UCO-UC-CICESE-Plenitas Team: Exploring the Automatic Leveling of ELE Student Texts Using Embeddings

Yoan Martínez-López[1,2,*], Mayte Guerra Saborit[3], Julio Madera[3], Luis Quintero Domínguez[5], Ansel Rodríguez-González[6], Kristell Aguilar Alarcón[1,2], Carlos de Castro Lozano[1,2] and Jose Miguel Ramirez Uceda[1,2]

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

Abstract

This paper describes our participation in the NivELE 2026 shared task at IberLEF, which addresses the automatic classification of texts produced by learners of Spanish as a Foreign Language (ELE) according to the six proficiency levels defined by the Common European Framework of Reference for Languages (CEFR): A1, A2, B1, B2, C1 and C2. We approach the problem as an ordinal multiclass classification task and propose a hybrid architecture that combines two complementary sources of information: dense semantic representations obtained from pretrained multilingual language models and a vector of twenty explicit linguistic features derived from CEFR descriptors and from the literature on Complexity, Accuracy and Fluency (CAF). The linguistic features capture length and structural density, lexical diversity, morphosyntactic complexity, discursive cohesion and genre conventions, and are computed exclusively through regular expressions and counting operations to keep the pipeline lightweight and reproducible. Both representations are concatenated through early fusion and standardized prior to training five classical supervised classifiers (Logistic Regression, RBF-SVM, Linear SVM, XGBoost and Random Forest), which are subsequently ranked by Macro F1 on a stratified 60/40 train/validation split. We systematically compare six multilingual embedding backbones — Longformer, BAAI/bge-m3, intfloat/multilingual-e5-large, sentence-transformers/LaBSE, microsoft/mdeberta-v3-base and nomic-ai/nomic-embed-text-v1.5 — combined with the proposed classifiers. The best configurations substantially outperform the random baseline (F1 $\approx$ 0.21), reaching 0.42063 on the private partition with multilingual-e5-large + SVM and 0.43948 on the public partition with bge-m3 + XGBoost. The experimental results suggest that linear classifiers combined with strong multilingual sentence encoders provide the most stable behaviour under the evaluated conditions, while tree-based ensembles exhibit higher sensitivity to the geometry of the embedding space.

Keywords

multilingual embeddings, linguistic complexity, automatic proficiency classification, CEFR

1. Introduction

The automatic classification of student-produced texts is a critical challenge in the field of Spanish as a Foreign Language (ELE) [1]. At the start of every academic term, language centers face the logistical burden of placing large volumes of students into appropriate proficiency levels within extremely tight deadlines. This placement process is essential, as the correct leveling of a student directly correlates with their future academic success. While many parts of language assessment can be streamlined, evaluating written expression remains the most arduous and time-consuming task for faculty, as each

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
yoan.martinez@plenitas.com (Y. Martínez-López); mayte.guerra@reduc.edu.cu (M. G. Saborit); julio.madera@reduc.edu.cu (J. Madera); laqdominguez@gmail.com (L. Q. Domínguez); ansel@cicese.mx (A. Rodríguez-González); kristell.aguilar@plenitas.com (K. A. Alarcón); carlosdecastrolozano@gmail.com (C. d. C. Lozano); p52raucj@uco.es (J. M. R. Uceda)
0000-0002-1950-567X (Y. Martínez-López); 0000-0002-9556-5869 (M. G. Saborit); 0000-0001-5551-690X (J. Madera); 0000-0002-3527-0516 (L. Q. Domínguez); 0000-0001-9971-0237 (A. Rodríguez-González); 0000-0001-8913-8388 (K. A. Alarcón); 0000-0001-6603-843X (C. d. C. Lozano); 0000-0002-5027-7521 (J. M. R. Uceda)

text must currently be leveled individually by a human evaluator. To address this challenge, the NivELE shared evaluation task was established with the goal of streamlining teaching efforts and enhancing the objectivity of student placement through automation [2]. The task focuses on the development and evaluation of Natural Language Processing (NLP) [3, 4, 5] models designed to automatically identify the specific level of the Common European Framework of Reference for Languages (CEFR)—ranging from A1 to C2—to which a student's text belongs. The NivELE proposal introduces three significant novelties to the field:

