NE-LID: A Fast and Accurate Language Identification System for Northeast Indian Languages

Badal Nyalang
MWire Labs, Shillong, Meghalaya, India
nyalang@mwirelabs.com

Abstract

Language identification (LID) is crucial for natural language processing systems, yet Northeast Indian languages remain severely underserved by existing multilingual LID models. We present NE-LID, a fast and accurate language identification system specifically designed for eleven languages of Northeast India. Built using character n-gram features with fastText, NE-LID achieves 99.09% accuracy on a balanced test set, significantly outperforming existing multilingual systems including GlotLID (73.12%), OpenLID (42.03%), IndicLID (39.30%), and LangDetect (24.33%). Our model processes predictions in 0.084 milliseconds on average, enabling real-time applications. We demonstrate that character-level modeling outperforms transformer-based approaches for script-diverse, low-resource languages.

Keywords: language identification, low-resource languages, Northeast India, fastText, multilingual NLP

1. Introduction

Northeast India is home to rich linguistic diversity with over 200 languages from multiple language families including Tibeto-Burman, Austroasiatic, and Indo-Aryan. Despite this diversity, these languages remain critically underserved by modern natural language processing tools, including language identification systems.

Language identification is the foundational task of automatically determining which language a given text is written in. While significant progress has been made in multilingual LID systems covering hundreds of languages (5), we demonstrate through empirical evaluation that existing systems perform poorly on Northeast Indian languages, with many systems completely failing to detect several languages in the region.

In this work, we present NE-LID, a language identification system specifically designed for eleven languages of Northeast India: Assamese, Bodo, English, Garo, Hindi, Khasi, Kokborok, Meitei, Mizo, Nagamese, and Nyishi. Our contributions are:

- A curated dataset of 22,000 sentences across 11 Northeast Indian languages for training and evaluation
- NE-LID, achieving 99.09% accuracy with 0.084ms inference time, significantly outperforming existing multilingual systems
- Comprehensive benchmark of four existing LID systems on Northeast Indian languages, revealing critical gaps in current multilingual models
- Empirical evidence that character n-gram models outperform transformer-based approaches for script-diverse, low-resource language identification

2. Related Work

2.1. Language Identification Systems

Modern language identification has evolved from early statistical methods to sophisticated neural approaches. GlotLID (5) covers over 2000 languages using fastText, while OpenLID (1) supports 201 languages. IndicLID (6) focuses specifically on 22 Indic languages including both native-script and romanized text. LangDetect (7), based on character n-gram profiles, supports 55 languages. However, coverage does not equal accuracy for low-resource languages: a system may list a language as supported while failing to correctly identify it in practice, particularly when training data for that language is scarce or absent.

2.2. Low-Resource Language Processing

Earlier attempts at language identification for Northeast Indian languages were primarily acoustic and prosodic in nature (2), leaving text-based LID severely underexplored. Northeast Indian languages face unique challenges including limited digital corpora, script diversity, and lack of standardized orthography. Recent efforts have begun addressing these gaps: ILID (3) created a dataset for 22 official Indian languages but excludes most Northeast languages (Khasi, Garo, Kokborok, Nyishi, Nagamese), and Tonja et al. (10) developed the first parallel corpora for 13 Northeast Indian languages. Terhiija et al. (9) surveyed spoken language technologies for Northeast Indian languages and highlighted the near-zero-resource status of most Tibeto-Burman varieties, underscoring the prerequisite role of accurate text-based language identification. The NE-BERT project (8) trained multilingual encoder models for nine Northeast Indian languages, providing the foundational corpus from

Proceedings of WILDRE-8, pages 104–108

which NE-LID draws its training data. However, language identification specifically for Northeast India has received limited attention despite being a prerequisite for other NLP tasks.

3. Languages and Data

3.1. Target Languages

We focus on eleven languages spanning four language families (Table 1). Hindi and English are included as anchor languages given their widespread use across Northeast India in administrative, educational, and digital contexts.

Family	Languages
Austroasiatic	Khasi
Tibeto-Burman	Garo, Bodo, Kokborok, Meitei, Mizo, Nagamese, Nyishi
Indo-Aryan	Assamese, Hindi
Germanic	English

Table 1: Target languages by language family

These languages exhibit significant orthographic diversity, using Latin script (Khasi, Garo, Mizo, Kokborok, Nyishi, Nagamese), Bengali-Assamese script (Assamese, Meitei when written in Bengali script), and Devanagari script (Bodo, Hindi). This script diversity poses challenges for LID systems that rely primarily on character-level features.

3.2. Dataset Construction

We constructed our dataset by sampling from the NE-BERT corpus (8), which contains web-scraped and curated text from various publicly available sources including news articles, social media, and digitized documents. We extracted 2,000 sentences per language, totaling 22,000 sentences. The dataset comprises approximately 90% publicly sourced content, with the remaining 10% drawn from curated and digitized materials.

