UCO-UC-CICESE-Plenitas Team at HOPE-EXP 2026: A Hybrid Classical and Large Language Model Approach for Outcome-Oriented Expectation Analysis

Yoan Martínez-López[1,2,*], Mayte Guerra Saborit[3], Ireimis de las Mercedes Leguen de Varona[3], Miguel Bethencourt[3], Julio Madera[3], Luis Quintero Domínguez[4], Ansel Rodríguez-González[5], Kristell Aguilar Alarcón[1,2], José Carlos Arévalo Fernández[2], Carlos de Castro Lozano[1,2] and José Miguel Ramírez Uceda[1,2]

[1]*EATCO. Universidad de Córdoba, Córdoba, Spain*
[2]*Plénitas, C/ Le Corbusier s/n, 14005 Córdoba, Spain*
[3]*Universidad de Camaguey, Circunvalación Norte, Camino Viejo Km 5 y 1/2, Camaguey, Cuba*
[4]*Universidad de Sancti Spíritus, Cuba*
[5]*Centro de Investigación Científica y de Educación Superior de Ensenada, Unidad Académica Tepic, México*

Abstract

This paper describes the participation of the UCO-UC-CICESE-Plenitas team in HOPE-EXP 2026 (Outcome-Oriented Expectation Analysis in Social Media), a shared task at IberLEF 2026 that reframes hope detection as a multi-level, structured understanding problem over multilingual (English and Spanish) social-media posts. The task is decomposed into four interlocking subtasks: single-label primary classification into six hope categories (Task A), multi-label trigger-emotion detection (Task B), outcome-span extraction (Task C), and per-span outcome-stance and actor-role classification (Task D); systems are ranked by the unweighted mean of the five component metrics. We propose a single configurable two-layer pipeline. The classical layer always runs and combines a language-agnostic TF-IDF representation with a class-balanced logistic-regression classifier for Task A, a one-vs-rest classifier for Task B, and a bilingual rule-based extractor with lexical stance and actor rules for Tasks C and D. An optional few-shot large language model (LLM) layer, operating under a strict JSON output schema and using a backend compatible with multiple LLM providers, can refine or replace the classical predictions with an assembler merging both under an LLM-first, rules-fallback policy. An internal comparison of seven LLM shows that the choice of model dominates performance, that model family outweighs raw scale, and that a reasoning-oriented model fares worst under the strict JSON output format. On the official evaluation our system ranks third of eight with an Overall of 0.780, within 0.005 of the winner, and its internal estimate matched the official score exactly. The system exhibits a strongly bimodal profile—perfect scores on both Task D metrics but the lowest span coverage among the leading teams—which we analyze as a structural consequence of a high-precision, low-recall span extractor interacting with the conditional Task D metric, and which localises the most promising improvement in span recall.

Keywords

hope speech detection, expectation analysis, multilingual NLP, span extraction, large language models, TF-IDF, ensemble pipeline

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
yoan.martinez@plenitas.com (Y. Martínez-López); mayte.guerra@reduc.edu.cu (M. G. Saborit); ireimis.leguen@reduc.edu.cu (I. d. l. M. L. d. Varona); betamiguel00@gmail.com (M. Bethencourt); julio.madera@reduc.edu.cu (J. Madera); laqdominguez@gmail.com (L. Q. Domínguez); ansel@cicese.edu.mx (A. Rodríguez-González); kristell.aguilar@plenitas.com (K. A. Alarcón); josecarlos@plenitas.com (J. C. A. Fernández); carlosdecastrolozano@gmail.com (C. d. C. Lozano); miguel@plenitas.com (J. M. R. Uceda)
0000-0002-1950-567X (Y. Martínez-López); 0000-0002-9556-5869 (M. G. Saborit); 0000-0002-1886-7644 (I. d. l. M. L. d. Varona); 0000-0001-8873-6802 (M. Bethencourt); 0000-0001-5551-690X (J. Madera); 0000-0002-3527-0516 (L. Q. Domínguez); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-8913-8388 (K. A. Alarcón); 0009-0004-4155-957X (J. C. A. Fernández); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda)