- Pioneering Framework: To the best of our knowledge, it represents the first task within the IberLEF framework to propose automatic multi-class leveling specifically tailored for ELE.
- Interdisciplinary Integration: The task successfully integrates deep linguistic knowledge derived from pedagogy and official certification standards with state-of-the-art NLP techniques.
- Complexity-Based Analysis: Unlike traditional models that may focus on general text quality, NivELE prioritizes specific indicators of grammatical, lexical, and discursive complexity to determine proficiency.

By focusing on these structural indicators, the NivELE task aims to provide a robust computational solution to a long-standing pedagogical problem, facilitating more efficient and standardized assessment processes in Spanish language education.

Our main contribution is the integration of multilingual semantic embeddings with lightweight CEFR-oriented linguistic indicators designed to capture structural and morphosyntactic properties associated with learner proficiency.

2. Methodology

The NivELE 2026 task is formulated as an ordinal multiclass classification problem in which each text produced by a learner of Spanish as a Foreign Language (ELE) must be assigned to one of the six proficiency levels defined by the Common European Framework of Reference for Languages (CEFR): A1, A2, B1, B2, C1 and C2. The official evaluation metric is Macro F1, which forces the system to treat all proficiency levels in a balanced manner regardless of their frequency in the training corpus [2]. The proposal builds on the hypothesis that CEFR level determination depends simultaneously on two complementary types of information: (i) the distributional semantics of the text, which can be captured through dense representations produced by pretrained language models, and (ii) a set of explicit linguistic indicators —lexical, morphosyntactic, discursive and formal— that the literature on automatic assessment of communicative competence has shown to correlate with proficiency level. On this basis, we design a three-stage architecture that fuses both families of features into a single representation space and trains on it a set of supervised classical classifiers, which are subsequently compared by their Macro F1 performance.

2.1. Data Preprocessing and Partitioning

The training corpus (`training_corpus_caes.csv`) is loaded in tabular format with two columns, label and text. A minimal and conservative preprocessing is applied to each document, consisting in the removal of redundant whitespace and the normalization of CEFR labels to uppercase. We deliberately avoid more aggressive transformations —such as lowercasing, lemmatization or removal of punctuation— because many discriminative features of learner proficiency reside precisely in punctuation, connectives and surface morphological markers[2]. The dataset is partitioned through a stratified 60/40 train/validation split, with the random seed fixed at 42 to ensure reproducibility. Stratification is performed over the CEFR label in order to preserve the original distribution of the six levels in both subsets, which is particularly important given the imbalanced nature of the corpus. The official test set is loaded separately and processed through the same pipeline, but without labels.

2.2. Text Representation

2.2.1. Dense Embeddings

For the semantic representation of texts we adopt a pretrained language model capable of processing sequences of considerable length, a relevant requirement for a task in which written productions can exceed the 512 tokens typical of standard BERT architectures [6, 7, 8].

As an initial baseline configuration, we employed Longformer (allenai/longformer-base-4096), which combines sliding-window local attention with selective global attention and supports contexts of up to 4,096 tokens, thereby avoiding the information loss associated with truncation.

During the experimental phase, we systematically compared several multilingual alternatives, including BAAI/bge-m3, intfloat/multilingual-e5-large, sentence-transformers/LaBSE, microsoft/mdeberta-v3-base and nomic-ai/nomic-embed-text-v1.5 [9, 10], and retained the best-performing configurations for the final evaluation.

Embeddings are computed with L2 normalization, batch size 32 and a maximum sequence length of 512 tokens for standard models, extended in the case of Longformer.

