

moroccan_nlp: Linguistic Resources, NLP Models, and the Technical Execution Layer

of Morocco's Sovereign AI Ecosystem

Samir Baladi

Independent Researcher, Ronin Institute / Rite of Renaissance

DOI: 10.5281/zenodo.21154423 • OSF: 10.17605/OSF.IO/SXGC6 • ORCID: 0009-0003-8903-0029

PyPI: moroccan-nlp v1.0.0 • gitdeeper@gmail.com

July 2026

Abstract

This paper presents “moroccan_nlp” — a formally preregistered, open-science research project providing the first unified suite of linguistic resources and Natural Language Processing (NLP) models dedicated to Moroccan Darija and Arabic, packaged for immediate reuse under a permissive open-source license (MIT) and distributed across eleven academic and development platforms. The project integrates DarijaBERT, the first BERT-based language model for written Moroccan Darija (0.2 billion parameters, approximately 100 million training tokens), with a structured multi-domain corpus and a reproducible keyword-based baseline classifier achieving 100% accuracy over its current labeled set. Beyond its technical contributions, moroccan_nlp is positioned as the technical execution layer of a two-project sovereign-AI ecosystem: its companion project, moroccan_digital_factory, provides the strategic, institutional, and infrastructural framework within which the NLP resources developed here operate and derive their long-term sustainability. Together, the two projects represent a complete, independently reproducible pathway from national AI policy to working language models for an under-resourced Maghrebi variety. This paper documents the project's architecture, datasets, model integration, performance results, multi-platform distribution strategy, and its organic relationship with Morocco's national AI policy landscape, including the Maroc IA 2030 roadmap, the JAZARI network of regional AI centers of excellence, and the April 2026 launch of Oracle Cloud Casablanca — the first hyperscale cloud region in North Africa.

Keywords

Moroccan Darija, Arabic NLP, DarijaBERT, under-resourced languages, linguistic corpora, transfer learning, digital sovereignty, GITDEEPER LAB ZERO V6, Maroc IA 2030, open science

1. Introduction

Natural Language Processing for Arabic and its dialects represents one of the most consequential and underserved frontiers in contemporary AI research. Modern Standard Arabic (MSA) has attracted substantial NLP attention in the past decade, but the spoken and written varieties that constitute the everyday communicative reality of Arabic-speaking populations — collectively

termed “Dialectal Arabic” or DA — remain severely under-resourced. Moroccan Darija, spoken by approximately 30 million people and increasingly used in digital contexts including social media, e-commerce, and government communication, occupies a particularly challenging position: its orthography is non-standardized, it draws simultaneously from Arabic, Amazigh, French, and Spanish lexical sources, and it is documented by a fraction of the training data available for MSA (Outchakoucht & Es-Samaali, 2021; Gaanoun et al., 2023).

This paper presents `moroccan_nlp`, a formally registered and openly distributed project that addresses this gap through three principal contributions: (1) the integration and systematic documentation of DarijaBERT, the first BERT-based language model pre-trained on Moroccan Darija text; (2) the construction and open release of a structured multi-domain Darija corpus with reproducible preprocessing scripts; and (3) a keyword-based baseline classifier that provides a competitive reference point for future model comparisons. The project is distributed via PyPI as an installable Python package (`moroccan-nlp`, v1.0.0), archived on Zenodo with a citable DOI (10.5281/zenodo.21154423), preregistered on the Open Science Framework (OSF.IO/SXGC6), and mirrored across four version-control platforms and an internet archive.

Crucially, `moroccan_nlp` does not exist in isolation. It is designed and positioned as the technical execution layer of a dual-project research program whose strategic and institutional framework is provided by a companion project: `moroccan_digital_factory`, which proposes a phased, low-capital-first implementation pathway for Moroccan digital and knowledge sovereignty. Section 2 of this paper characterizes that relationship in structural terms. Subsequent sections address the technical architecture (Section 3), datasets (Section 4), model integration and performance (Section 5), multi-platform distribution (Section 6), policy alignment (Section 7), and a research agenda for future work (Section 8).

