LOW-RESOURCE NLP • SINDHI QA

SdQuAD

A Benchmark Question Answering Dataset for Low-resource Sindhi Language

Wazir Ali • Muhammad Rafay • Nadia Ali • Amar Rehman

Quaid-e-Awam University of Engineering, Science & Technology • Aror University, Sukkur

OUTLINE

01 **The Gap**

Why QA for Sindhi has been missing

02 **QA Landscape**

Key existing QA datasets — English & multilingual

03 **SdQuAD Dataset**

Data collection, domains, and annotation

04 **Quality Assessment**

Inter-annotator agreement metrics

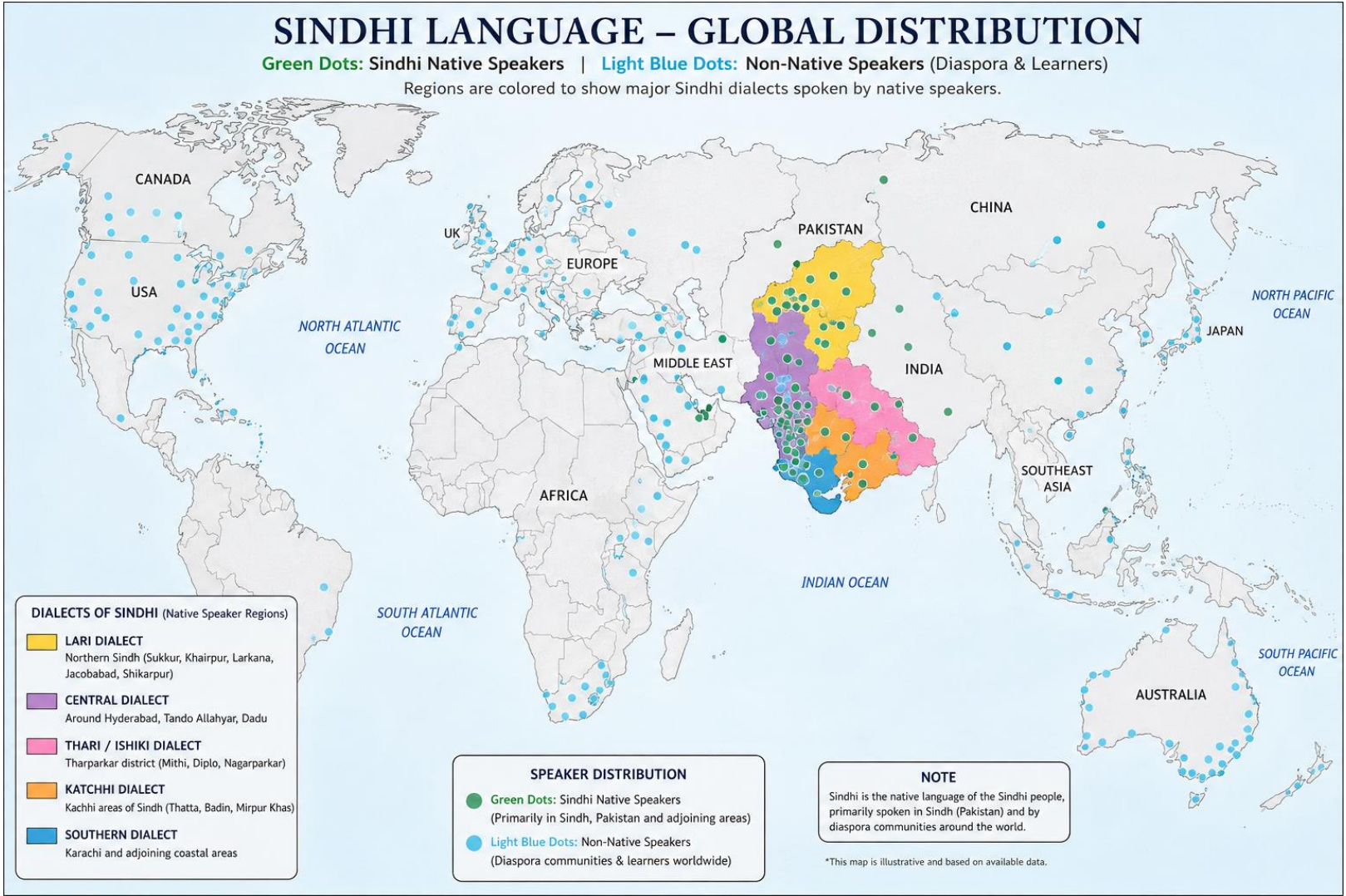
05 **Experiments & Models**

TF-IDF, mBERT, XLM-RoBERTa, mT5

06 **Results & Conclusions**

Performance analysis and future directions

Sindhi Language Speakers [Globally]