2.2.2. Explicit Linguistic Features

In parallel to the embeddings, for each text we extract a vector of 20 linguistic features designed manually from CEFR descriptors and from the literature on complexity, accuracy and fluency (CAF) in applied linguistics [2].

Table 1 summarizes the categories of the twenty explicit linguistic features used in our system. The features were manually designed from CEFR descriptors and from the Complexity, Accuracy and Fluency (CAF) framework, and were selected to capture lexical, morphosyntactic, discursive and genre-related aspects associated with learner proficiency.

Table 1
Categories and examples of the explicit linguistic features used in the hybrid representation.

Category	Feature
Length and Structural Density	Number of words
Length and Structural Density	Number of sentences
Length and Structural Density	Number of characters
Length and Structural Density	Number of paragraphs
Lexical Diversity	Type-Token Ratio (TTR)
Lexical Diversity	Mean word length
Lexical Diversity	Mean sentence length
Lexical Diversity	Longest sentence length
Morphosyntactic Complexity	Frequency of subjunctive triggers
Morphosyntactic Complexity	Frequency of conditional markers
Morphosyntactic Complexity	Frequency of relative clauses
Morphosyntactic Complexity	Past tense marker frequency
Morphosyntactic Complexity	Future tense marker frequency
Discursive Cohesion	Connector density
Discursive Cohesion	Punctuation density
Genre Conventions	Presence of greeting formulas
Genre Conventions	Presence of closing formulas
Genre Conventions	Formal register marker frequency
Discursive Cohesion	Argumentative connector frequency
Morphosyntactic Complexity	Complex sentence marker frequency

These features are computed through regular expressions and simple counting operations, with no dependency on external morphosyntactic analyzers, which keeps the pipeline lightweight and reproducible.

2.2.3. Representation Fusion

For each text t_i, we obtain an embedding vector $\mathbf{e}_i \in \mathbb{R}^{d_e}$ (with $d_e = 768$ for Longformer base) and a linguistic-feature vector $\mathbf{f}_i \in \mathbb{R}^{20}$.

The final representation is built by horizontal concatenation:

$$\mathbf{x}_i = \left[\mathbf{e}_i \,\|\, \mathbf{f}_i\right] \in \mathbb{R}^{d_e+20} \tag{1}$$

This early-fusion strategy allows classifiers to jointly weight latent semantic information and explicit linguistic indicators. Prior to training, the resulting matrix is standardized with StandardScaler fitted exclusively on the training set, thus avoiding information leakage into validation and test.

2.3. Supervised Classification

On top of the fused representation space we train a set of five classical classifiers covering different families of inductive bias:

- Multinomial Logistic Regression with L-BFGS solver and `class_weight="balanced"` [11].
- SVM with RBF kernel [12] (C=10, γ = "scale") under a one-vs-rest strategy.
- Linear SVM (C=1.0) as a linear baseline in high-dimensional space [13].
- XGBoost [14] with 300 trees, maximum depth 6, learning rate 0.1, multi:soft prob objective and balanced `sample_weight`computed via `compute_sample_weight("balanced")`.
- Random Forest [15] with 300 trees and `class_weight="balanced"`.

All models, except XGBoost, receive class weights inversely proportional to the frequency of each CEFR level in order to mitigate class imbalance. In XGBoost, equivalently, a balanced sample weight vector is applied. The random seed is fixed at 42 for all stochastic components. Final model selection is performed by ranking the six classifiers by Macro F1 on the validation set. Additionally, prior to training we run a stratified 5-fold cross-validation on the training data in order to verify that each partition preserves representation of all six CEFR classes, explicitly flagging any fold with incomplete class coverage.

2.4. Prediction Generation

Once the classifier with the highest validation Macro F1 has been selected, it is reused to predict on the official test set. For each text the system retrieves both the predicted label and the per-class probabilities ($n_{\text{samples}} \times 6$), which opens the door to subsequent model-combination strategies (*soft voting*, *stacking*). The final submission is generated by mapping the predicted integer identifiers back to CEFR labels via the ‘id2label‘ dictionary and exporting them in CSV format with two columns (‘id‘, ‘label‘).