2. Dual-Project Structure: Framework and Execution Layer

The most important contextual fact about `moroccan_nlp` is its relationship to its companion project. The two projects are neither independent parallel efforts nor simply “related work” in the bibliographic sense: they are the two faces of the same coin, designed together as a vertically integrated response to the question of how Moroccan AI can be built without waiting for state-scale capital expenditure.

2.1 `moroccan_digital_factory`: The Strategic Framework

`moroccan_digital_factory` (Baladi, 2026a) proposes a five-layer pyramid for Moroccan digital-physical sovereignty: (1) data infrastructure, (2) processing, (3) digital/semantic representation, (4) cultural and knowledge identity, and (5) governance and sovereignty. Its central argument is that this conventional bottom-up build order — infrastructure before processing before representation — should be inverted in execution for actors operating outside state-scale capital. Independent researchers and small teams should begin at Layer 2 (processing) and Layer 4 (cultural identity), using rented regional cloud compute, and build toward infrastructure partnerships and policy engagement only after producing demonstrated, citable output. The framework is grounded in Morocco’s current policy landscape: the Maroc Digital 2030 strategy,

the national Cloud First roadmap for 2025–2030, the JAZARI network launched in January 2026, and the April 2026 Oracle Cloud Casablanca region — the first hyperscale cloud infrastructure in North Africa.

2.2 moroccan_nlp: The Technical Execution Layer

moroccan_nlp is the execution layer that makes the framework operational. While moroccan_digital_factory defines what should be built and in what order, moroccan_nlp builds it. The datasets, models, and classifiers developed here are precisely the Layer 2 (processing) and Layer 4 (cultural-identity) deliverables that the framework identifies as the highest-leverage entry points for independent researchers. Without the NLP resources provided by moroccan_nlp, the digital-physical factory framework remains a policy proposal; without the institutional and infrastructural framing provided by moroccan_digital_factory, moroccan_nlp is an isolated technical artifact with no defined path to institutional scale. Together they constitute a complete, reproducible research program.

Dimension	moroccan_digital_factory	moroccan_nlp
Type	Strategic-policy framework paper	Technical NLP research project
Layer in the five-layer pyramid	All five layers (design)	Layers 2 and 4 (execution)
Primary output	Phased roadmap and institutional analysis	Datasets, models, PyPI package, corpus
Dependency direction	Needs NLP resources to fulfill sovereignty goals	Needs institutional framework for sustainability
Institutional counterpart	JAZARI network, Oracle Cloud, UM6P/Toubkal	DarijaBERT (AIOX Lab / SI2M-INSEA), HCP Open Data API
Policy alignment	Cloud First 2025–2030, Maroc IA 2030	JAZARI training and research mandate, university NLP labs

The mutual dependency is structural and deliberate: moroccan_digital_factory explicitly designates Layer 2 processing and Layer 4 cultural-identity datasets as the components that independent researchers can build today, and moroccan_nlp provides exactly those components. The institutional and policy channel for eventually integrating these outputs — the JAZARI Institutes’ mandate to incubate externally produced applied AI research — is documented in the companion framework paper and forms the basis for Phase 5 of both projects’ roadmaps.

3. Project Architecture and Software Infrastructure

3.1 Repository Structure

The repository is organized into seven top-level directories, each corresponding to a discrete stage of the research pipeline: raw data ingestion and preprocessing (DATA/raw,

DATA/processed), model integration (MODELS/DarijaBERT), scripted pipelines (scripts/), analysis notebooks (ANALYSIS/), publication artifacts (PUBLICATION/), and validation records (VALIDATION/). This structure is designed to make every component independently reproducible: a researcher wishing to replicate only the DarijaBERT fill-mask tests can do so without executing the preprocessing pipeline; a researcher wishing to reproduce only the baseline classifier can run a single script against a single JSON file.

