Building a Lightweight Classifier to Distinguish Closely Related Language Varieties with Limited Supervision: The Case of Catalan vs Valencian

Raúl García Cerdá **María Miró Maestre** **Miquel Canal**
University of Alicante
{raul.gc, maria.miro, mikel.canal}@ua.es

Abstract

Dialectal variation among closely related languages poses a major challenge in low-resource NLP, as their linguistic similarity increases confusability for automatic systems. We introduce the first supervised classifier to distinguish standard Catalan from its regional variety Valencian. Our lightweight approach finetunes a RoBERTa-base model on a manually curated corpus of 20 000 sentences—without any Valencian-specific tools—and achieves 98 % accuracy on unseen test data. In a human evaluation of 90 mixed-variety items per reviewer, acceptance rates reached 96.7 % for Valencian and 97.7 % for Catalan (97.2 % overall). We discuss limitations with out-of-distribution inputs and outline future work on confidence calibration and dialect-aware tokenization. Our findings demonstrate that high-impact dialect classification is feasible with minimal resources.

1 Introduction

Linguistic Background. Valencian is the variety of Catalan spoken in the Valencian Community and is officially recognized as one of its co–official languages. Linguistically, the Acadèmia Valenciana de la Llengua (AVL) "formally acknowledges Valencian as one variant of the common Catalan language" (European Language Equality (ELE), 2022; Acadèmia Valenciana de la Llengua, 2022). However, due to historical and political factors—such as the repression of Catalan during the Franco regime—Valencian has often occupied a minoritized position, surviving mainly in informal domains and facing strong pressure from Spanish (European Language Equality (ELE), 2022). This sociopolitical context is reflected in technology: **Google Translate** and most commercial voice assistants do not distinguish Valencian (they offer only "Catalan") (European Language Equality (ELE), 2022), and Microsoft Office requires a separate "Catalan (Valencian)" pack maintained by Softcatalà (Softcatalà, 2018). The Valencian variety of Catalan is commonly perceived as a regional dialect rather than a distinct linguistic entity, which has led to its underrepresentation in natural language processing (NLP) resources. Despite being an official language in the Valencian Community, Valencian lacks dedicated tools such as lemmatizers, spell checkers, or machine translation systems that treat it independently from standard Catalan. This scarcity of resources positions Valencian as a low-resource language variant in practical computational terms. Although Catalan has benefited from recent advances in language modeling and the availability of large-scale corpora, similar efforts for Valencian are virtually nonexistent.

In this paper, our aim is to contribute to the development of dialect-specific resources by presenting a lightweight binary text classifier capable of distinguishing between standard Catalan and Valencian. We train our model using manually curated data from official public sources and demonstrate that it is possible to obtain accurate and robust results even with limited supervision and minimal preprocessing tools. *To the best of our knowledge, this is the first work to formulate and evaluate the task of discriminating between standard Catalan and Valencian as a supervised classification problem.* Our work is framed within the broader context of dialectal NLP and highlights both the technical challenges and sociolinguistic implications of computationally differentiating closely related language varieties.

2 Related Work

Linguistic features. Catalan and Valencian are Romance languages derived from Vulgar Latin (Martines, 2024), sharing many features with Spanish and French but also showing systematic differences. One example is the present subjunctive: Catalan uses endings in *-i* (e.g., *canti*) (d'Estudis Catalans, 2022), whereas Valencian prefers *-e*

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 7–11, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-002

(e.g., *cante*) (Acadèmia Valenciana de la Llengua, 2006). Another contrast is in feminine possessive pronouns: Catalan *meva/teva/seva* vs. Valencian *meua/teua/seua* (Institut d'Estudis Catalans, 2022; Real Acadèmia de Cultura Valenciana, Secció de Llengua i Lliteratura Valencianes, 2025). Beyond morphology, numerous studies have documented lexical divergences between the two varieties (Wheeler, 2005; Marzà et al., 2006; Lledó, 2011), and ongoing online projects aim to compile them systematically (Idioma Valenciano, 2025; Acadèmia Valenciana de la Llengua, 2025).