2.5. System Architecture

The system is organized as a sequential pipeline of six decoupled modules, which makes it possible to replace individual components without affecting the remaining stages.

Module 1 — Data ingestion and preparation. Reads the training corpus and the test set from their respective CSV files, normalizes labels and texts, and produces a stratified 60/40 train/validation split. This module also encapsulates the CEFR label encoder (A1=0, A2=1, B1=2, B2=3, C1=4, C2=5), preserving the ordinal order of the proficiency levels.

Module 2 — Embedding extraction. Loads the pretrained Longformer model and produces dense representations for training, validation and test sets. The module is model-agnostic: a conditional branch handles the special case of BGE-M3 (which requires FlagEmbedding and returns dictionaries with dense-vecs), while all remaining models are processed through sentence-transformers with L2 normalization applied to the resulting vectors.

Module 3 — Linguistic-feature extraction. Applies the extract_linguistic_features function to each text, returning the corresponding 20-feature vector. The implementation relies exclusively on regular expressions and counting operations, with no heavy dependencies.

Module 4 — Fusion and standardization. Horizontally concatenates the embedding and linguistic-feature matrices, fits a StandardScaler on the training set, and applies the same transformation to validation and test sets. The resulting representation has dimensionality $d_e + 20$, where d_e is the dimensionality of the embedding model in use (768 for Longformer base).

Module 5 — Training, evaluation and prediction. Sequentially trains the evaluated classifiers, computes validation metrics, ranks models by Macro F1, and generates the final predictions.

Module 6 — Submission module. Exports the final predictions in the CSV format required by the task organizers.

The entire system configuration (paths, models, hyperparameters, random seed, and number of folds) is centralized in a dataclass named `Config`, making it possible to reproduce any experiment by modifying a single object. The pipeline is implemented in Python 3.12 using scikit-learn, xgboost, transformers, sentence-transformers, FlagEmbedding, and imbalanced-learn, with optional CUDA acceleration.

2.6. Evaluation Metrics

The official metric of NivELE 2026 is Macro F1, defined as the unweighted arithmetic mean of per-class F_1 scores:

$$\begin{aligned} \text{F1}_{\text{macro}} &= \frac{1}{C}\sum_{c=1}^{C} F_{1,c}, \\ F_{1,c} &= \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \end{aligned} \tag{2}$$

where C=6, is the number of CEFR levels and P_c and R_c are respectively the precision and recall of class c. This metric heavily penalizes bias toward majority classes, which motivates our systematic use of class weighting during training. Complementarily, each model is evaluated through the following metrics:

- Weighted F_1, which weights the F_1 of each class by its support in validation and allows us to quantify aggregate performance sensitive to the actual class imbalance of the corpus.
- Per-class Precision, Recall and F_1, reported through classification report with four-digit precision, in order to diagnose at which specific levels (typically the extremes $A1/C2$, since they are the most underrepresented) the system errors concentrate.
- 6 × 6 Confusion Matrix, computed with confusion matrix over labels encoded as integers, providing a qualitative view of confusion patterns between adjacent levels —particularly relevant in ordinal tasks such as CEFR classification, where higher overlap between B1/B2 or C1/C2 is expected than between A1/C2.

For final model selection, the six classifiers are ranked in decreasing order of Macro F1 on the validation set, with the resulting ranking printed explicitly to facilitate comparative analysis. The stratified 5-fold cross-validation on the training set is used exclusively as a diagnostic of per-fold class coverage, ensuring that generalization estimates are not compromised by the absence of any CEFR level in some partition

3. Results

The system was developed in Python using the PyTorch ecosystem (Transformers, sentence-transformers, scikit-learn, pandas, NumPy and PIL). Some embedding models were served locally through Ollama during experimentation and inference, while others were accessed through the Hugging Face Hub API depending on compatibility and deployment requirements. Execution was performed on CUDA-enabled GPUs when available, with automatic device detection and precision selection (float16 on GPU, float32 on CPU)

The following analysis presents the results of an internal comparison between different combinations of multilingual embedding models and classical machine learning classifiers, evaluated against a random submission baseline. The reported metrics correspond to the F1 score on both the private and public partitions of the evaluation set. See Table 2.

