Metric-Aware Calibration and Generate–Critic Reranking for Spanish Humor Analysis and Generation

Michael Ibrahim

Computer Engineering Department, Cairo University, 1 Gamaa Street, 12613, Giza, Egypt

Abstract

Spanish computational humor requires models that can separate satirical news from literal news, identify machine-authored jokes that may imitate human style, and generate short jokes that remain coherent with headline context. A metric-aware framework was developed for the HAHA 2026 shared task at IberLEF. In the two classification tasks, target-class F1 was treated as the central optimization criterion rather than as a post-hoc score. Sparse word and character TF–IDF representations with stylometric cues were combined with Spanish and multilingual Transformer encoders through out-of-fold convex ensembling. Final binary labels were assigned by validation-selected threshold and rank-based decision rules. In the generation task, a Qwen3-32B-AWQ model was used to sample multiple concise Spanish joke candidates for each headline, and a critic prompt selected the final candidate according to topicality, brevity, naturalness, and punchline quality. The developed framework achieved F1 scores of 0.9933 for satire detection and 0.8764 for machine-generated joke detection. The generated jokes obtained rank 1 with Elo rating 1110 and 95% confidence interval [1083, 1138]. These results indicate that short Spanish humor classification benefits from separating probability estimation from operating-point calibration, while generate–critic reranking provides a practical strategy for preference-oriented humor generation.

Keywords

computational humor, Spanish NLP, satire detection, ensemble calibration, humor generation

1. Introduction

Humor is a difficult object for computational modeling because it is not only a property of a text. It is also shaped by pragmatic inference, cultural background, audience expectations, genre conventions, and the relation between a setup and an implied or explicit punchline. Linguistic theories have often described verbal humor through incongruity, script opposition, and mechanisms that allow an apparent conflict to be resolved as a joke [1, 2]. These theories help explain why humor detection cannot be reduced to sentiment analysis or topic classification: the same vocabulary may be used in a serious headline, a satirical headline, a human joke, or a machine-generated joke, but the humorous interpretation depends on how the information is framed and how expectations are violated.

Computational humor has therefore been studied through several complementary tasks. Early work investigated automatic humor recognition using lexical, stylistic, and semantic cues [3]. Later evaluation campaigns introduced ranking and preference settings, including humorous hashtag completion [4], edited headline funniness [5], and joint humor/offense assessment [6]. Spanish humor analysis has been supported by the HAHA datasets and evaluations, where humor detection, rating, and analysis were treated as supervised NLP problems over social-media text [7, 8]. HAHA 2026 extends this line of work by joining Spanish humor classification with the assessment and generation of automatically produced humorous content [9, 10].

The 2026 setting is methodologically different from earlier humor-detection-only benchmarks for two reasons. First, machine-generated humor introduces an adversarially blurred boundary between human and automatic writing. A low-quality generator can be identified through repetition, over-explicitness, unnatural syntax, or generic punchlines, but stronger language models can produce concise and plausible jokes. Detection must therefore rely on weak distributional signals that can be unstable

IberLEF 2026, September 2026, León, Spain

michael.nawar@eng.cu.edu.eg (M. Ibrahim)

www.linkedin.com/in/michael-ibrahim-90 (M. Ibrahim)

0000-0003-2340-8917 (M. Ibrahim)

across generator families. Second, open-ended humor generation is evaluated through preference rather than exact matching. A generated joke can be grammatical and relevant while still being unfunny, and a joke can be funny to one evaluator but ineffective to another. The generation component must therefore be treated as a search and selection problem rather than as a deterministic text-to-text transformation. From a computational-creativity perspective, this makes the generation task a constrained creative search problem: the system must explore several plausible humorous continuations and select one under subjective criteria such as surprise, topical coherence, and naturalness.

The developed system was designed around these constraints. For the two classification tasks, probability estimation was separated from final decision selection. Complementary sparse and neural classifiers were trained under cross-validation, their out-of-fold predictions were blended, and the final operating point was selected to maximize positive-class F1. This design was used because F1 is sensitive not only to ranking quality but also to the number of examples predicted as positive. For the generation task, a large instruction-tuned language model was used to produce several candidate jokes for each headline, after which a critic step selected a single final output. This candidate generation and reranking strategy was intended to reduce the variance of single-sample decoding and to impose task-specific quality constraints.

