THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 6. Issue 1 (2026)
https://doi.org/10.62019/en6geh21

Low-Resource Balochi Named Entity Recognition: Corpus Construction and Multilingual Transformer Evaluation

Nazish Basir *, Rafique Ahmed Vighio, Beenish Ansari, Noorulain Patoli, Danish Nazir Arain, Aijaz Ali

Chronicle

Article history
Received: Jan 2, 2026
Received in the revised format: Jan 13, 2026
Accepted: Feb 17 2026
Available online: March 9, 2026

Nazish Basir & Noorulain Patoli, are currently affiliated with Department of Information Technology, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan.
Email: nazish.basir@usindh.edu.pk
Email: noorulain.patoli@usindh.edu.pk

Rafique Ahmed Vighio is currently affiliated with Dr. AHS Bukhari Postgraduate Centre of ICT, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan
Email: rafique.vighio@usindh.edu.pk

Beenish Ansari is currently affiliated with Department of Electronic Engineering, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan
Email: beenish.ansari@usindh.edu.pk

Danish Nazir Arain is currently affiliated with Dr. A. H. S. Bukhari Postgraduate Centre Of ICT, University of Sindh, Jamshoro, Pakistan, Pakistan
Email: danish.arain@usindh.edu.pk

Aijaz Ali is currently affiliated with Department of Software Engineering, Faculty of Engineering and Technology, University of Sindh, Jamshoro Pakistan
Email:
aijaz.laghari@students.usindh.edu.pk

Corresponding Author*

Abstract

Named Entity Recognition (NER) remains largely unexplored for Balochi, a morphologically complex, low-resource language spoken by approximately 8 to 10 million population across Pakistan, Iran, and Afghanistan. This paper makes three contributions to address this gap. First, we introduce a IOB2-annotated Balochi NER corpus, comprising 1,909 sentences, 48,920 tokens, and 4,359 named entity annotations across six semantic categories: Location (LOC), Person (PER), Nationality/Religion/Political Group (NORP), Number (NUM), Organization (ORG), and Date/Time (DAT). Second, we benchmark four multilingual pre-trained Transformer models, mBERT, XLM-R, mmBERT, and RemBERT, under identical fine-tuning conditions using the AdamW optimizer with cosine learning rate scheduling. Third, we provide a comprehensive evaluation encompassing entity-level F1-scores, confusion matrix analysis, and learning curve examination. RemBERT achieved the best overall performance with a micro-averaged F1-score of 0.85, outperforming XLM-R (F1 = 0.84), mmBERT (F1 = 0.83), and mBERT (F1 = 0.81). The entities like NORP and DAT were the most challenging categories across all models due to class imbalance and high lexical ambiguity, while mmBERT exhibited the most pronounced overfitting behaviour. These results establish the first quantitative NER baseline for Balochi and demonstrate that multilingual pre-trained models, particularly RemBERT, can be effectively adapted to extremely low-resource languages with limited annotated data.

Keywords: Balochi language, named entity recognition, low-resource NLP, multilingual BERT, RemBERT, XLM-RoBERTa, IOB tagging, South Asian languages, sequence labelling.

INTRODUCTION

Natural Language Processing (NLP) has witnessed a paradigm shift with the rise of deep learning and large-scale pre-trained language models. Named Entity Recognition (NER), the automatic identification and classification of named entities such as persons, locations, organizations, and dates in unstructured text , constitutes one of the foundational tasks of NLP pipelines, enabling downstream applications

including information extraction, question answering, knowledge graph construction, and machine translation (Li et al., 2020; Nasar et al., 2021). While NER has been extensively studied for high-resource languages such as English, Chinese, and German, most of the world's languages , including many spoken across South Asia and the Middle East , remain largely underrepresented in the research literature (Hedderich et al., 2021; Joshi et al., 2020).

Balochi is one of the critically low-resource language. Spoken by an estimated 8–10 million people primarily in the Balochistan regions of Pakistan, Iran and Afghanistan, it belongs to the northwestern branch of the Iranian language family. Despite its sizeable speaker community, Balochi lacks annotated linguistic resources, standardized orthographic conventions and computational tools (Jahani, 2013). The language is written in Perso-Arabic script, right-to-left and exhibits rich morphological complexity, including extensive agglutinative features and context-dependent diacritics. These characteristics make it challenging to apply NLP techniques designed for resource-rich languages (Sharan, 2024)

The shortage of annotated corpora for Balochi means that standard supervised training approaches cannot be applied directly. The pre-train and fine-tune paradigm, exemplified by multilingual transformer architectures such as Multilingual BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), RemBERT (Chung et al., 2020), and mmBERT (Marone et al., 2025) offers a compelling alternative. These models are pre-trained on massive multilingual corpora spanning over 100 languages and can be fine-tuned on comparatively small task-specific datasets, making them particularly well-suited for low-resource settings (Hedderich et al., 2021; Ruder et al., 2019).

NER for languages written in Arabic script presents additional challenges. The absence of capitalisation cues, abundant diacritics, and high lexical ambiguity demand models with robust contextual reasoning abilities (Shaalan & Raza, 2009). Prior work on closely related languages, including Urdu, Sindhi (Basir, Hakro, et al., 2025), and Pashto(Basir, Haider, et al., 2025), has demonstrated that multilingual transformers substantially outperform traditional methods such as Conditional Random Fields (CRF) and Hidden Markov Models (HMM) when training data is scarce (Khan et al., 2020; F. Ullah et al., 2024, 2025).

The present paper makes the following contributions:

- We construct the IOB-annotated NER dataset for Balochi
- We benchmark four state-of-the-art multilingual pre-trained language models , mBERT, XLM-R, mmBERT, and RemBERT , using a standard IOB fine-tuning framework.
- We provide a detailed entity-category-level comparative analysis to identify which architectures transfer most effectively to Balochi's unique linguistic profile.