MOTIVATION

The Challenge: Question Answering in Low-Resource Languages

SINDHI SPEAKERS WORLDWIDE

30M+

Native speakers primarily in Pakistan and the Indian state of Gujarat

Only 200 QA pairs existed in IndicQuest (Sindhi portion)

- QA datasets enable semantic parsing, reading comprehension & open-domain reasoning
- SQuAD, HotpotQA, Natural Questions have driven English NLP
- Sindhi NLP exists only for POS tagging, NER, sentiment — no QA resources
- Low-resource barrier: absence of data prevents model development and evaluation
- SdQuAD addresses this with 14K+ expert-annotated Sindhi QA pairs across 6 domains

RELATED WORK

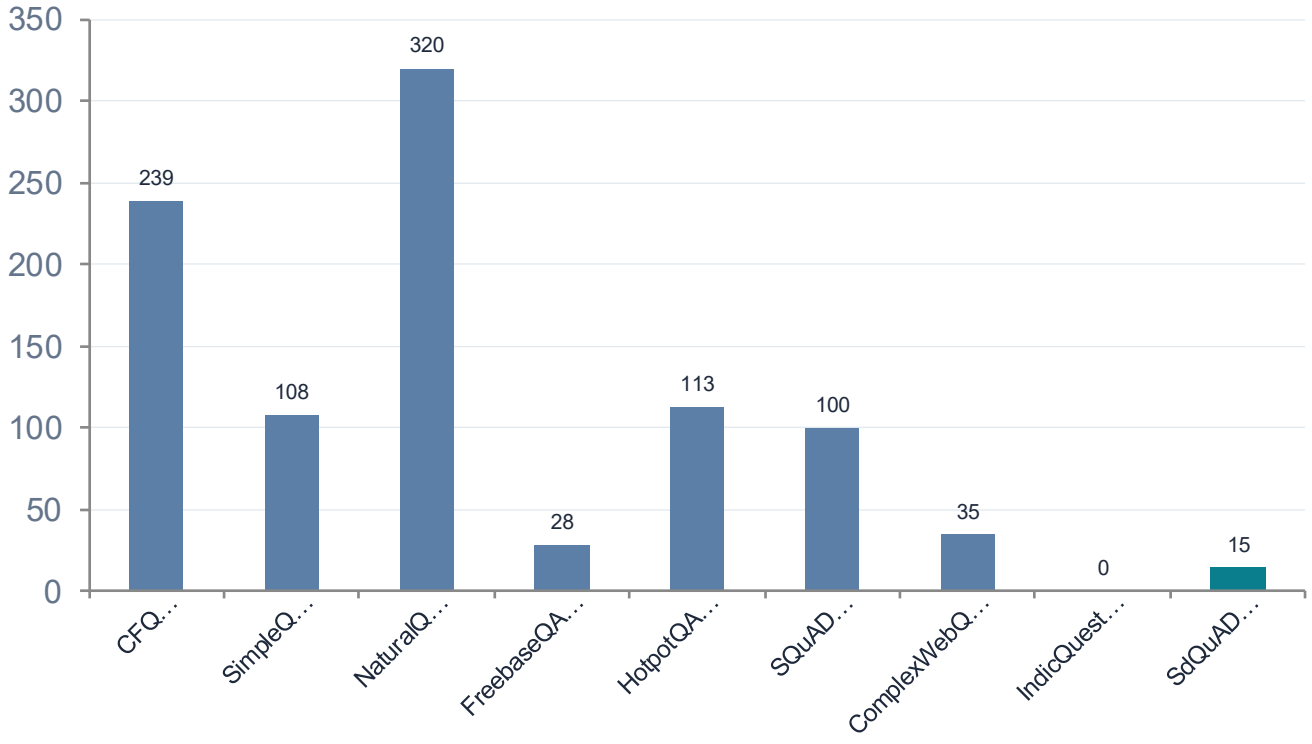
Key Question Answering Benchmarks: The Existing Landscape

WebQuestions 2013	SimpleQuestions 2015	SQuAD 1.0/2.0 2016/18	HotpotQA 2018	Natural Questions 2019
SIZE	**SIZE**	**SIZE**	**SIZE**	**SIZE**
5.8K Qs	108K Qs	100K+ Qs	113K Qs	320K Qs
FOCUS	**FOCUS**	**FOCUS**	**FOCUS**	**FOCUS**
Freebase answers Google Suggest API	Single-fact Freebase triple QA	Extractive RC from Wikipedia	Multi-hop, diverse explainable QA	Real Google queries open-domain

For low-resource languages: UQA (Urdu), UQuAD (Urdu) exist — but Sindhi had only 200 pairs in IndicQuest (2024). SdQuAD fills this gap.

