Under the Surface: Probing Tamil Paraphrase Intelligence

Viswadarshan R Raamiya[1] and Felicia Lilian J[2] and Mahalakshmi S[3]
[1,2,3]Department of Computer Science and Business Systems
[1,2,3]Thiagarajar College of Engineering, Madurai, Tamil Nadu, India
[1]viswadarshanrramiya@gmail.com, [2]jflcse@tce.edu, [3]mahalakshmisklu@gmail.com

Abstract

We present a systematic study of paraphrase detection in Tamil through the construction and evaluation of a unified benchmark dataset derived from three widely used English paraphrase corpora: QQP, PAWS, and MRPC. The dataset is generated using a translation and verification pipeline that combines semantic similarity analysis, round-trip consistency checks, classifier agreement evaluation, and human assessment to improve translation reliability and semantic preservation. Using this benchmark, we evaluate five multilingual transformer models (mBERT, XLM-R, IndicBERT, MuRIL, and DistilmBERT) alongside a compact Tamil-specific encoder, TLMR, pretrained on 525M Tamil tokens. We further examine the representational quality of these models through embedding-based classification using lightweight machine learning classifiers, including Logistic Regression, SVM, and XGBoost. In addition, we introduce a Tamil-oriented efficiency evaluation framework that analyzes trade-offs between masked language modeling performance, vocabulary utilization, and Tamil script fidelity. Experimental results show that multilingual models such as MuRIL achieve strong downstream paraphrase detection performance, while Tamil-specific pretraining in TLMR improves efficiency-oriented language modeling characteristics. Our findings provide a benchmark and analysis framework for future research on semantic understanding in Tamil and other low-resource languages.

1 Introduction

Paraphrase detection, which can be described as the ability to discern whether two sentences share the same semantic meaning, is an essential component of natural language understanding. This process is highly relevant to domains such as information retrieval, question answering, semantic search, duplicate identification, and summarization tasks. While the availability of large annotated corpora and pretrained transformers has led to breakthroughs in paraphrase identification within resource-rich languages such as English, comparable success in other languages, particularly low-resource languages such as Tamil, is still lagging behind. Given the unique characteristics of the Tamil language, which include complex morphological features, agglutinative word formation processes, and flexible syntax structures, semantic modeling poses many difficulties without ample task-relevant data (Thangarajan et al., 2016).

To support research in this direction, we construct a unified Tamil paraphrase detection benchmark by translating and validating three widely used English paraphrase datasets: QQP (Sharma et al., 2019), PAWS (Zhang et al., 2019), and MRPC (Dolan and Brockett, 2005). Rather than relying solely on machine translation outputs, we employ a structured quality verification pipeline that incorporates semantic similarity analysis, round-trip consistency evaluation, classifier agreement analysis, and human assessment to improve semantic fidelity and reduce translation noise. The resulting dataset combines multiple paraphrase styles and levels of semantic complexity, enabling broader evaluation of Tamil paraphrase understanding systems.

Using this benchmark, we evaluate several multilingual transformer models, including mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), IndicBERT-V2 (Doddapaneni et al., 2023), MuRIL (Khanuja et al., 2021), and DistilmBERT (Sanh et al., 2019), alongside a compact Tamil-specific encoder, TLM-DeBERTa (TLMR). TLMR is a lightweight 6-layer DeBERTa-based model pretrained on a large Tamil corpus using a Tamil-oriented BPE tokenizer. In addition to downstream paraphrase detection performance, we analyze representational quality

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 244–254

through embedding-based classification using traditional machine learning models such as Logistic Regression, SVM, and XGBoost.

Additional analysis of efficiency-related aspects of Tamil language models is conducted by evaluating them through the lens of masked language modeling, vocabulary usage, and consistency with the Tamil script fidelity. The results highlight complementary strengths across different models, where multilingual models such as MuRIL exhibit better performance on downstream tasks of paraphrase detection, while specific Tamil pre-training in TLMR increases efficiency in language modeling. Overall, these results provide an initial benchmark for future research into semantic comprehension and representation of Tamil.

2 Related Work

2.1 Paraphrase Detection Benchmarks

Paraphrase detection (PD) is a foundational task in natural language understanding, historically dominated by high-resource languages due to the availability of large-scale annotated corpora. Foundational English benchmarks have driven significant advancements in semantic equivalence tasks. The Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005) is a pioneering dataset comprising formal sentence pairs extracted from news sources. To capture more conversational and domain-diverse semantics, the Quora Question Pairs (QQP) dataset (Sharma et al., 2019) was introduced to identify semantically equivalent community questions. To fill this gap, Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019) provides structurally divergent sentences with high lexical overlap but different meanings to challenge the models. While these datasets have spurred significant advancements in English NLP, low-resource and morphologically complex languages such as Tamil suffer from a severe dearth of such task-specific annotated data.