We used a stratified split of 70% training (15,400 samples), 15% development (3,300 samples), and 15% test (3,300 samples). The larger development and test sets (15% each, yielding 300 samples per language) were chosen to ensure statistically reliable evaluation, which is especially important in low-resource settings where per-language sample sizes are small. Sentences range from short phrases (2-10 tokens) to longer passages (50+ tokens), ensuring diversity in text length.

4. Methodology

4.1. Model Architecture

We employ fastText (4) for supervised text classification. FastText learns vector representations of character n-grams and word n-grams, making it particularly suitable for morphologically rich and low-resource languages. Character n-grams allow the model to capture script-specific orthographic patterns without requiring large amounts of training data, making it well-suited for the script-diverse languages of Northeast India.

Our model configuration uses character n-grams of length 2-5 to capture subword patterns, word unigrams only, learning rate of 0.5, 25 training epochs, softmax loss function, and 8 training threads.

4.2. Training

Training was conducted on the 15,400-sample training set using the fastText supervised learning algorithm. The model converged after 25 epochs. FastText models are inherently compact and efficient, operating entirely on CPU without GPU requirements.

5. Experiments

5.1. Evaluation Setup

We evaluate on our held-out test set of 3,300 sentences (300 per language). We report overall accuracy and per-language accuracy. We additionally benchmark four existing multilingual LID systems: GlotLID, OpenLID, IndicLID, and LangDetect.

5.2. Main Results

Table 2 shows overall accuracy comparison across all five systems. NE-LID achieves 99.09% accuracy, outperforming the best competitor (GlotLID) by 25.97 percentage points, representing a 2.7$\times$ improvement. Development set accuracy is 99.00%, confirming that the model generalises well and is not overfitting to the test set.

Model	Accuracy
NE-LID (Ours)	**99.09%**
GlotLID	73.12%
OpenLID	42.03%
IndicLID	39.30%
LangDetect	24.33%

Table 2: Overall accuracy comparison

5.3. Per-Language Analysis

Table 3 shows per-language accuracy for all five systems. NE-LID achieves near-perfect accuracy ($>$95%) on all eleven languages, while competitor systems show significant gaps.

Language	NE-LID	GlotLID	OpenLID	IndicLID	LangDetect
Assamese	100.00	100.00	100.00	100.00	100.00
Bodo	98.67	89.33	0.00	96.67	0.00
English	96.00	79.00	96.00	83.00	94.33
Garo	99.67	0.00	0.00	0.00	0.00
Hindi	96.33	86.00	79.67	54.67	73.33
Khasi	99.67	95.33	0.00	0.00	0.00
Kokborok	99.33	99.33	0.00	0.00	0.00
Meitei	99.67	97.00	95.33	98.00	0.00
Mizo	99.00	92.67	91.33	0.00	0.00
Nagamese	100.00	0.00	0.00	0.00	0.00
Nyishi	99.33	65.67	0.00	0.00	0.00

Table 3: Per-language accuracy (%) for all systems

5.4. Inference Speed

We measured inference speed on CPU. NE-LID processes predictions extremely fast: short sentences ($<$50 chars) in 0.028ms, medium sentences (50-150 chars) in 0.065ms, and long sentences ($>$150 chars) in 0.160ms, with an overall average of 0.084ms ($\sim$12,000 predictions/second).

This speed enables real-time language identification for interactive applications and high-throughput batch processing. FastText operates entirely on CPU, making it accessible for deployment without specialized hardware.

6. Analysis

6.1. Coverage Gaps in Existing Systems

Our benchmark reveals critical gaps in existing multilingual LID systems:

- GlotLID fails completely on Garo and Nagamese (0% accuracy), despite claiming to support 2000+ languages
- OpenLID (Meta) only detects 5 of 11 Northeast Indian languages, completely missing Khasi, Garo, Bodo, Kokborok, Nagamese, and Nyishi
- IndicLID, despite focusing on Indic languages, covers only 4 of 11 languages. Notably, it achieves only 54.67% accuracy on Hindi, likely due to confusion with Maithili and Marathi which share the Devanagari script
- LangDetect performs worst overall (24.33%), essentially unusable for Northeast Indian languages

6.2. Why Character N-Grams Work

We initially attempted transformer-based approaches (NE-BERT, XLM-R) but found they performed poorly (9-37% accuracy) even when trained on our dataset. Character n-grams succeed because:

- **Script awareness**: Character n-grams directly capture script-specific patterns (Devanagari vs. Latin vs. Bengali-Assamese)
- **Orthographic distinctiveness**: Many Northeast languages have unique character combinations and diacritics
- **Low-resource robustness**: Character n-grams require less training data than transformer models to learn discriminative patterns
- **Efficiency**: FastText models are orders of magnitude faster than transformer inference

6.3. Error Analysis

The model produces 30 misclassifications (0.91% error rate) across the full test set. Table 4 shows development and test accuracy, confirming consistent performance across both splits.

