Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 251-263 recibido 05-12-2025 revisado 02-02-2026 aceptado 18-02-2026

Exploring the Impact of Linguistic Features on Pre-trained Models for Machine-generated Text Detection in Spanish

Explorando el Impacto de las Características Lingüísticas en Modelos Preentrenados para la Detección de Texto Generado por Máquinas en Español

Javier Alonso Villegas Luis[1], **Marco Antonio Sobrevilla Cabezudo**[12]
[1]Research Group on Artificial Intelligence, Pontificia Universidad Católica del Perú
[2]Aveni
{alonso.villegas, msobrevilla}@pucp.edu.pe

Abstract: Detecting machine-generated Spanish text remains challenging across domains and generators. Pre-trained models like RoBERTa provide strong contextual embeddings but often underperform on human-authored texts and are sensitive to domain shifts. In this work, we integrate linguistic features from PUCP-Metrix—covering lexical, syntactic, semantic, psycholinguistic, and cohesion properties—with pre-trained models. We evaluate feature-based classifiers, fine-tuned RoBERTa, hybrid models, and ensembles on the AuTexTification dataset. Hybrid models improve human-text detection (F1 65.49 vs. 60.74 for RoBERTa) and machine-text classification (F1 81.76), while a voting ensemble achieves the highest macro-F1 (74.75) and strongest robustness. Analyses indicate linguistic features provide stable, interpretable anchors, reducing overfitting and enhancing generalization across LLM outputs. Results demonstrate that combining linguistic and pre-trained models yields a robust solution for Spanish machine-generated text detection.
Keywords: Machine-generated text detection, linguistic features, pre-trained models.

Resumen: La detección de texto generado por máquinas en español sigue siendo un desafío. Modelos preentrenados como RoBERTa ofrecen representaciones contextuales potentes, pero a menudo tienen un desempeño inferior en textos humanos y son sensibles a cambios de dominio. En este trabajo, integramos características lingüísticas de PUCP-Metrix—que abarcan propiedades léxicas, sintácticas, semánticas, psicolingüísticas y de cohesión—con modelos preentrenados. Evaluamos clasificadores basados en características, RoBERTa afinado, modelos híbridos y ensamblajes en el dataset AuTexTification. Los modelos híbridos mejoran la detección de texto humano (F1 65.49 vs. 60.74 de RoBERTa) y la clasificación de texto máquina (F1 81.76), mientras que un ensamblaje por votación logra el macro-F1 más alto (74.75) y la mayor robustez. Los análisis muestran que las características lingüísticas actúan como anclas estables e interpretables, reduciendo el sobreajuste y mejorando la generalización sobre las salidas de múltiples LLMs. Los resultados demuestran que combinar características lingüísticas con modelos preentrenados ofrece una solución robusta para la detección de texto generado por máquinas en español.
Palabras clave: Detección de texto generado por máquina, características linguísticas, modelos pre-entrenados.

1 Introduction

Recent advances in large language models (LLMs) have significantly improved the fluency and realism of machine-generated text, enabling models to mimic human writing styles (Uchendu et al., 2020). This makes distinguishing human and machine text increasingly difficult and raises concerns across domains such as misinformation, academic integrity, and copyright (Zhou et al., 2024).

ISSN 1135-5948 DOI 10.26342/2026-76-17

Most research on machine-generated text detection has focused on English, but shared tasks such as AuTexTification 2023 (Sarvazyan et al., 2023) and IberAuTexTification 2024 (Sarvazyan et al., 2024) highlight challenges in Spanish. While transformer-based models are strong baselines, results from these tasks show that combining them with linguistic features—such as perplexity, lexical richness, morpho-syntactic cues, or stylometric patterns (Alonso Simón et al., 2023; Wu, Huang, and others, 2023; Espin-Riofrio, Ortiz-Zambrano, and Montejo-Ráez, 2024; Valdez-Valenzuela et al., 2024)—often improves performance by capturing structural and stylistic regularities that pure neural embeddings may overlook.

Beyond individual features, comprehensive repositories such as MultiAzterTest (Bengoetxea and Gonzalez-Dios, 2021) and PUCP-Metrix (Luis and Cabezudo, 2025) provide rich linguistic metrics covering syntax, morphology, lexical diversity, and style. In particular, PUCP-Metrix has already been applied to machine-generated text detection, demonstrating its practical value.

Despite this progress, the role of linguistic information within pre-trained model pipelines remains underexplored. Most shared-task systems fuse features superficially without analyzing which properties drive improvements or how stable they are across domains. This gap motivates a deeper evaluation of linguistic feature integration for Spanish, focusing on how they interact with pretrained representations, and when they improve generalization.

To address this gap, we study machine-generated text detection in Spanish, focusing on how to combine explicit linguistic knowledge with pre-trained models. We systematically evaluate hybrid architectures that integrate linguistic features and analyze their strengths and limitations. Results show consistent performance gains, especially for identifying human-authored text.

In summary, the main contributions of this work are:

- Exploratory analysis of linguistic features in combination with pre-trained models: We systematically investigate how to best combine linguistic knowledge with pre-trained models, assess the contribution of these features to model robustness and performance, and identify the limitations of the linguistic features.
- Hybrid and ensemble architectures for Spanish text detection: We propose and evaluate models that integrate explicit linguistic features with pre-trained language model representations, demonstrating consistent performance gains, particularly for identifying human-authored text.

2 Related Work

Research on machine-generated text detection in Spanish has grown rapidly, with the AuTexTification Shared Task (Sarvazyan et al., 2023) showcasing diverse approaches. While fine-tuned transformers dominated, several works showed that linguistic features remain valuable complements. For instance, Alonso Simón et al. (2023) integrated linguistic knowledge into transformers, Mikros et al. (2023) combined stylometric features with ensembles, and others explored syntax-aware learning (Wu, Huang, and others, 2023) or surprisal-based signals (Przybyła, Duran-Silva, and Egea-Gómez, 2023), highlighting the role of structural cues.

