Enhancing Portuguese Variety Identification with Cross-Domain Approaches

Hugo Sousa[1, 2]*, Rúben Almeida[1, 2, 3]*, Purificação Silvano[1, 2], Inês Cantante[1], Ricardo Campos[2, 4], Alípio Jorge[1, 2]

[1] University of Porto, Portugal[1] INESC TEC, Portugal [2] Innovation Point – dst group, Portugal [3] University of Beira Interior, Portugal[4]

Association for the Advancement of Artificial Intelligence

The 39th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, PENNSYLVANIA, USA

How to train a model to distinguish language varieties

Portuguese as a Case Study

Although Portuguese is not considered as a low-resource language, most available digital and textual resources are in Brazilian Portuguese.

Country	Population of country[75][76][77]	More information	Native language of the majority	Spoken by
Brazil	203,062,512	Portuguese in Brazil	✓	95% as a native language[78]
Angola	35,981,281	Portuguese in Angola	✗	40% as a native language, 60% total[79]
Mozambique	32,513,805	Portuguese in Mozambique	✗	17% as a native language, 44% total[78]
Portugal	10,467,366	Portuguese in Portugal	✓	95% as a native language[80]
Guinea-Bissau	2,078,820	Portuguese in Guinea-Bissau	✗	0.3% as a native language, 20% total[81]

This predominance skews the performance of natural language processing models, which often struggle to generalize to other varieties, particularly European Portuguese. The lack of balanced datasets limits the applicability of these models in domains where formal European Portuguese is required, such as legal and medical fields. Addressing this disparity is crucial to ensure fair and effective language technologies across all Portuguese-speaking communities

Approach

1. Collect corpora from European and Brazilian Portuguese sources
2. Use the source as the label
3. Fine-tune a pre-trained language model for the task

Dataset

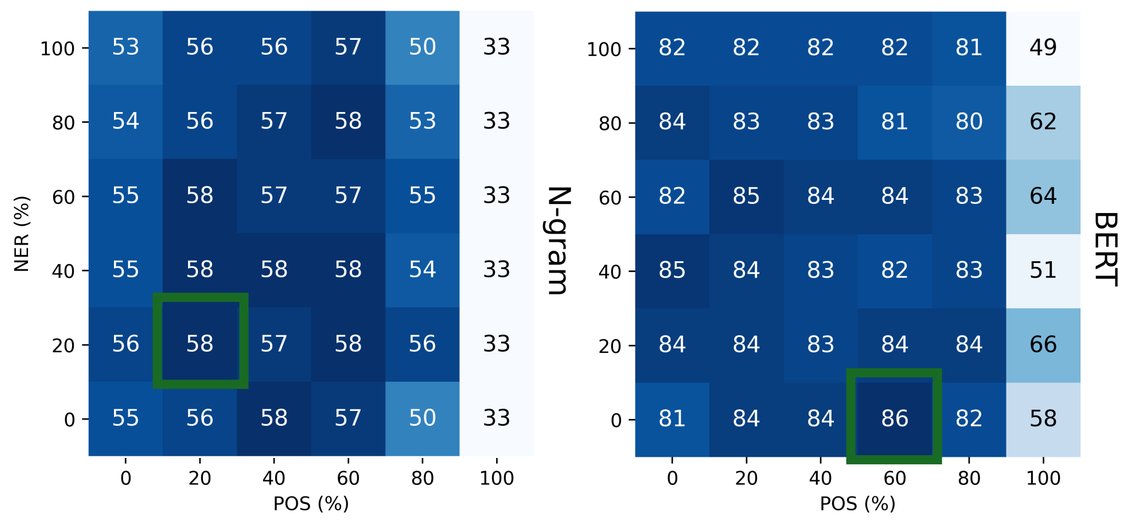
	Label	Tokens Count	Docs Count
Journalistic	🇵🇹	189,506,320	1,443,422
	🇧🇷	27,077,538	333,903
Literature	🇵🇹	1,859,660	24,090
	🇧🇷	3,805,896	52,458
Legal	🇵🇹	152,717,737	2,957,980
	🇧🇷	221,167	4,653
Politics	🇵🇹	7,203,739	27,887
	🇧🇷	1,012,586	3,656
Web	🇵🇹	22,598,587	43,630
	🇧🇷	23,913,771	44,313
Social Media	🇵🇹	44,758,304	2,363,261
	🇧🇷	94,177	5,504

Delexicalization

Our initial models relied on context terms like locations and person names to classify language variety, rather than linguistic differences. To mitigate this bias, we explored delexicalization as a hyperparameter, replacing named entities and part of speech elements with generic labels.

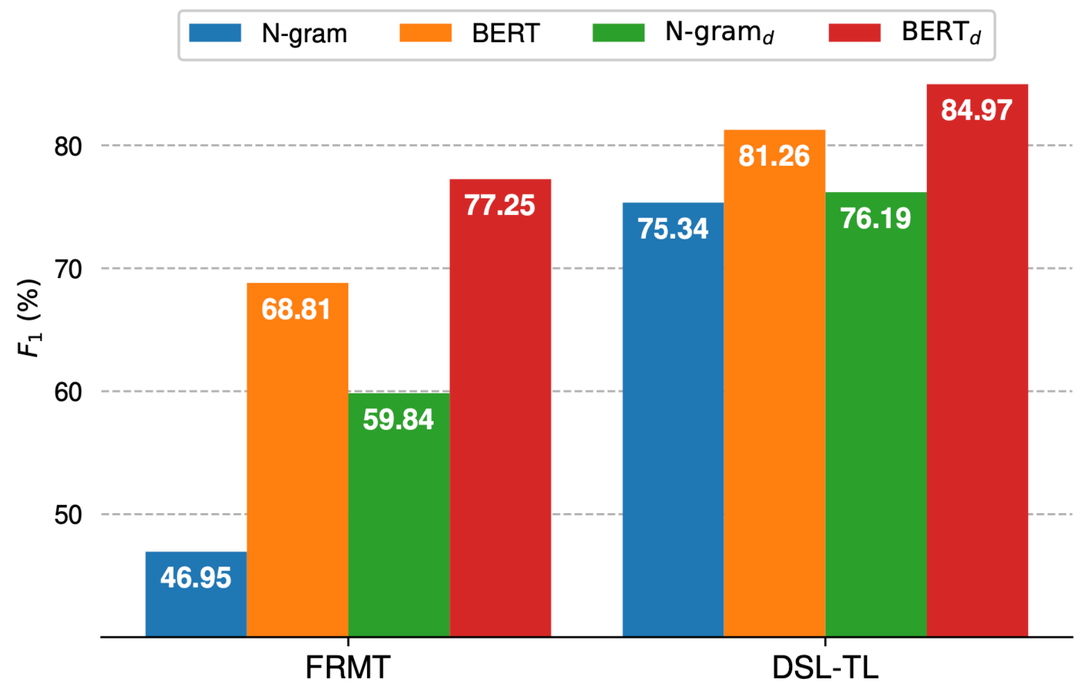
Ele foi de autocarro para Natal. ✖ 🇧🇷

Ele foi de autocarro para **[LOC].** → 🇵🇹



Results

Our results show that both N-gram and BERT models benefit from delexicalization, improving cross-domain generalization by reducing reliance on named entities and thematic content. Models trained with this approach achieved higher F1 scores, demonstrating its effectiveness in Portuguese variety classification.



PAPER MODEL CODE

U.PORTO INESCTEC TECHNOLOGY & SCIENCE innovationpoint dst group UNIVERSIDADE BEIRA INTERIOR fct Fundação para a Ciência e a Tecnologia