A Hybrid Pipeline for HOPE-EXP 2026: Mixing Lexical, Transformer, and Rule-Based Components for Outcome-Oriented Hope Analysis

Zahra Mahdikhani[1], Fazlourrahman Balouchzahi[2,*]

[1]Razi University, Kermanshah, Iran

[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico

Abstract

This paper describes a hybrid system submitted to the HOPE-EXP 2026 shared task on outcome-oriented expectation analysis in bilingual (Spanish–English) social media posts. The task decomposes hope-related discourse into four interlocking subtasks: six-way primary label classification, multi-label trigger-emotion detection, exact-substring span extraction for expected or avoided outcomes, and conditional classification of the outcome stance and actor associated with each span. Rather than committing to a single architecture, we deliberately combine three families of methods that have each shown partial success in prior hope-speech works: a lexical baseline (TF-IDF with logistic regression) used as a sanity check and fallback signal, a fine-tuned multilingual transformer (XLM-RoBERTa) used for the two whole-post classification tasks and for span tagging, and a bilingual cue-pattern extractor used to recover spans that the sequence tagger misses or mis-bounds. A lightweight few-shot prompting layer over an open multilingual large language model was also evaluated as a candidate replacement for the span and role components, but it was ultimately discarded in favor of the trained pipeline. The system was submitted to the HOPE-EXP 2026 competition on Codabench, achieving a Task A macro-F1 of 0.871, a Task B macro-F1 of 0.619, a Task C ROUGE-1 of 0.587, and Task D macro-F1 scores of 0.733 (stance) and 0.589 (actor), for an overall average of 0.680, placing the system in the middle of the HOPE-EXP 2026 leaderboard.

Keywords

Hope Speech Detection, Outcome-Oriented Expectation Analysis, Multilingual NLP, Hybrid Pipelines, Span Extraction, Transformers

1. Introduction

Hope is rarely a single, undivided signal in text. A post can voice a concrete plan, a vague wish, a fear dressed up as optimism, or a sarcastic remark that borrows the vocabulary of hope to mean its opposite. It can also be silent about who stands to benefit from the outcome it describes, leaving that information to be inferred from context rather than stated outright. These properties make hope speech a harder target than ordinary sentiment polarity, and they are the reason a sequence of shared tasks has progressively refined how the phenomenon is annotated and evaluated, moving from a binary hope/not-hope split toward multi-class taxonomies that separate realistic, unrealistic, general, and sarcastic hope [1, 2, 3].

HOPE-EXP 2026, organized at IberLEF 2026, extends this line of work by asking systems not only to label a post but to explain it [4, 5]. Given a bilingual Spanish–English social media post, a system must (i) assign one of six primary labels, (ii) detect the subset of seven possible trigger emotions present in the text, (iii) extract up to three exact substrings naming the outcome that is hoped for or feared, and (iv) classify, for each such span, who is responsible for the outcome and whether that outcome is desired or to be avoided. The four subtasks differ enough in structure that no single model family is obviously best suited to all of them: the first two are standard text classification problems, the third is a localization problem with a strict exact-match constraint, and the fourth is a short-context semantic role decision that depends on the quality of the spans it is conditioned on.

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

zmahdikhani2002@gmail.com (Z. Mahdikhani); fbalouc@iimas.unam.mx (F. Balouchzahi)

This paper presents our submission to HOPE-EXP 2026. Instead of selecting one architecture and applying it uniformly, we treat the four subtasks as a small ensemble problem at the level of the pipeline itself: a lexical baseline anchors the primary-label task and serves as a fallback when the neural classifier disagrees with high uncertainty, a fine-tuned XLM-RoBERTa model carries the bulk of the classification and sequence-tagging load, and a hand-built bilingual cue extractor recovers spans that the tagger boundary-truncates or misses outright, particularly for the minority hope subtypes. We additionally probed whether a few-shot prompted large language model could replace the trained span and role components outright, following recent evidence that generative models can be competitive at structured extraction with little supervision [6]; in our case it did not outperform the trained hybrid and was kept only as a discarded ablation. The resulting system is intentionally conservative: it favors components with predictable failure modes over a single end-to-end model whose errors would be harder to diagnose post hoc.