The IberAuTexTification 2024 shared task (Sarvazyan et al., 2024) extended this line of work to a multilingual Iberian setting. Several teams combined morpho-syntactic, syntactic, and stylometric features with multilingual transformers (Espin-Riofrio, Ortiz-Zambrano, and Montejo-Ráez, 2024; Guo et al., 2024; Valdez-Valenzuela et al., 2024), analyzing signals such as sentence length, punctuation, lexical richness, and perplexity. As in AuTexTification, results confirmed that hybrid approaches improve robustness in multilingual and cross-domain scenarios.

Beyond Spanish, recent multilingual and English benchmarks have explored hybrid detection approaches, including the GenAI Content Detection Task (Wang et al., 2025), which evaluates detectors across multiple languages and generators. These studies similarly report improvements when combining linguistic and neural signals, reinforcing the cross-lingual relevance of hybrid strategies.

Overall, machine-generated text detection methods fall into three groups: (i) neural detectors based on transformer embeddings, (ii) stylometric and linguistic feature models, and (iii) hybrid approaches. We contribute

to the third by systematically evaluating linguistic feature integration and robustness to domain shift in Spanish.

3 Experimental Setup

3.1 Dataset

We used the dataset from the *AuTexTification* challenge (Sarvazyan et al., 2023), designed to encourage models that generalize well by including human- and machine-generated texts across diverse domains and styles. The texts cover domains such as news, tutorials, and tweets; Detailed statistics are provided in Table 1.

Domain	Generated	Human	Total
Legal	4,846	4,358	9,204
News	5,514	5,223	10,737
Reviews	5,695	3,697	9,392
Tweets	5,739	5,634	11,373
How-to	5,690	5,795	11,485

Table 1: Number of texts per domain for the *AuTexTification* dataset.

Our experiments focused on the Machine-generated Text Detection task, which involves identifying whether a text was written by a human or generated by a machine. The dataset includes 24,707 human-generated and 27,484 machine-generated instances, totaling 52,191 examples.

Human texts were collected from publicly available datasets, including MultiEURLEX (Chalkidis, Fergadiotis, and Androutsopoulos, 2021) (legal), XLSUM/MLSUM (Hasan et al., 2021; Scialom et al., 2020) (news), COAR/COAH (Molina-González et al., 2014) (reviews), XLM-Tweets (Barbieri, Espinosa Anke, and Camacho-Collados, 2022) and TSD (Leis et al., 2019) (tweets), and WikiLingua (Ladhak et al., 2020) (how-to articles). Machine-generated texts were produced using six large language models: three from the BLOOM family (BLOOM-1B7[1], BLOOM-3B[2], BLOOM-7B1[3]) and three from the GPT-3 family (babbage, curie, text-davinci-003).

Another key aspect of this task is its emphasis on cross-domain generalization. Models are trained on texts from the Legal, Tweets, and How-to domains, but they are evaluated on two domains they never see during training—News and Reviews. This setup introduces a clear distribution shift and allows to assess how robust systems are when faced with new writing styles and contexts.

3.2 Extracted Linguistic Features

We extracted the full set of linguistic features available in `PUCP-Metrix` (Luis and Cabezudo, 2025) (182 features), a tool designed to compute a wide range of linguistic metrics for Spanish text analysis.[4] The extracted features are grouped into the following categories[5]:

- **Descriptives** (27): general text statistics, e.g., word, sentence, and paragraph counts, and sentence lengths.
- **Lexical Diversity** (22): vocabulary variety measures, e.g., type-token ratios, noun/verb/adjective/adverb density, MLTD, vocd (McCarthy Philip M, 2010).
- **Readability** (7): difficulty metrics, e.g., Flesch Grade, Brunet Index, Gunning Fog, SMOG, Szigriszt-Pazos.
- **Syntactic Complexity** (12): structural measures, e.g., sentence clause counts, minimal edit distances.
- **Psycholinguistics** (30): word-level human processing properties, e.g., concreteness, imageability, familiarity, age of acquisition, valence, arousal (Duchon et al., 2013; Stadthagen-Gonzalez et al., 2017).
- **Word Information** (24): counts of nouns, verbs, adjectives, adverbs, and content words.
- **Referential Cohesion** (12): text interconnections, e.g., noun/argument/stem/content word/anaphor overlap.
- **Textual Simplicity** (4): sentence length ratios (short/medium/long/very long).

[1]Available at `https://huggingface.co/bigscience/bloom-1b7`

[2]Available at `https://huggingface.co/bigscience/bloom-3b`

[3]Available at `https://huggingface.co/bigscience/bloom-7b1`

[4]The list of metrics and their definitions can be found at `https://github.com/iapucp/pucp-metrix`.

[5]For brevity, we list only representative metrics per category rather than the full set.

- **Semantic Cohesion** (8): semantic relatedness, e.g., LSA overlaps between sentences/paragraphs.
- **Word Frequency** (16): Zipf-based measures, e.g., counts of rare nouns/verbs/adverbs/content words, mean frequency.
- **Syntactic Pattern Density** (14): density of syntactic elements, e.g., noun/verb phrases, negative expressions, conjunctions.
- **Connectives** (6): incidence of logical, temporal, causal, adversative, and additive connectives.

3.3 Single Approaches

3.3.1 Feature-based models

In experiments reported by Luis and Cabezudo (2025), the usefulness of the collected features to detect machine-generated text was evaluated using several machine learning algorithms. Linguistic features were computed for both the training and test splits of the dataset, and applied directly to the raw texts without any preprocessing. The classifiers tested included `Support Vector Machines` (SVM), `Logistic Regression`, `Random Forest`, and `XGBoost`. Model training was performed using 5-fold cross-validation to determine optimal hyperparameters, with 20 % of the training data reserved for validation. We use the performance reported in the work for evaluation purposes.