The remainder of this paper is organised as follows. The Related Work section reviews prior NER research for multilingual, low-resource, and South Asian language settings.. The Methodology section details the corpus construction and annotation, the four models and experimental setup. The Results section presents quantitative findings, followed by Discussion and Conclusion.

LITERATURE REVIEW

Named Entity Recognition has evolved through three broad paradigms: rule-based systems, statistical machine learning, and deep neural approaches (Li et al., 2020).

Transformer-based models transformed NER fundamentally. (Devlin et al., 2019) demonstrated that fine-tuning BERT on NER benchmarks produced substantial improvements without task-specific architectural modifications. The Transformer architecture (Vaswani et al., 2017) established the pre-train and fine-tune paradigm that now underpins most NLP research. Multilingual BERT (mBERT) extended this to 104 languages using Wikipedia text with masked language modeling objectives. Despite receiving no explicit cross-lingual training signal, mBERT demonstrated surprising zero-shot cross-lingual transfer abilities (Pires et al., 2019), making it a widely adopted baseline for low-resource NER.

(Conneau & Lample, 2019) proposed XLM, which introduced cross-lingual language model pre-training with translation language modeling. Building on this, (Conneau et al., 2020) introduced XLM-RoBERTa (XLM-R), trained on over two terabytes of CommonCrawl text across 100 languages. XLM-R substantially outperformed mBERT on cross-lingual NER benchmarks, achieving up to 2.4% F1 improvement, particularly for low-resource languages. The model's advantage stems from its larger and more diverse training corpus and adoption of the RoBERTa training procedure(Liu et al., 2019), which removed the next sentence prediction objective.

RemBERT (Chung et al., 2020), which decouples input and output embeddings during pre-training. This architectural modification allows greater representational capacity to be allocated to contextual understanding rather than vocabulary reconstruction, leading to stronger downstream performance on multilingual benchmarks. RemBERT's advantage is particularly pronounced for morphologically complex and low-resource languages, directly motivating its inclusion in our comparative study. Systematic benchmarks such as XTREME (Hu et al., 2020) confirm that XLM-R and RemBERT exhibit the most consistent cross-lingual transfer for low-resource languages not well-represented in mBERT's Wikipedia-based corpus.

Among South Asian languages, Urdu has received the most NER attention, owing to its shared Perso-Arabic script and structural similarities with several regional languages. The adoption of Transformer-based architectures for Urdu NER has yielded significant gains. (Anam et al., 2024) achieved strong results using BiLSTM-GRU with Floret embeddings, while (F. Ullah et al., 2024) demonstrated that BERT-multilingual with contextual word embedding augmentation achieves F1 scores exceeding 98% on augmented Urdu NER datasets. A study by (Ahmed et al., 2024) further showed that enriching Urdu NER with BERT embeddings and hybrid encoder-CNN architectures consistently surpasses simpler baselines. For cross-lingual settings, (F. Ullah et al., 2025) explored multilingual NER on Arabic and Urdu tweets, illustrating the challenges of handling noisy and code-mixed content even with transformer-based models.

NER efforts have also been extended to other South Asian languages, (Mhaske et al., 2023) introduced Naamapadam, a large-scale NER corpus for 11 Indic languages, providing a reference framework for low-resource annotation. Another study by (Gorla et al., 2022) demonstrated BERT-based Telugu NER, while (Khalid et al., 2023) explored data augmentation with Transformer models for Punjabi NER, establishing that the fine-tuning paradigm generalises across South Asian scripts. A comprehensive survey by (Hedderich et al., 2021) identifies data augmentation, distant supervision, and cross-lingual transfer as the three most effective strategies for NLP in low-resource settings. Languages entirely absent from the pre-training corpora of mBERT and XLM-R represent an extreme low-resource scenario (Muller et al., 2021). For such languages, adapter-based fine-tuning frameworks such as MAD-X (Pfeiffer et al., 2020) and multilingual fine-tuning with language family proxies have emerged as promising

directions. Few-shot NER approaches (Moscato et al., 2023) are particularly relevant where annotation costs are high and domain expertise is scarce. Meta-learning frameworks and prototypical networks have been applied to enable NER from as few as five examples per entity type. The limited body of computational NLP work on Balochi provides important context for the present study. (Sharan, 2024) developed an N-gram language model for next-word prediction in Balochi, constructing a small unambiguous corpus and demonstrating that probabilistic n-gram approaches can capture basic sequential patterns in the language, though the study acknowledged the severe shortage of Balochi digital text as a fundamental constraint. (S. Ullah et al., 2024) extended Balochi NLP to Part-of-Speech tagging using Conditional Random Fields, establishing a data-driven annotation framework and confirming that Balochi's morphological complexity poses significant challenges for sequence labelling tasks even with statistical models.

More recently, (Hussain et al., 2025) performed sentiment analysis on Balochi text using deep learning and contributed a newly constructed Balochi sentiment dataset they evaluating LSTM and GRU models against traditional classifiers such as SVM and Random Forest. Their results showed LSTM achieving 83.57% accuracy, confirming that deep learning architectures generalise more effectively than conventional methods on Balochi text. Another work by (van Linschoten, 2023) documented an informal effort to collect and curate a small Balochi text dataset for NLP purposes, published as a technical blog post rather than a peer-reviewed study. Taken together, these studies establish that while foundational NLP tools for Balochi are beginning to emerge, no prior work has addressed the NER task undertaken in the present paper. This study directly addresses all three gaps by constructing a new annotated dataset and systematically evaluating four leading multilingual pre-trained models.