2.2 Pretrained Language Models for Indic Languages

The development of transformer-based Masked Language Models (MLMs) has propelled cross-lingual representation learning. Multilingual BERT (mBERT) (Devlin et al., 2019) has been a standard pretrained on 104 languages (including Tamil) with a shared WordPiece vocabulary. XLM-RoBERTa (XLM-R) (Conneau et al., 2020) scaled the model further by pretraining on the huge Common Crawl corpus over 100 languages with dynamic masking, and steadily improving cross-lingual transfer. For more constrained settings, DistilmBERT (Sanh et al., 2019) offers a distilled, lightweight version while retaining multilingual capabilities.

Recently, models specifically targeted at the linguistic diversity of the Indian subcontinent have emerged. IndicBERT (Doddapaneni et al., 2023) is a compact, ALBERT-based model pretrained on 12 Indian languages using a 9-billion-token monolingual corpus, including approximately 549 million Tamil tokens. Similarly, MuRIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021) also uses monolingual and parallel translated/transliterated data for 17 Indian languages and achieves good performance on Indian language benchmarks by bridging semantics across scripts.

2.3 Tamil Paraphrase Detection

Despite the progress in multilingual language models, representation-focused modeling for Tamil remains underexplored. Prior work in Tamil paraphrase detection has largely relied on traditional machine learning approaches and limited, domain-specific corpora (Thangarajan et al., 2016). Multilingual models often suffer from vocabulary fragmentation and suboptimal tokenization when processing agglutinative Dravidian languages. To address this, our work not only translates and unifies the QQP, PAWS, and MRPC benchmarks into a comprehensive Tamil dataset, but also introduces a Tamil-specific DeBERTa-based model (TLMR) to explicitly contrast the efficacy of large-scale multilingual pretraining against compact, language-specific representation learning.

3 Dataset

We construct a Tamil Paraphrase Detection dataset for training and evaluation of semantic similarity models in a low-resource setting. The dataset is obtained from existing English Paraphrase corpora and created via a translation pipeline of forward translation and back-translation using a multilingual translation model. To improve semantic consistency and reduce translation noise, the translated sentence pairs are further vali-

Dataset	Size	Train	Val	Test
QQP	116K	100K	8K	8K
PAWS	65.4K	49.4K	8K	8K
MRPC	5.8K	3.67K	408	1.73K
Total	187.2K	153.07K	16.4K	17.73K

Table 1: Statistics for the three English paraphrase datasets used in our study: QQP, PAWS, MRPC. Sizes shown in thousands of sentence pairs, reflecting quantities after initial preprocessing but before translation to Tamil.

dated using semantic similarity analysis, round-trip evaluation, and classifier-based consistency checks. The resulting corpus supports the study of paraphrase detection under cross-lingual and translation-induced variation.

3.1 Source Datasets

We use three publicly available English paraphrase datasets: Quora Question Pairs (QQP) (Sharma et al., 2019), Paraphrase Adversaries from Word Scrambling (PAWS) (Zhang et al., 2019), and Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005). These datasets were selected to capture different forms of paraphrase variation and semantic complexity.

QQP consists of question pairs with potential semantic equivalence and provides large-scale conversational paraphrases. The PAWS dataset consists of difficult paraphrases with many similarities at the lexeme level and variations in word order, which makes it useful for testing the ability to discriminate between subtle nuances in semantics. The MRPC dataset provides paraphrases sourced from news articles and comprises mostly formal structures.

Table 1 summarizes the statistics of the original English datasets used in our study before translation into Tamil.

4 Methodology

4.1 Modeling Setup for Translation

The NLLB-200 1.3B model released by Meta AI, was adopted for both forward and backward translation due to its broad multilingual coverage and demonstrated stability in English–Tamil translation. Unlike many English-centric or Indic-specific models, NLLB-200 is trained on dense parallel corpora that include Tamil, supporting consistency and bidirectional alignment. Preliminary comparisons with Indic-based models indicated that, although IndicTrans models are optimized for English-to-Tamil translation, their performance was less consistent when applied at scale, particularly in preserving syntactic fluency and producing reliable Tamil renderings. Employing a unified model for both directions also reduces architectural asymmetries and simplifies error attribution during round-trip evaluation.

4.2 Modeling Setup for Evaluation

To assess the quality and fidelity of the generated dataset, we incorporate the following evaluation models:

LaBSE (Language-Agnostic BERT Sentence Embedding) is used to compute semantic similarity between sentence pairs across languages(Feng et al., 2022). It enables embedding-based comparisons between original English, Tamil translations, and backtranslated English sentences.