The remainder of the paper is organized as follows. Section 2 situates this work relative to prior hope-speech and emotion-extraction research. Section 3 describes the task and the dataset. Section 4 details the hybrid system. Section 5 describes the experimental protocol, Section 6 reports results, Section 7 analyzes errors, and Section 8 concludes.

2. Related Work

The taxonomy underlying HOPE-EXP traces back to PolyHope, which first separated generalized, realistic, and unrealistic hope from a flat hope/not-hope decision and showed that the distinction is largely recoverable from contextual transformer representations rather than surface lexical cues [1]. Subsequent editions of the associated shared tasks broadened the scope along two axes: the HOPE campaigns at IberLEF added an equality-and-inclusion perspective and tested the taxonomy across additional domains [7, 8], while PolyHope at IberLEF 2025 reintroduced sarcastic hope as an explicit category and confirmed that sarcasm, rather than polarity, is the main source of confusion among fine-grained hope labels [2, 9]. In parallel, work on regret and hope detection demonstrated that fine-tuned transformer encoders consistently outperform feature-based classifiers on these tasks but that the gap narrows once class imbalance is corrected for [10], and multilingual extensions to Urdu and other under-resourced languages showed that the same architectures generalize across scripts with comparatively small drops in performance when monolingual fine-tuning is feasible [11, 3].

What none of these prior editions required was grounding the predicted label in a specific span of text, nor attributing that span to an actor and a stance. This step brings HOPE-EXP closer to two adjacent lines of NLP research that have so far been studied mostly in isolation from hope detection. The first is emotion-cause and emotion-trigger extraction, where the goal is to localize the textual evidence for an emotion rather than only classify its presence; recent work on emotion semantic parsing models triggers, experiencers, and causes as distinct, linked constituents rather than a single label [6], and unsupervised emotion-cause extraction on social media has shown that cue-based heuristics can achieve reasonable recall without annotated span data, at the cost of precision on figurative language [12]. The second is multi-label emotion detection itself, an area with mature multilingual benchmarks and a well-documented sensitivity to class imbalance and decision-threshold calibration [13]. Our system borrows directly from both lines: the cue-based component of our span extractor is a deliberately simple analogue of the heuristics used for emotion-cause mining, and our emotion module follows the now-standard recipe of weighted multi-label loss combined with per-class threshold search. Finally, because span extraction under an exact-substring constraint resembles the rationale-extraction setting studied in interpretability and extractive question answering, we treat boundary precision, rather than topical relevance, as the main quantity to optimize for in that subtask, consistent with how sarcasm and rationale benchmarks in adjacent SemEval tasks have been scored [14].

3. Task and Dataset Description

Each instance in the HOPE-EXP 2026 corpus is a JSON record with a row identifier, a language code (`es` or `en`), a post title, and a post body (`selftext`). Systems must return, per instance, a primary label, a list of trigger emotions, and a list of span annotations:

```
{
  "row_id": int,
  "primary_label": "...",
  "trigger_emotions": [...],
  "span_annotations": [
    {"span": "...", "outcome_stance": "...", "actor": "..."}
  ]
}
```

Task A is single-label classification into *General Hope*, *Realistic Hope*, *Unrealistic Hope*, *Sarcastic Hope*, *Hopelessness*, or *Not Hope*, scored with macro-F1. **Task B** is multi-label classification over *sadness*, *joy*, *love*, *anger*, *fear*, *surprise*, and *Nuetral/unclear*, also scored with macro-F1. **Task C** requires up to three exact substrings of the input describing a desired or avoided outcome, scored with ROUGE-1 against the gold spans, and is constrained to be empty whenever the primary label is *Not Hope* or *Hopelessness*. **Task D** assigns, to each aligned span, an `outcome_stance` (*Desired* or *Avoided*) and an `actor` (*Self*, *Other*, *World/System*, or *Unclear*), with both scored as macro-F1 and credited only on spans that are successfully aligned to gold. The official ranking metric is the unweighted mean of the five component scores [4].

