NIL-UCM at NivELE 2026: Automatic Leveling of ELE Student Texts

Raquel Jimeno[1], Alberto Díaz[2,*] and Sandra Conde[3]

[1]Facultad de Filología, Universidad Complutense de Madrid, Spain
[2]Facultad de Informática and ITC, Universidad Complutense de Madrid, Spain
[3]Facultad de Informática, Universidad Complutense de Madrid, Spain

Abstract

These working notes evaluate pre-trained transformer models — BETO and MrBERT-es —adapting them to a multi-class text classification setting, focused on the automatic classification of Spanish as a Foreign Language (ELE) learner texts according to the Common European Framework of Reference for Languages (CEFR) using the corpusELE dataset from CAES (Corpus de Aprendices de Español como Lengua Extranjera). This work also explored prompt engineering strategies with generative LLM, specifically Mistral-7B-Instruct. The results underline the need for fine-tuning and robust prompting for reliable educational NLP applications.

Keywords

text classification, multi-class text classification, ELE, CEFR, transformers models, MrBERT-es, BETO, NLP, zero-shot, few-shot, Mistral

1. Introduction

In recent years, the rapid evolution of large language models (LLMs) has transformed Natural Language Processing (NLP), driving significant progress in core tasks such as machine translation, text summarization, and text classification. This study focuses on the application of these models for the automatic classification of written texts produced by students of Spanish as a Foreign Language (ELE). Specifically, we explore how different types of language models—encoder-only versus generative—perform in this setting, and what challenges arise from task design, model architecture, and prompt formulation.

The motivation for this work is rooted in the increasing relevance of AI-assisted tools in education such as automated scoring or grammatical error detection, and learner proficiency classification. In the context of Spanish as a Foreign Language (ELE), the automatic prediction of CEFR levels from learner texts represents an important challenge due to the variability in lexical richness, syntactic complexity, discourse organization, and grammatical correctness present in learner productions. Understanding how current models perform in learner proficiency classification is key to developing effective and fair educational technologies. Our experiments compare the performance of MrBERT [1], BETO [2], and Mistral [3] on this task. We examine the impact of prompt design, model architecture, and output interpretation on model accuracy and reliability.

2. Task Description

2.1. Multi-Class classification

This work is part of the NivELE 2026 shared task [4], proposed as part of IberLEF 2026 [5]. This task focuses on the development and evaluation of Natural Language Processing (NLP) models for the automatic proficiency classification of learner texts written by students of Spanish as a Foreign

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
raquej06@ucm.es (R. Jimeno); adiazest@ucm.es (A. Díaz); saconde@ucm.es (S. Conde)
0009-0002-9704-9363 (R. Jimeno); 0000-0003-1966-3421 (A. Díaz); 0009-0009-3535-1330 (S. Conde)

Language (ELE). Proper student placement is critical, given its direct influence on academic achievement. However, assessing written expression can be time-consuming as texts must be leveled one by one. This evaluation task consists of identifying which level of the Common European Framework of Reference for Languages (CEFR) a text written by an ELE student belongs to:

- A1 : Beginner
- A2 : Elementary
- B1 : Intermediate
- B2 : Upper-intermediate
- C1 : Advanced

Systems received a learner text as input and had to predict the corresponding CEFR level.

2.2. Task Challenges

This task presents several significant challenges:

- **Overlapping Proficiency Features:** The CEFR standards for adjacent proficiency levels share overlapping linguistic features, making distinctions between levels such as A2 and B1 particularly difficult. For example, the learner production:

  > "Fue una buena experiencia para mi y para mi curriculum. Pero, también consegui ir una semana de vacaciones con mi familia para una ciudad fantastica. Empiezamos nuestra viaje en Barcelona, y yo creo que fue la mejor ciudad pues Barcelona es una ciudad muy moderna y bella que tiene muchos parques y plazas peatonales, las personas son muy educadas y animada."

  contains grammatical inconsistencies such as "nuestra viaje" and "animada", but simultaneously demonstrates discourse cohesion, descriptive vocabulary, and relatively complex sentence structures. Therefore, the text could reasonably be classified as either A2 or B1 depending on the evaluation criteria.
- **Interlanguage Noise:** Texts produced by language learners reflect the evolving state of their interlanguage. As a result, these texts often contain code-switching, spelling deviations, grammatical errors, and non-standard syntactic structures, which can severely degrade model performance. For instance, the beginner-level production:

  > "Porque estoy en una reuniao de trabajo e después iremos para una fiesta comemorar cumpleaños de Joao."

  shows clear interference from the learner's first language. Such cross-linguistic interference introduces noise that complicates automatic CEFR classification.
- **Multidimensional Modeling:** Automated proficiency classification demands the simultaneous modeling of multiple linguistic dimensions, including lexical richness, grammatical accuracy, fluency, and discourse organization. Learner texts, as illustrated in the first excerpt, may exhibit advanced vocabulary while still containing basic grammatical errors, making proficiency prediction particularly challenging. Consequently, models must jointly evaluate semantic content, syntax, cohesion, and error patterns rather than relying on a single linguistic feature.
- **Class Imbalance:** The dataset exhibits a class imbalance that reflects the actual distribution of students across proficiency levels. While representative, this imbalance poses a risk of biasing classification systems toward the majority categories.
- **Generative Model Inconsistencies:** Leveraging generative models introduces risks such as inconsistent output formats, mislabeling, and content hallucination.