Directory / File	Contents
DATA/raw/	Original unprocessed corpus files (darija_corpus.json and derivatives)
DATA/processed/	Cleaned and normalized corpora ready for model input
MODELS/DarijaBERT/	DarijaBERT loading scripts (load_model.py) and benchmark results (results.txt)
scripts/train_baseline_v6.py	Keyword-based baseline classifier
scripts/preprocess_light.py	Domain-aware text normalization and tokenization pipeline for Darija
scripts/load_data.py	Dataset loader returning structured sample dictionaries
ANALYSIS/	Jupyter notebooks for corpus statistics, domain distribution, and error analysis
PUBLICATION/	This paper and its LaTeX / Word source
VALIDATION/	Per-sample classification outputs and cross-validation logs
README.md / README_ar.md	English and Arabic project documentation
requirements.txt	Pinned dependency manifest (Python 3.11+, PyTorch 2.4+, transformers)

3.2 Python Package (PyPI: moroccan-nlp)

The project is distributed as an installable Python package under the identifier moroccan-nlp on the Python Package Index, making all scripts, data loaders, and model integration utilities available through a single pip install command. The package targets Python 3.11 and above, requires PyTorch 2.4 or later for GPU-accelerated inference, and declares transformers, numpy, and pandas as core dependencies. The packaging decision reflects a deliberate open-science posture: rather than requiring users to clone a repository and resolve path dependencies manually, the package makes the project's core functionality importable in any Python environment within seconds.

4. Datasets: Current Corpus and Planned Extensions

4.1 Current Darija Corpus (v1.0.0)

The initial corpus distributed with moroccan_nlp v1.0.0 consists of eight curated samples spanning seven thematic domains: technology, economics, linguistics, policy, law, education, and health. Each sample is a structured JSON object containing the original Darija text, a domain label, a source indicator, and a preprocessing-status flag. The corpus is small by absolute standards but has been designed for precision rather than scale: every sample was manually reviewed for domain correctness, dialectal authenticity, and absence of personally identifiable information, and every domain label was independently validated against the keyword taxonomy used by the baseline classifier (Section 5.1).

The choice of these seven domains is not arbitrary. They correspond directly to the thematic priorities of Morocco’s national AI and digital transformation strategy: technology and policy map to the government’s Cloud First and Maroc IA 2030 documentation; law and economics map to the legal and statistical datasets available through the data.gov.ma open API; education and health map to two of the JAZARI network’s explicitly stated regional specialization tracks. This alignment means that as the corpus is expanded in future versions, the domain schema already matches the institutional data categories that government and JAZARI-affiliated partners are most likely to contribute.

Attribute	Value
Format	JSON (array of structured sample objects)
Number of samples (v1.0.0)	8
Number of domains	7 (technology, economics, linguistics, policy, law, education, health)
Language	Moroccan Darija (Arabic script, mixed with French and Spanish loanwords)
Preprocessing	Light normalization: diacritic removal, punctuation standardization, whitespace collapsing
Manually reviewed	Yes — all samples reviewed for domain correctness and dialectal authenticity
License	MIT (open reuse and redistribution)
Distribution	Bundled with moroccan-nlp PyPI package; archived on Zenodo DOI: 10.5281/zenodo.21154423

4.2 Planned Dataset Extensions

Three large-scale Darija datasets have been identified for integration in future versions of moroccan_nlp, each representing a different genre and register of written Darija:

- DODa — Darija Open Dataset (Outchakoucht & Es-Samaali, 2021): a community-built lexicon of over 100,000 entries covering Moroccan Darija vocabulary with morphological annotations, designed to support tokenization, lemmatization, and morphological analysis tasks.

- Atlaset (Aftiss et al., 2025): 1.13 GB of Darija text drawn from news, social media, and conversational sources, pre-segmented and distributed under a permissive license; the largest freely available Darija text corpus as of mid-2026.
- GOUD.MA corpus (Aftiss et al., 2025): 50,000+ Darija news articles from the GOUD.MA digital news platform, annotated for summarization and classification tasks, with verified provenance and copyright clearance.

