Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6556–6562
Marseille, 20-25 June 2022

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]

[1]Thiagarajar College of Engineering, Madurai, India
[2]Indian Institute of Information Technology and Management - Kerala, India
[3]Mu Sigma Inc., [4]ULTRA Arts and Science College, Madurai, India
[5]National Institute of Technology, Tiruchirappalli, India
[6]Insight SFI Research Centre for Data Analytics, National University of Ireland Galway
shankarmahadevan12901@gmail.com, {rahul.mi20, prasanna.mi20}@iiitmk.ac.in,
prabakaran.chandran98@gmail.com, rubapriyadharshini.a@gmail.com,
sangeetha@nitt.edu, bharathi.raja@insight-centre.org

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.

Keywords: Thirumurai, Pann Classification, Tamil NLP

1. Introduction

NLP has advanced by leaps and bounds since the introduction of Deep Learning, particularly in the English language. This is mainly attributable to the vast amount of data for English that is available. Every second, massive volumes of data are generated from a variety of sources all across the internet. Companies with access to such large amounts of data are contributing significantly to the NLP community by training models with ever-increasing parameters, such as the trillion parameter Switch transformer (Fedus et al., 2021). Because Deep Learning is a data-driven learning approach, a large data corpus is required to build good NLP systems.

As the number of parameters in NLP models grows beyond trillions, they become increasingly resource and power hungry. Such models are trained on petabytes of data, such as the Common Crawl data, which has over 12 years of internet data. Due to the size of the data they are being trained on, this not only improves the accuracy of the models, but it also allows them to generalize to new, previously unknown tasks. There exists different kinds of datasets for different tasks like the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018) for question answering, Amazon Product Dataset (142.8 million reviews spanning May 1996 – July 2014) for sentiment analysis, WordNet database (Miller, 1995) for text classification, dataset of research papers from Arxiv which has been used to generate abstracts for research papers automatically and many other datasets like this. These types of datasets have aided us in preventing hate speech, obtaining effective product recommendations, and locating solutions to all questions using powerful NLP systems driven search engines. Not many languages have such large scale datasets. We lack such large scale datasets for Low-Resourced Languages (LRLs). Even though many of these languages have a large amount of population who speak it, they do not have standard datasets available to develop NLP based systems. NLP work in Low-Resource Languages is paralyzed due to the lack of such datasets and the difficulty of producing them. Tamil is one such language, with over 60 million native speakers. Even though increased attempts are being done to make Tamil data public, they focus on gathering data that is available from platforms like Wikipedia and Social Media websites such as Facebook, Twitter, YouTube etc. (Chakravarthi et al., 2020) These datasets have been used to do morphological analysis, identify hate speech, conduct sentiment analysis (Thavareesan and Mahesan, 2019) etc. Data of Tamil literature is basically non-existent. 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 (Subalalitha, 2019). We have developed a large dataset of Thirumurai poetry, as a first step towards improving the availability of Tamil literature corpora. We hope this work would stimulate other NLP practitioners,

scholars and students across the world to create such collections from their literature. Our significant contributions are as follows:

- Creation of a dataset of Thirumurai poems annotated with corresponding author, location where it was sung and its Pann type.
- Training baseline transformer models for Tamil Pann classification and author classification and discussing various factors that affect model performance in these tasks.

2. Background

The Tamil language has a treasure trove of poems starting from the works of Sangam age poets like Kabilar and Baranar to modern reformist-poets like Subramania Bharathiar (Venkataraman, 2012). Tamil poetry is classified into two types: Marabu Kavithai (Traditional form) and Puthu Kavithai (Modern form) . Traditional Tamil poetry follows the "Yappilakkanam", (Deshpande, 1995) which sets the rules of Tamil poetry writing. Poems from the Sangam literature and Bakthi literature are samples of traditional Tamil poetry. The rise of the independence struggle in India brought various alterations to the previous poetry structure. To make poems appeal to the general audience, poets like Subramania Bharathiar, Namakkal Kavingar and Barathidasan started stressing free-verse poetry. Bharathiar who lived during the height of the freedom fight instilled patriotism among millions of people with his poems. Poets like Kannadasan started writing Tamil songs for Tamil cinema. Modern Tamil poetry is also known as Vasana Kavithai or Yappilla Kavithai (poetry that does not follow the rules in the "Yappilakkanam"). Modern Tamil poetry features socially important subjects and the Haiku form of poems have also grown quite popular in Tamil Nadu. These are available in vast volumes through the internet. Traditional poems are seldom available, and mainly not in a format that can be fed to modern NLP systems. In order to bridge the gap between current and traditional poem datasets, we have built a dataset for Thirumurai.