1.1. Task Definition

HAHA 2026 contained three Spanish-language tasks. In Task 1, each instance corresponded to a news item represented by fields including country, date, headline, and context. The objective was binary classification between `satirical` and `real`, and the official metric was the F1 score of the `satirical` class. This formulation made satire the target class and required a classifier to distinguish deliberately humorous news from literal news in a domain where both classes may use current events, named entities, and journalistic style.

In Task 2, each instance contained a news headline and an associated joke. The objective was binary classification between `machine` and `human`, and the official metric was the F1 score of the `machine` class. This task was related to machine-generated text detection, but the setting was more constrained than document-level generation detection because the examples were short, humorous, and conditioned on the same kind of headline input used in the generation task.

In Task 3, a headline was provided and a short Spanish joke had to be produced. The evaluation was based on human preference judgments between system outputs, and systems were ranked with an Elo-style score. This task required the generated text to be concise, topical, and humorous without relying on a single reference answer. Consequently, the developed system used an explicit candidate-generation and candidate-selection stage.

The remainder of the paper is organized as follows. Section 2 reviews computational humor, generated-text detection, and representation learning for Spanish short texts. Section 3 presents the developed system. Section 4 reports and discusses the official results. Section 5 concludes and outlines future work.

2. Related Work

Research relevant to the developed system lies at the intersection of computational humor, generated-text detection, Spanish language modeling, sparse text classification, and preference-oriented generation. This section summarizes these areas and highlights how the HAHA 2026 setting combines them in a single evaluation campaign.

2.1. Computational Humor and Humor Benchmarks

Theoretical accounts of verbal humor have emphasized that jokes often rely on the compatibility of a text with two opposing scripts, together with a logical mechanism that enables reinterpretation [1, 2]. Although such theories were not encoded directly as symbolic rules in the developed system, they

motivate the use of contextual models and stylistic features. A headline or joke may appear literal until a particular word, phrase, or framing choice changes its interpretation; therefore, both local surface patterns and broader contextual representations can be informative.

Automatic humor recognition was investigated by Mihalcea and Strapparava [3], who treated humorous and non-humorous texts as a supervised classification problem. Subsequent evaluation campaigns broadened the field from binary humor recognition to ranking, regression, and subjective preference. SemEval-2017 Task 6 framed humor as the ranking of hashtag responses in the #HashtagWars setting [4]. SemEval-2020 Task 7 focused on edited news headlines, requiring systems to estimate funniness and compare alternative humorous edits [5]. SemEval-2021 Task 7 combined humor detection, humor rating, and offense-related labels, showing that humor analysis often interacts with social acceptability and subjective interpretation [6].

For Spanish, Castro et al. [7] introduced a crowd-annotated corpus for humor analysis, and later HAHA evaluations provided structured benchmarks for detecting, rating, and analyzing humor in Spanish [8]. HAHA 2026 continues this benchmark family while changing the data domain and adding automatic humor generation and generated-humor detection [9]. The shift from social-media humor to headline-conditioned humor is important because a headline provides a topical anchor. Humor may depend on how the joke reframes the event, whether the satire is plausible as journalism, and whether generated text sounds formulaic.

2.2. Machine-Generated Text Detection

The second HAHA 2026 task is connected to machine-generated text detection. The Giant Language Model Test Room (GLTR) showed that distributional properties of language-model outputs can expose generation artifacts and can support human inspection [11]. TuringBench formalized generated-text detection as a benchmark over outputs from multiple neural generators [12]. DetectGPT introduced a zero-shot criterion based on probability curvature under perturbations [13]. These studies demonstrate that generated text can be detected through statistical regularities, but they also show that detection is sensitive to the generator, decoding strategy, and domain.

HAHA 2026 differs from many generated-text detection settings because the target texts are jokes. Humor rewards surprise, compression, and sometimes unnatural associations, which can resemble generation artifacts. Conversely, modern generators may produce fluent short jokes that do not contain obvious lexical markers of automatic authorship. For this reason, the developed system treated the problem as supervised headline-conditioned short-text classification rather than as general-purpose likelihood analysis. Sparse features were retained because generator artifacts can be expressed through punctuation, character patterns, repeated formulas, and length; Transformer encoders were added because topical relation and pragmatic naturalness require contextual modeling.