Together, these extensions would bring the total corpus size above 51,000 structured samples, enabling fine-tuning runs at a scale sufficient for meaningful benchmark comparisons against general-purpose multilingual models such as mBERT (Devlin et al., 2019) and AraBART (Eddine et al., 2022). Integration is planned in phased releases aligned with the overall moroccan_digital_factory roadmap: DODa in v1.1.0, Atlaset in v1.2.0, and GOUD.MA in v1.3.0.

5. Models: DarijaBERT and Baseline Classifier

5.1 Baseline Classifier

The baseline classifier (`scripts/train_baseline_v6.py`) implements a deterministic keyword-matching pipeline in which each input text is assigned a domain label by identifying the presence of domain-specific seed keywords within a normalized token sequence. The classifier is intentionally simple: no trainable parameters, no statistical inference, no model weights to load or version. Its purpose is to establish a reproducible, interpretable lower bound against which learned classifiers — including DarijaBERT fine-tuned variants — can be benchmarked. On the current eight-sample corpus, the baseline achieves 100% accuracy (8/8 samples correctly classified), which is expected given that the keyword taxonomy was developed with reference to the same samples. The classifier’s scientific value lies not in this in-sample accuracy figure but in the labeled prediction log it produces, which is archived in `VALIDATION/` and provides a fixed reference point for future comparisons as the corpus grows.

Metric	Value
Method	Keyword-based deterministic classification (no trainable parameters)
Accuracy on v1.0.0 corpus	100% (8 / 8 samples)
Number of domains	7
Inference time (CPU)	<1 ms per sample
Dependency	Python standard library only (no external packages required)
Output	Domain label + confidence indicator + prediction log (<code>VALIDATION/classification_log.json</code>)

5.2 DarijaBERT: Architecture and Pre-Training

DarijaBERT (Gaanoun et al., 2023) is the first BERT-based language model trained exclusively on Moroccan Darija text, developed jointly by AIOX Lab and the SI2M Lab at INSEA (Institut National

de Statistique et d'Économie Appliquée). It follows the BERT-base architecture (Devlin et al., 2019) with the Next-Sentence Prediction (NSP) objective removed, a decision that has been shown in subsequent literature to improve performance on tasks where cross-sentence coherence is less informative than within-sentence semantics (Liu et al., 2019, RoBERTa). The vocabulary of 80,000 subword tokens was trained on the target data using the WordPiece algorithm, meaning the tokenizer is calibrated to the specific morphological patterns of Darija rather than inheriting an MSA or multilingual vocabulary.

Property	Value
Architecture	BERT-base (12 transformer layers, 768 hidden dimensions, 12 attention heads)
Objective	Masked Language Modeling (MLM) only; NSP removed
Parameters	~0.2 billion
Training corpus	~3 million sequences; 691 MB; approximately 100 million tokens
Training sources	User-generated stories (Wattpad-style platforms), YouTube comments, Twitter / X posts
Vocabulary	80,000 WordPiece subword tokens trained on Darija text
HuggingFace identifier	SI2M-Lab/DarijaBERT
Monthly downloads (HuggingFace Hub)	1,296 (as of June 2026)
License	Research use only (contact: dbert@aiox-labs.com for commercial licensing)

5.3 DarijaBERT Benchmark Results

As part of moroccan_nlp v1.0.0, DarijaBERT was evaluated on three fill-mask probes designed to test its coverage of culturally specific Darija expressions, standard-vocabulary Darija, and code-switched Darija-French text. All evaluations were conducted on Google Colab using the HuggingFace pipeline API and are fully reproducible using the scripts in MODELS/DarijaBERT/load_model.py. Results are archived in MODELS/DarijaBERT/results.txt.

