submission of the VCR team, which achieved first place in the shared task. The proposed approach employed a QLoRA-fine-tuned Qwen3.5-4B model for tasks 2 (multilabel detection), 3 (hope trigger span), 4 (actor) and 5 (stance), with a lightweight XLM-RoBERTa model for Task 1 (hope speech detection). Our system attained an overall score of 0.785, at 0.002 points ahead of the second-ranked submission.

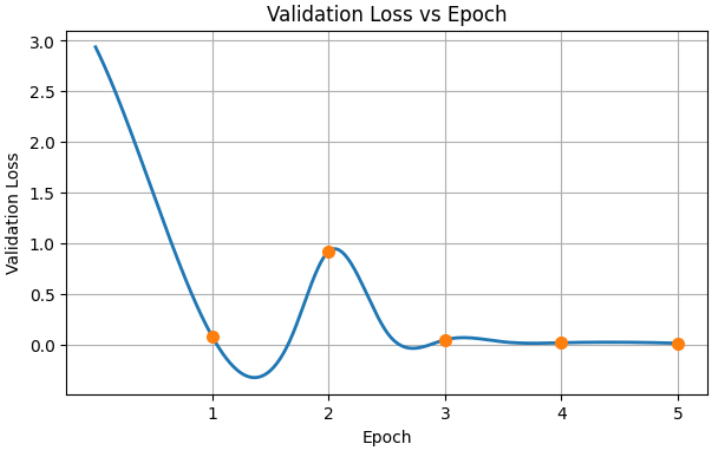
Beyond the obtained performance, the results provide interesting insights regarding the behavior of large language models in low-resource classification scenarios. In particular, the better performance achieved by XLM-RoBERTa on task 1 with respect of Qwen3.5 suggests that emergent properties may influence how such systems generalize when dealing with small datasets. Further investigation is required to better understand these phenomena and to determine the extent to which emergent capabilities contribute to performance in hope speech detection and, in general, related text classification tasks.

Declaration on Generative AI

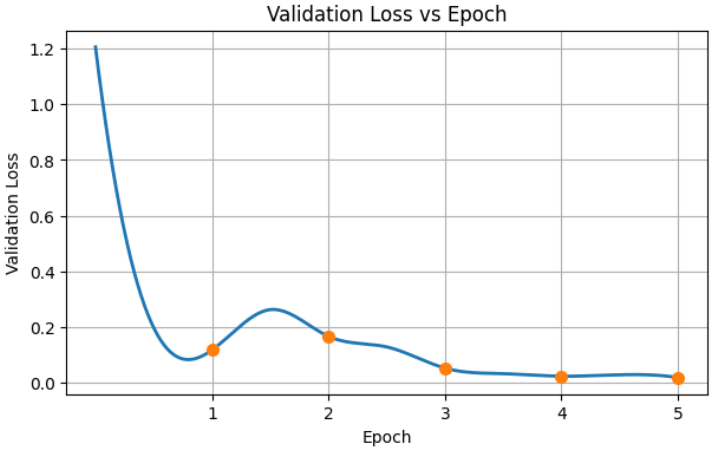
During the preparation of this work, the author(s) used GPT-5 in order to: Grammar and spelling check; figures 1 and 2 to generate an image out from a hand-made sketch; abstract drafting; citation management; Improve writing style. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Figure 3: Loss curves for all three experiments. As expected due their size, larger LLMs presented smoother and less chaotic losses curves than the toy RoBERTa model.

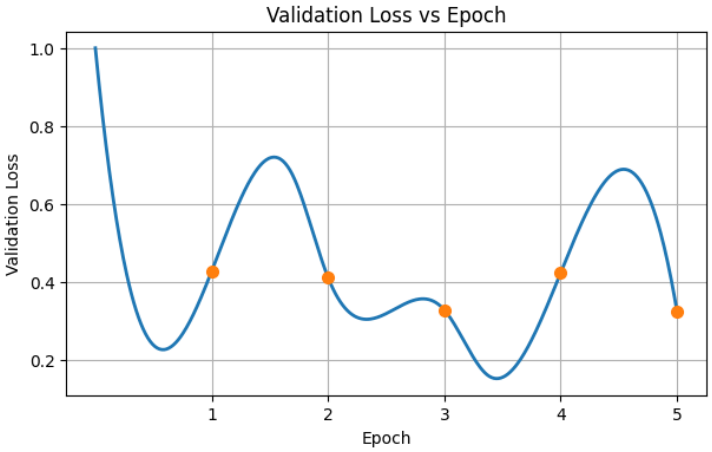
a) Qwen3.5 for emotions (task 2)



b) Qwen3.5 for spans, actor and outcome (tasks 3,4,5)



b) XLM-RoBERTa for hope (task 1)



References

[1] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078.