2.3. Sparse and Neural Representations for Spanish Short Texts

Sparse term weighting remains useful for short text classification because strong evidence can be concentrated in a small number of words or character patterns. TF–IDF weighting has long been used in text retrieval and classification [14]. Complement Naive Bayes was designed to address some of the assumptions that make standard Naive Bayes weak on imbalanced text classification [15]. In the developed system, sparse word and character TF–IDF models were combined with simple stylometric cues so that lexical, orthographic, and formatting information could be represented explicitly.

Transformer models provide the complementary ability to build contextual representations from pretrained language models [16]. BERT established bidirectional masked-language-model pretraining for language understanding [17], and RoBERTa showed that robust optimization and larger-scale pretraining can substantially strengthen this family of models [18]. XLM-R extended RoBERTa-style masked language modeling to large-scale multilingual pretraining [19]. For Spanish, MarIA introduced Spanish language models trained on large web corpora [20], while RoBERTuito was pretrained for

Spanish social-media text [21]. These models are relevant to HAHA 2026 because the data combine news style, headline style, informal joke wording, and cross-regional Spanish variation.

Open-ended generation has been transformed by instruction-tuned large language models. Qwen3 introduced a multilingual family of models with dense and mixture-of-experts variants and support for both thinking and non-thinking modes [22]. Local generation with large models is constrained by memory, and activation-aware weight quantization provides a practical method for reducing inference cost while preserving model behavior [23]. The developed system used this direction through Qwen3-32B-AWQ for candidate joke generation. Standard machine-learning and neural-model tooling was provided by scikit-learn and the Hugging Face Transformers library [24, 25].

3. Developed System

The developed system consisted of two discriminative pipelines and one generative pipeline. The discriminative pipelines shared the same structure: conservative normalization, task-specific text construction, sparse lexical and stylometric modeling, Transformer fine-tuning, out-of-fold prediction, convex ensembling, and metric-aware operating-point selection. The generative pipeline used an instruction-tuned model to sample multiple candidate jokes for each headline and a critic prompt to select the final candidate. The same design principle was used across the three tasks: a model score or generated candidate was not treated as final until a task-aligned selection step had been applied. In the classification tasks, logistic regression, Complement Naive Bayes, and fine-tuned Transformer encoders were trained as separate base learners; they were combined only through their positive-class probability estimates. No hard-label voting scheme and no second-level stacking classifier were used.

3.1. Text Construction and Label Encoding

Text normalization was intentionally limited. Non-breaking spaces were replaced with ordinary spaces, repeated whitespace was collapsed, and fields were stripped. Aggressive punctuation removal, case normalization, or accent removal was not applied because these signals may be informative in satire, joke style, and generated-text detection.

For Task 1, the model input concatenated country, date, headline, and context using explicit Spanish field markers:

$$x_i^{(1)} = \texttt{país=}c_i \quad \texttt{fecha=}d_i \quad \texttt{titular: } h_i \quad \texttt{contexto: } r_i. \tag{1}$$

The positive label was assigned to `satirical`. For Task 2, the input concatenated the headline and joke:

$$x_i^{(2)} = \texttt{titular: } h_i \quad \texttt{chiste: } j_i. \tag{2}$$

The positive label was assigned to `machine`. Headline groups were retained for Task 2 so that cross-validation could discourage leakage between jokes derived from the same headline.

3.2. Data and Validation

For the two classification tasks, supervised learning used the union of the official trial partition and the official development-gold partition for the corresponding task. The official test partitions were used only for producing final predictions, and no external labeled humor or generated-text examples were added. For Task 3, no supervised fine-tuning on HAHA labels was performed; each official test headline was used as a prompt for the generate–critic procedure.

Five-fold out-of-fold validation was used for all classification base learners. Task 1 used stratified folds over the target label. Task 2 used headline groups, so jokes associated with the same headline were assigned to the same fold. Consequently, every out-of-fold probability for a labeled example was produced by a model that had not trained on that example, and in Task 2 had not trained on another joke for the same headline. These out-of-fold probabilities were then used to select ensemble weights and final decision rules.

Table 1
Sparse classifiers used in both classification tasks.

Alias	Classifier	Main parameter	Features
LR-3	Logistic regression	$C = 3$	word, char, style
LR-12	Logistic regression	$C = 12$	word, char, style
CNB	Complement NB	$\alpha = 0.08$	word, char, style

Table 2
Transformer fine-tuning configuration. Accum. denotes gradient accumulation steps.