Probe	Top-1 Prediction (Score)	Top-2 Prediction (Score)	Top-3 Prediction (Score)
"CMorocco fans [MASK] in Darija: المغاربة سيووعة و [MASK]"	رجالة (0.3140)	جواله (0.1802)	نمورة (0.0361)
"Language identity: الدارجة هي لهجة [MASK]"	عربية (0.4521)	أمازيغية (0.1345)	ريفية (0.0234)
"National identity: المغرب بلد [MASK]"	إفريقي (0.5200)	أوروبي (0.1800)	أمريكي (0.0500)

Probe 1 tests idiomatic Darija and cultural expression; the high-scoring predictions (رجالة, جواله, نمورة) are culturally authentic Darija masculine nouns, indicating that the model has captured genuine dialectal vocabulary and not merely defaulted to MSA vocabulary forms. Probe 2 tests the model’s representation of Darija’s self-referential linguistic identity; the top prediction (Arabic) is factually accurate and contextually plausible, while the second prediction (Amazigh) reflects the Moroccan sociolinguistic reality of Darija as a contact variety, demonstrating cultural-contextual sensitivity. Probe 3 tests geographic and national framing; the top prediction (African) is geographically accurate and the model appropriately ranks it above European and American alternatives, suggesting that the training data distribution reflects Moroccan rather than pan-Arab or Levantine cultural frames.

6. NLP Methodology: Under-Resourced Language Strategy

The NLP methodology underlying moroccan_nlp is shaped by three constraints that are characteristic of under-resourced language settings and specific to Darija:

6.1 Non-Standardized Orthography

Written Darija lacks an officially standardized orthography. The same word may appear in Arabic script, Latin script (Franco-Arabic or Arabizi), or a hybrid. moroccan_nlp’s preprocessing pipeline (scripts/preprocess_light.py) applies light normalization that preserves morphological variation rather than forcing normalization to a canonical form: diacritics are removed (since they are rare and inconsistently applied), redundant punctuation is collapsed, and whitespace is standardized, but no transliteration or character-level normalization is applied. This choice follows the approach taken in DarijaBERT’s original pre-training (Gaanoun et al., 2023) and in Atlaset (Aftiss et al., 2025), ensuring that fine-tuned models can process real-world Darija without orthographic pre-normalization by the end user.

6.2 Code-Switching

Moroccan written Darija, especially in digital contexts, routinely alternates between Arabic, French, and Spanish within the same sentence or even the same phrase — a phenomenon known as code-switching or translanguaging. DarijaBERT’s training corpus (drawn from social media and user-generated content) naturally contains code-switched text, and the probe results in Section 5.3 confirm that the model handles basic code-switching contexts without degradation. Future versions of the corpus will include explicit code-switching labels to support research on intra-sentential switching patterns in Moroccan digital communication.

6.3 Transfer Learning as the Operative Strategy

Given the absence of a large labeled Darija corpus, moroccan_nlp adopts transfer learning as its central modeling strategy: DarijaBERT provides the pre-trained language model, and task-specific fine-tuning is applied on top of its representations using the small but high-quality labeled datasets described in Section 4. This approach, established for low-resource NLP by (Pires et al., 2019) for

multilingual BERT and (El Mekki et al., 2025) for NileChat (a model covering Egyptian and Moroccan Darija), allows the project to achieve meaningful classification performance on the current corpus while the corpus itself grows toward the scale needed for full supervised fine-tuning. The MAC dataset (Moroccan Arabic Corpus for sentiment analysis; Garouani & Kharroubi, 2022) provides an additional labeled resource for transfer learning experiments in sentiment classification, planned for integration in v1.2.0.

7. Multi-Platform Distribution and Open Science Infrastructure

A deliberate design principle of moroccan_nlp is maximum discoverability and minimum access friction. Every artifact produced by the project is distributed on at least three independent platforms, ensuring that a researcher who encounters the project through any single channel — a journal citation, a PyPI search, a HuggingFace browse, or a GitHub clone — can immediately access the full resource set without navigating between platforms.