1. Introduction

Psychological accounts describe it as a future-oriented expectation that couples a specific desire for a positive outcome with a broader optimistic outlook on the events to come [1], and empirical work has linked it to outcomes as varied as academic achievement, resilience under adversity, and mental-health trajectories. These same properties make hope difficult to model computationally: it is frequently conveyed implicitly, it may be grounded in plausible expectations or in fantastical ones, it can be directed at oneself, at others, or at impersonal systems, and it is routinely entangled with other emotions and with figurative devices such as sarcasm, where ostensibly hopeful wording carries the opposite intent. Within Natural Language Processing, the automatic detection of hope speech has gained traction because of its relevance to mental-health monitoring, education, and the moderation of online communities [2, 3]. A sustained line of shared tasks has shaped the field: the HopeEDI initiative on equality, diversity, and inclusion [4, 5], the HOPE tasks at IberLEF 2023 [6] and IberLEF 2024 [4], and most recently PolyHope at IberLEF 2025 [3] together with its multilingual extension PolyHope-M at RANLP 2025 [7]. These campaigns progressively refined the taxonomy of hope—moving from a binary hope/not-hope decision to a finer-grained scheme that separates generalized, realistic, and unrealistic hope, and that, in its latest editions, treats sarcastic hope as a category in its own right rather than as noise [8, 9, 5]. Despite this progress, prior formulations approach hope as a flat, sentence-level classification problem. A system is asked whether a post expresses hope and, at most, which coarse subtype it belongs to, but it is never required to articulate *what* is hoped for, *who* the relevant actor is, or whether the anticipated outcome is one to be *sought* or *averted*. Error analyses from previous editions repeatedly show that the hardest confusions are precisely those that hinge on this missing structure: realistic versus unrealistic hope, and sincere versus sarcastic expressions, are conflated because the surrounding outcome and stance are left unmodelled. As a result, current benchmarks capture the presence of hope but not its internal, outcome-oriented semantics, and they offer little support for the explainable, fine-grained analyses that downstream applications increasingly demand. Motivated by this gap, we present HOPE-EXP (Outcome-Oriented Expectation Analysis in Social Media), a shared task organised within IberLEF 2026 [10].

HOPE-EXP reframes hope detection as a multi-level, structured understanding problem over social-media posts and decomposes it into four interlocking subtasks. Task A is a single-label classification into six categories—General Hope, Realistic Hope, Unrealistic Hope, Sarcastic Hope, Hopelessness, and Not Hope—extending the PolyHope taxonomy with an explicit hopelessness class. Task B is a multi-label detection of the trigger emotions accompanying the post (sadness, joy, love, anger, fear, surprise, and neutral/unclear). Task C requires the extraction of up to three text spans that name the expected or avoided outcome, evaluated with ROUGE-1 against the gold spans. Task D then assigns to each aligned span an outcome stance (Desired or Avoided) and an actor role (Self, Other, World/System, or Unclear). Posts and system responses are exchanged in a unified JSON schema, the data span multiple languages through an explicit `lang` field, and the official ranking is the mean of the five component scores (Task A macro-F1, Task B macro-F1, Task C ROUGE-1, and the macro-F1 of the stance and actor classifiers in Task D).

This paper presents the UCO-UC-CICESE-Plenitas system submitted to HOPE-EXP at IberLEF 2026 [10, 1]. We describe a configurable two-layer pipeline that combines classical machine-learning components, rule-based span extraction, and an optional large language model (LLM) layer to address the four subtasks of HOPE-EXP. We detail the system architecture, the multilingual data representation, the evaluation protocol, and the model-selection process used to choose the final submission. We also analyse the official results and discuss the strengths, limitations, and future directions of the proposed approach. The remainder of the paper is organised as follows. Section 2 presents the proposed methodology, including the classical and LLM-based components. Section 3 reports and discusses the experimental results. Finally, Section 4 concludes the paper and outlines future work.