Overall Performance Against the Baseline

The baseline (random submission) achieves an F1 score of 0.21036 on the private set and 0.20980 on the public set. Most of the evaluated configurations substantially outperform this reference, which confirms that the proposed pipelines extract meaningful signal from the input data. However, a non-trivial subset of configurations performs at or below baseline level, indicating that the choice of embedding-classifier pairing is far from neutral and can degrade performance when poorly matched.

Best-Performing Configurations

The highest private F1 score is obtained by intfloat/multilingual-e5-large + SVM, reaching 0.42063, which represents approximately twice the baseline performance. On the public set, the top result corresponds to BAAI/bge-m3 + XGBoost with 0.43948, closely followed by nomic-ai/nomic-embed-text-v1.5 + SVM (0.42900). Notably, no single configuration dominates both partitions simultaneously, which suggests a degree of variance between the private and public distributions and highlights the importance of evaluating models across both splits before drawing final conclusions. Among the linear classifiers, intfloat/multilingual-e5-large + LR shows the most balanced behaviour, scoring 0.36670 (private) and 0.40995 (public), making it one of the most stable configurations across both partitions.

Effect of the Classifier

Logistic Regression (LR) emerges as the most consistent classifier across embedding backbones, with all five LR configurations clearly surpassing the baseline. SVM yields the single best private result but exhibits higher variance depending on the embedding used. XGBoost produces the strongest public score with bge-m3, yet collapses dramatically with LaBSE (0.09605 / 0.09797), mDeBERTa (0.07422 / 0.07253) and nomic (0.07422 / 0.07253), where it performs well below the random baseline. Random Forest, in turn, never reaches competitive scores: only the BAAI/bge-m3 + Random Forest combination clearly exceeds the baseline, while the remaining configurations fall close to or below it.

Effect of the Embedding Model

Intfloat/multilingual-e5-large and BAAI/bge-m3 are the most robust embedding backbones, producing the top scores on private and public sets respectively and behaving reliably with linear classifiers. nomic-ai/nomic-embed-text-v1.5 is competitive when paired with SVM but unstable with tree-based methods. sentence-transformers/LaBSE and microsoft/mdeberta-v3-base show the weakest overall behaviour, particularly when combined with XGBoost or Random Forest, where their performance degrades to near-zero F1 levels. This pattern suggests that the geometry of these embedding spaces is better suited to margin-based or linear decision boundaries than to axis-aligned splits used by tree ensembles. Private vs Public Discrepancies Several configurations show noticeable gaps between private and public scores, in both directions. For instance, BAAI/bge-m3 + XGBoost improves by more than 11 points from private to public (0.32393 $\rightarrow$ 0.43948), while intfloat/multilingual-e5-large + SVM moves in the opposite direction (0.42063 $\rightarrow$ 0.33455). These discrepancies point to a degree of overfitting or distributional shift between partitions, and reinforce the need to select final submissions based on robustness across both splits rather than peak performance on a single one.

The experiments suggest that linear classifiers (LR and SVM) combined with strong multilingual sentence encoders (E5-large, bge-m3) provided the most reliable performance under the evaluated experimental conditions, consistently outperforming the random baseline by a wide margin. Tree-