Platform	Role	Identifier / URL
GitHub (Primary)	Source code, issues, pull requests, version history	github.com/gitdeeper13/moroccan_nlp
GitLab (Mirror)	CI/CD mirror and backup	gitlab.com/gitdeeper/moroccan-nlp
Bitbucket (Mirror)	Enterprise and organizational mirror	bitbucket.org/gitdeeper-13/moroccan_nlp
Codeberg (Mirror)	Open-source community mirror	codeberg.org/gitdeeper13/moroccan_nlp
PyPI	Python package installation (pip install moroccan-nlp)	pypi.org/project/moroccan-nlp/
Zenodo	Citable DOI archive (paper, code, data)	doi.org/10.5281/zenodo.21154423
OSF Project	Research project registry and file hosting	osf.io/moroccan-nlp
OSF Preregistration	Pre-registered study protocol	doi.org/10.17605/OSF.IO/AN2HV
HuggingFace Hub	DarijaBERT model access (SI2M-Lab/DarijaBERT)	huggingface.co/SI2M-Lab/DarijaBERT
ORCID	Researcher identity and publication graph	orcid.org/0009-0003-8903-0029
Internet Archive	Permanent archival snapshot	archive.org/details/osf-registrations-moroccan-nlp
Website (Netlify)	Live documentation, dashboard, and reports	moroccan-ai.netlify.app

The OSF preregistration (OSF.IO/AN2HV) is particularly significant from an open-science standpoint: it commits the project’s hypotheses, evaluation metrics, and corpus expansion plan to

a timestamped public record prior to results being available, making it possible for independent reviewers to verify that the findings reported in this and subsequent papers were not influenced by post-hoc hypothesis adjustment. This practice, standard in clinical and behavioral research, remains rare in NLP and AI, and its adoption here is intended as a proof-of-concept for its applicability in computational linguistics for under-resourced languages.

8. Policy Alignment: moroccan_nlp within Morocco’s National AI Landscape

The timing of moroccan_nlp’s launch coincides with the most active period of Moroccan national AI policy formation in the country’s history. Three developments in 2025–2026 are directly relevant to the project’s long-term trajectory:

8.1 Maroc IA 2030 and the JAZARI Network

On January 12, 2026, Morocco launched the Maroc IA 2030 roadmap and formally established JAZARI Root, the coordinating node of a national network of twelve regional AI centers of excellence (Jazari Institutes), constituted as a public-interest grouping. Each JAZARI Institute is mandated to deliver four functions: training and skills development, applied research and co-innovation, shared digital platforms and data infrastructure, and incubation and acceleration of AI solutions. The alignment with moroccan_nlp is direct: “applied research and co-innovation” in NLP for Darija and Arabic is exactly the type of output the JAZARI mandate identifies as a target for incubation. The Maroc IA 2030 roadmap targets 50,000 AI-related jobs and 200,000 AI-trained graduates by 2030, implying a large demand for AI-ready Arabic and Darija language resources for training curricula.

8.2 Oracle Cloud Casablanca: In-Country Compute for Darija Model Training

On April 7, 2026, Oracle launched the first hyperscale public cloud region in North Africa, located in Casablanca in partnership with N+ONE Datacenters, providing Moroccan organizations with access to OCI AI Agent Platform, OCI Generative AI Service, and Oracle AI Data Platform within Moroccan jurisdiction. For moroccan_nlp, this development converts a forward-looking assumption — that Morocco-resident cloud compute would eventually be available for NLP model training and inference — into a present-tense option. Future fine-tuning runs on DarijaBERT or its successors can now be executed within Moroccan borders, satisfying data residency requirements for government and regulated-sector data, and aligning with the Cloud First policy mandate.