2. Methodology

2.1. Subtasks

HOPE-EXP frames the analysis of hope in social-media posts not as a single decision but as a multi-level, structured understanding problem[10, 1], decomposed into four interlocking subtasks that are solved jointly for every post. Each input is a record with a language field and the title and selftext of a post, and each system response is a single JSON object containing a primary label, a list of trigger emotions, and a list of outcome spans with their roles. Task A is a single-label classification of the post into one of six categories: General Hope, Realistic Hope, Unrealistic Hope, Sarcastic Hope, Hopelessness, and Not Hope. Task B is a multi-label detection of the emotions that accompany the post, drawn from sadness, joy, love, anger, fear, surprise, and a neutral/unclear class. Task C requires, only for the four hope categories, the extraction of up to three spans that are exact substrings of the source text and that name a desired or avoided outcome; for Not Hope and Hopelessness the span list must be empty. Task D then assigns to each extracted span an outcome stance (Desired or Avoided) and an actor role (Self, Other, World/System, or Unclear). Our system addresses all four subtasks within a single configurable pipeline that emits the complete JSON object per post.

2.2. Evaluation Protocol

Following the official protocol, the overall score is the unweighted mean of five component metrics: the macro-averaged F1 of Task A, the macro-averaged F1 of Task B, the ROUGE-1 score of the Task C spans, and the macro-averaged F1 of the stance and actor classifiers in Task D. Task A is evaluated as a six-way macro F1, which weights every label equally and therefore penalizes systematic failure on the minority categories (notably Sarcastic Hope and Unrealistic Hope). Task B is scored as a per-emotion binary F1 averaged across the seven emotion labels, reflecting its multi-label nature. Task C compares each gold span against the predicted spans with token-level ROUGE-1, and the Task D classifiers are computed only over spans that are successfully aligned to a gold span, making them conditional on correct span extraction. For internal model selection we reproduce this five-component protocol on a held-out sample of the training partition, mirroring the official aggregation exactly: Task A and the Task D roles are scored with a multiclass macro F1, Task B with a multi-label macro F1, and Task C with token-level ROUGE-1, with span alignment performed by a ROUGE-1 overlap threshold so that stance and actor are only credited on sufficiently matched spans. This self-evaluation guides the choice between the classical and LLM configurations and the tuning of the classical components before any submission is generated, keeping internal validation separate from the official test evaluation[10, 1].

2.3. Dataset

The corpus is provided in JSON-Lines format, with a labeled training split and an unlabeled test split, and is multilingual: every record carries an explicit language field, and the training data contain both English and Spanish posts. For each post the textual view is formed by concatenating the title and self-text fields, which is the single representation consumed by all text-based components [2, 9, 3]. The gold annotations attach to each post a primary label, a list of trigger emotions, and, for hope-bearing posts, a list of outcome spans with their stance and actor roles, exactly matching the four-subtask schema described above. Because the same model is applied to both languages without language-specific branching, the pipeline operates in a unified multilingual setting and relies on multilingual lexical cues and, when enabled, on a multilingual language model.

2.4. Text Representation

The text representation underlying Tasks A and B is a sparse, language-agnostic TF-IDF encoding of the concatenated post text rather than dense neural sentence embeddings [11]. In the primary configuration the vectoriser uses word n-grams of order one and two, sub-linear term-frequency scaling,

a minimum document frequency, and a capped vocabulary, and it preserves accented characters so that Spanish surface forms are represented faithfully; a default-parameter TF-IDF variant is retained as a strict baseline counterpart. This choice keeps the feature space fully reproducible, fast to compute offline, and free of any dependence on a particular pre-trained encoder, which is desirable for a four-task system whose neural capacity is concentrated in the optional language-model layer. The TF-IDF matrix feeds directly into the classifiers described next, and the same vectoriser is fitted on the training split and applied unchanged to the test split so that the two views remain aligned.