METHODOLOGY

A. Data Collection and Pre-processing

Online news portals and digital newspapers were used to collect raw text of Balochi-language, as they possess high named entity density and formal written style. After removing HTML markup, duplicate articles, and non-Balochi content, sentences were segmented using punctuation boundaries and tokenised at the whitespace level, producing a corpus of 1,909 sentences and 48,920 tokens. The dataset follows most commonly used IOB2 tagging scheme in which *B-* marks the beginning of an entity span, *I-* marks continuation, and *O* marks non-entity tokens. Six entity categories were annotated that are: Person (PER), Location (LOC), Organization (ORG), Nationality/Religion/Political Group (NORP), Number (NUM), and Date/Time (DAT). Table 1 illustrates a sample annotated sentence. Annotation was carried out by two trained annotators with native or near-native Balochi Language proficiency. A calibration session was held prior to annotation to align on category definitions and boundary conventions. Disagreements were resolved through discussion and adjudication by a senior annotator, a simple peer review technique was maintain the integrity.

Table 2 presents the entity distribution across the six categories. LOC is the most frequent entity type (34.18%), followed by PER (23.60%) and NUM (18.93%). NORP and DAT are the least frequent categories (6.61% and 4.08% respectively), creating a class imbalance that directly affects model performance on these categories.

Table 1
Sample IOB2 Annotation of a Balochi Sentence

English									
Amjad	**Memon**	**was**	**coming**	**from**	**Lahore**	**on**	**23**	**August**	**2025**
B-PER	I-PER	O	O	O	B-LOC	O	B-DAT	I-DAT	I-DAT
امجد	**میمن**	**چہ**	**لاہور**	**ءَ**	**آھگ**	**ءَ**	**اَت**	**اگست**	**بلوچی 2025**
B-PER	I-PER	O	B-LOC	O	O	O	O	B-DAT	I-DAT

Non-entity tokens account for 84.28% of all tokens, which is typical for news-domain NER corpora.

Table 2.
Named Entity Category Distribution

Entity Category	Count	Proportion (%)
LOC (Location)	1,490	34.18%
PER (Person)	1,029	23.60%
NUM (Number/Numerical)	825	18.93%
ORG (Organization)	549	12.59%
NORP (Nationality/Religion/Political)	288	6.61%
DAT (Date/Time)	178	4.08%
Total Named Entities	**4,359**	**100%**
Non-entity Tokens (O)	41,230	--
Total Tokens	**48,920**	--

The dataset was split into 80% training (1,527 sentences) and 20% validation (382 sentences) using a stratified random split with a fixed seed of 42, ensuring reproducibility across all four model experiments.

B. Overview of the Experimental Framework

This study adopts the pre-train and fine-tune paradigm, in which a multilingual Transformer model pre-trained on large-scale multilingual corpora is subsequently fine-tuned on the Balochi NER dataset described in the preceding section. Four state-of-the-art multilingual pre-trained language models are evaluated under identical experimental conditions to enable fair comparison: Multilingual BERT (mBERT), XLM-RoBERTa (XLM-R), Google Remade BERT (RemBERT) and Modern Multilingual Encoder with Annealed Language Learning (mmBERT)All models are accessed through the HuggingFace Transformers library (Ruder et al., 2019) and fine-tuned using the token classification head provided by the *ForTokenClassification* API, which appends a linear classification layer over the final hidden states of each token. Table 4 summarises the four models evaluated in this study. Each model represents a distinct point in the design space of multilingual pre-trained language models, varying in pre-training data volume, vocabulary size, architectural modifications, and parameter count.

Four multilingual pre-trained Transformer models were selected for this benchmark. mBERT serves as the primary baseline, demonstrating strong zero-shot cross-lingual transfer despite receiving no explicit cross-lingual training signal during its masked language modeling and next sentence prediction pre-training. XLM-R, training exclusively with the MLM objective on a substantially larger and more diverse corpus, which achieved stronger cross-lingual NER performance particularly for low-resource languages. RemBERT introduces a key architectural distinction by decoupling input and output embedding matrices, freeing the Transformer layers to develop richer contextual representations rather than allocating capacity to vocabulary

reconstruction make it suited for morphologically complex languages. Finally, mmBERT that is based on the ModernBERT architecture, uses a significantly larger pre-training corpus spanning over 1,800 languages, representing the most extensively pre-trained model in this benchmark.

Table 4.
Summary of Pre-trained Language Models Used in This Study

Model	Base Architecture	Pre-training Data	Parameters	Languages
mBERT	BERT	Wikipedia	~177M	104
XLM-R	RoBERTa	CommonCrawl	~278M	100
RemBERT	BERT variant (decoupled embeddings)	Wikipedia	~575M	110
mmBERT	ModernBERT-based encoder	~3T multilingual tokens	~307M	1800+

Note. Parameter counts are approximate and reflect the base model variants used for fine-tuning. M = million.

C. Token Classification with IOB2 Labels

All four models are fine-tuned for token-level sequence labelling using the IOB2 scheme. Each model processes an input sentence as subword tokens, with a linear classification head applied over the final hidden state of each token to predict its entity label.

A key challenge is the many-to-one mapping between subwords and word, a single word may be split into multiple subword tokens yet carries only one IOB label. To handle this, only the first subword of each word is assigned the word-level label; all subsequent subword tokens of the same word are masked with an ignore index (−100) and excluded from loss computation. Special tokens [CLS], [SEP], and padding are treated the same way.

D. Training Procedure and Hyperparameters

All four models were fine-tuned under identical hyperparameter settings, detailed in Table 5, ensuring that performance differences reflect model architecture rather than training configuration.

Table 5.
Hyperparameter Configuration for All Models

Hyperparameter	Value
Learning rate	3×10^{-5}
Optimizer	AdamW
Weight decay	0.01
LR scheduler	Cosine decay with linear warmup
Warmup proportion	10% of total steps
Batch size (train & validation)	4
Maximum sequence length	256 tokens
Number of training epochs	5
Loss function	Cross-entropy (subword tokens masked with −100)
Random seed	42
Hardware	Google Colab (NVIDIA T4 GPU)
Framework	PyTorch + HuggingFace Transformers

AdamW was used with a weight decay of 0.01 to mitigate overfitting on the small training set. A cosine learning rate schedule with 10% linear warmup was applied, improving convergence stability over constant or linear decay alternatives. Input sequences were padded or truncated to 256 tokens, well above the average

sentence length of 25.6 tokens, and cross-entropy loss was computed with subword and special tokens masked as described above.