Thirumurai is a collection of beautiful Shaivite devotional poems, in the traditional format, that were composed by several poets during the Bhakti movement in between the 6th and 10th centuries (Kodeeswari, 2021). Shaivism is one of the Shanmathas or the six major sects of Hinduism. Followers of Shaivism consider Lord Shiva as the supreme power. It emphasizes both bhakti (devotion towards the lord) and the yogic way of life for liberation from the materialistic world. It has been influenced a lot by southern Tamil Dravidian Shaiva Siddhanta traditions and philosophies. (Dyczkowski, 1987) The followers of this system are called "Shaivites" or "Saivas". The growth of Shaivism in Tamil Nadu along with Vaishnavism (devotion towards Lord Vishnu as the supreme power) saw the growth of a lot of Tamil literature. (Rathakrishnan, 2019) They were set to 'Tamil Pann', the music system followed during ancient times in Tamil Nadu.

Nos.	**Hymns**	**Author**
1,2,3	Tirukadaikkappu	Sambandhar
4,5,6	Tevaram	Thirunavukkarasar
7	Tirupaatu	Sundarar
8	Thiruvasakam, Thirukkovaiyar	Manikkavasagar
9	Thiruvisaippa, Thiruppallaandu	Various
10	Tirumandiram	Tirumular
11	Prabandham	Various
12	Periya Puranam	Sekkizhar

Table 1: Thirumurai - Division of poems.

Tamil is divided into three sections - Iyal Tamil (Prose), Isai Tamil (Poetry), and Nadaga Tamil (Drama) (Hoisington, 1853). As time progressed, the Tamil musical system went almost extinct due to the emergence of other musical systems and different rulers. The Carnatic music system, which is prevalent in the southern part of India, is considered the successor of the Tamil Pann system. (Balambegai, 2007) The Tamil Pann system has existed right from the Sangam age. "Pann" has evolved from the word **"Panna-Paduvathu"** which means doing something. There are various references to this ancient musical tradition found in the ancient Sangam books such as Ettuthogai and Pattupattu. Silappathikaram, written by Elongo Adigal, belonging to the post-Sangam period (5th or 6th century) also mentions various forms of music practiced by the Tamil people. It has many chapters that talk about music and dance. The chapter "Kanal Vari", a duet between the hero Kovalan and his lover Madhavi, can be considered a work of excellence in the musical field (Karunakaran, 2020). It also describes many technical details of music such as musical pitch and the smallest fraction of an audible sound distinguishable by the human ear.

Tholkapiyam, the oldest grammar book of Tamil, speaks about different Panns in detail, and also quotes from two musical books which are now extinct (Kanakaraj, 2016). It tells that the four original panns of Maruthappann, Kurinchippann, Sevvazhi, and Sadari have evolved into 103 panns with varying characteristics. It explains how the five lands in the Tamil tradition (Kurinji, Mullai, Marudham, Neithal, and Palai) had their Pann, musical (Yazh), and percussion (Parai) instruments associated with them based on their mood and qualities. The Neithal land, which has been used in literature to illustrate the grief of separation of the heroine from the hero, is associated with the

Sevvazhi Pann, which induces pathos when sung. Malaippadukadam refers to farmers relaxing after hard work by singing Marudappann in their rest time.

In the first 7 Thirumurais, Out of the 103 panns 24 have been handled by Thirugnana Sambandhar, Thirunavukkarasar, and Sundarar. The Tamil Isai Sangam has done extensive research on 103 panns, with many eminent musicians, musicologists, and oduvars (temple singers) participating in the research. In today's society, knowledge of the Pann system of music has practically vanished. Even for experts in the discipline, identifying the Pann in a poem is difficult. As a result, we propose developing Deep Learning based AI solutions for this task, as this will aid researchers in learning more about the poem.