To analyze these challenges, our work evaluates two distinct methodologies: fine-tuning encoder models (BETO and MrBERT) and implementing generative models via prompt engineering, allowing us to benchmark their robustness against interlanguage noise and class imbalance.

3. Methodology

3.1. Dataset: CorpusELE

For this study, we use the CorpusELE dataset [6] available on Hugging Face. This corpus is based on learner productions collected from Spanish language learners across multiple CEFR proficiency levels (A1–C1). The texts originate from a subset of files from the CAES (Corpus de Aprendices de Español como Lengua Extranjera) [7], which were downloaded from the Instituto Cervantes website.

Within the diverse taxonomy of existing corpora, CAES is classified as a "learner corpus," as it comprises texts generated by language learners with varied native backgrounds and distinct levels of target-language proficiency. In total, the dataset contains 46,787 texts distributed unevenly across five CEFR proficiency levels: A1 (12,284), A2 (13,926), B1 (10,848), B2 (6,601), and C1 (3,128). It includes examples of learner-produced data containing grammatical errors, lexical variation, syntactic simplifications, and discourse patterns characteristic of second language acquisition processes. Such properties make the task particularly challenging from a computational perspective.

3.2. Experimental Overview

We experimented with two different types of Transformer models to evaluate both their text classification and text generation capabilities:

- **Encoder-Only Models:** We fine-tuned these models end-to-end on the CorpusELE dataset, adapting their architecture specifically for the sequence classification task.
- **Decoder-Only Models:** These generative models are evaluated zero-shot and few-shot, utilizing designed prompt engineering templates to instruct them to generate the target proficiency labels.

3.3. Encoder Models for Multi-Class Classification

To evaluate the effectiveness of different pre-trained language models on the CEFR classification task, we tested two transformer models specialized for the Spanish language:

- **MrBERT-es** (`BSC-LT/MrBERT-es`)
- **BETO** (`dccuchile/bert-base-spanish-wwm-cased`)

Both encoder-based models were adapted for a multi-class classification setup and fine-tuned specifically on the CorpusELE dataset to predict the corresponding proficiency levels.

For the encoder-based experiments, the dataset was divided into training and validation sets following a 90/10 split. The resulting partitions consisted of 42,108 texts for training and 4,679 texts for validation. To preserve the original CEFR-level distribution and mitigate the impact of class imbalance, a stratified sampling strategy was applied during the split. Consequently, each partition maintained approximately the same proportion of texts for each proficiency level as the original dataset. The official NivELE 2026 test set was used exclusively for final evaluation and was never employed during training, validation, or model selection.

3.4. Prompting Strategy for Generative Models

In addition to encoder-based approaches, we explored decoder-based generative models using prompting strategies. Specifically, we employed Mistral-7B-Instruct-v0.2 and evaluated two prompting settings: zero-shot and few-shot classification. The model was instructed to assign one of the five CEFR levels (A1, A2, B1, B2, or C1) to each learner text. Prompts were written in Spanish to align with the language of the task and included explicit descriptions of the linguistic characteristics associated with each proficiency level.

For the zero-shot setting, the model received detailed CEFR descriptors and was asked to output only the corresponding proficiency label. A representative prompt template is shown below:

"Eres un experto evaluador del MCER (Marco Común Europeo de Referencia). Tu tarea es clasificar el nivel de dificultad de un texto en español. Categorías permitidas: [A1, A2, B1, B2, C1]. Analiza el siguiente texto y responde únicamente con la etiqueta del nivel (A1, A2, B1, B2 o C1). Texto: {texto}"

Additionally, the prompt included detailed descriptions of each CEFR level covering lexical, grammatical, and discourse-level characteristics.

For the few-shot setting, the prompt incorporated one representative example for each CEFR level before presenting the target text. The examples were selected to illustrate typical learner productions associated with each proficiency category. A simplified example of the prompt structure is shown below:

Texto: "Tiene los ojos negros y grandes y el pelo largo." Nivel: A1
Texto: "Nació en Damasco en 1983, de una familia pequeña." Nivel: A2
Texto: "Nuestra amiga Nadia me ha dicho que dejaste de trabajar..." Nivel: B1
Texto: "Es un problema que pone a todos en riesgo..." Nivel: B2
Texto: "Este triste evento fue el origen de una gran huelga..." Nivel: C1
Texto: {texto}
Nivel:

Although this zero-shot and few-shot approach allowed rapid experimentation without supervised fine-tuning, the generative systems showed significantly lower performance compared to encoder-based models. Nevertheless, the experiments provided valuable insights into the limitations of prompt-only text classification approaches.