E. Evaluation Methodology

Model performance is evaluated using strict span-level assessment; where a prediction is counted as correct only if both the entity boundary and the entity type exactly match the gold annotation and the partial matches are counted as errors. Primary metrics used are precision, recall, and F1-score computed at the entity span level. Micro-averaged F1 is the primary overall metric as it weights performance by entity frequency, while macro-averaged F1 is additionally reported to assess robustness across rare categories such as DAT and NORP. Entity-level confusion matrices are analysed to distinguish missed entities, false positives and type confusions. The learning curves are recorded per model to assess convergence and detect overfitting.

EXPERIMENTS AND RESULTS

A. Overall Model Performance

Table 6 presents the overall precision, recall, and F1-scores (micro-averaged, macro-averaged, and weighted) for all four models evaluated on the Balochi NER validation set. RemBERT achieves the highest overall performance with a micro-averaged F1 of 0.85, precision of 0.86, and recall of 0.84, followed by XLM-R (F1 = 0.84), mmBERT (F1 = 0.83), and mBERT (F1 = 0.81). Macro-averaged F1 follows the same ranking RemBERT (0.82) > XLM-R (0.80) > mmBERT (0.79) > mBERT (0.77). The largest micro-to-macro gap belongs to mBERT (0.04), indicating it struggles most on low-frequency categories. These results confirm that architectural innovations, particularly RemBERT's decoupled embedding design, yield meaningful gains even on a low-resource language.

Table 6.
Overall Evaluation Results Across All Four Models

Model	Precision	Recall	F1 (Micro)	F1 (Macro)	F1 (Weighted)
mBERT	0.80	0.81	0.81	0.77	0.81
XLM-R	0.83	0.85	0.84	0.80	0.84
mmBERT	0.83	0.82	0.83	0.79	0.83
RemBERT	**0.86**	**0.84**	**0.85**	**0.82**	**0.85**

Note. Best scores per column are highlighted. All scores computed using strict span-level IOB2 evaluation via seqeval.

B. Per-Entity Category Performance

Figure 1 present the corresponding evaluation report heatmaps visualising precision, recall, and F1 simultaneously for each model.

Location (LOC) is the best-performing category across all models, with F1-scores ranging from 0.87 (mBERT and mmBERT) to 0.91 (XLM-R). This is consistent with LOC being the most frequent entity type in the corpus (n = 322 in the validation set) and the observation that location names in Balochi news text tend to follow recognisable orthographic patterns. RemBERT achieves an F1 of 0.90 for LOC, reflecting strong boundary detection for this dominant category.

Figure 1 *Evaluation Report Heatmaps for All Four Models (A: mBERT, B: XLM-R, C: mmBERT, D: RemBERT). Precision, recall, and F1-score per entity category and aggregate averages on the Balochi NER validation set.*

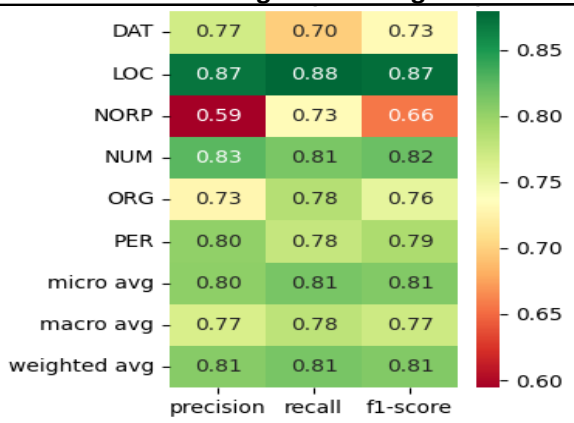

Figure A mBERT

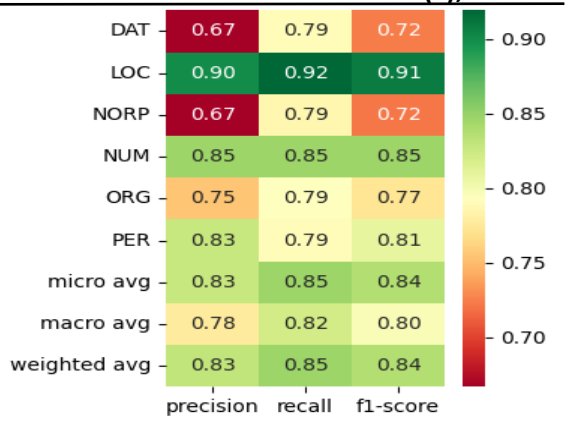

Figure 1B XLM-R

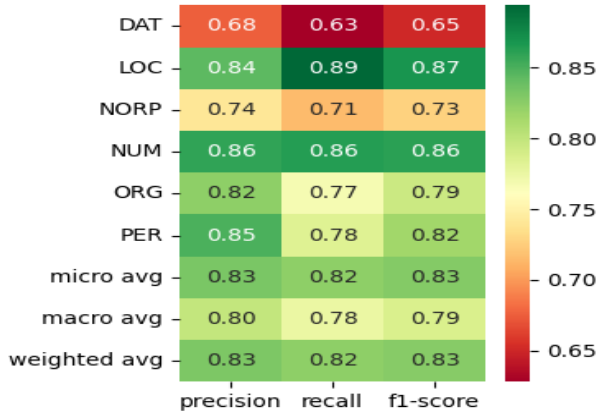

Figure 1C mmBERT

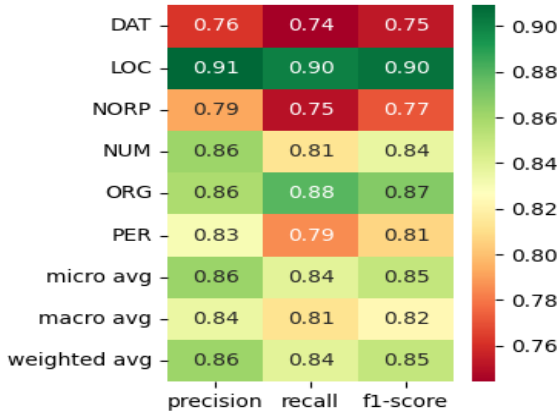

Figure 1D RemBERT

Note. Colour scale ranges from low (red) to high (green). Rows represent entity categories; columns represent precision, recall, and F1-score respectively.