Task	Encoder	Len.	Batch	Accum.	LR	Epochs
1	BNE-L	192	4	2	1.0×10^{-5}	7
1	XLM-R-L	192	4	2	8.0×10^{-6}	7
1	RoBERTuito	192	12	1	1.6×10^{-5}	8
2	RoBERTuito	224	12	1	1.4×10^{-5}	9
2	XLM-R-L	224	4	2	8.0×10^{-6}	8
2	BNE-L	224	4	2	1.0×10^{-5}	8

3.3. Sparse Lexical and Stylometric Models

Three sparse classifiers were trained for each classification task. Word and character TF–IDF representations were combined with a compact stylometric vector. The stylometric features included logarithmic character count, logarithmic word count, character-to-word ratio, uppercase ratio, digit ratio, punctuation ratio, counts of question and exclamation marks, ellipses, quote characters, dash indicators, and binary indicators for a small set of phrases that were judged likely to capture formulaic joke or generated-text style.

The sparse representation for Task 1 used word n-grams from 1 to 3 and character n-grams from 3 to 6. The sparse representation for Task 2 used word n-grams from 1 to 3 and character n-grams from 3 to 7. Logistic regression with balanced class weights was trained with two regularization settings, $C = 3$ and $C = 12$, and Complement Naive Bayes was trained with smoothing parameter $\alpha = 0.08$. Table 1 summarizes the sparse classifiers.

3.4. Transformer Encoders

Three pretrained encoders were fine-tuned for the classification tasks. RoBERTa-large-BNE was used as a Spanish news and web encoder, XLM-RoBERTa-large was used for multilingual robustness, and RoBERTuito was used for Spanish social-media and informal text. The task inputs described above were tokenized with truncation and dynamic padding. Training used a cosine learning-rate schedule, warmup ratio 0.08, weight decay 0.03, label smoothing 0.04, gradient checkpointing, mixed precision when available, best-model selection by validation F1, and early stopping with patience of two evaluation epochs. Five folds were used in the final configuration.

3.5. Out-of-Fold Ensembling and Operating-Point Selection

For each task, let $D = \{(x_i, y_i, g_i)\}_{i=1}^{n}$ denote the labeled data, where x_i is the constructed input text, $y_i \in \{0, 1\}$ is the target label, and g_i is an optional group. Each base model m produced an out-of-fold probability estimate $p_i^{(m)} = P_m(y_i = 1 \mid x_i)$. A convex ensemble score was then defined as

$$s_i = \sum_{m=1}^{M} w_m p_i^{(m)}, \qquad w_m \geq 0, \qquad \sum_{m=1}^{M} w_m = 1. \tag{3}$$

The weights were selected using out-of-fold predictions. Candidate weight vectors included every single-model solution, the uniform average, and 4000 Dirichlet-sampled convex mixtures with concentration

parameter 0.7. Therefore, the final classifier for each task could be an individual base learner, a uniform mean, or a validation-selected convex mixture, depending on which option maximized out-of-fold positive-class F1.

A candidate score vector was converted into binary labels by a decision rule $\mathcal{R}(s;\theta)$, and the final pair of ensemble weights and rule parameters was selected by

$$(w^\star, \theta^\star) = \arg\max_{w,\theta} F_{1,\mathrm{pos}}\left(y, \mathcal{R}(s(w);\theta)\right), \tag{4}$$

where $F_{1,\mathrm{pos}}$ denotes F1 for the official positive class. Three decision-rule families were considered. The first was a probability threshold searched from 0.02 to 0.98. The second was a global top-fraction rule, where the highest-scoring ρn examples were labeled positive for $\rho \in \{0.46, 0.48, 0.50, 0.52, 0.54\}$. The third was a group-fraction rule, where a fraction of examples was selected within each group and then adjusted to a global target count. Country groups were available for Task 1, and headline groups were used for Task 2.

This calibration step was included because a high-quality probability ranking does not automatically maximize F1. When the evaluation metric is target-class F1, threshold choice and predicted positive count can change the official score substantially. The rule selection was therefore treated as a model-selection component rather than as a minor post-processing step.