2.1. Tamil Music Scale

In Tamil music, the corresponding musical notes used in carnatic music were given by:

Carnatic Note	Pann Note
Sa	Kural
Ri	Tuttam
Ga	Kaikilai
Ma	Uzhai
Pa	Ili
Da	Vilari
Ni	Taram

Table 2: Musical Notes in Tamil Pann

There are notations in tamil for notes in different octaves also.

3. Previous Work

To the best of our knowledge, there is no prior work of classifying Panns in Tamil Sangam age literature. Learning the poetic and musical structure of a poem is a very novel area in NLP that has not been explored much by researchers. There has been some work in other languages such as Arabic (Yousef et al., 2019), Ottoman (Can et al., 2011) and Kurdish (Mahmudi and Veisi, 2021b).

(Yousef et al., 2019) have employed LSTMs, GRUs and their Bidirectional counterparts for categorizing 16 poetry meters in Arabic and 4 poetic meters in English (with corresponding accuracy measures of 96.38% and 82.31% respectively). It is one of the earliest articles to research about addressing this problem using Deep Learning based techniques. It also studies carefully into Arabic and English poetry form, their roots and poetic rules that govern them. The researchers demonstrated that Deep Learning based techniques exceed conventional grammatical rule-based techniques by a very significant margin, and do not require any human extracted features; The RNNs were able to learn the features by observing the distributions. One important difference from our work is that they do character-level prediction, whereas we do word-level classification and also employ State-of-the-Art transformer based architectures instead of Recurrent Neural Network Architectures.

(Mahmudi and Veisi, 2021b) Uses a rules-based approach to divide Kurdish Poems into three categories: quantitative verse, syllabic verse and free-verse. They employ a rule-based method of SCK grapheme to-phoneme conversion given by (Mahmudi and Veisi, 2021a), which converts the input text into a syllabified string of phonemes, and then classify this string of phonemes.

(Sahin et al., 2018) Aims to categorise poetry based on the poet. They extract text properties such as Term Frequency Inverse Document Frequency (TF-IDF), Bag-of-Words and feed these data to Support Vector Machines, K-Nearest Neighbour algorithms and Random Forest models to identify the corresponding poet.

4. Dataset

The dataset contains a total of 8416 individual poems cleaned manually. Out of these, the Pann of 5548 poems is currently available, as information on the Pann of many poems was lost as time passed. Each poem contains 6 lines and 23 words on average. The dataset also includes the poet and the location where the poem was sung. All the poems were scraped from the Tamil Virtual Academy[1] website, after obtaining prior permission. It is an organization established by the Government of the State of Tamil Nadu, to provide Tamil education through internet for the Tamil Diaspora and to promote Tamil Computing. Despite the existence of alternative sources such as Project Madurai[2], this website provided high-quality material that was easy to scrape and did not contain archaic Tamil letters.

4.1. Text Preprocessing

The raw dataset contains a lot of noise. Since a lot of words used in the poems are archaic, their meanings are given in the website. These types of unrelated data are removed first. Unnecessary characters like asterisks etc, which have been included as part of the digitization process, are removed. Extra spaces are also removed.

Thevaram poetry employs a variety of poetic structures, including "Venba", "Asiriyappa", "En-Seer Kazhinediladi Asiriyappa", and others (Rajam and Rajam, 1992). All these structures are classified according to the number of words in a line and the number of lines in the poem. So, it is highly critical

[1] http://www.tamilvu.org/

[2] https://www.projectmadurai.org/

'நீரின்மல்குசடை யன்விடையன்னடை யார்தம்அரண் மூன்றுஞ்
சீரின்மல்குமலை யேசிலையாக முனிந்தன்றுல குய்யக்(*)
காரின்மல்குகடல் நஞ்சமதுண்ட கடவுள்ளிட மென்பர்
ஊரின்மல்கிவளர் செம்மையினாலுயர் வெய்தும்புக லூரே.