RELATED WORK

QA Dataset Scale: Where SdQuAD Stands



SDQUAD

14,565
QA pairs

6
Domains

3
Native annotators

200×
vs. IndicQuest Sindhi

DATASET

SdQuAD: Data Collection & Domain Distribution

Science	Geography	News	Tourism	Business	History
5,080	2,153	2,132	1,886	1,695	1,619

News
Awami Awaz newspaper — politics, events, society

Science
Sindhi textbooks — physics, biology, chemistry

Geography
Geography books & educational resources

History
Books from Sindh Salamat (open access)

Tourism
Sindh Tourism Dept. & travel blogs

Business
Associated Press of Pakistan — economic affairs

DATASET

Annotation Process & Dataset Structure

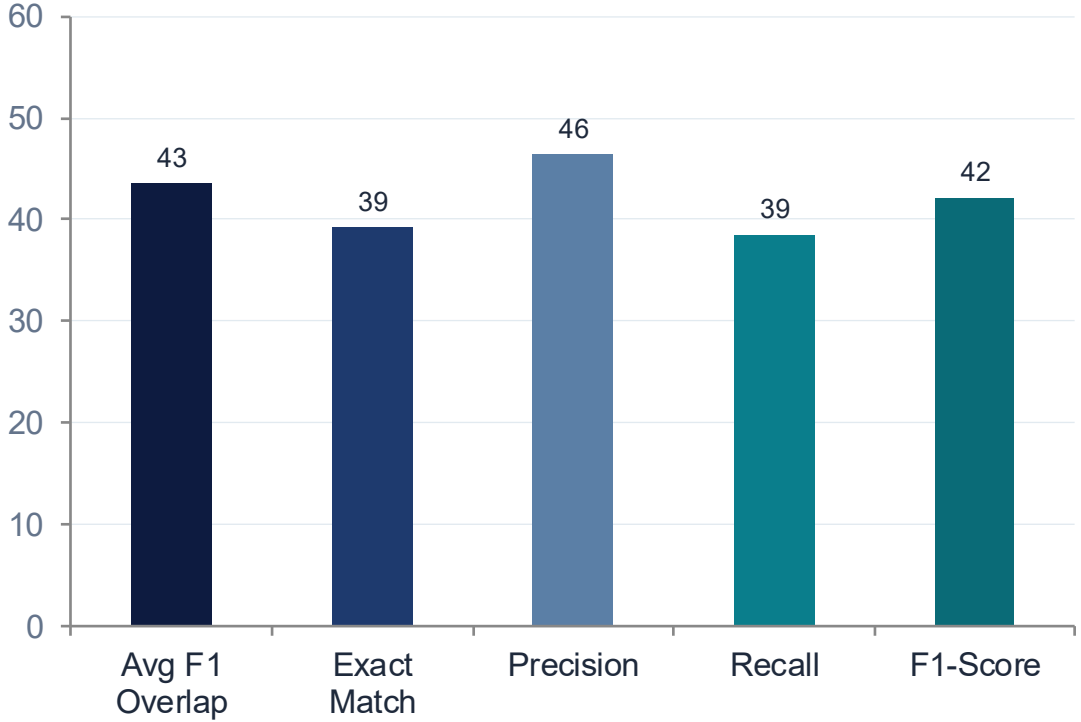
STEP 1	STEP 2	STEP 3	STEP 4	STEP 5
Web Scraping	**Preprocessing**	**Label Studio**	**Span Marking**	**JSON Export**
20K+ raw documents collected from 7 source categories	Filtering, deduplication, OOV removal, Sindhi normalisation	3 native Sindhi speakers annotate context-question-answer triples	Exact spans for extractive QA; paraphrased for abstractive	Train (12,052) / Test (3,013) split in SQuAD-format JSON

Example from SdQuAD

QUESTION	CONTEXT	ANSWER
What is the capital of Sindh? (سنڌ جو راڄڌاني ڪهڙو آهي؟)	Sindh is a province of Pakistan whose capital is Karachi. (سنڌ پاڪستان جو هڪ صوبو آهي جنهن جو راڄڌاني ڪراچي آهي)	Karachi (ڪراچي)