3.3.2 Transformer-based model

In addition to feature-based models, we also report the transformer approach by fine-tuning a pre-trained RoBERTa model (Liu et al., 2019) on the *AuTexTification* dataset .

More specifically, the `RoBERTa` model used was pre-trained on data extracted from the *National Library of Spain*, the largest Spanish corpus known to date (Fandiño et al., 2022)[6]. This pre-trained model was also a baseline in the AuTexTification shared-task (named `FT-RoBERTa-BNE` in experiments) and the authors of `PUCP-Metrix` replicated the same idea with some slightly differences. The authors used the `[CLS]` token's final hidden state output from the `RoBERTa` model as the input to a subsequent fully connected neural network. Figure 1 shows the complete architecture of this model. To enhance regularization and prevent overfitting during training, a dropout layer was incorporated before the hidden layer.

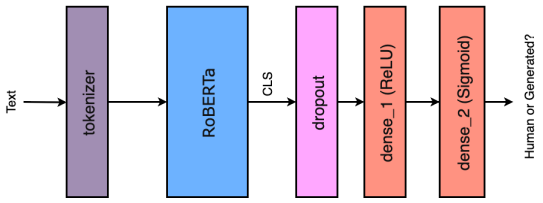

Figure 1: `FT-RoBERTa-BNE` architecture: `RoBERTa` followed by a fully connected neural network.

Hyperparameter optimization was performed using a cross-validation strategy leveraged by the Optuna library[7]. The model was then trained for 3 epochs using the Adam optimizer (Kingma and Ba, 2014), with 20% of the training dataset reserved for validation. Table 2 shows the optimal hyperparameters used to fine-tune the model.

Parameter	`FT-RoBERTa-BNE`	`FT Features + RoBERTa-BNE`
Batch size	32	16
Epochs	3	5
Optimizer	Adam	Adam
Learning rate	0.00002474	0.00002767
Dropout	0.3	0.1
Dense sizes	768	16/64/64

Table 2: Hyperparameters used in the fine-tuning of the `FT-RoBERTa-BNE` model and `FT Features + RoBERTa-BNE`.

3.4 Hybrid models

3.4.1 Linguistic features + RoBERTa classifier

Combining linguistic features with contextualized word embeddings from transformer language models is a prevalent approach to tackling challenges of this nature (Tyo, Dhingra, and Lipton, 2022). Adopting this strategy, we combined the linguistic features extracted in the Section 3.2 with a pre-trained `RoBERTa` model to improve its classification performance.

[6]Available at `https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne`

[7]Available at `https://github.com/optuna/optuna`.

To achieve this integration, both the linguistic features and the `RoBERTa` embeddings were first projected into a lower-dimensional space using dense layers. The resulting transformed representations were then concatenated into a single vector. This combined vector served as input to a classification network consisting of two dense layers. In addition, dropout was applied to this intermediate vector to mitigate overfitting during training. The architecture of the hybrid model is shown in Figure 2.

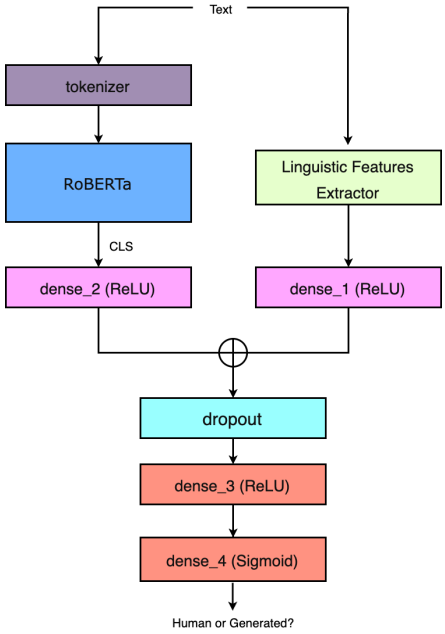

Figure 2: Proposed hybrid model using a pre-trained `RoBERTa` model and linguistic features.

The model was trained using the Adam optimizer and the hyperparameters listed in Table 2 and all the parameters, including `RoBERTa` parameters, were fine-tuned (the model is named `FT Features + RoBERTa-BNE` in experiments). These parameters were also tuned using Optuna's cross-validation.

3.4.2 Linguistic Features + RoBERTa logits

In this approach (named `Features + RoBERTa logits`), we leverage the probabilistic output of the `FT-RoBERTa-BNE`. This output is concatenated with the set of linguistic features previously developed into a single vector. This combined vector is then used as input to an XGBoost classifier. Similarly to the training process of previous feature-based models, cross-validation is used to optimize the hyperparameters of the XGBoost model.

3.5 Ensemble models

Ensemble models leverage the collective knowledge of individual models to achieve improved classification performance and robustness compared to using a single model. For this reason, we explored the combination of our feature-based classifiers with the fine-tuned `RoBERTa` model (`FT-RoBERTa-BNE`) from the previous section. To accomplish this, we employed two popular ensemble methods:

3.5.1 Stacking Classifier

The `Stacking Classifier` involves concatenating the feature representations derived from our feature-based classifiers with the output of `FT-RoBERTa-BNE`'s final layer. This combined feature vector serves as input to an additional classifier, which learns to make the final prediction based on the combined information. In our case, a simple Logistic Regression model was employed as this final classifier. The models used for this classifier were Logistic Regression, Random Forest, XGBoost and the fine-tuned RoBERTa. We did not include SVM in the voting classifier because it had the lowest cross-validation performance.