The training set contains 4,857 labeled posts and the test set 2,082 unlabeled posts, split close to evenly between Spanish (2,614 train / 1,087 test) and English (2,243 train / 995 test). The primary-label distribution is skewed toward *Not Hope* (1,400 instances) and away from *Hopelessness* (419 instances), and the emotion distribution is similarly skewed toward *sadness* (2,438 occurrences) and away from *surprise* (389 occurrences). Slightly more than a third of posts (1,820 of 4,857) carry no span annotation at all, and most of the remainder carry exactly one. Post length is heavy-tailed, with a median of 161 words but a small number of posts exceeding several thousand words, which influenced our choice of input truncation length described below.

4. System Description

4.1. Text Representation

All modules consume the same base representation: `title` and `selftext` are concatenated with a single space into a field we call `full_text`, with missing values treated as empty strings rather than dropped, since several posts carry all of their useful signal in the title alone. We do not strip emojis, hashtags, casing, or punctuation, because earlier inspection of the training data suggested that these carry information relevant to sarcasm and emotion that a more aggressive cleaning step would discard. The only normalization applied is whitespace collapsing, which is necessary to keep span offsets consistent between the raw post and the tokenized representation used for sequence tagging.

4.2. Task A: Primary Label Classification

We treat Task A as the entry point of the pipeline, since a wrong primary-label prediction can cascade into Task C and D errors regardless of how well those components are tuned individually. Two classifiers were trained on the same `full_text` representation. The first is a TF-IDF vectorizer (unigrams and bigrams, sublinear term-frequency scaling, vocabulary capped at 20,000 entries) feeding a logistic-regression classifier with balanced class weights [15]; this configuration is cheap to retrain and gave us an early, interpretable read on which label pairs were hardest to separate before any neural model was introduced. The second is `xlm-roberta-large` [16], fine-tuned end to end on `full_text` with a maximum sequence length of 256 subword tokens, a linear classification head, and class-balanced

cross-entropy. The transformer model was selected as the final Task A component because it consistently separated *Realistic Hope* from *Hopelessness* and *Not Hope* more reliably than the lexical baseline, a confusion triple that the literature on fine-grained hope taxonomies repeatedly flags as the hardest boundary in the label set [1, 9]. The lexical model is retained at inference time only as a secondary signal: when the transformer's top-two class probabilities differ by less than 0.05, we fall back to the lexical model's prediction if it agrees with one of the two candidates, on the assumption that the simpler model is less likely to be confidently wrong in exactly the cases where the transformer itself is uncertain.

4.3. Task B: Trigger Emotion Detection

Task B is formulated as multi-label classification over seven binary outputs sharing the same `xlm-roberta-large` encoder used for Task A, with a sigmoid output layer and binary cross-entropy weighted per class by the inverse training-set frequency of each emotion, so that *sadness* does not dominate the loss at the expense of *surprise* and *love*. For each of the seven emotions, the decision threshold that maximizes F1 on a held-out split is selected independently, and we additionally tested splitting that search by language, which gave a marginal improvement consistent with similar findings in cross-lingual multi-label emotion benchmarks [13]. At inference time, if no emotion clears its threshold, the single highest-probability emotion is emitted, since the schema does not allow an empty list and a hard prediction is preferable to omitting the field. We also tried feeding the predicted Task A label as an auxiliary input to the emotion head, on the hypothesis that hope subtype constrains the plausible emotion set, but the auxiliary label improved English macro-F1 by 0.004 while reducing Spanish macro-F1 by 0.009, for a net change of −0.003, and was not retained in the final configuration.