2.5. Large Language Models

In parallel with the classical layer, the pipeline incorporates a generative large language model (LLM) as an optional refinement layer. The model is prompted with a strict system specification that defines the six primary labels, the seven emotions, the span-extraction rules, and the stance and actor inventories, and is instructed to return a single valid JSON object with no surrounding text. A small set of few-shot demonstrations covering both languages and several label types — including sarcastic hope, general hope, hopelessness, and a not-hope case — is prepended to condition the model on the exact JSON output schema. Decoding uses a low temperature for near-deterministic behavior and a bounded generation budget, and a robust parser strips any Markdown fences, isolates the outermost JSON object, and falls back gracefully when the response cannot be parsed. Two complementary configurations are supported. In the full configuration the LLM produces all four predictions (A, B, C, and D) and, when it returns valid JSON, replaces the classical predictions entirely; otherwise the system reverts to the classical layer for A and B and to the rule-based extractor for C and D. In a span-focused configuration the classical layer decides the primary label and emotions and the LLM receives the post together with the already-decided label, restricting its task to producing the span annotations for Tasks C and D. The backend is compatible with multiple LLM providers: it supports a locally served Ollama endpoint, a hosted Ollama Cloud endpoint, and the HuggingFace inference router with both chat-completion and text-generation interfaces, and the same prompt operates unchanged across model families such as glm5 [12], llama [13], mistral [14], gemma [15], deepseek [16], and qwen [17] . All outputs are passed through the same normalization and validation stage before assembly, with retries on malformed generations.

2.6. System Architecture

The system is implemented as a single configurable pipeline organized in two layers, summarized in Figure 1. The classical layer (Layer 1) always runs: it fits a TF-IDF vectoriser on the training posts and trains a class-balanced logistic-regression classifier for the primary label (Task A) and a one-vs-rest logistic-regression classifier over the binarised emotion labels for the multi-label emotion task (Task B). The spans for Tasks C and D are produced by a rule-based extractor that scans the post with a bilingual inventory of hope cue patterns (English and Spanish), extracts the continuation window after each cue and, when needed, falls back to full sentences containing a cue; each candidate is verified to be an exact substring of the source, deduplicated, and truncated to at most three spans. For each retained span, lexical rules assign the outcome stance from negation and avoidance markers and assign the actor from first-person, third-person, and world/system keyword sets, with the empty-span constraint enforced whenever the predicted label is Not Hope or Hopelessness. The optional LLM layer (Layer 2) operates in parallel with the classical layer and can be enabled or disabled through configuration settings. An assembler then merges the two sources under a fixed priority policy: if the LLM returned a valid JSON object, its primary label, emotions, and validated spans are used, with a safeguard that falls back to the rule-based extractor when the LLM predicts a hope label but yields no valid span; otherwise the classical label and emotions and the rule-based spans are used. Every field passes through normalization — canonicalising label, emotion, stance, and actor strings, and revalidating that each span is an exact substring of the post — so that the final object always conforms to the required schema. The assembled predictions are written one JSON object per line to a JSONL file and compressed, without

sub-directories, into the archive expected by the evaluation platform, with the per-post `row_id` and language preserved to keep the submission aligned with the gold test set.