COMET is used to evaluate translation quality through reference-based scoring(Rei et al., 2020). We compute COMET scores in two configurations (i) using English originals as reference and backtranslations as machine outputs; (ii) reversing source and MT to align with Tamil input as source.

A task-specific **RoBERTa-based English paraphrase classifier**, trained on combined MRPC, QQP, and PAWS data is used to estimate paraphrase label consistency. It predicts labels on both original and backtranslated English pairs to quantify label agreement and flip rates.

Each model is chosen for its fit with a particular evaluation dimension, such as fluency, semantic retention, paraphrase preservation, or translation quality.

4.3 Quality Assurance Filtering

To ensure the reliability and cleanliness of the translated dataset before training and evaluation, we added a quality assurance step to filter out noisy or erroneous translations. Through analysis of the QQP, PAWS, and MRPC translations, we found common patterns of translation failure. These included overly long sentences, often caused by repetition loops or poor constructions, along with unnatural word repetitions. As a cautious measure, we excluded all translated sentence pairs where either sentence exceeded a specific token length. At that length, we consistently saw a drop in fluency or meaning accuracy.

In addition to this automatic filtering, we manually reviewed the data to find and fix structural

errors, confusing fragments, and alignment inconsistencies that could create noise during model training. After these filtering steps, we used the final dataset to create a 75,000-sample automatic evaluation subset and a 6,000-sample human evaluation subset, both chosen randomly. These measures ensured that the dataset used for training and evaluation was high-quality, structurally consistent, and minimally affected by translation errors, reducing the risk of bias or misleading patterns from noise.

4.4 Evaluation Methodology and Setup

We evaluate on 75,000 sentence pairs sampled from QQP (40,000), PAWS (32,500) and MRPC (2,500) by automatic metrics. We randomly select 6,000 sentence pairs for human evaluation. This combined setup enables us to evaluate the translation quality, semantic preservation, and paraphrase consistency from both automatic and human perspectives.

4.4.1 Automatic Evaluation

Our automatic evaluation protocol is designed to capture both surface-level and semantic properties of the translated data.

Surface-Level Analysis: We computed Self-BLEU scores to measure lexical diversity between sentence pairs. A lower Self-BLEU score indicates minimal redundancy, which is preferred in paraphrase datasets. Additionally, we calculated the length ratio between the translated Tamil sentences (T_{ta}) and their English versions (S_{en}). This serves as a proxy for adequacy. It captures whether the content is preserved without too much expansion or omission.

Semantic Similarity: We computed cosine similarity scores between the English sentence and its corresponding Tamil translation EN→TA using LaBSE(Feng et al., 2022) in order to evaluate semantic alignment between the original and translated content. Through the computation of similarity between the original English sentence and its backtranslated version obtained via TA→EN→TA, we also assessed round-trip semantic preservation.

Round-Trip Consistency Evaluation: We used the COMET(Rei et al., 2020) framework, which scores a machine translation by comparing it to a source and a reference sentence, to assess meaning preservation across translation directions. To evaluate round-trip consistency, we created two complementary setups.

Setup	Source (src)	Hypothesis (mt)	Reference (ref)
OR → TA → BT	Palm Beach County is considering adding up to $ 200 million more in incentives.	பாம்ம் பீச் கவுண்டி 200 மில்லியன் டாலர் கூடுதல் ஊக்கத்தொகை வழங்குவதை பரிசீலித்து வருகிறது.	Palm Beach County is considering an additional $200 million in incentives.
TA → BT → OR	எனது IQ ஐ ஆன்லைனில் சரிபார்க்க சிறந்த வழி எது?	What's the best way to check my IQ online?	What is the best way to get my IQ checked online?

Figure 1: Examples from the round-trip COMET evaluation configurations. The terms OR, TA, and BT stand for the original English sentence, Tamil translation, and backtranslated English sentence, respectively. The Tamil translation's ability to retain enough semantic content for accurate recovery into English is assessed in the first row (OR → TA → BT). The second row (TA → BT → OR) checks to see if the backtranslation accurately conveys the Tamil sentence's meaning.

In the first configuration, we used the original English sentence as the source, the Tamil translation as the hypothesis, and the backtranslated English as the reference. This setup evaluates whether the Tamil output retains sufficient semantic information for accurate reconstruction of the original English sentence.

In the second configuration, the original English was used as the reference, the backtranslated English as the hypothesis, and the Tamil sentence as the source. This setup evaluates how well the backtranslation captures the Tamil sentence's semantic intent.

Together in Figure 1, these configurations offer a bidirectional perspective on translation quality. This strategy is particularly important in low-resource settings, where direct references in the target language are unavailable. It provides a robust, reference-based signal to validate semantic integrity in the absence of native Tamil gold standards.