3.5.2 Voting Classifier

The `Voting Classifier` combines the predictions of multiple base classifiers. We utilized a *Hard Voting* strategy, where the final prediction is determined by the class that receives the majority of votes from the individual base models. the models used for this classifier were the same as the ones used for the `Stacking Classifier`.

4 Results and Discussion

Table 3 presents the performance of the previously described models on the Machine-generated text detection task. More specifically, we report the Precision, Recall and F1 for each category (Human and Machine) and Macro F1. In addition, we report the performance of the `FT-RoBERTa-BNE` baseline that was used in the AuTexTification shared-task (`FT-RoBERTa-BNE-AuTexTification` in the Table) and the best model reported in the

Model	Human			Machine			F1
	P	R	F1	P	R	F1	
Single Feature-based Models							
`Logistic Regression`	71.09	55.93	62.61	70.02	81.90	75.49	69.05
`XGBoost`	71.34	61.36	65.97	72.33	80.38	76.14	71.06
`Support Vector Machines`	71.04	56.05	62.66	70.06	81.82	75.48	69.07
`Random Forest`	63.57	58.24	60.79	68.85	73.44	71.07	65.93
Pre-trained Models							
`FT-RoBERTa-BNE`	86.28	46.87	60.74	68.99	94.07	79.60	70.17
`FT-RoBERTa-BNE-AuTexTification*`	-	-	-	-	-	-	68.52
Best model*	-	-	-	-	-	-	70.77
Feature + Pre-trained Models							
`FT Features + RoBERTa-BNE`	91.14	51.11	65.49	71.17	96.05	81.76	73.63
`Features + RoBERTa logits`	91.99	44.56	60.04	68.72	96.91	80.42	70.23
`Stacking Classifier`	70.02	58.96	64.01	70.99	79.91	75.18	69.60
`Voting Classifier`	74.54	67.43	70.81	75.91	81.68	78.69	74.75

Table 3: Results on AuTexTification test set. *The authors of the shared-task only provide F1 in the report. McNemar's test on paired predictions using `FT-RoBERTa-BNE` as the baseline. Underlined values indicate statistically significant improvements ($p < 0.01$).

shared-task.[8]

The best overall performance is achieved by the `Voting Classifier`, with a macro F1 score of 74.75. This model demonstrates balanced effectiveness across both human and machine categories, obtaining the highest F1 for human-generated texts (70.81) and a strong F1 for machine-generated ones (78.69). This robust performance highlights the potential of hybrid strategies that combine diverse model outputs rather than relying on a single architecture.

Although our study focuses on Spanish, it is useful to compare it with multilingual research. Recent benchmarks such as the GenAI Content Detection Task (Wang et al., 2025) report macro-F1 scores of 75–85 for hybrid detectors. Our result (macro-F1 74.75) is comparable despite resource and cross-domain challenges, which may suggest that hybrid approaches generalize across languages, although linguistic features remain language-dependent.

When comparing machine learning approaches using linguistic features (`Logistic Regression`, `XGBoost`, `SVM`, and `Random Forest`) to single RoBERTa-based models, we may see that XGBoost achieves a competitive F1 score of 71.06, even surpassing both `FT-RoBERTa-BNE` and `FT-RoBERTa-BNE-AuTexTification` and the best model, showing the relevance of linguistic features for this task. Besides, the general trend shows that the `RoBERTa` models tend to perform better on machine-generated texts (79.60 F1) than on human-generated ones (60.74 F1). Even more, *FT-RoBERTa-BNE* presents the lowest recalL on Human-authored texts (46.87). In contrast, feature-based models exhibit less instability between performance on both human and machine texts.

How do linguistic features contribute to performance improvement? Some hybrid systems that combine linguistic features with `RoBERTa` representations outperform models that rely on a single family of signals. While fine-tuned `RoBERTa` (`FT-RoBERTa-BNE`) achieves a strong Machine F1 of 79.60, it suffers from a low Human Recall of 46.87, resulting in a modest Human F1 of 60.74. This imbalance indicates a bias toward predicting the machine class. However, incorporating linguistic features effectively mitigates this issue: the `FT Features + RoBERTa-BNE` model increases Human F1 to 65.49 (+4.75 over `FT-RoBERTa-BNE`) while simultaneously improving Machine F1 from 79.60 to 81.76. These improvements demonstrate that handcrafted stylistic cues help recover human texts that `RoBERTa` alone mis-

[8]Comparisons with prior systems are constrained by the reporting format of the shared-task overview, where only aggregate metrics (e.g., macro-F1) are available for the top-performing system. This limits fine-grained comparisons but ensures consistency with the standard benchmark evaluation.

classifies, while still preserving—and even enhancing—RoBERTa's strong ability to detect machine-generated content.

Beyond correcting class-specific weaknesses, hybrid approaches may also yield more stable performance. Ensemble-based hybrids such as the `Voting Classifier` achieve the strongest overall results, reaching 74.75 F1, outperforming both the best feature-only system (`XGBoost`, 71.06 F1) and the `FT-RoBERTa-BNE` baseline (70.17 F1). This gain of +4.58 F1 over `FT-RoBERTa-BNE` suggests that hybridization acts as a form of regularization: linguistic features anchor the model in interpretable stylistic patterns, while `RoBERTa` embeddings capture richer semantic and contextual signals. As a result, hybrid models avoid the tendencies observed in `RoBERTa` only systems and achieve more balanced, reliable discrimination across human and machine classes—an essential property for authorship verification scenarios where asymmetric errors can lead to systematic biases.

It is worth noting that, while the inclusion of linguistic features improves human-text detection, the absolute performance remains moderate. This reflects a broader pattern observed in prior detection benchmarks (Sarvazyan et al., 2023), where models often exhibit a bias toward predicting machine-generated text. Our findings suggest that linguistic features partially mitigate this tendency, improving class balance without sacrificing machine-text detection performance.