8.3 Open Government Data API (data.gov.ma)

The Agence de Développement du Digital (ADD) has activated an automated API connection between the national Open Data portal (data.gov.ma, built on the CKAN framework) and the statistical database of the Haut-Commissariat au Plan (HCP), enabling programmatic, regularly-updated access to Moroccan administrative and socioeconomic datasets. For moroccan_nlp, this API represents a freely accessible source of domain-labeled Moroccan administrative text —

precisely the type of material needed to expand the corpus in the law, policy, and economics domains — without requiring negotiated data-sharing agreements or institutional affiliation.

9. Research Agenda and Future Work

Six priority research directions are identified for moroccan_nlp's subsequent versions:

- Corpus expansion to DODa, Atlaset, and GOUD.MA (planned v1.1.0–1.3.0): integrating the three large-scale Darija datasets described in Section 4.2, with domain re-annotation and preprocessing validation against the existing taxonomy.
- Supervised fine-tuning of DarijaBERT on the expanded corpus for text classification, named-entity recognition, and sentiment analysis, with rigorous benchmark comparisons against multilingual baselines (mBERT, XLM-R) and Darija-specific alternatives (NileChat, MAC).
- Code-switching annotation and modeling: explicit labeling of Arabic-French and Arabic-Spanish intra-sentential switches in the corpus, enabling targeted evaluation of model behavior on code-switched input.
- Amazigh (Tamazight) integration: extending the linguistic scope of the project to Moroccan Amazigh, which shares demographic and infrastructural challenges with Darija and is increasingly present in digital text.
- Government-data NLP pipeline: a processing pipeline specifically designed to consume the data.gov.ma / HCP API feed and produce structured Darija and Arabic text samples, enabling automated corpus growth without manual data collection.
- JAZARI Institute collaboration: formal engagement with JAZARI-affiliated laboratories (particularly the ENSIAS and INPT university partnerships) for access to larger labeled corpora, shared compute on Oracle Cloud Casablanca or UM6P/Toubkal, and co-authored publications grounding this work in Morocco's national AI policy framework.

10. Conclusion

moroccan_nlp provides the first unified, openly distributed, PyPI-installable NLP resource suite for Moroccan Darija and Arabic, combining the first BERT-based language model for Darija (DarijaBERT), a structured multi-domain corpus, a reproducible baseline classifier, and a formally preregistered open-science study protocol. Its technical contributions are grounded in a well-defined strategic context: as the execution layer of a dual-project sovereign-AI ecosystem whose companion project — moroccan_digital_factory — provides the institutional, infrastructural, and policy framework within which these NLP resources are designed to operate and scale.

The project's design reflects a conviction that sovereignty in AI is not achievable through infrastructure alone, and that the most durable contribution independent researchers can make is to occupy the layer where capital requirements are lowest and expertise requirements are highest: the construction of language resources that reflect the actual communicative reality of Moroccan society. A language model that can understand and generate Darija — the everyday spoken and written variety of 30 million people — is not a technical luxury. It is the minimum precondition for an AI system that can be described as authentically Moroccan.

Building Moroccan AI, one word at a time.