3.6. Generate–Critic Humor Generation

For Task 3, each headline was processed by a generate–critic pipeline. The same local instruction-tuned model, `Qwen/Qwen3-32B-AWQ`, was used to generate candidates and to act as the critic/reranker. The critic was not a separately trained HAHA classifier; it was a low-temperature prompt to the same model. The generation prompt instructed the model to write Spanish jokes with a concise punchline, direct topical relation to the headline, and a natural tone. Explanations, hashtags, emojis, list formatting, discriminatory insults, attacks on protected groups, and generic moralizing were discouraged. The prompt also discouraged the common "*Por qué...? Porque...*" template unless the wordplay was unusually strong.

Five candidates were requested for each headline. Candidate sampling used temperature 0.92, top-$p = 0.85$, top-$k = 40$, repetition penalty 1.07, and a maximum length sufficient for one-line jokes. The generated text was cleaned by removing numbering and bullet markers, stripping surrounding quotation marks, retaining only one line, shortening overlong outputs, and discarding duplicate candidates. If fewer than two valid candidates were obtained, additional sentence-level splitting of the raw output was attempted; if no valid candidate remained, a deterministic fallback template was used.

When several valid candidates were available, a critic prompt was applied. The critic received the headline and candidate list and was instructed to choose the most suitable joke for human preference evaluation. The criteria were punchline clarity, topicality, naturalness, brevity, and avoidance of obvious artificial templates. Critic decoding used a lower temperature than candidate generation. The selected text was matched back to one of the cleaned candidates whenever possible, so that the final output was a candidate joke rather than a newly generated explanation.

4. Results and Discussion

The official classification results are presented in Table 3. The developed system reached F1 0.9933 on Task 1. Precision, recall, and accuracy were also 0.9933. The same score was obtained by three submitted runs, and the first run identifier is reported in the table. The result indicates that the available evidence was sufficient to distinguish satirical and real news items with very high accuracy in the official test distribution. However, it should not be interpreted as a domain-independent solution to satire detection. Satirical style, news source conventions, political context, and temporal cues may shift substantially outside the evaluated data.

Table 3
Official classification results for the best evaluated runs of the developed system.

Task	Run	Precision	Recall	F1	Accuracy
1	776365	0.9933	0.9933	0.9933	0.9933
2	776619	0.8764	0.8764	0.8764	0.8764

Table 4
Official generation result for Task 3.

Run	Rank	Elo	CI low	CI high
776934	1	1110	1083	1138

The Task 2 result was closer to the other leading entries. The best evaluated run reached F1 0.8764. This smaller separation is consistent with the nature of the problem: short human jokes may contain formulaic wording, while machine-generated jokes may be fluent, brief, and topical. The distinction between `human` and `machine` therefore depends on subtle artifacts rather than on an obvious semantic contrast. Grouped validation by headline was important in this task because jokes attached to the same headline can share topical lexical cues; without grouping, validation may overestimate generalization by rewarding headline-specific memorization.

The official generation result is shown in Table 4. The developed system was assigned rank 1 with Elo rating 1110 and 95% confidence interval $[1083, 1138]$ over 344 votes. Another participant was also listed with rank 1 and a higher point estimate, but the confidence intervals overlapped. The conservative interpretation is therefore that the developed generate–critic pipeline was competitive at the top of the human preference evaluation rather than statistically separated from all other systems.

The results support the use of different selection mechanisms for the discriminative and generative parts of the evaluation. For classification, sparse models captured character patterns, punctuation, lexical markers, and formulaic phrasing, while Transformer encoders supplied contextual information. The final calibration stage aligned the predicted positive count with the F1 objective. For generation, multiple stochastic candidates increased the chance that at least one useful joke was produced, and the critic step provided an additional filter for topicality and naturalness.

Several limitations should be noted. First, class-count calibration can be sensitive to test-set prior shifts. If the proportion of satirical headlines or machine-generated jokes changes, top-fraction rules may require recalibration. Second, machine-generated text detection is inherently generator-dependent; artifacts learned from one set of models or prompts may not transfer to other generators, decoding settings, or future model families. Finally, the generation score reflects aggregate human preference and does not identify which qualities—funniness, topicality, surprise, fluency, or harmlessness—were most responsible for pairwise victories.