Different from `FT Features + RoBERTa-BNE` and `Voting Classifier`, the model `Features + RoBERTa logits` performs less strongly (70.23 F1), nearly matching the `RoBERTa` baseline. This suggests that shallow fusion of logits may underutilize the complementary information provided by features, whereas approaches that allow deeper interaction—either during fine-tuning or through ensemble aggregation—unlock more robust improvements. The `Stacking Classifier` shows the lowest performance (69.60 F1), indicating potential overfitting in the meta-learner and highlighting that hybrid architectures must carefully balance model complexity with the stability of feature–representation interactions.

How do pre-trained models contribute to hybrid models? Given that `Voting Classifier` gets the highest score and it combines machine learning models and the fine-tuned `RoBERTa` model, we try to estimate the contribution of the `RoBERTa` model in the voting schema. More specifically, we estimate its relevance by removing the `RoBERTa` model from the ensemble. Table 4 shows the performance on the validation set and also on the test set. Overall, the full `Voting Classifier` achieves the highest validation and test F1 score (85.67 and 74.75, respectively) and outperforms the machine learning model ensemble variant (83.43 and 71.33 F1).

In particular, removing `FT-RoBERTa-BNE` induces opposite shifts in human–machine precision/recall. Human recall decreases from 90.34 to 83.97 (–6.37), while machine precision also drops from 90.02 to 84.73 (–5.29). At the same time, the metrics on the opposite axes—human precision and machine recall—change minimally (Human Precision: from 81.85 to 82.12, +0.27; Machine Recall: from 81.31 to 82.94, +1.63).

This pattern may suggest that `FT-RoBERTa-BNE` contributes primarily to capturing harder positive signals: it helps the ensemble recover more true human texts (high human recall) and avoid mislabeling machine-generated content as human (high machine precision). When this component is removed, the classifier compensates by slightly improving the complementary axes (human precision and machine recall), but these gains are small and insufficient to offset the substantial recall/precision losses on the dominant axes.

Another aspect that we analyzed was the effect of keeping `RoBERTa-BNE` frozen in the `FT Features + RoBERTa-BNE` pipeline to understand how pre-trained representations influence cross-domain generalization. To do this, we kept the fine-tuned weights of the `FT-RoBERTa-BNE` model frozen.

Table 5 shows the results on both validation and test sets. Results in the validation set show that freezing RoBERTa slightly improves F1 scores for both human (93.95 vs. 92.30) and machine (94.22 vs. 92.69), leading to a modest increase in overall macro-F1 (94.09 vs. 92.49). This indicates that frozen `FT-RoBERTa-BNE` helps stabilize in-domain performance.

On the test set, however, freezing `RoBERTa` significantly harms performance on human-authored texts. Human recall drops from 51.11 to 33.24, and F1 falls from 65.49 to

	Validation							Test
	Human			Machine			F1	F1
	P	R	F1	P	R	F1		
`Voting Classifier`	81.85	90.34	85.89	90.02	81.31	85.45	85.67	74.75
`Voting Classifier - FT-RoBERTa-BNE`	82.12	83.97	83.04	84.73	82.94	83.83	83.43	71.33

Table 4: Results for `Voting Classifier` with/without `FT-RoBERTa-BNE` on the validation and test sets.

49.19, showing that the model struggles to generalize to unseen writing styles. Machine text detection sees a small increase in recall (from 96.05 to 98.48) but a decrease in precision (from 71.17 to 64.96), resulting in a lower overall macro-F1 (78.28 vs. 81.76). These results suggest that while freezing RoBERTa can preserve in-domain stability, it increases sensitivity to domain-specific patterns and reduces robustness to unseen domains.

How robust are hybrid models under domain shift between validation and test? As discussed earlier, the Machine-generated Text detection task introduces a deliberately challenging cross-domain scenario, making it an effective test for overfitting. Because validation and test data come from different domains, models trained on AuTexTification are forced to rely on domain-invariant signals rather than superficial patterns tied to specific writing styles. Figure 3 shows the validation–test performance gaps for both human- and machine-generated text classes.

`FT RoBERTa-BNE` exhibits the largest deterioration under domain shift, with the greatest performance drops observed for human-authored texts ($\Delta = 47.67$ for Human Recall and 32.18 for Human F1) and substantial reductions in overall macro-F1 ($\Delta = 22.87$). This pattern reflects a systematic miscalibration: the model becomes overly confident in predicting machine-generated texts and fails to generalize to unseen human-writing styles. Interestingly, Recall performance for machine-generated texts slightly improves under the shift (Δ = -2.42), highlighting that fine-tuning pushes `RoBERTa` towards a machine-biased decision boundary that does not transfer to new domains.

This behavior may also suggest that `FT-RoBERTa-BNE` struggles to capture stylistic and content variations inherent in different human-written domains. However, when texts are partially converted into machine-generated content—as in AuTexTification, where segments of human text are completed by a Large Language Models—they inherit consistent generation patterns. In these cases, `FT-RoBERTa-BNE` is able to generalize effectively, successfully detecting machine-generated style even in domains it has never seen before.

In the case of hybrid models, we can see that they mitigate these overfitting effects to varying degrees. Adding features during fine-tuning (`FT Features + RoBERTa-BNE`) reduces the macro-F1 gap ($\Delta = 18.86$), improving by nearly 4 over `FT-RoBERTa-BNE`, and yields a smaller Human F1 drop ($\Delta = 26.81$). This indicates that linguistic features act as domain-stable anchors, helping the model avoid over-specializing to the training domain. However, simply concatenating features with `RoBERTa` logits and training XGBoost (`Features + RoBERTa logits`) on top amplifies instability: this configuration shows the largest Human recall loss ($\Delta = 50.53$) and the worst macro-F1 delta (24.18), suggesting that tree models can overfit to already-biased logits and thus exacerbate domain mismatch.

