# Thirumurai: A Large Dataset of Tamil Shaivite Poems and Classification of Tamil Pann

Shankar Mahadevan[1], Rahul Ponnusamy[2], Prasanna Kumar Kumaresan[2], Prabakaran Chandran[3], Ruba Priyadharshini[4], Sangeetha Sivanesan[5], Bharathi Raja Chakravarthi[6]

## ABSTRACT

Thirumurai, also known as Panniru Thirumurai, is a collection of Tamil Shaivite poems dating back to the Hindu revival period between the 6th and the 10th century. These poems are par excellence, in both literary and musical terms. They have been composed based on the ancient, now non-existent Tamil Pann system and can be set to music. We present a large dataset containing all the Thirumurai poems and also attempt to classify the Pann and author of each poem using transformer based architectures.

Our work is the first of its kind in dealing with ancient Tamil text datasets, which are severely under resourced. We explore several Deep Learning based techniques for solving this challenge effectively and provide essential insights into the problem and how to address it.

## INTRODUCTION

Tamil, being a great source of ancient poems, can be extensively researched using modern NLP techniques to get new insights that would be hard to get through conventional research methods. We have created a dataset of Thirumurai poems annotated with corresponding author, location where it was sung and its Pann type.

We have also trained baseline transformer models for Tamil Pann classification and author classification and discussed various factors that affect model performance in these tasks.

## Benchmarking Methodology

We experimented with one RNN-based model and 3 transformer-based models. They are:

Bi-LSTM, Language-Agnostic Bert Sentence Embedding (LaBSE) Model (Feng et al., 2020), XLM-RoBERTa Model (Conneau et al., 2020), Multilingual BERT model (Devlin et al., 2018)

We used the Precision score, Weighted Recall Score and the Weighted F1 score as measures to analyze the performance of our selected models on the Thirumurai dataset.

## RESULTS

We randomly split the dataset which has Pann annotated (nearly 5548 poems out of the 8416 poems available) into three parts: 70% for training, 10% for validation and rest of the 20% for testing the models.

Among the mentioned models, LaBSE produced the best precision, weighted recall and weighted average F1-score in both classification problems. The LaBSE model marginally outclassed BiLSTM, mBERT and XLM-RoBERTa models.

**Table 1.** Pann Classification Results

| Model Name | Precision | Recall | F1 Score |
|---|---|---|---|
| **LaBSE** | **0.91** | **0.91** | **0.91** |
| m-BERT | 0.90 | 0.90 | 0.90 |
| XLM-RoBERTa | 0.90 | 0.89 | 0.89 |
| BiLSTM | 0.78 | 0.74 | 0.75 |

**Table 2.** Author Classification Results

| Model | Precision | Recall | F1 Score |
|---|---|---|---|
| **LaBSE** | **0.94** | **0.94** | **0.94** |
| m-BERT | 0.93 | 0.93 | 0.93 |
| BiLSTM | 0.93 | 0.93 | 0.93 |
| XLM-RoBERTa | 0.92 | 0.92 | 0.92 |

## DISCUSSION

The results clearly shows the supremacy of transformers in settings where the data available is low and there exists a class imbalance in the dataset.

It also illustrates the efficacy of Transformer architectures, for low resource languages like Tamil. They achieve a precision which is comparable to human-level classification ability.

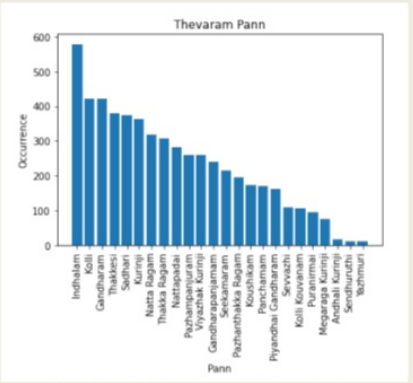


**Fig 1. Distribution of Pann in the dataset.**

## CONCLUSIONS

This work can be extended to all Sangam age literatures, and analyzing even more type of Panns found in those ancient Tamil literature can be done. It can also be used to predict the Pann type of old poems for which the Pann was lost. Moreover, different poetry composing styles of the authors can also be analyzed through this dataset. We leave all these possibilities to interested researchers to work in the future.