QUALITY ASSESSMENT

BASELINE RETRIEVAL



DATASET SPLIT	
12,052	Training pairs
3,013	Test pairs
14,565	Total QA pairs

EXPERIMENTS

Experimental Setup: Baseline & Model Configurations

TF-IDF
Lexical Baseline

CONFIG

scikit-learn vectorizer
Custom Sindhi tokenizer
Term-frequency matching

METRICS

Exact Match (EM)
Token-level F1-score

mBERT
Encoder-only

CONFIG

100+ language pretraining
lr=2×10^{-4}, batch=16
5 epochs, Colab T4 GPU

METRICS

Exact Match (EM)
Token-level F1-score

XLM-RoBERTa
Encoder-only

CONFIG

Large multilingual corpus
lr=2×10^{-4}, batch=16
5 epochs, Colab T4 GPU

METRICS

Exact Match (EM)
Token-level F1-score

mT5
Encoder-Decoder

CONFIG

Seq2seq multilingual
lr=2×10^{-4}, batch=16
5 epochs, Colab T4 GPU

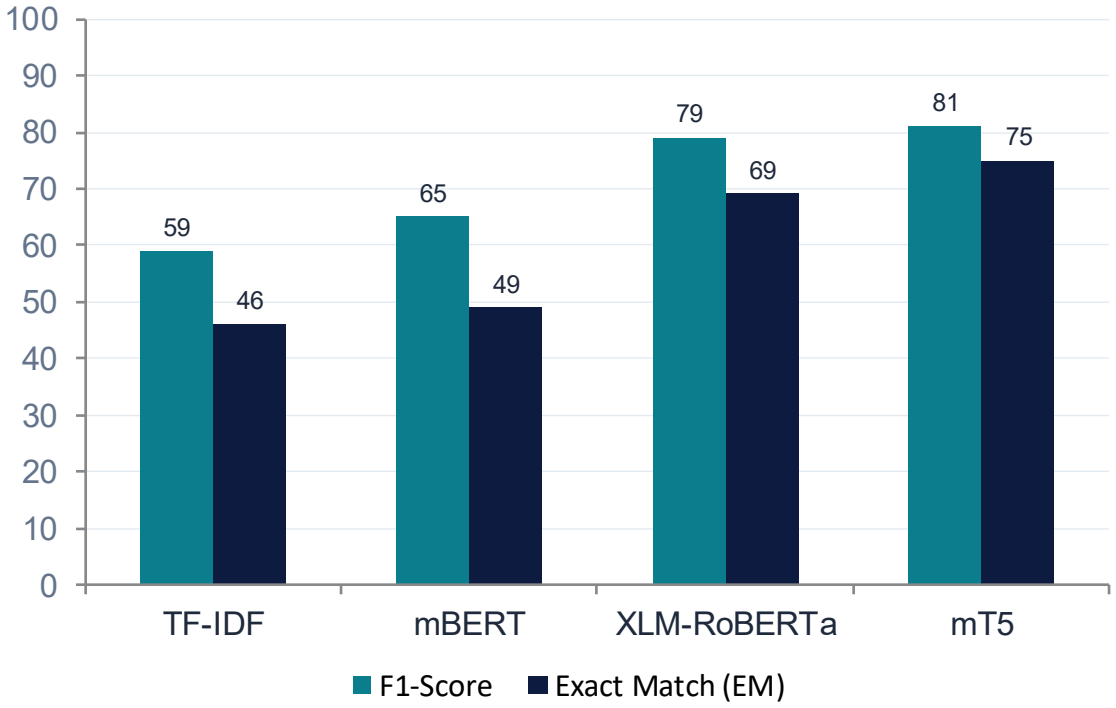
METRICS

Exact Match (EM)
Token-level F1-score

All transformer models use pretrained multilingual tokenizers — no new tokenizer trained (100+ language subword coverage already includes Sindhi script)

RESULTS

Model Performance on SdQuAD



mT5
F1: 81.47
Best overall — seq2seq architecture + large multilingual pretraining

XLM-R
F1: 79.28
Competitive encoder — strong multilingual representations

mBERT
F1: 64.89
Weakest transformer — less precise span boundary detection

TF-IDF
F1: 59.46
Lexical baseline — retrieves keywords but misses exact spans

CONCLUSIONS

What SdQuAD Contributes & Future Directions

C1 First Sindhi QA Dataset

14,565 high-quality QA pairs from 6 domains — the only open-source Sindhi QA benchmark (200× larger than IndicQuest Sindhi portion).

C2 Multi-domain Coverage

Science, News, Geography, Tourism, Business, History — captures Sindhi's linguistic and cultural richness for generalisation.

C3 Extractive & Abstractive

Supports both QA paradigms with exact span annotations and paraphrased answers — flexible for future research directions.

C4 Open Release on HuggingFace

Dataset publicly available at Aliwj/SdQuAD — fine-tuned models for mBERT, XLM-RoBERTa, and mT5 also released.

Thank You

Key Takeaway

SdQuAD is the first open-source benchmark QA dataset for Sindhi — 14,565 expert-annotated pairs across 6 domains — enabling NLP research for 30M+ Sindhi speakers. mT5 achieves the best performance with F1: 81.47 and EM: 74.58.

14,565	6	81.47	200×
QA Pairs	Domains	mT5 F1-Score	vs. prior Sindhi QA

aliwazirjam@gmail.com • HuggingFace: Aliwj/SdQuAD • Quaid-e-Awam University & Aror University