By contrast, ensemble-based hybrids models offer the strongest robustness under domain shift. The `Stacking Classifier` reduces the macro-F1 delta to 14.18, while the `Voting Classifier` achieves the smallest observed degradation ($\Delta = 10.92$) and the lowest Human F1 drop ($\Delta = 15.08$). These ensembles work because they aggregate diverse inductive biases and reduce variance: when facing unseen textual styles, individual model overfitting does not dominate the final decision. Overall, hybrid approaches—particularly ensembling—meaningfully alleviate overfitting and provide substantially more robust generalization to out-of-domain, unseen classes.

To what extent are linguistic features beneficial for models of varying sizes/capabilities? We evaluate the recall of different strategies for detecting machine-

	Validation							Test						
	Human			Machine			F1	Human			Machine			F1
	P	R	F1	P	R	F1		P	R	F1	P	R	F1	
`FT Features + RoBERTa-BNE`	91.52	93.09	92.30	93.45	91.95	92.69	92.49	91.14	51.11	65.49	71.17	96.05	81.76	73.63
`FT Features + RoBERTa-BNE` (frozen RoBERTa)	92.79	95.15	93.95	95.37	93.10	94.22	94.09	94.58	33.24	49.19	64.96	98.48	78.28	63.74

Table 5: Effect of keeping `RoBERTa-BNE` frozen in `FT Features + RoBERTa-BNE`.

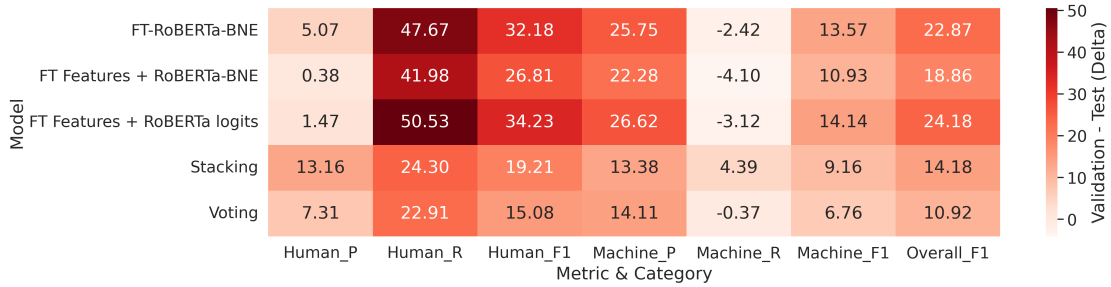

Figure 3: Difference between validation and test sets for `FT-RoBERTa-BNE`, hybrid and ensemble models.

generated text across several Large Language Models (LLMs) used in the task, including Bloom variants (1B7, 3B, 7B1), OpenAI Babbage, Curie, and text-davinci-003. The strategies considered include XGBoost on the linguistic features (`XGBoost`), fine-tuned `RoBERTa` (`FT-RoBERTa-BNE`), and hybrid combinations of RoBERTa and linguistic features (`FT Features + RoBERTa-BNE` and `FT Features + RoBERTa logits`), and ensemble methods (`Stacking` and `Voting classifiers`). Results are shown in Figure 4.

The `XGBoost` model achieves high recall for smaller models such as Bloom-1B7 (96.22) and Babbage (85.76), but performance drops significantly on larger LLMs like Bloom-7B1 (75.54) and text-davinci-003 (66.86), indicating that linguistic feature-based models are effective for smaller or mid-sized models but struggle to generalize to outputs from larger LLMs. `FT-RoBERTa-BNE` improves generalization, yielding consistently high recall across most models, from 88.55 (Bloom-1B7) to 97.10 (text-davinci-003), and outperforming feature-only models on larger or more complex LLMs.

Hybrid models such as `FT features + RoBERTa-BNE` and `FT Features + RoBERTa logits` further improve both recall and stability. In particular, `FT Features + RoBERTa logits` achieves recalls between 96.45 and 97.95, with the highest value on Bloom-3B. These results suggest that integrating complementary information from pre-trained embeddings and linguistic features produces robust detection across all LLMs. In contrast, `Stacking` and `Voting classifiers` show high recall on some smaller models, such as Bloom-1B7 and Babbage, but suffer on larger or more complex models, with recalls dropping to 65.15 (`Stacking`) and 67.89 (`Voting`), suggesting that naive ensembles may fail to capture model-specific patterns.

Finally, to illustrate the difficulty of the task, Figure 5 presents examples from the AuTexTification dataset. The top two instances were correctly classified by both `FT-RoBERTa-BNE` and the `voting classifier`, while the bottom two were misclassified. These cases highlight how challenging it can be to distinguish between human-written and machine-generated text, as both often exhibit similar fluency and stylistic patterns.

5 *Conclusion*

We investigated the integration of linguistic features from PUCP-Metrix with pre-trained models for Spanish machine-generated text detection. Our experiments show that hybrid models consistently outperform feature-only and RoBERTa-only approaches, improving both human-text and machine-text classification, while voting classifier achieves the highest macro-F1 and robustness under domain shifts. Analyses indicate that linguistic features serve as stable, interpretable anchors that reduce overfitting and enhance generalization across outputs from multiple LLMs.