Classifier Agreement Evaluation: To check if paraphrase relationships were maintained after round-trip translation, we used a RoBERTa-based paraphrase classifier. This model was fine-tuned on MRPC, QQP, and PAWSX. It achieved 83.33% accuracy on English test sets and served as a baseline for checking semantic consistency.

For each of the 75,000 evaluation sentence pairs, we compared the model's predictions on the original English inputs and their backtranslated versions in Tamil. We measured two key metrics: (i) **Prediction Agreement Rate**, which is the share of examples with consistent labels before

and after translation, and (ii) **Paraphrase Label Flip Rate**, which is the share of cases where the label changed after backtranslation, even though it initially matched the gold label. This change suggests possible semantic drift.

This evaluation offers another way to check if semantic equivalence stays intact during translation. It supports our other automatic assessments, including round-trip COMET and LaBSE-based semantic similarity.

4.5 TLMR (Tamil Language Model - DeBERTa)

The model adopts a compact 6-layer DeBERTa-V3 encoder architecture(He et al., 2021) with 768 hidden dimensions, 8 self-attention heads, and GELU activation. It employs disentangled self-attention, separating content and positional representations, along with enhanced relative position embeddings, enabling richer contextual modeling compared to RoBERTa-style encoders. Where required for implementation compatibility, absolute positional encoding is retained. We apply gradient checkpointing, dropout (0.1), and untied embeddings to efficiently capture Tamil's morphological richness under a compact model footprint.

A carefully selected multilingual corpus of 525 million Tamil tokens and 70 million English tokens sourced from Tamil Wikipedia, Project Madurai, IndicCorp v2, OSCAR, and domain-specific news text was used to pretrain TLMR from scratch. A cosine scheduler with warm restarts and the Adam optimizer, which has a peak learning rate of 2e-4, were used to train the model over five epochs. It demonstrated strong language modeling ability with a validation loss of 2.21 and a final perplexity of 5.17.

A custom Byte-Pair Encoding (BPE) tokenizer with a 32K vocabulary size is included with the model. It was trained on the same cleaned corpus after normalization, noise reduction, and script consistency filters were applied. When compared to conventional multilingual tokenizers, the tokenizer demonstrated noticeably better vocabulary efficiency, coverage of Tamil script, and token compactness.

In this study, we evaluate the effect of Tamil-centric pretraining on paraphrase detection performance using TLMR in conjunction with multilingual baselines like XLM-R(Conneau et al., 2020), IndicBERT(Doddapaneni et al., 2023), MuRIL(Khanuja et al., 2021), and mBERT(Devlin et al., 2019).

4.6 Tamil MLM Efficiency-Oriented Evaluation

To better analyze Tamil-compatible masked language models beyond downstream paraphrase accuracy, we introduce a Tamil-oriented efficiency evaluation framework designed for low-resource and morphologically rich language settings. Metrics such as perplexity or downstream F1 alone do not fully capture differences in tokenizer behavior, Tamil script consistency, or language-specific vocabulary utilization.

Our evaluation combines four complementary factors: Top-5 Accuracy, Perplexity, Vocabulary Efficiency, and Tamil Script Purity. Top-5 Accuracy measures contextual token prediction capability during masked language modeling. Perplexity estimates overall language modeling uncertainty. Vocabulary Efficiency reflects how effectively Tamil vocabulary tokens are utilized during inference, while Tamil Script Purity measures the proportion of generated Unicode characters belonging to the Tamil script.

Since these metrics operate on different numerical scales, all higher-is-better metrics are normalized using min-max normalization:

$$\tilde{x}_{ik} = \frac{x_{ik} - \min_j x_{jk}}{\max_j x_{jk} - \min_j x_{jk}} \quad (1)$$

For perplexity, which is a lower-is-better metric, we apply an inverse logarithmic utility transformation:

$$u_i = \frac{\frac{1}{\log(1+\mathrm{PPL}_i)} - \min_j \frac{1}{\log(1+\mathrm{PPL}_j)}}{\max_j \frac{1}{\log(1+\mathrm{PPL}_j)} - \min_j \frac{1}{\log(1+\mathrm{PPL}_j)}} \quad (2)$$

The final Tamil MLM Efficiency Score is computed as:

$$S_i = 0.40\tilde{A}_i + 0.25\tilde{P}_i + 0.20\tilde{V}_i + 0.15u_i \quad (3)$$

where $\tilde{A}_i$, $\tilde{P}_i$, and $\tilde{V}_i$ denote the normalized Top-5 Accuracy, Tamil Script Purity, and Vocabulary Efficiency respectively.

