Ensembling BETO, Linguistic SVM, and Qwen2.5 with LoRA for Automatic Leveling of Spanish Learner Texts

Alvaro Infante[1], Marina Castillo[1]

[1]Department of Computer Science, Universidad Carlos III de Madrid, Madrid, Spain

Abstract

This paper describes our participation in the NivELE shared task, organized within IberLEF 2026, whose goal is to classify texts written by Spanish as a Foreign Language (ELE) students into one of the five CEFR levels present in the data: A1, A2, B1, B2, and C1. Our proposal combines three model families through a weighted probability ensemble: a BETO model (bert-base-spanish-wwm-cased) fine-tuned using 5-fold stratified cross-validation, an SVM classifier over 48 hand-crafted linguistic features (readability, syntactic complexity, lexical density, and discourse markers), and a Qwen2.5-1.5B-Instruct model adapted with LoRA. The ensemble weights are optimized with Nelder-Mead by maximizing Macro-F1 on the *out-of-fold* predictions, obtaining 0.576 for BETO, 0.212 for the SVM, and 0.212 for the generative model. In cross-validation, the system achieves a Macro-F1 of 0.9891, compared with 0.9861 for the transformer alone, 0.8759 for the classical model, and 0.2967 for the generative model fine-tuned with LoRA. We also report an error analysis and discuss why the generative model, despite its low individual performance, contributes marginally to the ensemble.

Keywords

Natural Language Processing, Text classification, Spanish as a foreign language, CEFR, BETO, Model ensembling

1. Introduction

Leveling texts written by students of Spanish as a Foreign Language (ELE) is an essential task in language centers: at the beginning of each academic term, it is necessary to classify large cohorts of students within very tight deadlines, and assessing written production is the most costly part of the process, since each text must be leveled individually by instructors.

The NivELE shared task [1], organized within IberLEF 2026 [2], proposes to automate this process using Natural Language Processing (NLP) techniques. Specifically, it consists of assigning each text produced by an ELE student to one of the six levels of the Common European Framework of Reference for Languages (CEFR): A1, A2, B1, B2, C1, or C2. In the edition and the data available for this participation, only examples up to C1 appear. Unlike other text complexity classification tasks, NivELE focuses on indicators of grammatical, lexical, and discursive complexity rather than on the overall quality of the text.

In this work we present a system that combines three deeply complementary model families: a Spanish pre-trained transformer (BETO) that learns contextual representations of the full text, a classical classifier (SVM) that relies on 48 hand-crafted linguistic features, and a generative model (Qwen2.5-1.5B) adapted to the task via LoRA. The predictions of the three models are combined in a probability ensemble whose weights are optimized with Nelder-Mead on the validation set.

The main contributions of this work are:

- A final weighted-ensemble system that achieves an *out-of-fold* Macro-F1 of 0.9891 over 21,220 training texts.
- A direct comparison of three paradigms (fine-tuned transformer, linguistic feature engineering, and LoRA-based LLM) under the same 5-fold stratified cross-validation split.

IberLEF 2026, September 2026, León, Spain

alvaroinfantee@gmail.co (A. Infante); 100569970@alumnos.uc3m.es (M. Castillo)

0009-0002-0680-9612 (A. Infante); 0009-0001-6678-3568 (M. Castillo)

- A quantitative and qualitative error analysis showing that most mistakes concentrate between adjacent levels (B1↔B2) and that the generative model, although performing far below the other two components, provides a small but consistent improvement to the ensemble.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the task and the data. Section 4 details the system architecture. Section 5 describes the experimental setup, and Section 6 presents the results. Section 7 contains the error analysis, and Section 8 concludes.

2. Related work

Automatic assessment of language proficiency according to the CEFR has received increasing attention, especially for English through corpora such as EFCAMDAT or CEFR-LM. For Spanish, the most widely used learner resources are CAES [3] and CEDEL2 [4], and there are prior works on readability and simplification. In the 2025 edition of IberLEF, the PROFE task addressed Spanish proficiency assessment using multiple-choice questions rather than free written production [5].

On the other hand, Spanish pre-trained models such as BETO [6] and MarIA [7], together with recent generative families such as Qwen2.5 [8], provide a solid basis for building classification systems adapted to the educational domain. Efficient adaptation with LoRA [9] also makes it possible to fine-tune LLMs at reduced computational cost.