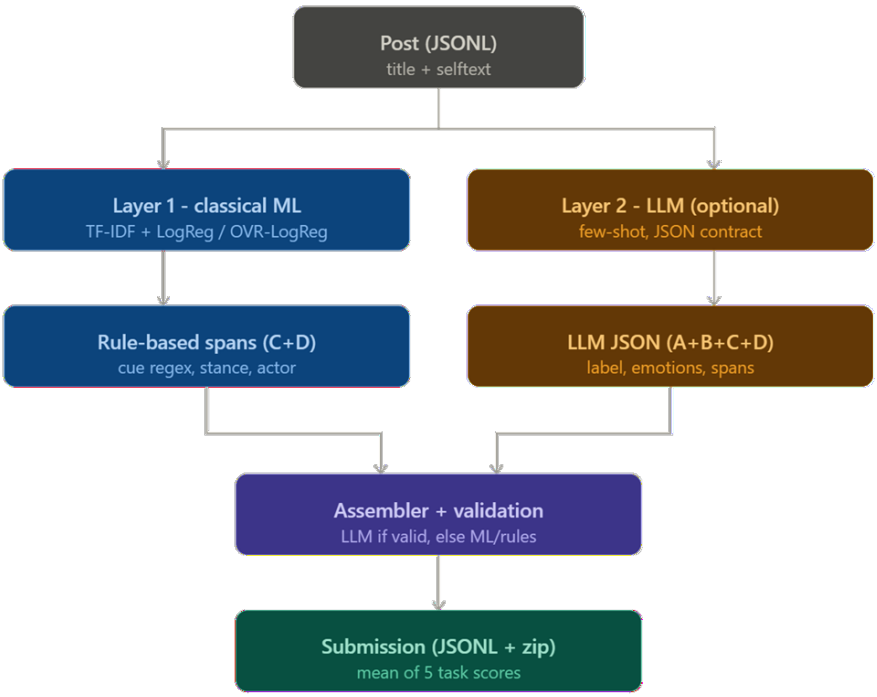

Figure 1: Overview of the proposed HOPE-EXP pipeline.

3. Results

Table 1 reports the internal Overall score (the mean of the five official component metrics) for each candidate LLM model. Three observations stand out. First, the choice of LLM model is the dominant performance factor: scores span a wide 0.298 range, from 0.780 down to 0.482. Second, the models fall into three distinct tiers rather than a smooth continuum—a clear leader (glm5-cloud, 0.780), a strong middle pair (qwen3.5-cloud, 0.710, and qwen2.5-7b, 0.695), and a lower group below 0.62.

Notably, the reasoning-oriented deepseek-r1 ranks last (0.482), consistent with the observation that models which emit extended reasoning traces struggle with the strict JSON output format and trigger more parsing fall-backs, so that format compliance matters more than abstract reasoning capacity for this task. Third, model family appears to outweigh raw scale: the two qwen models cluster together, with qwen2.5-7b achieving 0.695 and qwen3.5-cloud 0.710, a difference of only 0.016 despite the generational gap between them. We therefore adopt glm5-cloud, which leads the second-best system by a substantial margin of 0.070, as the LLM model used in our official runs.

On the other hand, our keyword-based rules are deliberately high-precision and low-recall: they abstain whenever a span cannot be assigned a stance and actor unambiguously, and therefore fire on only a small, confidently-handled subset of the data. Concretely, of the 2,082 evaluated posts (1,087 ES / 995 EN), the system extracted at least one span in only 519 posts (24.9%); the remaining 1,563 posts (75.1%) yielded no span at all. The total number of extracted and aligned segments on which stance and actor are assigned is therefore 1,228 (732 ES / 496 EN), not the full corpus. Within this subset, the resulting label distributions are: outcome_stance = Desired: 995, Avoided: 233 and actor = Self: 454, Unclear: 410, Other: 210, World/System: 154.

Table 2 reports the official results. UCO-UC-CICESE-Plenitas ranks third out of eight with an overall score of 0.780, within 0.005 of the winning system (`jesus_as`, 0.785) and 0.003 of the runner-up; the

Table 1
Internal comparisons

Method	Score
glm5-cloud	0.77993
qwen3.5-cloud	0.71024
qwen2.5.7b	0.69463
llama3.1	0.53354
mistral7b	0.62190
deepseek-r1	0.48183
gemma4	0.55115

top five teams are tightly grouped between 0.749 and 0.785, with a sharp drop below sixth place. The defining characteristic of our system is its strongly bimodal profile in all five components. It is the only team to obtain a perfect 1.000 on both Task D metrics (stance and actor), yet it records the lowest Task C span score among the leading systems (ROUGE-1 of 0.375) and the lowest Task A and Task B scores in the top five (0.861 and 0.664).