Higher weights are assigned to Top-5 Accuracy and Tamil Script Purity because prior work treats tokenization as a core modeling decision rather than a preprocessing detail, noting that subword choices can misalign with linguistic structure and waste capacity across languages (Alqahtani et al.,

2026). Token granularity has also been shown to influence language-model behavior substantially, and recent low-resource studies report that improvements in perplexity do not necessarily translate into better understanding-based performance (Oh et al., 2025; Luitel et al., 2025). Vocabulary Efficiency is therefore given a moderate weight as a complementary signal of tokenizer compactness, while Perplexity receives a smaller weight because it is an indirect uncertainty measure that is sensitive to vocabulary design and does not fully capture downstream quality (Chen and Goodman, 1999).

The greater weights for Top-5 Accuracy and Script Purity result from their direct relationship to the accuracy of context prediction and Tamil orthography. The Vocabulary Efficiency metric measures the efficiency of the tokenizer, as well as the usage of Tamil tokens. Perplexity receives a relatively lower weight because it is sensitive to the tokenizer.

The proposed score is intended to provide a comparative estimate of Tamil-oriented language modeling efficiency rather than replace downstream paraphrase detection metrics.

4.7 Paraphrase Detection Task

We create a unified training framework based on the high-quality Tamil PD dataset mentioned in Section 3.1 in order to assess the performance of multilingual and Tamil-specific models for paraphrase detection (PD). This dataset is made up of sentence pairs that have been classified as paraphrases or non-paraphrases. It was compiled from both machine-translated and human-verified versions of three common resources: MRPC(Dolan and Brockett, 2005), PAWS(Zhang et al., 2019), and QQP(Sharma et al., 2019). Enabling model evaluation is made possible by the resulting corpus, which captures a variety of linguistic phenomena and Tamil sentence structures.

We approach the detection of paraphrases as if it were a binary classification task. Predicting whether a pair of sentences conveys the same meaning is the aim. To preserve context and prevent truncation artifacts, all pairs were tokenized using the corresponding model-specific tokenizers, with a maximum sequence length of 256. For all transformer-based models, inputs were formatted using standard classification heads over the [CLS] token representation.

We experimented with two evaluation strategies: (1) fine-tuning pretrained language models and (2) training lightweight machine learning classifiers using fixed sentence embeddings. This two-stage setup provides a comprehensive perspective on the representational quality and downstream adaptability of each model.

4.7.1 Fine-Tuning Pretrained Models

We fine-tuned five multilingual transformer models: mBERT (Pires et al., 2019), XLM-R (Conneau et al., 2020), IndicBERT (Doddapaneni et al., 2023), MuRIL (Khanuja et al., 2021), and distil-mBERT (Sanh et al., 2019), along with our Tamil-specific TLM-DeBERTa (TLMR). These models were selected to represent different multilingual pretraining strategies, parameter scales, and levels of Tamil language specialization.

For each model, we appended a linear classification head over the [CLS] representation and optimized the network using binary cross-entropy loss. Fine-tuning was performed on the combined Tamil paraphrase dataset containing 152,464 sentence pairs derived from QQP, PAWS, and MRPC. Validation and test splits were constructed by proportionally combining the corresponding subsets from all three datasets.

Among the multilingual baselines, mBERT and XLM-R serve as general-purpose multilingual encoders, IndicBERT and MuRIL provide stronger Indic-language specialization, and distil-mBERT represents a lightweight compressed multilingual baseline. TLMR is included to evaluate the effect of Tamil-specific pretraining and tokenization on paraphrase detection performance under a compact model setting.

Training took place over three epochs using the AdamW optimizer, with a learning rate of 2e-5 and a batch size of 16. All training and evaluation were conducted on Google Colab Pro with A100 GPUs, enabling efficient experimentation with large-scale models. Evaluation was performed on the held-out test set using key metrics, including accuracy, F1 score, and AUC.

4.8 Embedding-Based Classification with ML Models

To better understand the sentence representations learned by each finetuned model, we extracted the [CLS] token embeddings from the final encoder layer. We used these embeddings as fixed features for downstream classifiers. We generated embeddings for a subset of 53,666 sentence pairs. To

maintain balance across domains, this subset contained 25K from QQP, 25K from PAWS, and 3.6K from MRPC.

Let $x = (s_1, s_2)$ represent two Tamil sentences. Using the CLS token, we extract the fixed representation of a pre-trained model M as follows:

$$h_x = \text{CLS}\big(M(\text{Tokenize}(x))\big) \tag{4}$$

A downstream classifier C then predicts the paraphrase label $\hat{y} \in \{0, 1\}$ as:

$$\hat{y} = \arg\max_{c \in \{0,1\}} f_C(\mathbf{h}_x)_c \tag{5}$$

where $f_C(\cdot)$ denotes the probability distribution output by the classifier.