3. Task and data description

3.1. Task formulation

The NivELE 2026 task is formulated as a multiclass classification problem. Given an essay written by an ELE student, the system must predict a single CEFR level. In the training and test data released by the organizers, only examples up to C1 appear, so our system produces predictions over five labels: A1, A2, B1, B2, and C1.

3.2. Datasets

We combine two sources for training: the Corpus of Spanish Learners (CAES) [3], which forms the bulk of the corpus, and a complementary collection extracted from the HablaCultura website [10], used to enrich the intermediate and higher levels. The data cleaning process involved removing HTML tags, normalizing Unicode characters (e.g., standardizing quotation marks and apostrophes), and stripping excessive whitespace, while preserving line breaks that denote paragraph boundaries. We also retained original spelling and grammatical errors, as they provide valuable signals for lower proficiency levels. After this cleaning, the training corpus contains 21,220 texts. The official test set contains 316 unlabeled texts. Table 1 shows the distribution by level and source.

Table 1
Distribution of the training corpus by CEFR level and source, and size of the test set.

Level	CAES	HablaCultura	Total train	Test (official)
A1	5.553	0	5.553	—
A2	5.708	33	5.741	—
B1	4.483	245	4.728	—
B2	3.100	273	3.373	—
C1	1.726	99	1.825	—
Total	20.570	650	21.220	316

The distribution is clearly imbalanced (A1 and A2 are the most frequent levels; C1 is the least represented). Average text length increases with level, from 103 words in A1 to 286 words in C1, which already provides a signal that can be exploited (Table 2).

Table 2
Text length (number of words per text) by CEFR level.

Level	Mean	Std. dev.	Min.	Max.
A1	103.2	52.1	1	438
A2	139.8	53.5	1	420
B1	163.1	65.2	2	913
B2	222.6	81.9	6	511
C1	285.9	122.2	5	771

3.3. Evaluation metric

We report Macro-F1 as the main metric (in line with the imbalanced nature of the task) and Accuracy and Quadratic Weighted Kappa (QWK) as secondary metrics. QWK is especially informative because of the ordinal nature of the CEFR classes, since it penalizes errors between distant levels more heavily than errors between adjacent levels.

4. Proposed system

4.1. Overview

Our final system is a weighted ensemble of three independent classifiers: (i) a BETO **transformer** fine-tuned for sequence classification, (ii) a **classical classifier** based on SVM over 48 hand-crafted linguistic features, and (iii) a **generative model** Qwen2.5-1.5B-Instruct adapted to the task with LoRA. Each model produces a probability vector over the five classes; the final ensemble is a weighted average of these vectors, with weights optimized to maximize *out-of-fold* Macro-F1.

4.2. Preprocessing

Preprocessing includes light normalization of spacing and punctuation, preservation of spelling errors (a useful signal for lower levels), and truncation to 512 tokens for BETO with a sliding window overlap of 64 for longer texts (especially B2 and C1).

4.3. Model 1: Fine-tuned BETO

We use `dccuchile/bert-base-spanish-wwm-cased` [6], chosen for its proven performance in Spanish classification tasks. On top of the `[CLS]` representation we apply a linear classification head with five outputs. Training is performed with class-frequency weighted cross-entropy to mitigate imbalance.

4.4. Model 2: SVM over linguistic features

We extract 48 features grouped into five families:

- *Surface statistics*: number of characters, tokens, words, sentences, paragraphs, mean and standard deviation of sentence and word length, syllables per word, ratio of long words.
- *Readability*: Fernández-Huerta index.
- *Lexical diversity*: type-token ratio (TTR), corrected TTR, Guiraud index, hapax ratio, ratio of unique words, ratio of repeated words, ratio of basic vocabulary, lexical density.

- *Morphosyntax*: POS proportions (noun, verb, adjective, adverb, adposition, determiner, pronoun, conjunction, subordinating conjunction), number of distinct verb tenses, ratio of subjunctive, number of subordinate clauses and subordinate clauses per sentence, mean and maximum dependency depth.
- *Discourse and error*: discourse markers by category (additive, adversative, causal, temporal, reformulative), total markers and markers per sentence, ratio of consecutive repetitions, missing accent marks (count and ratio), and ratio of out-of-vocabulary words.