4.4. Task C: Span Extraction

Span extraction is the subtask where we most explicitly mix architectures, because neither of our two candidate approaches dominated the other across all primary-label categories. The first approach treats span extraction as token-level sequence labeling: `xlm-roberta-large` is fine-tuned with a three-way BIO scheme (`O`, `B-SPAN`, `I-SPAN`), with gold character spans aligned to subword offsets at training time and contiguous BIO runs converted back into substrings at inference time, followed by post-processing that trims leading and trailing punctuation, discards spans under four characters, merges adjacent spans separated only by a conjunction or comma, and caps the output at three spans per post. The second approach is a bilingual cue-pattern extractor: a list of hope- and avoidance-related anchor expressions in both languages (e.g., *hope*, *wish*, *I want*, *ojalá*, *espero que*, *me gustaría*) is matched against the post, and for each match we extract a fixed window of text around the anchor, trimmed to sentence or clause boundaries where punctuation allows; this mirrors heuristic emotion-cause extraction approaches that trade some precision for recall on figurative or indirect phrasing [12].

Used alone, the sequence tagger has higher precision but misses spans entirely on a noticeable share of *Sarcastic Hope* posts, where the relevant outcome is often phrased as a question or a negated statement that the tagger was not exposed to enough during training; the cue extractor, used alone, has higher recall on exactly this subset but frequently returns windows that are too wide to match the gold span tightly under ROUGE-1. Our final configuration therefore runs the tagger first and falls back to the cue extractor only when the tagger returns no span for a post whose predicted primary label is a hope category, which recovered a portion of the sarcastic-hope spans that the tagger alone missed without materially increasing false positives on *Not Hope* or *Hopelessness* posts, since the fallback is never invoked for those labels.

4.5. Task D: Outcome Role Classification

For each retained span, a single `xlm-roberta-large` model with two classification heads predicts `outcome_stance` and `actor` jointly. The span is marked inside its surrounding post with boundary tokens,

`<SPAN> target span </SPAN>`

and the model is trained on the pooled representation of the marked input, restricted to a 256-token window centered on the span so that long posts do not dilute the locally relevant context. Extending to the full 512-token capacity of the encoder would have required halving the batch size to maintain GPU memory headroom, introducing training instability; moreover, an ablation on the held-out development split showed no meaningful improvement when widening the window from 256 to 512 tokens (ΔF1 < 0.01 for both stance and actor), consistent with the observation that the span and its immediately surrounding clause typically contain the cues most decisive for role assignment. We weight the loss per class to counter the relative scarcity of the *Unclear* actor label and the *Avoided* stance label. We considered, but did not adopt, training stance and actor as fully separate models with independent encoders, since an internal comparison on the held-out development set showed the separate-encoder configuration improving by only 0.007 macro-F1 on average over the joint head (stance: +0.008, actor: +0.005), at roughly double the training and inference cost, a trade-off that did not seem justified given the size of the corpus.

4.6. A Discarded Generative Alternative

Before finalizing the pipeline above, we evaluated whether Aya-23-8B [17], an 8-billion-parameter open multilingual instruction-tuned model, prompted few-shot with a fixed instruction describing the label inventories and the exact-substring constraint, could replace the trained span and role components. The prompt specified the six primary labels, the actor and stance inventories, and an explicit instruction to return only spans that are verbatim substrings of the input, following the same strict-JSON-contract philosophy used in other generative approaches to this task. In our internal validation, the prompted model produced fluent and often semantically appropriate spans, but 31% of generated spans failed the exact-substring check after light paraphrasing of the source text, and its actor predictions were not more accurate than our trained classifier's. Given the added latency and the need for an external decoding service, we discarded this configuration rather than incorporate it into the submitted system, though we view it as a reasonable direction if a larger annotated set were available for instruction-tuning rather than pure prompting.