Three traditional machine learning models(Pedregosa et al., 2011) were trained: XGBoost (XGB)(Chen and Guestrin, 2016), Support Vector Machine (SVM) with a linear kernel, and Logistic Regression (LR). Using the combined training set, grid search with three-fold cross-validation was used to tune the hyperparameters for SVM and XGB. To ensure consistent comparison, all models were assessed using the same test set that was used in the fine-tuning experiments.

We evaluate the adaptability of pretrained models to the Tamil paraphrase detection task and the inherent quality of their sentence embeddings by combining the results of fine-tuned classifiers and embedding-based shallow models. Section 4 presents evaluation findings and comparative analysis.

5 Results & Analysis

5.1 Dataset Evaluation and Quality Analysis

To validate the reliability and utility of the constructed Tamil Paraphrase Detection dataset, we present both corpus-level statistics and multilingual quality evaluation results. The dataset comprises translated and filtered sentence pairs from three widely-used English paraphrase benchmarks: QQP(Sharma et al., 2019), PAWS(Zhang et al., 2019), and MRPC(Wang et al., 2019). While the methodology section details the construction pipeline, here we focus on the empirical analysis that demonstrates the semantic integrity and linguistic quality of the resulting Tamil data.

5.1.1 Corpus Composition

Table 2 summarizes the size of the dataset for each of the three sources. For each dataset, the original count of English sources is shown next to the corresponding Tamil sentence pairs kept after translation, quality checking, and Tamil-only filtering. In total, the final training set includes about 152.5K Tamil sentence pairs, with balanced validation and test splits taken proportionally from each sub-dataset. Notably, the QQP and PAWS subsets make up most of the corpus, while MRPC offers a small but high-quality selection of challenging paraphrases.

Dataset	Original	Retained	Train	Val	Test
QQP	116K	115K	99.6K	7.96K	7.99K
PAWS	65.4K	65.1K	49.2K	7.98K	7.96K
MRPC	5.8K	5.8K	3.67K	0.41K	1.72K
Total	**187.2K**	**185.9K**	**152.5K**	**16.3K**	**17.7K**

Table 2: Composition of the final Tamil paraphrase dataset. All counts in thousands (K). 'Original' refers to English source count; 'Train', 'Val' and 'Test' refer to Tamil sentence pairs used for fine-tuning and evaluation.

5.1.2 Translation Quality Assessment

To assess the semantic alignment between the original English sentence pairs and their Tamil translations, we use a variety of automated metrics. The dataset demonstrates strong semantic fidelity across a number of metrics, as indicated in Table 3. Good cross-lingual meaning preservation is indicated by the LaBSE(Feng et al., 2022) similarity scores, which average over 87% for Tamil translations and over 91% after back-translation to English.

Metric	Sentence 1	Sentence 2
Self-BLEU	0.33	0.33
Length Ratio	94.60%	95.10%
LaBSE ($OR_{en} \rightarrow TA$)	87.10%	87.10%
LaBSE ($OR_{en} \rightarrow BT_{en}$)	91.32%	91.50%
COMET ($OR_{en} \rightarrow TA \rightarrow BT_{en}$)	0.668	0.669
COMET ($TA \rightarrow BT_{en} \rightarrow OR_{en}$)	0.869	0.877
Classifier Agreement Rate	82.59%	
Paraphrase Label Flip Rate	14.73%	

Table 3: Quantitative evaluation metrics comparing translation quality and semantic preservation across different configurations. OR_{en}: Original English, TA: Tamil translation, BT_{en}: Backtranslated English.

The COMET scores, assessed in two directions (original → Tamil → back-translated and reverse), show good translation quality (0.66 to 0.87). This matches accepted standards for multilingual MT evaluation. Additionally, the Self-BLEU score of 0.33 per sentence suggests significant syntactic diversity within the paraphrase

Evaluation Metric	Average Score
Sentence 1 Fluency Score	98.2
Sentence 2 Fluency Score	97.8
Sentence 1 Translation Adequacy	92.6
Sentence 2 Translation Adequacy	93.4
Semantic Similarity Preservation	98.0

Table 4: Human evaluation results on the translated Tamil paraphrase dataset. Scores reflect fluency, translation adequacy, and semantic similarity preservation across the evaluated sentence pairs. Scores were assigned on a 0–100 scale, where higher scores indicate better fluency, adequacy, and semantic preservation.

class. The classifier agreement rate, determined by passing both original and translated sentence pairs through a task-specific paraphrase detector, is 82.6% agreement. This further confirms semantic stability. Finally, we see a paraphrase label flip rate of 14.73%. This rate reflects the proportion of meaning-altered pairs post-translation, consistent with prior cross-lingual paraphrasing findings.