The features are standardized with `StandardScaler`. We compare SVM (RBF), Random Forest, and Logistic Regression through cross-validation and select the best one (SVM), with an *out-of-fold* Macro-F1 of 0.8759.

4.5. Model 3: Qwen2.5-1.5B with LoRA

As a third component, we fine-tune `Qwen/Qwen2.5-1.5B-Instruct` [8] with LoRA [9] (r=16, α=32, *dropout* = 0.05) on the `q_proj`, `k_proj`, `v_proj`, and `o_proj` projections. The model is trained to emit the CEFR level as a single output token, conditioned with an instruction prompt that includes five representative examples (one per class). To obtain probabilities over the five classes, we extract the logits of the first response token and apply softmax over the identifiers `A1`, `A2`, `B1`, `B2`, and `C1`. The exact prompt template used for the generative model was: *"Given the following text written by a Spanish learner, classify its CEFR level (A1, A2, B1, B2, or C1). Only output the label. Text: {text} Level: "*, followed by five few-shot examples (one per class).

4.6. Ensemble

The *out-of-fold* predictions of each model are combined through a weighted average. The weights are optimized with Nelder-Mead by maximizing Macro-F1, using multiple initializations to mitigate local minima and reparameterizing via softmax to guarantee positive weights that sum to 1. The final weights are $(w_{\text{BETO}}, w_{\text{SVM}}, w_{\text{Qwen}}) = (0.576,\ 0.212,\ 0.212)$.

4.7. Additional data and technologies

Beyond CAES and the HablaCultura texts, no synthetic data or data augmentation were used. The system was implemented in Python with PyTorch, the Hugging Face `transformers` library, `peft` (for LoRA), `scikit-learn`, `xgboost`, `lightgbm`, and `spaCy` (model `es_core_news_lg`) for morphosyntactic and dependency tagging.

5. Experimental setup

5.1. Splits and validation

We use 5-fold *StratifiedKFold* with fixed seed (42). All reported metrics are computed on concatenated *out-of-fold* predictions, covering the full set of 21,220 texts for BETO and SVM. Due to computational constraints, the generative model is trained on a stratified subsample of 2,000 texts (400 per class), preserving the same fold split.

5.2. Hyperparameters

Table 3 summarizes the main hyperparameters of each model.

Table 3
Hyperparameters of the final system.

Hyperparameter	Value
BETO	
Base model	dccuchile/bert-base-spanish-wwm-cased
Maximum sequence length	512 tokens
Sliding-window overlap	64 tokens
Learning rate	2×10^{-5}
Batch size	8 (effective 32 with grad. accum. = 4)
Maximum epochs	10
Early stopping (patience)	3
Optimizer	AdamW
Weight decay	0.01
Warm-up ratio	0.1
Scheduler	linear with warm-up
Loss	Class-weighted cross-entropy
Classical SVM	
Number of features	48
Scaling	StandardScaler
Model selection	SVM vs. RF vs. Logistic (CV) $\rightarrow$ SVM
Qwen2.5-1.5B + LoRA	
Base model	Qwen/Qwen2.5-1.5B-Instruct
LoRA r, α, dropout	16, 32, 0.05
Target modules	q_proj, k_proj, v_proj, o_proj
Prompt examples	5 (one per class)
Common	
No. of folds	5 (StratifiedKFold)
Seed	42

6. Results

6.1. Cross-validation results

Table 4 summarizes the *out-of-fold* metrics of each component and of the final ensemble. The fine-tuned transformer is by far the strongest component, followed by the linguistic SVM and the generative model. The ensemble surpasses the best individual component in Macro-F1 (0.9891 versus 0.9861).

The results from the generative component (Qwen2.5-1.5B-Instruct + LoRA) warrant special attention. Its standalone *out-of-fold* Macro-F1 (0.2967) and Accuracy (0.3385) are roughly equivalent to a random classifier in a 5-class setting, falling significantly short of even the traditional rule-based SVM (0.8746). We hypothesize this poor performance stems from several factors: the small parameter count (1.5B) struggling with nuanced stylistic evaluation, the architectural mismatch between next-token generation and structured multiclass classification, and the inherent difficulty of fine-tuning LLMs on highly imbalanced ordinal data with a limited subset of 2,000 texts. Instead of learning underlying linguistic complexity features, the model often hallucinated labels or defaulted to the most frequent classes.