Table 2
Internal comparison vs Baseline

Method	Private F1 Score	Public F1 Score
BAAI/bge-m3 + LR	0.34418	0.34410
intfloat/multilingual-e5-large + LR	0.36670	0.40995
sentence-transformers/LaBSE + LR	0.24870	0.36493
microsoft/mdeberta-v3-base + LR	0.34215	0.29533
nomic-ai/nomic-embed-text-v1.5 + LR	0.36190	0.30182
BAAI/bge-m3 + SVM	0.34068	0.30491
intfloat/multilingual-e5-large + SVM	**0.42063**	0.33455
sentence-transformers/LaBSE + SVM	0.31879	0.33903
microsoft/mdeberta-v3-base + SVM	0.36099	0.32997
nomic-ai/nomic-embed-text-v1.5 + SVM	0.36759	0.42900
BAAI/bge-m3 + XGBoost	0.32393	**0.43948**
intfloat/multilingual-e5-large + XGBoost	0.25464	0.27862
sentence-transformers/LaBSE + XGBoost	0.09605	0.09797
microsoft/mdeberta-v3-base + XGBoost	0.07422	0.07253
nomic-ai/nomic-embed-text-v1.5 + XGBoost	0.07422	0.07253
BAAI/bge-m3 + Random Forest	0.35182	0.30381
intfloat/multilingual-e5-large + Random Forest	0.20676	0.19039
sentence-transformers/LaBSE + Random Forest	0.20676	0.19039
microsoft/mdeberta-v3-base + Random Forest	0.19364	0.22404
nomic-ai/nomic-embed-text-v1.5 + Random Forest	0.22612	0.21266
Baseline(Random submission)	0.21036	0.20980

based ensembles, by contrast, were highly sensitive to the choice of embedding model and, in some configurations, exhibited substantial performance degradation. The absence of a single configuration consistently dominating both the private and public partitions further suggests that ensemble strategies or more robust model-selection procedures could yield additional improvements over any individual pipeline evaluated in this work.

4. Conclusions

In this work we have presented a hybrid system for the automatic classification of CEFR proficiency levels in texts produced by learners of Spanish as a Foreign Language, developed in the framework of the NivELE 2026 shared task at IberLEF. Our proposal is built on the hypothesis that proficiency-level determination depends jointly on the distributional semantics of the text and on a set of explicit linguistic indicators that the applied linguistics literature has shown to correlate with communicative competence. The implemented architecture integrates both sources of information through an early-fusion strategy and feeds the resulting representation into a battery of classical supervised classifiers, which are systematically compared by Macro F1 on a stratified validation partition. The experimental analysis allows us to draw several conclusions of methodological and practical relevance. First, the proposed pipeline consistently outperforms the random baseline (F1 $\approx$ 0.21) in the vast majority of configurations, which confirms that the combination of multilingual embeddings and explicit linguistic features extracts a meaningful signal for the task of CEFR-level discrimination. The best results — 0.42063 F1 on the private partition with intfloat/multilingual-e5-large + SVM, and 0.43948 F1 on the public partition with BAAI/bge-m3 + XGBoost — approximately double baseline performance, although they also reveal that there is still ample room for improvement before the ceiling of the task is reached. Second, the choice of embedding-classifier pairing is far from neutral.

Linear classifiers (Logistic Regression and SVM) showed the most consistent behaviour across the evaluated embedding backbones, while tree-based ensembles (XGBoost and Random Forest) display markedly higher variance and, in some combinations (LaBSE, mDeBERTa and Nomic with XGBoost),

collapse to F1 values well below random performance. This pattern suggests that the geometry of the embedding spaces explored is better suited to margin-based or linear decision boundaries than to the axis-aligned splits exploited by tree ensembles, an observation aligned with prior findings on sentence-level representations from contrastively trained encoders. Third, no single configuration dominates simultaneously on the private and public partitions, and several pipelines exhibit substantial gaps between them in both directions. This discrepancy points to a degree of distributional shift or partition-dependent overfitting and reinforces the need to select final submissions based on cross-partition robustness rather than on peak performance on a single split. In this respect, intfloat/multilingual-e5-large + Logistic Regression stands out as one of the most stable pipelines, with comparable performance on both subsets.