5.2 Human Evaluation

Besides the automated evaluation technique, we also conducted human evaluation on the translated dataset. Five Tamil language experts and researchers from MuthirAI Global Research Center for Tamil AI have evaluated a curated subset of 6,000 sentence pairs. The evaluation was based on three aspects, (i) fluency and grammatical accuracy of the translated Tamil sentences, (ii) adequacy of translation with respect to the original English sentences, and (iii) preservation of semantic similarity between the paired sentences.

Both sentences in each pair were individually assessed by the evaluators through a score-based evaluation process manually. The fluency score was determined according to the grammatical accuracy, readability, and naturalness of the Tamil sentences. The adequacy of translation was determined by how well the translated Tamil sentence was able to maintain the meaning of the original English sentence. Scores for semantic similarity were determined by checking if the semantic relation of paraphrasing between the two sentences still maintained semantic consistency upon translation.

Overall, the human evaluation results support the reliability of the dataset and indicate that the translated sentence pairs preserve semantic consistency at a level suitable for Tamil paraphrase detection tasks.

5.3 Evaluation of Tamil MLM Efficiency

Table 5 summarizes the Tamil MLM Efficiency results across multilingual and Tamil-specific language models. The evaluation jointly considers contextual prediction accuracy, Tamil script fidelity, vocabulary utilization, and language modeling uncertainty.

TLMR achieves the highest efficiency-oriented score, indicating strong Tamil-focused language modeling characteristics, particularly in vocabulary utilization and script consistency. Its Tamil-specific tokenizer and compact pretraining setup contribute to improved token efficiency and orthographic fidelity compared to broader multilingual baselines.

Among the multilingual models, MuRIL and IndicBERT obtain competitive scores due to their stronger Indic-language specialization. XLM-R demonstrates strong contextual prediction capability but comparatively lower Tamil vocabulary efficiency, while DistilmBERT performs substantially worse across most metrics, highlighting the limitations of aggressive multilingual compression for morphologically rich Tamil text.

Importantly, the efficiency-oriented evaluation captures tokenizer behavior and Tamil-specific language modeling characteristics rather than downstream paraphrase detection performance directly. While MuRIL achieves the strongest downstream paraphrase detection results, TLMR demonstrates stronger efficiency-oriented characteristics under Tamil-focused evaluation constraints. These findings suggest that multilingual transfer and Tamil-specific pretraining provide complementary advantages depending on the target objective.

5.4 Model Performance on Tamil Paraphrase Detection

To benchmark the effectiveness of multilingual and Tamil-specific models with our dataset on the paraphrase detection task, we fine-tuned six transformer-based encoders of mBERT(Devlin et al., 2019), XLM-R(Conneau et al., 2020), MuRIL(Khanuja et al., 2021), IndicBERT(Doddapaneni et al., 2023), DistilmBERT(Sanh et al., 2019), and our proposed TLM-DeBERTa (TLMR) on the combined dataset of 152,464 sentence pairs. Each model was trained end-to-end with a classification head using binary cross-entropy loss. With an F1 score of

Model	Top-5 Accuracy (%)	Perplexity	Vocab Efficiency (%)	Tamil Purity (%)	Efficiency Score
mBERT	14.12	1.16	0.72	80.98	0.3549
XLM-R	43.00	**1.04**	0.94	92.40	0.6482
IndicBERT	51.00	1.95	1.93	91.20	0.6588
MuRIL	49.00	1.62	2.24	93.20	0.6727
TLM-DeBERTa	**62.00**	1.76	**14.76**	**94.80**	**0.9504**
DistilmBERT	8.40	1347.09	0.72	49.00	0.000021

Table 5: Model performance comparison showing Top-5 Accuracy, Perplexity, Vocabulary Efficiency, Tamil Script Purity, and the composite Efficiency Score. TLM-DeBERTa (TLMR) achieves the highest score, demonstrating the effectiveness of Tamil-specific pretraining and tokenization. Vocabulary Efficiency is expressed in raw percentage form.

Model	Fine-tuned			Logistic Regression		SVM		XG Boost	
	F1	Acc	AUC	F1	Acc	F1	Acc	F1	Acc
mBERT	82.35	82.34	90.69	82.63	82.56	82.66	82.64	82.69	82.60
XLM-R	84.42	84.39	92.46	85.36	85.29	85.28	85.37	85.18	85.10
MuRIL	86.03	85.99	93.65	86.53	86.46	86.45	86.37	86.53	86.48
IndicBERT	83.72	83.68	91.77	84.62	84.54	84.80	84.96	85.03	84.96
DistilmBERT	75.44	75.33	83.90	76.08	75.97	76.01	75.92	76.22	76.16
TLMR	85.00	84.80	92.85	84.60	85.50	85.90	85.75	85.10	85.00

Table 6: Performance of Transformer-Based and Embedding-Based Models on Tamil Paraphrase Detection. All values are percentages. Fine-tuned results show end-to-end model performance, while other columns show classifier performance using model embeddings. AUC is only reported for fine-tuned models.