[2] F. Balouchzahi, S. Butt, H. Gómez-Adorno, S. M. Jiménez-Zafra, H. G. Ceballos, G. Sidorov, Overview of hope-exp at iberlef 2026: Outcome-oriented expectation analysis in social media,

Procesamiento del Lenguaje Natural 77 (2026).
[3] D. García-Baena, F. Balouchzahi, S. Butt, M. Á. García-Cumbreras, A. L. Tonja, J. A. García-Díaz, others, S. M. Jiménez-Zafra, Overview of hope at iberlef 2024: Approaching hope speech detection in social media from two perspectives, for equality, diversity and inclusion and as expectations, Procesamiento del Lenguaje Natural 73 (2024) 407–419.
[4] F. Balouchzahi, S. Butt, M. Amjad, G. Sidorov, A. Gelbukh, Urduhope: Analysis of hope and hopelessness in urdu texts, Knowledge-Based Systems 308 (2025) 112746.
[5] G. Sidorov, F. Balouchzahi, S. Butt, A. Gelbukh, Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets, Applied Sciences 13 (2023) 3983.
[6] S. Butt, F. Balouchzahi, M. Amjad, S. M. Jiménez-Zafra, H. G. Ceballos, G. Sidorov, Overview of polyhope at iberlef 2025: Optimism, expectation or sarcasm?, Procesamiento del Lenguaje Natural 75 (2025) 461–474.
[7] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of iberlef 2026: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, Leon, Spain, 2026, pp. –.
[8] Q. Team, Qwen3.5: Accelerating productivity with native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.
[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[10] S. Butt, F. Balouchzahi, A. I. Amjad, M. Amjad, H. G. Ceballos, S. M. Jiménez-Zafra, Optimism, expectation, or sarcasm? multi-class hope speech detection in spanish and english, arXiv preprint arXiv:2504.17974 (2025).
[11] G. Sidorov, F. Balouchzahi, L. Ramos, H. Gómez-Adorno, A. Gelbukh, Multilingual identification of nuanced dimensions of hope speech in social media texts, Scientific Reports 15 (2025) 26783.
[12] J. Armenta-Segura, G. Sidorov, Ometeotl at hope2024@iberlef: Custom bert models for hope speech detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), Valladolid, Spain, 2024, pp. –.
[13] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023) 10088–10115.
[14] M. H. Daniel Han, U. team, Unsloth, 2023. URL: http://github.com/unslothai/unsloth.
[15] L. Tian, J. R. Trippas, M.-A. Rizoiu, Mario at exist 2025: a simple gateway to effective multilingual sexism detection, arXiv preprint arXiv:2507.10996 (2025).
[16] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, arXiv preprint arXiv:2206.07682 (2022).
[17] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, arXiv preprint arXiv:2001.08361 (2020).

A. Prompts for Qwen3.5

This appendix presents both employed prompts. Variables are states between brackets {} with a description/possible values inside.

Prompt for Task 2:

```
You are a professional psychologist specializing in the underlying emotions expressed on {english / spanish}.
Your task is to analyze the follow text written in language, and predict the list of emotions present on it:
```

{sample to be analised by the model}

Instructions:
1. Allowed emotions to the list: anger, fear, joy, love, sadness, surprise, Nuetral/unclear
2. Multi-label allowed except Nuetral/unclear
3. If no emotion is detected, return Nuetral/unclear
4. Do not duplicate labels
5. Output a list written in the format <LIST> emotion_1, emotion_2, ..., emotion_last </LIST> and nothing else

Your list of emotions:

Prompt for Tasks 3,4 and 5:

You are a helpful assistant highly interested in spans triggering hope on text written in {english / spanish}.

You know the following types of hope: Realistic Hope, Unrealistic Hope, General Hope, Hopelessness and Sarcastic Hope.

Hope is structured with an actor responsible for the outcome:
- Self: the speaker is responsible.
- Other: another person or group.
- World/System: external systems or forces.
- Unclear: cannot be determined.

Actors also have stance:
- desired: wants outcome
- avoided: does not want outcome

PROCEDURE:

1. Hope span extraction:
Extract exact spans supporting the hope class.
For each span:
outcome_stance: desired or avoided
actor:
Self / Other / World/System / Unclear

OUTPUT RULES:
No markdown
No explanation
No extra text
Return a list of dictionaries for each span corresponding hope:[
{
'span': ["the span 1 corresponding to hope"],
'outcome_stance': ["Self"] or ["Other"] or ["World/System"] or ["Unclear"],
'actor': ["desired"] or ["avoided"]

```
},
{
'span': ["the span 2 corresponding to hope"],
'outcome_stance': ["Self"] or ["Other"] or ["World/System"] or ["Unclear"],
'actor': ["desired"] or ["avoided"]
}
]
 Can include 1 or more spans, or no span (return "[]" on that case).
  DO NOT invent tokens.
  DO NOT output partial words.
  DO NOT output characters.
  DO NOT return "..."

   Which list of spans fits this text written in {english / spanish}?
```