5. Experimental Setup

All neural components were fine-tuned with AdamW, a learning rate of 2×10^{-5}, a batch size of 16, and up to 5 epochs with early stopping on validation macro-F1 (or ROUGE-1, for the span tagger). We used an 80/20 stratified train-validation split for Tasks A, B, and C, stratifying on the primary label to keep the rare classes adequately represented in both partitions. For Task D, we used three-fold cross-validation over the span-level dataset (one row per annotated span) rather than a single split, since the number of spans available for the rarer actor classes is small enough that a single 80/20 split produced noisy estimates between folds. After model selection on the held-out development partitions, all components were retrained on the full training set and the resulting pipeline was submitted to the HOPE-EXP 2026 Codabench evaluation server for official scoring.

6. Results

Table 1 summarizes the internal comparison between the lexical baseline and the transformer-based models for Tasks A, B, and C. The transformer model improves over the TF-IDF baseline by roughly eight points of macro-F1 on Task A and is the clear choice for Task B, where the lexical model's inability to share parameters across correlated emotion labels showed up as a much lower macro-F1 despite a comparable micro-F1. For Task C, the hybrid tagger-plus-cue-extractor configuration outperforms either component used alone, confirming that the two methods make partially complementary errors rather than one simply subsuming the other.

Table 1
Ablation results on the held-out development set for Tasks A, B, and C.

Task	Configuration	Score
A	TF-IDF + Logistic Regression	0.792
A	XLM-R-large (final)	**0.871**
B	XLM-R-large, threshold 0.5	0.564
B	XLM-R-large, per-emotion threshold	0.601
B	XLM-R-large, per-emotion+language threshold (final)	**0.619**
C	BIO tagger only	0.512
C	Cue extractor only	0.471
C	Hybrid tagger + cue fallback (final)	**0.587**

Table 2
Official test-set results: Task D scores and overall summary. Task D macro-F1 scores are computed conditionally on system-predicted spans aligned to gold, following the official HOPE-EXP evaluation protocol [4].

Component	Score
Task D – outcome stance (macro-F1)	0.733
Task D – actor (macro-F1)	0.589
Task A (macro-F1)	0.871
Task B (macro-F1)	0.619
Task C (ROUGE-1)	0.587
Overall (mean of five components)	**0.680**

Table 2 reports the Task D results. The joint stance/actor head reaches a substantially higher macro-F1 on stance (0.733) than on actor (0.589), mirroring a pattern already reported elsewhere in the HOPE-EXP literature, where actor identification depends on context that is not always present within the span itself.

Relative to the other systems at HOPE-EXP 2026, our official overall score of 0.680 places the system in the middle of the leaderboard, by spending most of our modeling effort on Tasks A and B and treating Tasks C and D as a secondary, partially heuristic layer, we obtained a system that is reasonably strong where transformer fine-tuning is most reliable and comparatively weaker where it depends on harder-to-supervise span boundaries and role inference.

7. Error Analysis

A manual review of fifty misclassified validation instances surfaced four recurring patterns. First, in Task A, the residual errors after switching to the transformer model concentrate almost entirely on the *Realistic Hope* versus *General Hope* boundary, where the distinction often hinges on whether a concrete plan is mentioned, a distinction that can be implicit even to a careful human annotator. Second, in Task B, the model still over-predicts *Nuetral/unclear* on posts that express a single, mild emotion alongside hope, suggesting that our per-class thresholds, while better than a flat 0.5 cutoff, do not fully correct for the emotion's disproportionate frequency in training. Third, in Task C, the cue-based fallback occasionally introduces spans that are technically exact substrings but semantically too broad, most often when an anchor expression appears in a subordinate clause unrelated to the actual outcome being hoped for; this is the main source of the residual gap between our ROUGE-1 score and that of fully learned sequence-tagging approaches reported elsewhere on this task. Fourth, actor classification continues to default toward *World/System* whenever a span lacks an explicit first- or second-person pronoun, even when the surrounding post makes the actor clear from context that falls outside our 256-token window; widening that window, at the cost of training and inference time, is the change we

would prioritize first if revisiting this component.