Split	Accuracy
Development	99.00%
Test	99.09%

Table 4: Development vs. test accuracy

Table 5 shows accuracy broken down by sentence length. Short sentences ($<$50 characters) are the most challenging, consistent with the intuition that very short inputs provide fewer character n-gram features for discrimination.

Length	Count	Accuracy
Short ($<$50 chars)	1,107	98.37%
Medium (50–150 chars)	1,370	99.49%
Long ($>$150 chars)	823	99.39%

Table 5: Accuracy by sentence length on test set

Table 6 shows mean confidence scores per language on correctly classified samples. Languages with script-distinctive orthographies (Meitei, Assamese, Kokborok, Nyishi) yield the highest confidence, while Hindi and English show the lowest

confidence, consistent with their script overlap with other languages in the dataset.

Language	Mean Confidence
Hindi	0.9839
English	0.9941
Mizo	0.9949
Bodo	0.9949
Khasi	0.9956
Naga	0.9968
Garo	0.9975
Kokborok	0.9980
Nyishi	0.9985
Assamese	0.9985
Meitei	0.9986

Table 6: Mean prediction confidence per language (correct predictions only)

Manual inspection of the 30 misclassified samples reveals that several cases involve label noise in the source corpus, for example fully Hindi sentences labeled as Bodo, or fully English sentences labeled as Mizo, suggesting the true model error rate may be lower than 0.91%. The remaining errors follow interpretable patterns: Bodo-Hindi confusion due to shared Devanagari script; Khasi-Garo confusion as both use Latin script with similar phonological patterns; and Nyishi misclassifications on extremely short inputs with ambiguous character patterns.

7. Limitations

NE-LID has several limitations: (1) The model is designed for monolingual sentences and may struggle with code-mixed text; (2) While overall accuracy is high, performance may degrade on extremely short inputs ($\leq$2 tokens); (3) The model relies heavily on script patterns, which may fail when languages are transliterated to non-standard scripts; (4) Many other Northeast Indian languages (e.g., Ao, Tangkhul, Dimasa) are not covered.

8. Conclusion

We present NE-LID, a fast and accurate language identification system for eleven Northeast Indian languages. With 99.09% accuracy and 0.084ms inference time, NE-LID significantly outperforms existing multilingual systems and addresses critical gaps in language technology for Northeast India.

Our comprehensive benchmark reveals that "multilingual" LID models often fail to adequately support low-resource languages, with many systems completely missing 6-7 Northeast Indian languages. We demonstrate that character n-gram models are more effective than transformer-based approaches for script-diverse, low-resource language identification. The model and dataset are publicly available at https://huggingface.co/MWirelabs/ne-lid.

9. Future Work

Future work includes extending coverage to additional Northeast Indian languages, handling code-mixed text, improving robustness on very short inputs, and integrating NE-LID into downstream NLP pipelines for the region.

10. Bibliographical References

[1] Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

[2] Chuya China Bhanja, Mohammad Azharuddin Laskar, and Rabul Hussain Laskar. 2019. A pre-classification-based language identification for northeast indian languages using prosody and spectral features. *Circuits Syst. Signal Process.*, 38(5):2266–2296.

[3] Yash Ingle and Pruthwik Mishra. 2025. ILID: Native script language identification for Indian languages. *arXiv preprint arXiv:2507.11832*.

[4] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

[5] Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12028–12051, Singapore. Association for Computational Linguistics.

[6] Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023. Bhasa-Abhijnaanam:

Native-script and romanized language identification for 22 Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–826, Toronto, Canada. Association for Computational Linguistics.

[7] Shuyo Nakatani. 2010. Language detection library for java.

[8] Badal Nyalang. 2026. NE-BERT: A multilingual language model for nine Northeast Indian languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 1–12, Rabat, Morocco. Association for Computational Linguistics.

[9] Viyazonuo Terhiija, Samudra Vijaya, and Priyankoo Sarmah. 2019. Spoken language technology for north-east indian languages. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 182–185, Paris, France. European Language Resources Association (ELRA).

[10] Atnafu Lambebo Tonja, Hellina Hailu Nigatu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. First attempt at building parallel corpora for machine translation of Northeast India's very low-resource languages. *arXiv preprint arXiv:2312.04764*.

11. Language Resource References

[1] MWire Labs. (2025). NE-LID: Language identification model for Northeast Indian languages. HuggingFace. https://huggingface.co/MWirelabs/ne-lid

[2] MWire Labs. (2025). NE Multilingual Corpus: Web-scraped text for Northeast Indian languages. HuggingFace. https://huggingface.co/datasets/Badnyal/ne-multilingual-corpus