Organization (ORG) recognition shows the clearest advantage of RemBERT, which achieves an F1 of 0.87 , the highest among all models for this category. In contrast, mBERT (F1 = 0.76) and XLM-R (F1 = 0.77) perform notably lower, suggesting that ORG names , which often consist of multi-token abbreviations and formal institutional names , benefit from the richer contextual representations provided by RemBERT's decoupled embedding architecture.

Nationality, Religion, and Political Group (NORP) is consistently the most challenging category, recording the lowest F1-scores across three of the four models: mBERT (0.66), XLM-R (0.72), mmBERT (0.73). RemBERT achieves the best NORP performance at 0.77. The problem with NORP is attributable to the small support count (n = 56 in validation), the high lexical overlap between NORP terms and common nouns or adjectives in Balochi, and the inherent ambiguity in distinguishing political from ethnic or religious group references.

Date and Time (DAT) presented a moderate challenge for all models (F1 range: 0.65–0.75), with the smallest support count (n = 43). mBERT (0.73) and RemBERT (0.75) outperform XLM-R (0.72) and mmBERT (0.65) on this category. The variability in DAT performance likely reflects inconsistencies in how Balochi temporal expressions are rendered in news text, including mixed use of Arabic-script numerals, written-out month names, and Islamic calendar references.

Numerical expressions (NUM) are well-handled by all models (F1 range: 0.82–0.86), benefiting from their relatively high frequency (n = 176) and the fact that digits and number words are visually and contextually distinctive in Balochi text. Person names (PER) likewise show stable performance across models (F1 range: 0.79–0.82), with XLM-R and mmBERT both achieving 0.81–0.82.

C. Entity-Level Confusion Matrix Analysis

Figure 2 present the entity-level confusion matrices for each model. These matrices capture not only correct predictions along the diagonal but also the distribution of boundary misses (entities predicted as O) and type confusions (entities predicted with an incorrect category label).

Figure 2 *Entity-Level Confusion Matrices for All Four Models (A: mBERT, B: XLM-R, C: mmBERT, D: RemBERT). Rows represent actual entity categories; columns represent predicted categories. Diagonal values indicate correct predictions.*

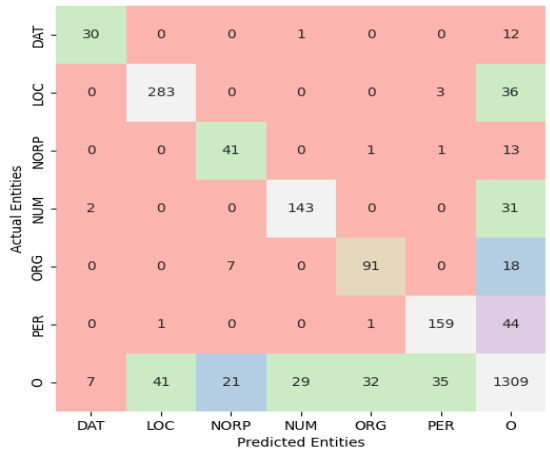

Figure 2A mBERT

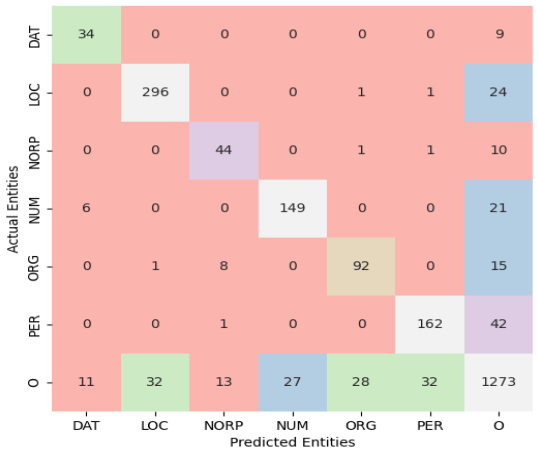

Figure 2B XLM-R

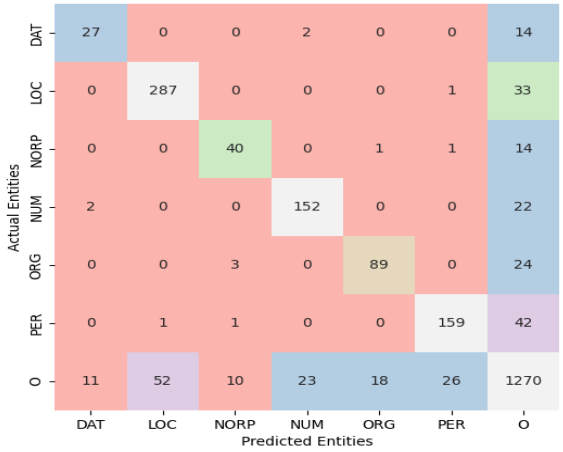

Figure 2C mmBERT

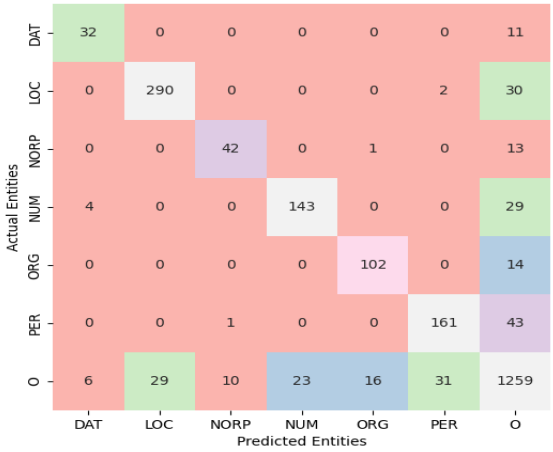

Figure 2D remBER

Note. O = non-entity tokens. Off-diagonal values indicate misclassifications or boundary errors.