Existing resources. Most NLP resources for Catalan do not explicitly handle Valencian. Spellcheckers such as Softcatalà provide a unified Catalan dictionary with a "Valencià" variant (Softcatalà, 2018), and LanguageTool or the SALT platform offer only basic configurations. Open-source MT systems (Apertium, Politraductor) adapt Valencian lexicon, while commercial engines (DeepL, Google Translate) collapse Valencian into Catalan (European Language Equality (ELE), 2022). Morphological analyzers like FreeLing or spaCy are trained for Catalan and must be reused for Valencian, which can miss regional features. State-of-the-art PLMs such as CALBERT and RoBERTa (Projecte AINA) are trained on broad Catalan data with no explicit Valencian component (VIVES, 2025), and available Valencian corpora (DOGV, À Punt Mèdia) remain relatively small for standalone training (European Language Equality (ELE), 2022; VIVES, 2025).

Dialect identification and lightweight models. Dialect identification has been studied extensively in other language pairs but not for Catalan/Valencian. The DSL shared tasks included Czech vs. Slovak and Brazilian vs. European Portuguese, achieving high accuracy on newswire (Zampieri et al., 2014, 2015). More recently, (Preda et al., 2024) revisited pt-BR vs. pt-PT with updated methods, and (Zampieri et al., 2020) provide a survey of techniques and pitfalls in similar-language discrimination. Lightweight fine-tuning has also proven effective in low-resource dialectal NLP: BERT on Arabic tweets (Mansour et al., 2020), AfriBERTa on African languages (Ogueji et al., 2021), and small multilingual models like mBERT or XLM-R that often outperform larger LLMs in limited-data regimes (Gurgurov et al., 2025).

3 Corpus

Because no labeled dataset exists to distinguish Catalan and Valencian, we compiled a new balanced corpus of 20 000 sentences (10 000 per variety). Sources included the Valencian government gazette (DOGV) and the À Punt Media portal for Valencian, and the Catalan government portal (gencat.cat) and the 3Cat/324 news site for Catalan.

We preserved all original tokens (including dates, headers, codes) to retain contextual cues, and only applied lowercasing. Sentence segmentation was carried out with regex rules plus manual review. Each sentence received a binary label: Valencian (1) or Catalan (0). The corpus was split into 80% training (16 000 sentences) and 20% test (4 000), ensuring class balance.

Data collection. We assembled 20 000 sentences (10 000 per class) from public institutional and media sources: the DOGV (`https://dogv.gva.es`) and À Punt Mèdia (`https://www.apuntmedia.es`) for Valencian, and the Catalan government's public portal gencat.cat (`https://web.gencat.cat`) (the Catalan equivalent of DOGV) and the 3Cat/324 news site (`https://www.3cat.cat/324`) for Catalan. All texts retained original metadata (dates, headers, codes) to leverage contextual cues.

Prior work has shown that case-sensitive models retain useful signals from capitalization and diacritics without loss in accuracy (e.g., BETO vs. lowercase BETO in Spanish; (Cóster and Martínez, 2021)), and that preserving punctuation and numerals maintains structural cues crucial for text classification (HaCohen-Kerner and Levin, 2020).

Preprocessing and labeling. Preprocessing follows Section 3: we only apply lowercasing. Sentence segmentation uses regex rules with manual review. Labels and the 80/20 split are as in Section 3.

4 Methodology

Model. We fine-tune RoBERTa-base (Liu et al., 2019), pre-trained on Catalan (Projecte AINA), for binary classification.

Training Setup.

- **Optimizer:** AdamW (weight decay 0.01).
- **Learning rate:** 2×10^{-5}, linear warmup (500 steps), total 3 epochs.
- **Batch size:** 16.
- **Scheduler:** Linear decay to zero.
- **Early stopping:** validation loss, patience = 1 epoch.

Input Representation. We tokenized with the standard RoBERTa tokenizer and truncated or padded sentences to 128 tokens. All other preprocessing followed the corpus description in Section 3.

