TECHNICAL RESEARCH REPORT / PREPRINT

Benchmarking Intent Classification in Extremely Low-Resource Indo-Aryan Dialects: An Empirical Study on Dhatki Agricultural Telemetry

Lokesh Kumar (Sole Author & Lead Researcher)
Department of Software Engineering, Information Technology Centre, Sindh Agriculture University (SAU), Tando Jam, Pakistan
ORCID: 0009-0006-6510-6393 | Email: lokeshsoni653@gmail.com | Dataset: huggingface.co/datasets/lokeshsonara/dhatki-agri-intents

ABSTRACT

Natural Language Processing (NLP) models achieve state-of-the-art benchmarks in high-resource global languages, yet regional vernacular dialects spoken by millions of marginalized pastoralists remain entirely excluded from digital conversational interfaces. Dhatki, an Indo-Aryan transitional language spoken by over 2.5 million inhabitants across the Thar Desert of southeastern Sindh, Pakistan, lacks standardized tokenizers, orthographic corpora, and annotated evaluation benchmarks. In this paper, we introduce the first domain-specific, curated Intent Classification Dataset for Dhatki agricultural telemetry, comprising 1,000 verified acoustic and textual utterances spanning 10 operational agricultural classes. We empirically benchmark baseline lexical models against multilingual transformer representations (XLM-RoBERTa, mBERT, and IndicBERT) under severe resource constraints. Our experimental findings demonstrate that custom subword tokenization tuned to phonetic Dhatki transcriptions achieves a Micro-F1 score of 89.4%, proving the viability of low-latency, edge-deployable voice interfaces.

1. Introduction & The Linguistic Vacuum in Global Agritech

Digital agriculture platforms in developing economies rely predominantly on text-based English or standard national languages (such as Urdu in Pakistan). However, over 65% of rural pastoralists in the Thar Desert are non-literate in standard Urdu, relying exclusively on Dhatki - an oral-first Indo-Aryan language belonging to the Rajasthani-Sindhi transitional sub-branch. Consequently, critical alerts regarding pest outbreaks, soil salinity, micro-irrigation schedules, and extreme weather events fail to reach the farmers who need them most.

Bridging this systemic divide necessitates speech-to-intent voice interfaces that operate directly on mobile handsets without requiring high-bandwidth cloud APIs. However, off-the-shelf NLP foundation models fail catastrophically when presented with Dhatki phonology due to out-of-vocabulary (OOV) rates exceeding 78% on standard tokenizers.

2. The Dhatki Agricultural Corpus Construction

To establish an empirical baseline, we constructed the 'dhatki-agri-intents' dataset (hosted openly on Hugging Face). The corpus consists of 1,000 verified acoustic transcripts collected from smallholder farmers across Mithi, Chachro, Islamkot, and Umerkot. The dataset covers 10 primary intent classes:

1. Crop Irrigation Timing (Pani Jo Waqt) | 2. Pest Attack Identification (Kiro / Bimari) | 3. Fertilizer Dosage (Khaad Ji Miqdar)
4. Livestock Disease Diagnosis (Maweshi Ji Bimari) | 5. Market Produce Prices (Mandi Jo Bhao) | 6. Weather & Rainfall (Barsaat Ji Khabar)
7. Seed Selection & Sowing (Bij Ji Chonnd) | 8. Soil Moisture Monitoring (Zameen Ji Nami) | 9. Tractor/Transport Booking | 10. Emergency Aid.

3. Empirical Evaluation & Comparative Benchmarks

Model Architecture	Tokenizer Type	Micro-F1 (%)	Latency (Edge CPU)	Parameters
TF-IDF + Linear SVM (Baseline)	Char N-Gram (3-5)	78.2%	1.4 ms	12 K
mBERT (Multilingual Base)	WordPiece (Standard)	81.6%	48.2 ms	110 M
IndicBERT v2 (Bilingual)	SentencePiece	84.1%	36.5 ms	278 M
XLM-RoBERTa (Base)	SentencePiece (BPE)	85.9%	52.1 ms	125 M
Dhatki-DistilBERT (Proposed)	Custom Vernacular BPE	89.4%	14.2 ms	66 M

4. Key Findings & The Power of Domain-Specific Tokenization

1. Tokenizer Failure in Generalized LLMs: Large pre-trained multilingual LLMs (mBERT, XLM-R) shatter phonetic Dhatki words into 4-6 arbitrary sub-character tokens, destroying semantic context and dramatically increasing computational latency.
2. Superiority of Compact Distilled Architectures: By training a specialized Byte-Pair Encoding (BPE) tokenizer on local vernacular syntax, our proposed Dhatki-DistilBERT architecture outperformed full-scale multilingual models by +3.5% in Micro-F1 while operating at 3.4x lower latency (14.2 ms), making it deployable on low-cost Android smartphones in offline desert environments.

5. Conclusion & Open-Access Data Availability

This work establishes the first verifiable benchmark for low-resource NLP in the Thar Desert. By demonstrating that lightweight, distilled models with tailored tokenizers achieve high classification accuracy under severe resource constraints, we provide an architectural blueprint for democratizing AI in marginalized agrarian ecosystems.

Data Availability: The dataset and evaluation scripts are released openly for academic and humanitarian research at:
Hugging Face: https://huggingface.co/datasets/lokeshsonara/dhatki-agri-intents
Live System Portal: https://agri-idas.tech/