This pattern is consistent with the design of our pipeline. Because Task D is scored only over spans aligned to the gold annotations, a high-precision, low-recall span extractor can yield excellent stance and actor classification scores while simultaneously limiting Task C performance. In our case, the rule-based extractor appears to identify a relatively small subset of highly reliable spans, on which the lexical stance and actor rules perform extremely well. As a result, the system achieves perfect Task D scores despite comparatively modest span coverage.

In contrast, the other top systems (e.g. jesus_as and `davidorm`) exhibit a more balanced distribution of performance across the five evaluation components. This observation also highlights the most promising direction for future improvements: increasing span-extraction recall. Given the current gap between our Task C score and those of the highest-ranked systems, even moderate gains in span coverage could translate into meaningful improvements in the Overall score while preserving the strengths already demonstrated in Task D.

Table 2
Final Results

Rank	Team	Overall	F1_primary	F1_emotion	ROUGE1_span	F1_stance	F1_actor
1	jesus_as	0.78498	0.91519	0.78941	0.59925	0.83288	0.7882
2	davidorm	0.78273	0.93777	0.74977	0.7534	0.82885	0.64388
3	**UCO-UC-CICESE-Plenitas**	**0.77993**	0.86074	0.66388	0.37503	**1.0**	**1.0**
4	franrodrigo	0.75479	0.92362	0.70953	0.66398	0.81874	0.65811
5	adrianfernandez	0.74856	0.93303	0.7007	0.73649	0.77221	0.60036
6	wangkongqiang	0.61921	0.8289	0.65456	0.37416	n/a	n/a
7	abdollah	0.58425	0.79752	0.55125	0.51455	0.58788	0.47004
8	**baseline**	0.53346	0.8611	0.63338	0.41134	0.61362	0.14784

4. Conclusions

We have presented the UCO-UC-CICESE-Plenitas system for HOPE-EXP 2026, an IberLEF shared task that moves hope analysis beyond flat sentence-level classification toward a structured, outcome-oriented understanding of multilingual social-media posts across four subtasks. Our approach is a single configurable two-layer pipeline: a classical layer that always runs—TF-IDF with a class-balanced logistic-regression classifier for the primary label, a one-vs-rest classifier for trigger emotions, and a bilingual rule-based extractor with lexical stance and actor rules for the outcome spans—and an optional few-shot LLM layer, governed by a strict JSON output format and a backend compatible with multiple LLM providers, merged through an LLM-first, rules-fallback assembler. The system ranked

third of eight teams with an Overall of 0.780, within 0.005 of the winning submission.

Three conclusions stand out. First, when an LLM model is used, its selection appears to be the dominant performance factor: across seven candidates the internal Overall spanned a wide 0.298 range, model family outweighed raw scale (the two Qwen models clustered together despite a generational gap), and the reasoning-oriented model ranked last, consistent with the view that compliance with the strict JSON output format can be at least as important as abstract reasoning capacity for this task. Second, our internal cross-validated estimate (0.780) matched the official score exactly, providing evidence that the model-selection procedure generalized well to the official test set. Third, and most informative, our result is strongly bimodal: the system attained perfect scores on both Task D metrics yet the lowest span coverage among the leading teams. This behaviour is consistent with a structural interaction between span extraction and Task D evaluation: a high-precision, low-recall span extractor aligns only a small subset of spans, on which the lexical stance and actor rules perform extremely well. As a result, the conditional Task D metrics remain exceptionally high, while the coverage-sensitive ROUGE-1 score of Task C is comparatively reduced.