Across all four models, the most prevalent error type is boundary miss , predicted as O when a true entity was present. For mBERT, 44 PER tokens, 36 LOC tokens, and 31 NUM tokens were falsely predicted as O, reflecting a tendency to under-detect entities at

the boundaries of longer spans. RemBERT shows the fewest boundary misses overall, with 43 PER, 30 LOC, and 29 NUM tokens predicted as O , a modest but consistent improvement.

Type confusion , where an entity span is correctly detected but assigned the wrong label , is most common between NORP and ORG categories. Across all models, a non-trivial number of ORG entities are misclassified as NORP (mBERT: 7; XLM-R: 8; mmBERT: 3; RemBERT: 0), reflecting the lexical overlap between political party names (ORG) and political group references (NORP) in Balochi news. RemBERT is the only model to record zero ORG→NORP confusions, suggesting that its deeper contextual representations better distinguish institutional from group-level references. False positive predictions , where the model predicts an entity on a true O token , are most frequent for LOC across all models (mBERT: 41 O→LOC; XLM-R: 32; mmBERT: 52; RemBERT: 29). The notably high O→LOC false positive rate for mmBERT (52) is the largest such error in the entire benchmark, suggesting that mmBERT has a higher tendency to over-predict location boundaries. This pattern may reflect mmBERT's exposure to location-dense training data during pre-training, causing it to generalise location patterns aggressively on the Balochi corpus.

D. Learning Curves and Convergence Analysis

All four models demonstrate consistent convergence behaviour: training loss decreases monotonically across all five epochs, while validation loss decreases rapidly in the first three epochs and then stabilises or shows marginal increase in epochs four and five. This pattern is characteristic of effective fine-tuning with cosine learning rate decay and indicates that the models have largely converged by epoch three. Figure 3 present the corresponding learning curve plots.

Training time per epoch varies considerably, mBERT and XLM-R complete each epoch in approximately 1.5–1.8 minutes, mmBERT in 2.3 minutes, and RemBERT in 7.8 minutes, consistent with their respective parameter counts. Despite the computational cost, RemBERT's superior generalisation justifies its use on this limited dataset.

Figure 3 *Learning Curves for All Four Models (A: mBERT, B: XLM-R, C: mmBERT, D: RemBERT). Training and validation cross-entropy loss across five fine-tuning epochs on the Balochi NER dataset.*