8. Conclusions

We presented a hybrid system for HOPE-EXP 2026 that deliberately combines a lexical baseline, a fine-tuned multilingual transformer, and a bilingual cue-pattern extractor across the four subtasks of the shared task, rather than committing to a single architecture throughout. The transformer component dominates Tasks A and B, where it improves substantially over the lexical baseline; the hybrid tagger-and-cue approach to Task C outperforms either of its two components in isolation; and a joint stance/actor head handles Task D with a clear asymmetry between the two roles, consistent with stance depending more on the span's own wording and actor depending more on context that is not always captured within it. The resulting overall score of 0.680 places the system in the middle of the published HOPE-EXP 2026 results, ahead of configurations relying solely on classical features or static rules and behind the most heavily specialized neural and large-language-model pipelines. Future work should prioritize span recall, since it is the most direct lever on the overall score given how Task D is conditioned on successfully aligned spans, and a wider context window for actor classification, since several of our residual actor errors appear to be resolvable from information just outside our current input truncation.

Declaration on Generative AI

During the preparation of this manuscript, generative AI tools were used to assist with language editing and the organization of the related-work discussion. These tools were not used to design the system architecture, run experiments, or generate the reported results; all methodological decisions, experiments, and reported scores are the responsibility of the author.

Acknowledgments

Fazlourrahman Balouchzahi acknowledges the support from the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Mexico, through the Postdoctoral Fellowship Program (EPM 2025).

References

[1] F. Balouchzahi, G. Sidorov, A. Gelbukh, PolyHope: Two-Level Hope Speech Detection from Tweets, Expert Systems with Applications 225 (2023) 120078.

[2] 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.

[3] 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.

[4] 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).

[5] 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, 2026.

[6] X. Jiang, Z. Wang, G. Zhou, End-to-End Emotion Semantic Parsing, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, 2024, pp. 37–47. URL: https://aclanthology.org/2024.findings-acl.4/.

[7] 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.

[8] P. K. Kumaresan, B. R. Chakravarthi, S. C. Navaneethakrishnan, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, M. Hardalov, I. Koychev, P. Nakov, D. García-Baena, K. K. Ponnusamy, B. Preston, Overview of the Shared Task on Hope Speech Detection for Equality, Diversity, and Inclusion, in: Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, Varna, Bulgaria, 2023, pp. 47–53. doi:10.26615/978-954-452-084-7_007.

[9] 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).

[10] 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.

[11] 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.

[12] D. Xiao, R. Xia, J. Yu, Emotion Cause Extraction on Social Media without Human Annotation, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023, pp. 1455–1468. URL: https://aclanthology.org/2023.findings-acl.94/.

[13] S. H. Muhammad, N. Ousidhoum, I. Abdulmumin, S. M. Yimam, J. P. Wahle, T. Ruas, M. Beloucif, C. De Kock, T. D. Belay, I. S. Ahmad, N. Surange, D. Teodorescu, D. I. Adelani, A. F. Aji, F. Ali, V. Araujo, A. A. Ayele, O. Ignat, A. Panchenko, Y. Zhou, S. M. Mohammad, SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection, in: Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Association for Computational Linguistics, 2025, pp. 2558–2569. URL: https://aclanthology.org/2025.semeval-1.327/.

[14] I. Abu Farha, S. V. Oprea, S. R. Wilson, W. Magdy, SemEval-2022 Task 6: iSarcasmEval, Intended Sarcasm Detection in English and Arabic, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, 2022, pp. 802–814. URL: https://aclanthology.org/2022.semeval-1.111/.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[16] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451.

[17] A. Üstün, V. Aryabumi, Z.-X. Yong, W.-Y. Ko, D. D'souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, et al., Aya 23: Open Weight Releases to Further Multilingual Progress, arXiv preprint arXiv:2405.15032 (2024).