6.2. Stability across folds

Table 5 shows the best Macro-F1 in each fold for BETO and Qwen. BETO is very stable (range 0.984–0.990); Qwen shows high variance across folds (range 0.369–0.528), consistent with the difficulty of inducing ordinal multiclass classification with a small LLM and few examples.

Table 4
Out-of-fold metrics on the training corpus (5-fold CV, seed 42). The generative model is evaluated on the stratified subsample of 2,000 texts; the ensemble is also reported on that subsample for comparability.

System	N	Macro-F1	Accuracy	QWK
Linguistic SVM (48 features)	21.220	0.8759	0.8782	0.9201
Fine-tuned BETO	21.220	0.9861	0.9874	0.9888
Qwen2.5-1.5B + LoRA	2.000	0.2967	0.3385	0.2240
BETO (subsample n=2,000)	2.000	0.9905	0.9905	0.9945
SVM (subsample n=2,000)	2.000	0.8746	0.8745	0.9244
Ensemble (BETO+SVM+Qwen)	2.000	**0.9895**	**0.9895**	**0.9939**
Optimal ensemble (Nelder-Mead, OOF)	—	0.9891	—	—

Table 5
Best Macro-F1 by fold for each component.

Model	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Mean
BETO	0.9850	0.9844	0.9845	0.9901	0.9865	0.9861
Qwen2.5-1.5B+LoRA	0.5282	0.3692	0.4430	0.4302	0.4440	0.4429

6.3. Optimal ensemble weights

Nelder-Mead assigns almost 58 % of the weight to the transformer and splits the rest equally between the SVM and the generative model (Table 6). Despite Qwen's poor individual performance, removing it from the ensemble slightly degrades Macro-F1, suggesting that it contributes a complementary signal on a small but useful subset of examples.

Table 6
Optimal ensemble weights (Nelder-Mead, maximizing OOF Macro-F1).

Component	Weight
Fine-tuned BETO	0.576
Linguistic SVM	0.212
Qwen2.5-1.5B + LoRA	0.212

6.4. Per-class results

Table 7 shows precision, recall, and F1 by class for BETO (the dominant component) and for the SVM. BETO obtains F1 above 0.98 in all classes; the SVM degrades on B1 and C1 (the classes with fewer examples and greater length overlap).

Table 7
Precision / Recall / F1 by class (OOF, n=21,220).

	Fine-tuned BETO			Linguistic SVM		
Level	P	R	F1	P	R	F1
A1	0.990	0.993	0.992	0.915	0.937	0.926
A2	0.993	0.988	0.991	0.854	0.871	0.862
B1	0.984	0.984	0.984	0.843	0.824	0.834
B2	0.981	0.985	0.983	0.899	0.896	0.897
C1	0.981	0.981	0.981	0.892	0.830	0.860

6.5. Official Competition Results

On the official test set provided by the organizers, our ensemble system achieved a Macro-F1 of 0.55460. With these results, our team ranked in the 5th position out of 9 participating teams in the shared task.

7. Error analysis

Table 8 shows the confusion matrix of the BETO component on the 21,220 training texts (OOF). Of the 264 total errors, the vast majority concentrate in adjacent-level pairs, which is the expected and desirable behavior given the ordinal nature of the CEFR labels. The most frequent error is the B1↔B2 confusion (74 cases counting both directions), followed by A1↔A2 (49 cases).

Table 8
Confusion matrix of BETO (OOF, rows = true level, columns = predicted).

	A1	A2	B1	B2	C1
A1	**5.516**	18	7	3	9
A2	31	**5.674**	20	7	9
B1	15	16	**4.650**	41	6
B2	4	3	33	**3.322**	11
C1	3	4	14	13	**1.791**

After a qualitative inspection of the errors, three recurring patterns emerge:

- *Short texts at higher levels:* some very brief B2 and C1 essays (near the minimum length in the set) are classified as B1, probably because the signal from length and syntactic complexity is diluted.
- *Systematic spelling errors:* texts with systematic omission of accent marks or opening punctuation push the prediction toward lower levels, even when syntax and vocabulary are typical of a higher level.
- *Sophisticated vocabulary with simple syntax:* short texts with infrequent vocabulary tend to be over-leveled as C1.

