Indic Sentiment Models

This report details a survey of transformer-based language models on Hugging Face, specifically those designed or adapted for Indic languages, crucial for NLP in this linguistically diverse region. The focus is on their suitability for sentiment analysis, particularly in Konkani.

Models Surveyed:

- google-bert/bert-base-multilingual-cased
- google/muril-base-cased
- ai4bharat/indic-bert
- ibraheemmoosa/xlmindic-base-uniscript
- ibraheemmoosa/xlmindic-base-multiscript

1. google-bert/bert-base-multilingual-cased

- Function: Multilingual BERT-based model (104 languages) for nuanced language understanding (masked language modeling, next sentence prediction). Adaptable for downstream tasks like sentiment analysis.
- Size: ~714 MB (PyTorch), 12 layers, 110M parameters.
- Architecture: BERT base.
- Key Features: Multilingual training, case-sensitive, WordPiece tokenization, bidirectional representations.

2. google/muril-base-cased

- Function: BERT-based model for Indian languages (17), trained on native scripts and transliterations. Effective for translation, generation, and understanding (including sentiment analysis).
- Size: ~700+ MB (estimated).
- Architecture: BERT base, trained from scratch on Indic text.
- Key Features: Pre-trained on 17 Indic languages and English, native script and transliteration training, whole word masking, strong Indic NLP performance.

3. ai4bharat/indic-bert

- Function: Multilingual model (12 Indian languages + English) based on the efficient ALBERT architecture. Enhances Indic NLP tasks.
- Size: ~70-100 MB (estimated).
- Architecture: ALBERT base.
- Key Features: Large-scale Indic and English corpus (8.9B tokens), ALBERT efficiency, competitive performance, accent-preserving tokenizer option.

4. ibraheemmoosa/xlmindic-base-uniscript

- Function: ALBERT base v2 model for 14 Indo-Aryan languages (Brahmic scripts), trained on ISO-15919 transliterated text to improve cross-lingual understanding. Suitable for text classification and QA.
- Size: ~11M parameters.
- Architecture: ALBERT Base v2.
- Key Features: Pre-trained on 14 Indo-Aryan languages, ISO-15919 transliteration training (Aksharamukha), SentencePiece tokenization (50k vocab), enhanced cross-lingual understanding for Brahmic script languages.

5. ibraheemmoosa/xlmindic-base-multiscript

- Function: Similar to the uniscript model (ALBERT v2, 14 Indo-Aryan languages) but trained on original scripts to assess transliteration impact on cross-lingual learning. Applicable to NLP tasks in these languages.
- Size: ~11M parameters.
- Architecture: ALBERT Base v2.
- Key Features: Pre-trained on the same 14 languages (native scripts), SentencePiece tokenization (50k vocab), allows comparison with uniscript model for transliteration benefit analysis.

Comparison Table

Feature	google-bert/bert-base-multilingual-cased	google/muril-base-cased	ai4bharat/indic-bert	ibraheemmoosa/xlmindic-base-uniscript	ibraheemmoosa/xlmindic-base-multiscript
Base Architecture	BERT base	BERT base	ALBERT base	ALBERT Base v2	ALBERT Base v2
Number of Parameters	~110 million	~110 million (estimated)	~11 million (estimated)	~11 million	~11 million
Number of Languages	104	17 (Indian languages + English)	12 (Indian languages + English)	14 (Indo-Aryan)	14 (Indo-Aryan)
Konkani Support	Yes (Goan Konkani included)	Yes (Goan Konkani included)	No (not explicitly listed)	Yes (Goan Konkani included)	Yes (Goan Konkani included)
Training Data	Wikipedia (104 languages)	Wikipedia, Common Crawl, PMINDIA, Dakshina	IndicNLP corpus (8.9B tokens)	ISO-15919 transliterated text	Original scripts
Tokenization	WordPiece	Not explicitly specified	Not explicitly specified	SentencePiece	SentencePiece
Key Focus	General multilingual understanding	Indian language understanding	Efficient Indian language understanding	Cross-lingual understanding (uniscript)	Cross-lingual understanding (multiscript)
Transliteration Handling	No	Yes (trained on transliterated text)	No	Yes (trained on transliterated text)	No

References

1. *google-bert/bert-base-multilingual-cased · Hugging Face*. (2024, March 11). Huggingface.co. https://huggingface.co/google-bert/bert-base-multilingual-cased
2. *google/muril-base-cased · Hugging Face*. (2018). Huggingface.co. https://huggingface.co/google/muril-base-cased
3. *ai4bharat/indic-bert · Hugging Face*. (n.d.). Huggingface.co. https://huggingface.co/ai4bharat/indic-bert
4. *ibraheemmoosa/xlmindic-base-uniscript · Hugging Face*. (2024, March 21). Huggingface.co. https://huggingface.co/ibraheemmoosa/xlmindic-base-uniscript
5. *ibraheemmoosa/xlmindic-base-multiscript · Hugging Face*. (2024, March 21). Huggingface.co. https://huggingface.co/ibraheemmoosa/xlmindic-base-multiscript