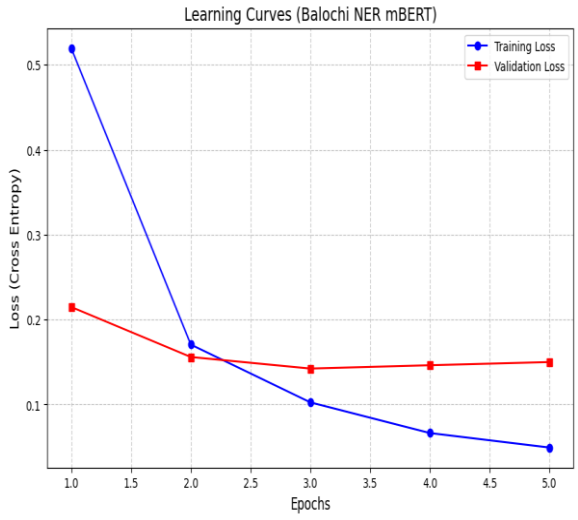

***Figure 3A* mBERT**

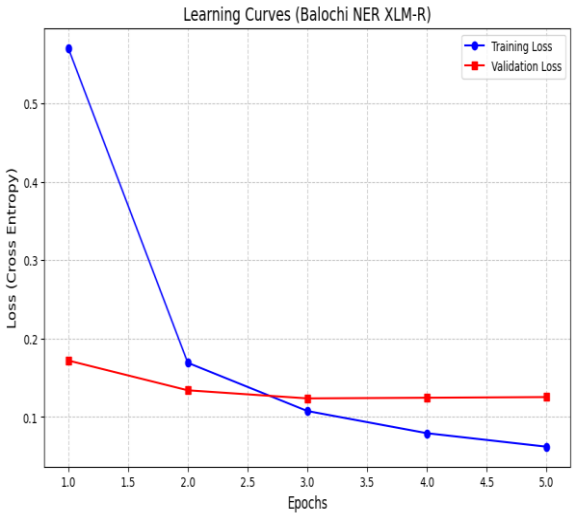

***Figure 3B* XLM-R**

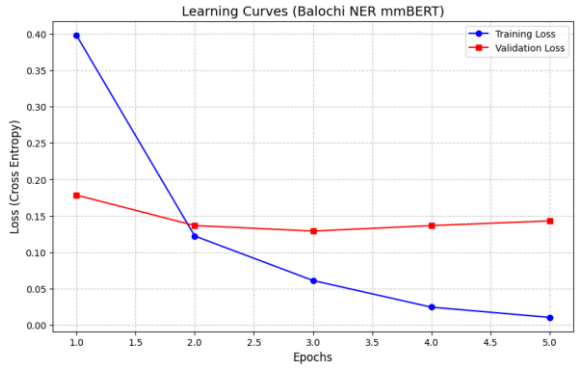

Figure 3C mmBERT

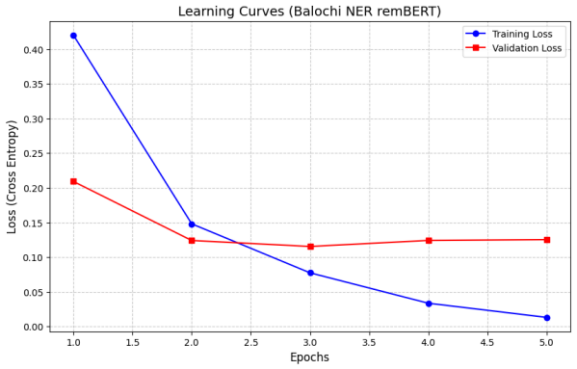

Figure 3D remBERT

Note. Blue line = training loss; Red line = validation loss. Divergence between the two lines indicates overfitting.

DISCUSSION

A. *Model Performance*

RemBERT achieved the best results across all aggregate metrics and performed best on the three most ambiguous categories: ORG, NORP, and DAT. Its core architectural advantage is the decoupling of input and output embedding matrices, which frees the Transformer layers to develop deeper contextual representations rather than allocating capacity to vocabulary reconstruction. For Balochi, where correct entity identification depends heavily on surrounding context rather than surface form, this depth is particularly valuable. The 0.11 F1 gap between RemBERT and mBERT on ORG entities illustrates this clearly: organization names in Balochi news are multi-token, institutionally specific, and rarely repeated across sentences, making context the primary recognition cue. XLM-R achieved the highest LOC F1 (0.91) due to its geographically rich CommonCrawl pre-training data, but this advantage on the most frequent category only partially offsets its weaker performance on rarer ones.

B. Class Imbalance and Challenging Categories

NORP and DAT are the lowest-performing categories across all models, directly reflecting their low frequency in the corpus 288 and 178 instances respectively. NORP presents an additional linguistic challenge: many nationality and political group descriptors in Balochi are morphological derivations of location or person roots, creating systematic ambiguity that requires discourse-level context to resolve. DAT performance is further complicated by orthographic variability dates appear as Eastern Arabic numerals, Western Arabic numerals, written-out month names, and Islamic calendar references within the same corpus. Both issues point to targeted data augmentation for these two categories as the most impactful near-term improvement.

C. Overfitting in mmBERT

mmBERT shows the most severe overfitting in this benchmark. Its training loss reaches 0.0153 by epoch five, the lowest of any model yet its validation loss of 0.1380 and training-validation gap of 0.1227 are both the highest. Validation loss begins rising after epoch two, earlier than any other model. This is further reflected in the confusion matrix, where mmBERT produces 52 false positive LOC predictions the highest across

all models indicating that it has memorised surface-level location cues from the training set and over-applies them on unseen data. Early stopping at epoch two or three would likely improve mmBERT's generalisation on this corpus size.

D. Limitations

Four limitations should be noted. First, the dataset of 1,909 sentences is small by NLP standards, amplifying class imbalance effects and increasing overfitting risk. Second, a single 80/20 split was used rather than k-fold cross-validation, so performance estimates may vary across different data partitions. Third, no formal inter-annotator agreement score was computed, meaning label consistency cannot be quantitatively verified, particularly for ambiguous categories like NORP and ORG. Fourth, all data comes from a single domain news articles and generalisation to other text types such as social media or literature remains untested.

CONCLUSION

This paper presented the first systematic NER study for the Balochi language. A new IOB2-annotated corpus of 1,909 sentences, 48,920 tokens, and 4,359 named entity annotations across six categories was constructed from Balochi news articles. Four multilingual pre-trained Transformer models; mBERT, XLM-R, mmBERT, and RemBERT were benchmarked under identical fine-tuning conditions, and their performance was evaluated using strict span-level metrics, confusion matrices, and learning curve analysis.

RemBERT achieved the best overall results with a micro-averaged F1 of 0.85, outperforming XLM-R (0.84), mmBERT (0.83), and mBERT (0.81). Its decoupled embedding architecture provided deeper contextual representations that were particularly effective for ORG, NORP, and DAT — the three most ambiguous and context-dependent entity categories in the corpus. XLM-R led on LOC recognition (F1 = 0.91) owing to its geographically rich pre-training data. NORP and DAT remained the most challenging categories across all models, primarily due to low corpus frequency and high lexical ambiguity. mmBERT exhibited the most severe overfitting, with the widest training-validation loss gap and the highest false positive rate, highlighting the risks of fine-tuning large models on small datasets without sufficient regularisation.

The most impactful next steps are corpus expansion and targeted data augmentation for minority entity categories, particularly DAT and NORP. Adapter-based fine-tuning and cross-lingual transfer from typologically related languages such as Persian and Pashto offer promising low-resource modelling alternatives. In the longer term, training a dedicated monolingual Balochi language model on large-scale web-collected text would provide a transformative foundation for all downstream NLP tasks for this language.

Closing Remarks

Balochi is spoken by millions of people yet remains one of the most computationally underserved languages in the world. The corpus and benchmark introduced in this paper constitute a first step toward building the NLP infrastructure that Balochi speakers deserve, and it is hoped that they will serve as a foundation for continued research in this underexplored field.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor to the research and their different affiliations.
Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.
Availability of data and material: In the approach, the data sources for the variables are stated.
Authors' contributions: Each author participated equally in the creation of this work.
Conflicts of Interest: The authors declare no conflict of interest.

Consent to Participate: Yes
Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Ahmed, A., Huang, D., & Arafat, S. Y. (2024). Enriching Urdu NER with BERT Embedding, Data Augmentation, and Hybrid Encoder-CNN Architecture. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(4). https://doi.org/10.1145/3648362

Anam, R., Anwar, M. W., Jamal, M. H., Bajwa, U. I., Diez, I. de la T., Alvarado, E. S., Flores, E. S., & Ashraf, I. (2024). A deep learning approach for Named Entity Recognition in Urdu language. Plos One, 19(3), e0300725.