Implementation. The experiments were run with HuggingFace Transformers v4.5.1 and PyTorch v1.10.1 on a single NVIDIA T4 GPU (16GB) provided via Google Colab, with 25GB host RAM available. We freeze the first 6 encoder layers for the first epoch to stabilize training, then unfreeze all layers.

Evaluation Protocol and Human Setup. Automatic metrics (accuracy, precision, recall, F_1) are computed on the held-out test set of 4 000 sentences (20% of the 20k corpus). In addition, we performed a human evaluation on a separate pool of 6 000 sentences (3 000 per variety). Following a statistically representative sampling procedure (Barros et al., 2021; Vázquez et al., 2010), we sampled up to 90 sentences per reviewer for each variety. We aimed for a balanced mix of error and correct cases (up to 45 each), though the exact proportions varied due to random sampling. To increase cross-variety exposure, some items came from the opposite class. A native Valencian speaker (bilingual in Spanish) annotated the Valencian-focused set, and a native Catalan speaker (also bilingual in Spanish) annotated the Catalan-focused set. We report per-class acceptance rates (96.7% for Valencian, 97.7% for Catalan) as the proportion of model predictions confirmed by humans.

5 Experiments and Results

Following the training setup described in Section 4, we fine-tuned RoBERTa-base for three epochs. Table 1 reports automatic test metrics and human acceptance rates.

Table 1: Automatic test metrics (n=4,000) and human acceptance (n=6,000).

	Acc (%)	Prec	Rec	F1
Automatic (overall)	98.0	0.978	0.976	0.977
Valencian (auto)	–	0.980	0.982	0.981
Catalan (auto)	–	0.982	0.980	0.981
Human acceptance				
Valencian		96.7%		
Catalan		97.7%		
Overall		97.2%		

The automatic confusion matrix (Table 2) remains unchanged, showing false positives and negatives below 2%.

	Predicted Catalan	Predicted Valencian
True Catalan	1,960	40
True Valencian	35	1,965

Table 2: Confusion Matrix (n=4,000).

6 Discussion

While our model achieves 98% automatic accuracy, human acceptance rates confirm high reliability across both varieties: it correctly labels 96.7% of Valencian sentences and 97.7% of Catalan ones, for an overall 97.2% acceptance. This consistency suggest robust performance, though further analysis is needed to ensure the model does not over-rely on contextual metadata and to better handle challenging or ambiguous cases.

7 Conclusions and Future Work

We have presented a lightweight classifier capable of distinguishing between standard Catalan and Valencian using minimal data and without dialect-specific tools. Trained in 20,000 sentences with contextual metadata, our model achieves automatic accuracy 98%. Future directions include:

- Incorporating human acceptance rates for confidence calibration.
- Extending training to informal varieties (e.g., social media dialects).
- Developing a dialect-aware tokenizer to better handle metadata and numerals.

All trained model checkpoints and associated code will be released upon acceptance. The resources will be accessible at `https://github.com/leurzz/modelo-catalan-valenciano`.

Limitations

Our study is limited to formal, institutional sources; generalisation to informal or noisy domains (e.g., social media) remains untested. In addition, we did not run ablations to disentangle linguistic features from metadata cues, and comparisons to alternative classifiers (e.g., SVMs or multilingual BERT) remain for future work. These aspects should be addressed to fully understand the robustness and portability of our approach.

Acknowledgments

This research has been conducted within the project *The limits and future of data-driven approaches: a comparative study of deep learning, knowledge-based and rule-based models and methods in Natural Language Processing*, funded by the Generalitat Valenciana under the CIDEGENT programme (Exp. CIDEXG/2023/13).

All authors are supported through this grant and are members of the Research Group on Natural Language Processing (GPLSI), University of Alicante. We thank the anonymous reviewers for their constructive feedback.

References

Acadèmia Valenciana de la Llengua. 2006. Gramàtica normativa valenciana. https://www.avl.gva.es/va/gramatica-normativa-valenciana.