Future work will extend the analysis to more Spanish LLMs and text genres, explore feature selection methods, and better char-

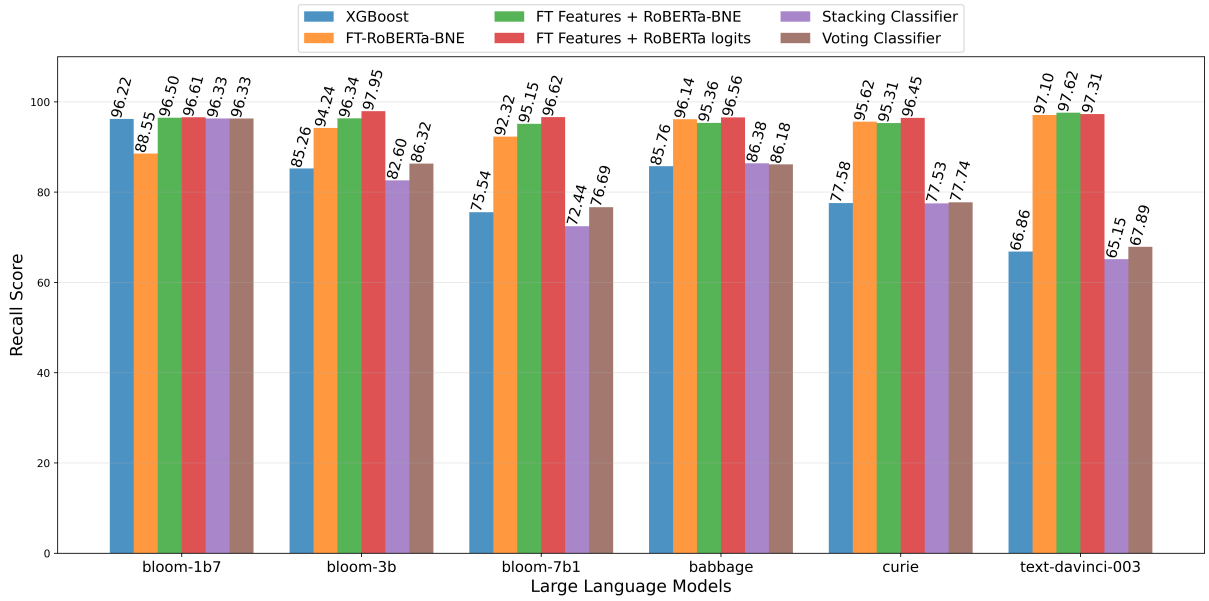

Figure 4: Recall scores for `FT-RoBERTa-BNE`, hybrid, and ensemble models across evaluations on different large language models.

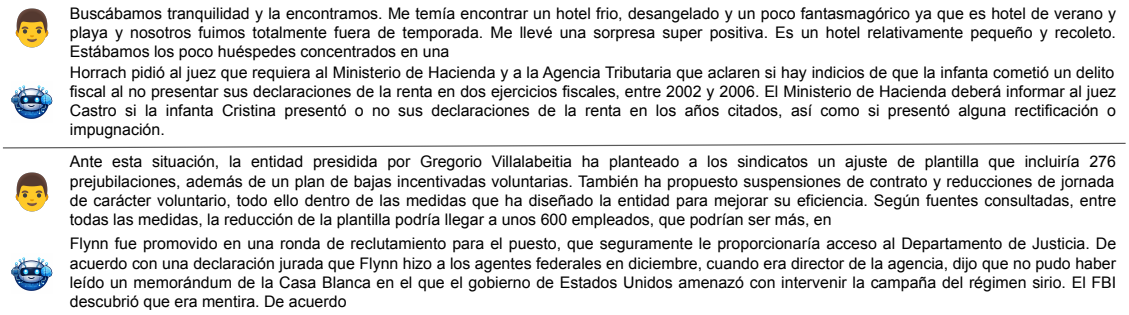

Figure 5: Examples from the AuTexTification dataset. The top two were correctly classified by both `FT-RoBERTa-BNE` and the `voting classifier`, while the bottom two were misclassified.

acterize the role of individual linguistic features. We also plan to investigate multilingual extensions, study how linguistic features can be integrated into LLM-based systems, and conduct error analyses to identify common failure modes and potential improvements.

Limitations

Despite the contributions of this work, several limitations remain:

- As large language models become more human-like, the discriminative value of explicit linguistic features may decrease. We evaluate their current utility and defer longitudinal analyses to future work.
- This work focuses on Spanish, where resources remain limited. Many linguistic metrics depend on language-specific resources, making cross-lingual transfer non-trivial; multilingual extensions remain an open direction.
- Our setup follows the AuTexTification protocol, which evaluates systems by model rather than text length. While length may affect difficulty, a systematic length-based analysis is beyond the scope of this study.
- Feature-level importance was analyzed in prior work (Luis and Cabezudo, 2025). Here, we study how aggregated linguistic signals interact with pretrained models in hybrid settings. Frequency, readability, and cohesion appear most influential, while fine-grained hybrid analysis warrants deeper investigation.

References

Alonso Simón, L., J. Á. Gonzalo Gimeno, A. M. Fernández-Pampillón Cesteros, M. Fernández Trinidad, and M. V. Escandell Vidal. 2023. Using linguistic knowledge for automated text identification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) — AuTexTification Shared Task*, volume 3496, Jaén, Spain, September. CEUR Workshop Proceedings. Paper describing the use of morpho-syntactic, lexical and discourse features for human vs. machine text detection.

Barbieri, F., L. Espinosa Anke, and J. Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June. European Language Resources Association.

Bengoetxea, K. and I. Gonzalez-Dios. 2021. Multiaztertest: a multilingual analyzer on multiple levels of language for readability assessment.

Chalkidis, I., M. Fergadiotis, and I. Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Duchon, A., M. Perea, N. Sebastián-Gallés, A. Martí, and M. Carreiras. 2013. Espal: One-stop shopping for spanish word properties. *Behavior Research Methods*, 45(4):1246–1258.