Basir, N., Haider, G., Arain, D. N., & Nizamani, S. B. (2025). Enhancing Pashto NER Using Machine-Labeled Data and Transformer-Based Models. 2025 20th International Conference on Emerging Technologies (ICET). https://doi.org/10.1109/ICET66147.2025.11321232

Basir, N., Hakro, D. N., Khoumbati, K. U. R., & Bhatti, Z. (2025). Leveraging machine-labeled data and cross-lingual transfer for NER in Urdu and sindhi. J. Inf. Commun. Technol.—(JICT), 19, 1–8.

Bioscience Research, 2(1), 25-31. https://doi.org/10.70749/ijbr.v2i1.2286

Chung, H. W., Fevry, T., Tsai, H., Johnson, M., & Ruder, S. (2020). Rethinking embedding coupling in pre-trained language models. ArXiv Preprint ArXiv:2010.12821.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. Advances in Neural Information Processing Systems, 32.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

Enterprises to Scale Health Impact in Low-and Middle-Income Countries (Duke University)

Gorla, S., Tangeda, S. S., Neti, L. B. M., & Malapati, A. (2022). Telugu named entity recognition using bert. International Journal of Data Science and Analytics, 14(2), 127–140.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2545–2568.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. International Conference on Machine Learning, 4411–4421.

Hussain, S., Bazaib, S. U., Qadir, S., Marjan, S., Ghafoor, M. I., & Pervaiz, P. (2025). Sentiment Analysis of Balochi Text Using Deep Learning. VAWKUM Transactions on Computer Sciences, 13(1). https://doi.org/10.21015/vtcs.v13i1.2081

Jahani, C. (2013). The Balochi language and languages in Iranian Balochistan. The Journal of the Middle East and Africa, 4(2), 153–167.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6282–6293.

Khalid, H., Murtaza, G., & Abbas, Q. (2023). Using data augmentation and bidirectional encoder representations from transformers for improving Punjabi named entity recognition. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6), 1–13.

Khan, W., Daud, A., Alotaibi, F., Aljohani, N., & Arafat, S. (2020). Deep recurrent neural networks with word embeddings for Urdu named entity recognition. ETRI Journal, 42(1), 90–100. https://doi.org/10.4218/etrij.2018-0553

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, 34(1), 50–70. https://doi.org/10.1109/TKDE.2020.2981314

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. ArXiv Preprint ArXiv:1907.11692.

Marone, M., Weller, O., Fleshman, W., Yang, E., Lawrie, D., & Van Durme, B. (2025). mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. ArXiv Preprint ArXiv:2509.06888. https://arxiv.org/abs/2509.06888

Mhaske, A., Kedia, H., Ranade, S., Khapra, M. M., Kumar, A., Murthy, R., & Bhattacharyya, P. (2023). Naamapadam: A large-scale named entity annotated data for Indic languages. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), 582–599. https://doi.org/10.18653/v1/2023.acl-long.582

Moscato, V., Postiglione, M., Sperlí, G., & Picariello, A. (2023). Few-shot named entity recognition: Definition, taxonomy and research directions. ACM Transactions on Intelligent Systems and Technology, 14(4), 1–46. https://doi.org/10.1145/3609483

Muller, B., Anastasopoulos, A., Sagot, B., & Seddah, D. (2021). When being unseen from mBERT is just the beginning: Handling new languages with low-resource multilingual NER. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2021), 380–396. https://doi.org/10.18653/v1/2021.naacl-main.38

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. ACM Computing Surveys, 54(1), 1–39. https://doi.org/10.1145/3445965

Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020). MAD-X: An adapter-based framework for multi-task cross-lingual transfer. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), 7654–7673. https://doi.org/10.18653/v1/2020.emnlp-main.617

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), 4996–5001. https://doi.org/10.18653/v1/P19-1493

Rahman, M. (2023). Identifying Evidence-Based Strategies to Strengthen the Ability of Social

Rahman, M. (2024). Molecular Epidemiology of Uranium Exposure: Omics Approaches in Cancer Research. Indus Journal of

Rahman, M., Faisal, M., & Inam, R. Effect Of Visual Aids on The Motor Development Of Children With Down

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, 15–18. https://doi.org/10.18653/v1/N19-5004

Shaalan, K., & Raza, H. (2009). NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 60(8), 1652–1663. https://doi.org/10.1002/asi.21090

Sharan, S. (2024). Prediction of Next Word in Balochi Language Using N-gram Model. Sukkur IBA Journal of Computing and Mathematical Sciences, 7(2), 48–63. https://doi.org/10.30537/sjcms.v7i2.1273 Syndrome

Ullah, F., Ahmad, M., Sidorov, G., Batyrshin, I., Riverón, E. M. F., & Gelbukh, A. (2025). Multilingual Named Entity Recognition in Arabic and Urdu Tweets Using Pretrained Transfer Learning Models. Computers, 14(8), 323.

Ullah, F., Gelbukh, A., Zamir, M. T., Riveron, E. M. F., & Sidorov, G. (2024). Enhancement of named entity recognition in low-resource languages with data augmentation and BERT models: A case study on Urdu. Computers, 13(10), 258. https://doi.org/10.3390/computers13100258

Ullah, S., Ali, N. I., Chandio, S. M., Brohi, I. A., & Laghari, B. A. (2024). Part-Of-Speech Tagging for Balochi Language: A Data Driven Application of Conditional Random Fields. Asian Bulletin of Big Data Management, 4(1). https://doi.org/10.62019/abbdm.v4i1.111

van Linschoten, A. (2023). Building a Balochi language dataset for NLP applications.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 5998–6008.