3.5. Technical Implementation

All experiments were implemented in Python using Google Colab as the execution environment. The dataset was loaded from Google Drive, and predictions were saved back in .csv format. The following tools and libraries were used:

- Hugging Face Transformers for pretrained transformer architectures.
- Standard Python libraries: pandas and os for file and data handling, scikit-learn was for evaluation metrics.

Model fine-tuning and evaluation were performed via the Hugging Face ecosystem. The test dataset was processed through direct model inference to compute class probabilities, with the final predicted CEFR levels ultimately exported to a CSV file for analysis.

4. Results and Discussion

4.1. Final Evaluation Results

As shown in Table 1, encoder-based transformer models substantially outperformed prompt-based generative approaches on the development set. Both encoder models, BETO and MrBERT, achieved similar macro-averaged F1 scores on the development set (81% and 80.5%, respectively) and on the test set (57.6% and 58.2%).

In contrast, the Mistral-based configuration obtained substantially lower scores for both zero-shot and few-shot prompting strategies. On the development set, the zero-shot configuration achieved an F1-score of 22% and the few-shot only 10%. Similarly, on the test set, the zero-shot and few-shot approaches achieved F1-scores of 26% and 34%, respectively.

4.2. Prompting Limitations for Generative Models Issues

This performance difference can be explained by the architectural distinction between encoder-based transformers (MrBERT and BETO), which are optimized for classification tasks, and decoder-based

Table 1
Results from the exploratory experiments.

Method	Main model	F1-score	Observations
Fine-tuning	BETO	81.0%	Most stable configuration across CEFR levels
Fine-tuning	MrBERT-es	80.5%	Strong semantic adaptation to learner Spanish
Zero-shot prompting	Mistral-7B-Instruct	22.0%	Frequent invalid labels and inconsistent outputs
Few-shot prompting	Mistral-7B-Instruct	10.0%	Highly sensitive to prompt wording and examples

transformers such as Mistral, primarily designed for text generation rather than discriminative NLP classification tasks.

Additionally, the decoder-based experiments frequently generated outputs that did not correspond to the target proficiency labels (A1–C1), instead producing free-form textual responses such as “It is”, “I would”, “please”, or “no sé”, which negatively affected classification accuracy and overall robustness.

Furthermore, the model was highly sensitive to the prompt. Small changes in the wording or adding examples caused completely different and unpredictable behaviors. Such variability was observed across both the zero-shot and few-shot prompt configurations described in Section 3.3, suggesting that prompt engineering alone was insufficient to reliably constrain the model outputs to the predefined CEFR label set. Therefore, prompt-based approaches may not provide sufficient task-specific adaptation for evaluating linguistic competence.

4.3. Interpretation and Future Directions

The experimental results confirm the effectiveness of encoder-only transformer architectures for CEFR classification tasks. Spanish-specific pretrained models such as BETO appear particularly suitable for learner language analysis because they capture lexical and syntactic patterns specific to Spanish.

The limited performance of prompt-only generative systems does not necessarily imply that decoder architectures are unsuitable for the task. Instead, the results suggest that generative models require task-specific fine-tuning and better alignment strategies. Hence, future work will focus on two main research directions. First, we plan to fine-tune generative decoder-based models using supervised training rather than prompt engineering alone. Fine-tuning may allow generative models to learn more stable mappings between learner texts and CEFR labels. Second, we plan to develop generative architectures with classification heads capable of directly predicting CEFR levels. Such architectures may combine the contextual understanding abilities of large language models with the stability of discriminative classifiers.

5. Conclusions and Future Work

The main objective of this study was to evaluate the performance of different NLP approaches in automated CEFR proficiency classification in Spanish. Our experiments show a clear advantage for fine-tuned encoder-only models, such as BETO and MrBERT, which achieved stable and robust performance across both development and test sets. However, decoder models achieved lower results due to their generative nature and high sensitivity to the zero-shot and few-shot prompts.

To address these limitations, our future efforts will prioritize transitioning from open-ended prompting to supervised and hybrid decoder training strategies to ensure stable classification. Beyond these experimental refinements, this research has broader implications for educational applications. Accurate automated proficiency classification systems could support large-scale language assessment, adaptive learning environments, and intelligent tutoring systems for Spanish as a foreign language.

Declaration on Generative AI

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

References

[1] D. Tamayo, I. Lacunza, P. Rivera-Hidalgo, S. Da Dalt, J. Aula-Blasco, A. Gonzalez-Agirre, M. Villegas, Mrbert: Modern multilingual encoders via vocabulary, domain, and dimensional adaptation, 2026. URL: https://arxiv.org/abs/2602.21379. arXiv:2602.21379.

[2] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: Proceedings of the 8th ICLR Workshop on Representation Learning for NLP, 2020. URL: https://arxiv.org/abs/2308.02976.

[3] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, W. El Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[4] 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).

[5] A. Bonet-Jover, J. 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] FJC, Corpusele: A dataset of texts from learners of spanish as a foreign language, 2024. [Accessed: 2026-05-27].

[7] Instituto Cervantes, Universidad de Santiago de Compostela, Corpus de aprendices del español (caes), 2012. URL: https://galvan.usc.es/caes/, [Versión 2.1].