86.03%, MuRIL outperformed XLM-R (84.42%) and IndicBERT (83.72%) in terms of fine-tuning performance, as indicated in Table 6. Despite its lightweight 6-layer design, the Tamil-specific TLMR model produced an F1 score of 85.00%, demonstrating competitive performance. DistilmBERT performed poorly on every metric, overall highlighting how crucial representational depth and high-quality tokenization are for Tamil.

To further assess the embedding quality learned during fine-tuning, we extracted $[CLS]$ token embeddings from each model and trained three shallow classifiers of Logistic Regression (LogReg), Support Vector Machine (SVM), and XGBoost (XGB) on 53,666 labeled sentence pairs. The majority of the discriminative information obtained during fine-tuning was retained by the embedding-based models, according to the results in Table 6.

MuRIL attains the better embedding-based F1 score (86.53% with Logistic Regression and 86.37% with XGBoost); however, the compact DeBERTa-V3–based TLMR exhibits consistently strong performance, achieving a fine-tuned F1 of 85.00% and maintaining closely aligned scores across all embedding-based classifiers (LR 84.60, SVM 85.90, XGBoost 85.10). This narrow performance gap, despite TLMR's reduced model capacity, highlights the effectiveness of targeted pretraining for Tamil and the robustness of its learned representations. Overall, the results demonstrate that TLMR generalizes well across lightweight downstream setups while preserving performance consistency.

6 Conclusion

In this work, we present a new Tamil paraphrase detection datasets constructed from translated and verified versions of existing English benchmarks, along with a unified training and evaluation framework. We assess the effectiveness of several multilingual transformer models and our lightweight Tamil-specific encoder, TLM-DeBERTa, using both fine-tuning and embedding-based classification strategies were MuRIL was overperforming all. Additionally, we propose an efficiency-oriented evaluation metric to capture trade-offs between language modeling accuracy, vocabulary usage, and Tamil script fidelity. Our experiments provide a comparative analysis of model performance under different settings, offering insights into representational quality and generalization for Tamil paraphrase detection.

7 Acknowledgement

This research work was carried out under the MuthirAI Global Research Center for Tamil AI,

a Center of Excellence at Thiagarajar College of Engineering (TCE), Madurai, India, and was supported through the Thiagarajar Research Fellowship (TRF).

The authors also acknowledge the contributions of the MuthirAI research team for supporting the human evaluation and validation process of the translated Tamil paraphrase dataset. The evaluation was conducted by a group of Tamil language professionals and researchers who reviewed translation quality, semantic consistency, and paraphrase preservation across the dataset.

Limitations

This work focuses on Tamil paraphrase detection by providing a curated dataset and a comparative evaluation of multilingual and Tamil-specific models. The study is limited to sentence-level paraphrasing and does not examine transferability to other downstream tasks or discourse-level semantic phenomena. To ensure fairness, all models were trained under uniform configurations; task-specific hyperparameter optimization may yield further improvements. These constraints reflect deliberate design choices for controlled benchmarking and define clear directions for future work.

References

Sawsan Alqahtani, Vivek Gupta, and Hinrich Schütze. 2026. Tokenization matters: The impact of tokenization on language modeling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 6423–6438. Association for Computational Linguistics.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Fuli Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei-Cheng Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Deberta-v3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Simran Khanuja, Sandipan Dandapat, Divyanshu Kakwani, Naveen Arivazhagan, Pushpak Bhattacharyya, Mitesh Khapra, and Irshad Bhat. 2021. Muril: Multilingual representations for indian languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 620–633.

Prakash Luitel, Rohan Sharma, and Abhinav Mishra. 2025. Perplexity is not enough: Evaluating language models beyond uncertainty metrics. *arXiv preprint arXiv:2502.04561*.

Byung-Jun Oh, Sangmin Lee, and Jihwan Choi. 2025. The impact of token granularity on language model performance. *arXiv preprint arXiv:2501.01234*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Pro-*

ceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001.

Ricardo Rei, Ana Farinha, Alon Lavie, and André F. T. Martins Silva. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS)*.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset.

R. Thangarajan, S. R. Balasundaram, and M. Aramudhan. 2016. Tamil paraphrase detection using machine learning approaches. In *Working Notes of FIRE 2016 - Forum for Information Retrieval Evaluation*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Yichong Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1298–1308.