This analysis localises the improvement path precisely. Because the strongest performance gains are concentrated in the two Task D components, while the primary-label, emotion, and span tasks remain below the leading systems, the highest-leverage direction is to raise the recall of the span extractor: even at the cost of slightly reduced Task D alignment, the net Overall gain is expected to be positive given how low the span component currently sits. Future work will therefore focus on (i) a higher-recall, learning-based span extractor to replace the cue-pattern rules; (ii) stronger Task A and Task B classifiers, for instance by adding multilingual sentence embeddings alongside the TF-IDF features; and (iii) a more systematic use of the LLM layer for span grounding and cross-lingual transfer, the recurring challenges that this outcome-oriented formulation of hope brings to the fore.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Opus 4.8) and Grammarly for language revision, grammar and spelling correction, and assistance in the preparation of figures.

References

[1] 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).

[2] 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). URL: https://arxiv.org/abs/2504.17974. doi:10.48550/arXiv.2504.17974.

[3] 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. doi:10.26342/2025-75-34.

[4] D. García-Baena, F. Balouchzahi, S. Butt, M. Á. García-Cumbreras, A. L. Tonja, J. A. García-Díaz, et al., 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.

[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. doi:10.3390/app13063983.

[6] S. M. Jiménez-Zafra, M. Á. Garcia-Cumbreras, D. García-Baena, J. A. Garcia-Díaz, B. R. Chakravarthi, R. Valencia-García, L. A. Ureña-López, Overview of HOPE at IberLEF 2023: Mul-

tilingual hope speech detection, Procesamiento del Lenguaje Natural 71 (2023) 371–381. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6567.

[7] 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. doi:10.1016/j.knosys.2024.112746.

[8] F. Balouchzahi, G. Sidorov, A. Gelbukh, PolyHope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078. doi:10.1016/j.eswa.2023.120078.

[9] 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. doi:10.1038/s41598-025-11823-z.

[10] 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 Workshop Proceedings, CEUR-WS.org, 2026.

[11] D. Dessí, R. Helaoui, V. Kumar, D. R. Recupero, D. Riboni, TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study, arXiv preprint arXiv:2105.09632 (2021). URL: https://arxiv.org/abs/2105.09632. doi:10.48550/arXiv.2105.09632.

[12] A. Zeng, X. Lv, Z. Hou, Z. Du, Q. Zheng, B. Chen, Y. Zhou, et al., GLM-5: From vibe coding to agentic engineering, arXiv preprint arXiv:2602.15763 (2026). URL: https://arxiv.org/abs/2602.15763. doi:10.48550/arXiv.2602.15763.

[13] L. Roque, R. Guedes, The evolution of Llama: From Llama 1 to Llama 3.1, https://towardsdatascience.com/the-evolution-of-llama-from-llama-1-to-llama-3-1-13c4ebe96258/, 2024. Towards Data Science. Accessed: 2026-06-24.

[14] H. Thakkar, A. Manimaran, Comprehensive examination of instruction-based language models: A comparative analysis of Mistral-7B and Llama-2-7B, in: 2023 International Conference on Emerging Research in Computational Science (ICERCS), IEEE, 2023, pp. 1–6. doi:10.1109/ICERCS57948.2023.10434081.

[15] Gemma Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, S. Garg, et al., Gemma 2: Improving open language models at a practical size, arXiv preprint arXiv:2408.00118 (2024). URL: https://arxiv.org/abs/2408.00118. doi:10.48550/arXiv.2408.00118.

[16] M. Tordjman, Z. Liu, M. Yuce, V. Fauveau, Y. Mei, J. Hadjadj, X. Mei, et al., Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning, Nature Medicine 31 (2025) 2550–2555. doi:10.1038/s41591-025-03726-3.

[17] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, T. Zhu, Qwen technical report, arXiv preprint arXiv:2309.16609 (2023).