5. Conclusion and Future Work

A metric-aware system for Spanish humor analysis and generation was developed for HAHA 2026 at IberLEF. In the two classification tasks, sparse lexical-stylometric models and Transformer encoders were combined through out-of-fold convex ensembling. Final predictions were produced by decision rules selected for positive-class F1, thereby aligning operating-point calibration with the official metric. In the generation task, an instruction-tuned Qwen3-32B-AWQ model produced several headline-conditioned joke candidates, and a critic prompt selected the final output.

The official evaluation produced F1 0.9933 for satire detection, F1 0.8764 for machine-generated joke detection, and a rank-1 generation result with Elo rating 1110. These results suggest that short Spanish humor classification benefits from combining surface-level and contextual representations, especially when probability estimation is separated from final decision calibration. The generation result

further suggests that generate–critic reranking is a useful strategy when humor quality is evaluated by preference rather than by reference matching.

For future work, robustness should be evaluated under shifted class priors, new news sources, different Spanish-speaking regions, and generator families not represented in the development data. For humor generation, more detailed human evaluation is needed so that funniness, topicality, originality, naturalness, offensiveness, and memorability can be analyzed separately rather than collapsed into a single Elo score.

6. Declaration on Generative AI

During the preparation of this work, the author used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Attardo, V. Raskin, Script theory revis (it) ed: Joke similarity and joke representation model (1991).

[2] S. Attardo, C. F. Hempelmann, S. Di Maio, Script oppositions and logical mechanisms: Modeling incongruities and their resolutions., Humor: International Journal of Humor Research 15 (2002).

[3] R. Mihalcea, C. Strapparava, Making computers laugh: Investigations in automatic humor recognition, in: Proceedings of human language technology conference and conference on empirical methods in natural language processing, 2005, pp. 531–538.

[4] P. Potash, A. Romanov, A. Rumshisky, Semeval-2017 task 6:# hashtagwars: Learning a sense of humor, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 49–57.

[5] N. Hossain, J. Krumm, M. Gamon, H. Kautz, Semeval-2020 task 7: Assessing humor in edited news headlines, in: Proceedings of the fourteenth workshop on semantic evaluation, 2020, pp. 746–758.

[6] J.-A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 105–119.

[7] S. Castro, L. Chiruzzo, A. Rosá, D. Garat, G. Moncecchi, A crowd-annotated spanish corpus for humor analysis, in: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, 2018, pp. 7–11.

[8] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. Meaney, R. Mihalcea, Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish., Procesamiento del Lenguaje Natural 67 (2021) 257.

[9] L. Chiruzzo, I. Sastre, G. Rey, A. Martínez, A. Rosá, S. Góngora, S. Castro, V. Amoroso, J. Prada, J. Conde, G. Moncecchi, Overview of HAHA at IberLEF 2026: Humor Analysis based on Human Annotation and Automatic Humor Generation, Procesamiento del Lenguaje Natural 77 (2026).

[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-WS.org, 2026.

[11] S. Gehrmann, H. Strobelt, A. M. Rush, Gltr: Statistical detection and visualization of generated text, in: Proceedings of the 57th annual meeting of the association for computational linguistics: system demonstrations, 2019, pp. 111–116.

[12] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, Turingbench: A benchmark environment for turing test in the age of neural text generation, in: Findings of the association for computational linguistics: EMNLP 2021, 2021, pp. 2001–2016.

[13] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, in: International conference on machine learning, PMLR, 2023, pp. 24950–24962.
[14] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information processing & management 24 (1988) 513–523.
[15] J. D. Rennie, L. Shih, J. Teevan, D. R. Karger, Tackling the poor assumptions of naive bayes text classifiers, in: Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 616–623.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, volume 30, 2017.
[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019, pp. 4171–4186.
[18] 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, 2019.
[19] 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, 2020, pp. 8440–8451.
[20] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, M. Villegas, Maria: Spanish language models, arXiv preprint arXiv:2107.07253 (2021).
[21] J. M. Pérez, D. A. Furman, L. A. Alemany, F. M. Luque, Robertuito: a pre-trained language model for social media text in spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 7235–7243.
[22] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).
[23] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, S. Han, Awq: Activation-aware weight quantization for on-device llm compression and acceleration, Proceedings of machine learning and systems 6 (2024) 87–100.
[24] 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, the Journal of machine Learning research 12 (2011) 2825–2830.
[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.