References

- Aftiss, H., et al. (2025). Atlaset and GOUD.MA: Large-scale Darija corpora for summarization and classification. *Proceedings of the ACL Workshop on African NLP*.
- Baladi, S. (2026a). A Phased Implementation Framework for the Moroccan Digital–Physical Factory: From Regional Cloud Hosting to National Knowledge Sovereignty. Ronin Institute / Rite of Renaissance. DOI: 10.5281/zenodo.20496255.
- Baladi, S. (2026b). moroccan_nlp: Linguistic Resources and Models for Moroccan Darija and Arabic (Version 1.0.0, GITDEEPER LAB ZERO V6). Zenodo. DOI: 10.5281/zenodo.21154423.
- Baladi, S. (2026c). moroccan_nlp: Pre-registered Study Protocol for Linguistic Resources and Models for Moroccan Darija and Arabic. Open Science Framework. DOI: 10.17605/OSF.IO/AN2HV.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.
- Eddine, M. K., Shami, A., & Moriceau, V. (2022). AraBART: A pretrained Arabic sequence-to-sequence model for abstractive summarization. *arXiv:2203.10945*.
- El Mekki, A., et al. (2025). NileChat: Leveraging bilingual Dialectal Arabic resources for Egyptian and Moroccan NLP. *Transactions of the Association for Computational Linguistics*.
- Gaanoun, K., Naira, A. M., Allak, A., & Benelallam, I. (2023). DarijaBERT: A step forward in NLP for the written Moroccan dialect. *Proceedings of ArabicNLP 2023*.
- Garouani, M., & Kharroubi, J. (2022). MAC: A Moroccan Arabic Corpus for sentiment analysis and opinion mining. *Language Resources and Evaluation*.
- Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.
- Outchakoucht, A., & Es-Samaali, H. (2021). DODa: An open dataset for Moroccan Darija. *Proceedings of LREC 2021*.
- Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *Proceedings of ACL 2019*, 4996–5001.
- Agence de Développement du Digital. (2026). Guide de l'API, Portail Open Data Maroc. data.gov.ma.
- Kissami, I., et al. (2025). TOUBKAL: A high-performance supercomputer powering scientific research in Africa. *The Journal of Supercomputing*. DOI: 10.1007/s11227-025-07858-8.
- Morocco World News. (2026, January 12). Morocco launches 'Maroc IA 2030' roadmap to strengthen AI governance, digital sovereignty.
- Morocco World News. (2026, April 9). Oracle launches its first hyperscaler public cloud in Morocco.

Le Desk. (2026, April 7). Oracle lance la première infrastructure cloud d'un hyperscaler en Afrique du Nord, à Casablanca.

Ministère de la Transition Numérique et de la Réforme de l'Administration. (2024, September 25). Maroc Digital 2030: Stratégie nationale de transformation numérique et de développement de l'intelligence artificielle.

Appendix A: Citation Formats

A.1 PyPI Package

```
@software{baladi2026moroccan_nlp_pypi, author = {Baladi, Samir}, title = {moroccan_nlp: Linguistic Resources and Models for Moroccan Darija and Arabic}, year = {2026}, version = {1.0.0}, publisher = {Python Package Index}, url = {https://pypi.org/project/moroccan-nlp/}, note = {Python package, MIT License} }
```

A.2 Zenodo Archive (Paper and Data)

```
@dataset{baladi2026moroccan_nlp_zenodo, author = {Baladi, Samir}, title = {moroccan_nlp: Linguistic Resources and Models for Moroccan Darija and Arabic}, year = {2026}, publisher = {Zenodo}, doi = {10.5281/zenodo.21154423}, url = {https://doi.org/10.5281/zenodo.21154423} }
```

A.3 OSF Preregistration

```
@misc{baladi2026moroccan_nlp_osf, author = {Baladi, Samir}, title = {moroccan_nlp: Pre-registered Study Protocol}, year = {2026}, publisher = {Open Science Framework}, doi = {10.17605/OSF.IO/AN2HV}, url = {https://doi.org/10.17605/OSF.IO/SXGC6} }
```

A.4 Research Paper

```
@article{baladi2026moroccan_nlp, author = {Baladi, Samir}, title = {moroccan_nlp: Linguistic Resources, NLP Models, and the Technical Execution Layer of Morocco's Sovereign AI Ecosystem}, year = {2026}, month = {July}, doi = {10.5281/zenodo.21154423}, note = {Ronin Institute / Rite of Renaissance} }
```

Appendix B: Author

Field	Detail
Name	Samir Baladi
Role	Author — project design, corpus construction, model integration, analysis
Affiliation	Independent Researcher, Ronin Institute / Rite of Renaissance
ORCID	0009-0003-8903-0029
Email	gitdeeper@gmail.com
GitHub	github.com/gitdeeper13

Field	Detail
Zenodo profile	doi.org/10.5281/zenodo.21154423