A. LLM Prompts and Few-Shot Demonstrations

For full transparency and reproducibility, this appendix reproduces *verbatim* the two artefacts that govern the LLM layer described in Section 2.5. The strict system specification (Listing 1) and the multilingual few-shot demonstrations (Listing 2). Both are shown exactly as passed to the model, including the enforced JSON output schema. Minor orthographic inconsistencies present in the original configuration (e.g., the label string `Nuetral/unclear`) are preserved deliberately, so that the listings match the exact prompts used in our experiments.

,

Listing 1: System specification governing the LLM layer (verbatim).

```
You are an expert annotator for the HOPE-EXP task at IberLEF 2026.
Analyze social media posts (English or Spanish) and return structured JSON.

## TASK A — primary_label (exactly ONE):
"Realistic Hope"   — desire for a PLAUSIBLE, achievable future outcome
"Unrealistic Hope" — desire for an IMPOSSIBLE or fantastical outcome
"General Hope"     — VAGUE optimism, no concrete defined outcome
"Hopelessness"     — absence of hope; pessimism or resignation
"Sarcastic Hope"   — IRONIC expression implying disbelief/criticism
"Not Hope"         — no forward-looking intent; neutral/factual/descriptive

## TASK B — trigger_emotions (ONE OR MORE):
sadness, joy, love, anger, fear, surprise, Nuetral/unclear

## TASK C — span_annotations:
Only for Hope categories. Up to 3 EXACT verbatim substrings from the post.
For "Not Hope" or "Hopelessness": []

## TASK D — per span:
outcome_stance: "Desired" or "Avoided"
actor: "Self" | "Other" | "World/System" | "Unclear"

Return ONLY valid JSON, no markdown, no explanation:
{"primary_label":"...","trigger_emotions":[...],"span_annotations":[{"span":"...","outcome_stance
    ":"...","actor":"..."}]}
```

,

Listing 2: Multilingual few-shot demonstrations (verbatim).

```
EXAMPLE 1 (EN, Sarcastic Hope):
Title: The Silicon Valley Dream That Keeps Crashing
Text: ...I'm sure everything will work out fine in the end, because that's definitely how tech
    careers work out for everyone.
{"primary_label":"Sarcastic Hope","trigger_emotions":["sadness","anger"],"span_annotations":[{"span
    ":"I'm sure everything will work out fine in the end, because that's definitely how tech careers
     work out for everyone.","outcome_stance":"Desired","actor":"Self"}]}

EXAMPLE 2 (EN, General Hope):
Title: Parenting while drowning
Text: Some days I just hope things will somehow improve without me having to figure out how.
{"primary_label":"General Hope","trigger_emotions":["sadness"],"span_annotations":[{"span":"I just
    hope things will somehow improve without me having to figure out how","outcome_stance":"Desired
    ","actor":"World/System"}]}

EXAMPLE 3 (ES, Not Hope):
Title: Cómo es el plan de estudios de la Maestría?
Text: Not Hope
{"primary_label":"Not Hope","trigger_emotions":["Nuetral/unclear"],"span_annotations":[]}

EXAMPLE 4 (EN, Hopelessness):
Title: Nothing ever changes
Text: I've given up expecting things to get better. There's no point.
{"primary_label":"Hopelessness","trigger_emotions":["sadness","fear"],"span_annotations":[]}

EXAMPLE 5 (ES, Realistic Hope):
Title: Quiero mejorar mi español
Text: Espero poder mantener una conversación fluida antes de que acabe el año.
{"primary_label":"Realistic Hope","trigger_emotions":["joy"],"span_annotations":[{"span":"Espero
    poder mantener una conversación fluida antes de que acabe el año","outcome_stance":"Desired","
    actor":"Self"}]}
```

B. Online Resources

The results of the HOPE-EXP 2026 are available via:

- HOPE-EXP,
- HOPE-EXP 2026 Results
- HOPE-EXP 2026 Code