Approaches we tried and that did not work. NivELE explicitly asks participants to report failed attempts. During development, we experimented with three additional lines that were not included in the final system:

- *Zero-shot LLM via API* (a fourth component, *approach3*): inference cost and instability at the intermediate levels made it uncompetitive against an already highly accurate BETO.
- *Ensemble with temperature calibration* and *TTA* on BETO sliding windows: gains were marginal (second decimal) and increased variance across *seeds*.
- *Bias from the test-set prior distribution*: we tested several versions of *prior shift* toward the CAES distribution or toward a uniform distribution; the effect was positive in some levels but negative in others, with no consistent net benefit.

Lessons learned. In this corpus, the bottleneck is no longer the main model but the few dozen genuinely difficult examples (short texts, systematic errors, out-of-distribution vocabulary). Any substantial improvement would likely require additional data specifically designed for these edge cases.

8. Conclusions and future work

We have presented a system for the automatic leveling of texts written by ELE students within the NivELE shared task of IberLEF 2026. The system combines, through weighted ensembling, a fine-tuned

BETO, a 48-feature linguistic SVM, and a Qwen2.5-1.5B model with LoRA, achieving an *out-of-fold* Macro-F1 of 0.9891. While this ensemble improves the Macro-F1 score from 0.9861 (BETO alone) to 0.9891—a statistically valuable gain in a competition framework—we acknowledge a significant computational trade-off. The added inference latency and extraction cost of running an SVM with 48 linguistic features and an LLM alongside a transformer represents a heavy burden for a marginal improvement (∼0.3%). For real-world educational deployments, deploying the standalone BETO model would offer a far superior efficiency-performance balance.

As future work, we consider: (i) replacing the SVM with an ordinal regression model that explicitly captures the ordinal structure of the CEFR, (ii) expanding the corpus with synthetic data generated by larger LLMs for the least represented levels (C1), and (iii) exploring larger generative models (Qwen2.5-7B, Llama-3.1 in Spanish) with discriminative heads instead of generative output.

Code and data availability

The source code and evaluation scripts are available at https://github.com/Alvaroinfantee/nivele-2026-experiments.git under the MIT license. The data are those provided by the organizers (CAES and HablaCultura) and are subject to their terms of use.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI Codex in order to draft and revise paper text, improve grammar and wording, structure the CEURART LaTeX source, and check formatting against the submission requirements. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. V. Cantero Romero, J. Collado Montañez, I. Cabrera de Castro, J. Fruns Giménez, A. Arjonilla Sampedro, J. Cruz, A. Montejo-Ráez, S. M. Jiménez-Zafra, Resumen de nivele en iberlef 2026: Nivelación automática de textos de estudiantes de ele (español como lengua extranjera), Procesamiento del Lenguaje Natural 77 (2026).

[2] 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.

[3] N. T. S.L., CAES — galvan.usc.es, https://galvan.usc.es/caes, 2026.

[4] N. T. S.L., CEDEL2: Corpus Escrito del Español L2 (version 3) — cedel2.learnercorpora.com, https://cedel2.learnercorpora.com, 2026.

[5] Á. Rodrigo, A. Peñas, A. Pérez, S. Moreno-Álvarez, J. Fruns, I. Soria Pastor, R. Agerri, Overview of PROFE at IberLEF 2025: Language Proficiency Evaluation, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4098/.

[6] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: PML4DC at ICLR 2020, 2020. URL: https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf.

[7] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, MarIA: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. doi:10.26342/2022-68-3.

[8] Qwen Studio — qwen.ai, https://qwen.ai/home, 2026.

[9] LoRA: Low-Rank Adaptation of Large Language Models — doi.org, https://doi.org/10.48550/ar{X}iv.2106.09685, 2021.
[10] Textos para aprender español gratis | HABLACULTURA — hablacultura.com, https://hablacultura.com/cultura-textos-aprender-espanol/, 2026.