Acadèmia Valenciana de la Llengua. 2022. Statement on valencian as a variant of catalan. https://avl.gva.es.

Acadèmia Valenciana de la Llengua. 2025. L'avl i la universitat d'alacant col·laboren en l'atles lingüístic valencià. https://www.avl.gva.es/lacademia-valenciana-de-la-llengua-i-la-universitat-dalacant-collaboren-en-el-projecte-atles-linguistic-valencia/.

Cristina Barros, Manuel Vicente, and Elena Lloret. 2021. To what extent does content selection affect surface realization in the context of headline generation? *Computer Speech & Language*, 67:101179.

Institut d'Estudis Catalans. 2022. Gramàtica essencial de la llengua catalana. https://geiec.iec.cat/text/5.3.3.

Adam Cóster and Paula Martínez. 2021. Evaluating the impact of lowercasing on spanish bert (beto). In *IberLEF 2021*.

European Language Equality (ELE). 2022. Report on the catalan language in the digital age. Technical report, ELE Project, Deliverable D1.6.

Stefan Gurgurov, Anna Ivanov, and Pavel Petrov. 2025. Small models, big impact: Adaptation of small multilingual language models for low-resource languages. *arXiv*, 2502.10140.

Yaron HaCohen-Kerner and Boris Levin. 2020. On the role of punctuation in text classification. *Computational Linguistics*, 46(1):1–22.

Idioma Valenciano. 2025. Listado de palabras diferentes en valenciano y catalán. https://www.idiomavalenciano.com/listado-palabras-diferentes-valenciano-catalan.html.

Institut d'Estudis Catalans. 2022. Gramàtica bàsica i d'ús de la llengua catalana. https://gbu.iec.cat/text/14.4.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv*, 1907.11692.

Miquel Àngel Lledó. 2011. The independent standardization of valencian: From official use to underground. *Handbook of language and ethnic identity: The success-failure continuum in language and ethnic identity efforts*, 2:336–348.

Walid Mansour, Rafid Fakiyurd, and Ammar Farahat. 2020. Arabic dialect identification using bert finetuning. In *Proceedings of the 6th Workshop on Arabic Natural Language Processing (WANLP)*.

Josep Martines. 2024. History of the catalan lexicon. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Anna Marzà, Frederic Chaume, Glòria Torralba, and Ana Alemany. 2006. The language we watch: an approach to the linguistic model of catalan in dubbing. In *Mercator Media Forum*, volume 9, pages 14–25. University of Wales Press.

Collins Ogueji, Rena Mika, and Temitope Adebowale. 2021. Small data? no problem! exploring the viability of pretrained multilingual models for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Diana Preda et al. 2024. Distinguishing between european and brazilian portuguese. In *Proceedings of PROPOR 2024*.

Real Acadèmia de Cultura Valenciana, Secció de Llengua i Lliteratura Valencianes. 2025. Diccionari general de la llengua valenciana. https://diccionari.llenguavalenciana.com/general/consulta/possessius.

Softcatalà. 2018. Catalan spell-checker dictionaries, including valencian variant. https://softcatala.org.

Yoselyn G. Vázquez, Francisco A. Orquín, Antonio M. Guijarro, and Sergio V. Pérez. 2010. Integración de recursos semánticos basados en wordnet. *Procesamiento del Lenguaje Natural*, 45:161–168.

Projecte VIVES. 2025. Plan for valencian language technologies. Data-gathering for speech and text corpora.

Max W Wheeler. 2005. *The phonology of Catalan*. OUP Oxford.

Marcos Zampieri, Preslav Nakov, Nikola Ljubešić, Jörg Tiedemann, and Shervin Malmasi. 2020. Natural language processing for similar languages, varieties and dialects: a survey. *Natural Language Engineering*, 26(6):695–717.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the dsl shared task 2014. In *Proceedings of VarDial Workshop @ COLING*.

Marcos Zampieri et al. 2015. Overview of the dsl shared task 2015. In *Proceedings of LT4VarDial Workshop @ RANLP*.