Espin-Riofrio, C., J. Ortiz-Zambrano, and A. Montejo-Ráez. 2024. Sinai at iberautextification in iberlef 2024: Perplexity metrics and text features for classifying automatically generated text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) — Iber AuTexTification Shared Task*, volume 3756. CEUR Workshop Proceedings. Uses perplexity, phrasing, and punctuation/connector frequency features in supervised classifiers for human vs. machine-generated text detection.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Guo, M., Z. Han, X. Wang, and J. Peng. 2024. Multidimensional text feature analysis: Unveiling the veil of automatically generated text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) — Iber AuTexTification Shared Task*, volume 3756. CEUR Workshop Proceedings. Combines word-level perplexity features and sentence-level features (readability, lexical richness, punctuation) with a pre-trained language model for machine-generated text detection (Macro-F1 = 0.7663 on Subtask 1).

Hasan, T., A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August. Association for Computational Linguistics.

Kingma, D. P. and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ladhak, F., E. Durmus, C. Cardie, and K. McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online, November. Association for Computational Linguistics.

Leis, A., F. Ronzano, M. A. Mayer, L. I. Furlong, and F. Sanz. 2019. Detecting signs

of depression in tweets in spanish: Behavioral and linguistic analysis. *J Med Internet Res*, 21(6), Jun.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Luis, J. A. V. and M. A. S. Cabezudo. 2025. Pucp-metrix: A comprehensive open-source repository of linguistic metrics for spanish.

McCarthy Philip M, J. S. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. In *Behavior Research Methods*, pages 381–392.

Mikros, G., A. Koursaris, D. Bilianos, and G. Markopoulos. 2023. Ai-writing detection using an ensemble of transformers and stylometric features. In *Proceedings of the IberLEF 2023 Workshop — AuTexTification Shared Task*, volume 3496, Jaén, Spain. CEUR Workshop Proceedings. Detection of AI-generated texts using a hybrid of transformer classifiers and stylometric feature ensembles.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. 2014. Cross-domain sentiment analysis using spanish opinionated words. In E. Métais, M. Roche, and M. Teisseire, editors, *Natural Language Processing and Information Systems*, pages 214–219, Cham. Springer International Publishing.

Przybyła, P., N. Duran-Silva, and S. Egea-Gómez. 2023. I've seen things you machines wouldn't believe: Measuring content predictability to identify automatically-generated text. In *Proceedings of the IberLEF 2023 Workshop — AuTexTification Shared Task*, volume 3496, Jaén, Spain. CEUR Workshop Proceedings. Uses predictability metrics (LM likelihood, entropy) plus linguistic features for machine-generated text detection.

Sarvazyan, A. M., J. Á. González, M. Franco-Salvador, F. Rangel, B. Chulvi, and P. Rosso. 2023. Overview of autextification at iberlef 2023: Detection and attribution of machine-generated text in multiple domains. *arXiv preprint arXiv:2309.11285*.

Sarvazyan, A. M., J. Ángel González, F. Rangel, P. Rosso, and M. Franco-Salvador. 2024. Overview of iberautextification at iberlef 2024: Detection and attribution of machine-generated text on languages of the iberian peninsula. *Procesamiento del Lenguaje Natural*, 73(0):421–434.

Scialom, T., P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano. 2020. MLSUM: The multilingual summarization corpus. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November. Association for Computational Linguistics.

Stadthagen-Gonzalez, H., C. Imbault, M. A. Pérez Sánchez, and M. Brysbaert. 2017. Norms of valence and arousal for 14,031 spanish words. *Behavior Research Methods*, 49(1):111–123.

Tyo, J., B. Dhingra, and Z. C. Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *arXiv preprint arXiv:2209.06869*.

Uchendu, A., T. Le, K. Shu, and D. Lee. 2020. Authorship attribution for neural text generation. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395, Online, November. Association for Computational Linguistics.

Valdez-Valenzuela, A., R. L. Zavala-Reyes, V. G. Morales-Murillo, and H. Gómez-Adorno. 2024. The iimasnlp team at iberautextification 2024: Integrating graph neural networks, multilingual llms, and stylometry for automatic text identification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) — Iber AuTexTification Shared Task*, volume 3756. CEUR Workshop Proceedings. Hybrid approach combining graph-based representations, stylometric features, and multilingual LLM embeddings for machine-generated text detection.

Wang, Y., A. Shelmanov, J. Mansurov, A. Tsvigun, V. Mikhailov, R. Xing, Z. Xie, J. Geng, G. Puccetti, E. Artemova, J. Su, M. N. Ta, M. Abassy, K. A. Elozeiri, S. E. D. A. El Etter, M. Goloburda, T. Mahmoud, R. V. Tomar, N. Laiyk, O. Mohammed Afzal, R. Koike, M. Kaneko, A. F. Aji, N. Habash, I. Gurevych, and P. Nakov. 2025. GenAI content detection task 1: English and multilingual machine-generated text detection: AI vs. human. In F. Alam, P. Nakov, N. Habash, I. Gurevych, S. Chowdhury, A. Shelmanov, Y. Wang, E. Artemova, M. Kutlu, and G. Mikros, editors, *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 244–261, Abu Dhabi, UAE, January. International Conference on Computational Linguistics.

Wu, S.-H., H.-Y. Huang, et al. 2023. Syntax-based contrastive learning for automated text identification. In *Proceedings of the IberLEF 2023 Workshop — AuTexTification Shared Task*, volume 3496, Jaén, Spain. CEUR Workshop Proceedings. Introduces a syntax-aware contrastive learning (SCL) method to improve machine-generated text detection.

Zhou, J., H. Müller, A. Holzinger, and F. Chen. 2024. Ethical chatgpt: Concerns, challenges, and commandments. *Electronics*, 13(17).