Finally, the inclusion of twenty explicit linguistic features computed through regular expressions — covering length, lexical diversity, morphosyntactic complexity, discursive cohesion and genre conventions — provides an interpretable and lightweight complement to dense embeddings, without introducing dependencies on heavy morphosyntactic analyzers. This design keeps the pipeline reproducible and easy to deploy in real placement scenarios, where computational constraints and processing time are decisive factors for language centers. However, since only the complete hybrid architecture was evaluated in this study, the individual contribution of the linguistic feature set could not be isolated. Future work will therefore include dedicated ablation experiments comparing embedding-only, feature-only and hybrid configurations.

As future work, we plan to (i) extend the system with fine-tuned Transformer encoders specifically adapted to ELE corpora, (ii) explore ordinal-aware loss functions that explicitly exploit the natural ordering of CEFR levels and penalize non-adjacent confusions more strongly, (iii) enrich the linguistic-feature module with morphosyntactic information derived from dependency parsers and from error annotations typical of learner corpora, and (iv) investigate calibration and stacking strategies over the per-class probability outputs of the best individual pipelines.

Beyond the specific results of NivELE 2026, we believe that the proposed framework — combining multilingual sentence embeddings with lightweight CEFR-oriented linguistic indicators derived from complexity, accuracy and fluency (CAF) analysis — constitutes a reproducible and practically deployable baseline for future research on automatic proficiency assessment in Spanish as a Foreign Language.

Declaration on Generative AI

During the preparation of this work, the authors used Claude(Opus 4.7) and Grammarly in order to: Grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. Bonet-Jover, J. A. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[2] 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).

[3] S. C. Fanni, M. Febi, G. Aghakhanyan, E. Neri, Natural language processing, in: Introduction to Artificial Intelligence, Springer International Publishing, Cham, 2023, pp. 87–99.

[4] P. Shetty, R. Udhayakumar, A. Patil, M. Manwal, P. S. Vadar, Application of natural language

processing (NLP) in machine learning, in: 2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE), IEEE, 2023, pp. 949–957.
[5] A. K. Joshi, Natural language processing, Science 253 (1991) 1242–1249.
[6] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in: International Conference on Machine Learning, PMLR, 2015, pp. 957–966.
[7] Q. Liu, M. J. Kusner, P. Blunsom, A survey on contextual embeddings, arXiv preprint arXiv:2003.07278 (2020).
[8] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 878–891.
[9] Z. Nussbaum, J. X. Morris, B. Duderstadt, A. Mulyar, Nomic embed: Training a reproducible long context text embedder, arXiv preprint arXiv:2402.01613 (2024). URL: https://arxiv.org/abs/2402.01613. doi:10.48550/arXiv.2402.01613.
[10] J. P. Gujjar, H. P. Kumar, Text similarity technique using BGE-M3, in: 2025 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE), IEEE, 2025, pp. 1–7.
[11] A. S. Berahas, J. Nocedal, M. Takáč, A multi-batch L-BFGS method for machine learning, Advances in Neural Information Processing Systems 29 (2016).
[12] S. Han, C. Qubo, H. Meng, Parameter selection in SVM with RBF kernel function, in: World Automation Congress 2012, IEEE, 2012, pp. 1–4.
[13] T. Joachims, Training linear SVMs in linear time, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2006, pp. 217–226. doi:10.1145/1150402.1150429.
[14] S. Ramraj, N. Uzir, R. Sunil, S. Banerjee, Experimenting XGBoost algorithm for prediction and classification of different datasets, International Journal of Control Theory and Applications 9 (2016) 651–662.
[15] A. Parmar, R. Katariya, V. Patel, A review on random forest: An ensemble classifier, in: International Conference on Intelligent Data Communication Technologies and Internet of Things, Springer International Publishing, Cham, 2018, pp. 758–763.

A. Online Resources

The results of the NIVELE 2026 are available via:

- NIVELE 2026,
- NIVELE 2026 Kaggle,
- NIVELE 2026 Code