(*) முனிந்தானுலகுய்ய என்றும் பாடம்.'

Figure 1: An Example poem from the Thirumurai dataset. The starred line contains information extrinsic to the poem. Thirumurai 1.2.4, Sambandhar.

to preserve this information while training the model on this dataset. So, we maintained the newline character while tokenizing the poem. This made sure the structural context was not lost while training the model. The distribution of various Panns in Thevaram is shown in Figure 2.

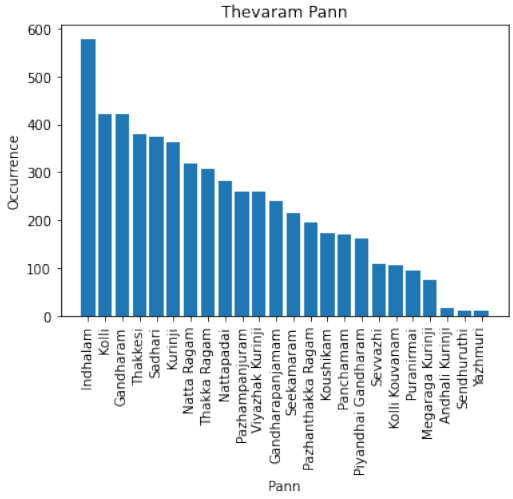

Figure 2: Distribution of Pann in the dataset.

5. Methodology/Benchmarking

Recently, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) based models have achieved great success by pre-training on large scale corpora and then fine-tuning on downstream tasks. When fine-tuning on classification tasks, BERT first uses the specific token [CLS] to obtain task-specific representation. Then, the CLS token is applied along with one additional output layer for classification. Since this task requires dealing with a Tamil dataset, we were not able to use common transformer models that have been trained on huge English corpora. We needed models that are able to adapt to different languages (in this case, Tamil). Multilingual models are powerful models that encode text from different languages into a shared embedding space, enabling them to be applied to various tasks like text classification, parts-of-speech tagging, clustering, etc. For this task, we use 3 very popular Multilingual transformer models:

- **Language-Agnostic Bert Sentence Embedding (LaBSE) Model** (Feng et al., 2020) :
  This model is trained on 17 billion monolingual sentences and 6 billion bilingual sentence pairs using Masked Language Modelling and Translation Language Modelling pre-training on 109 languages, resulting in a model that is effective even on low-resource languages for which there is no data available during training.
- **XLM-RoBERTa Model** (Conneau et al., 2020): It is a multilingual transformer model, based on the XLM (Lample and Conneau, 2019) model and RoBERTa (Liu et al., 2019) model, trained on 100 different languages. XLM-RoBERTa has shown great performance in various multilingual NLP tasks.
- **Multilingual BERT model** (Devlin et al., 2018): m-Bert was one of the first transformer models to show promise on multilingual NLP tasks. It is based on the BERT transformer model.

5.1. Task Formalization

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. A particular model is identified as best if it has the highest weighted average of these three metrics.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

True positives occur when the model correctly classifies the positive class, True Negatives occur when the model correctly classifies the negative class, False Positives occur when the model incorrectly classifies the negative class as positive class, and False Negatives occur when the model incorrectly classifies the positive class as negative class. This Precision measure tends to reduce the number False Positives. The F1 score gives a balance between precision and recall and is also very useful when there is an uneven distribution of classes in a dataset, which is the case here. In the context of Pann classification, the precision of a Pann such as Kurinji is the number of correctly predicted Kurinji Pann poems out of all the predicted Kurinji Pann poems. The recall for the class is the number of correctly predicted Kurinji Pann poems out of the number of actual Kurinji Pann poems.

6. Results and Discussion

6.1. Experiment Setup

We randomly split the dataset which has Pann annotated (nearly 5548 poems out of the 8416 poems avail-

able) into two parts: 70% for training, 10% for validation and rest of the 20% for testing the models. There are 24 types of Pann available in total, and the dataset is partitioned so that both the train and test sets contain all of them. Initial experimentation with RNN based Architectures did not give much precision scores. We trained a model with 3 layers of Bi-LSTM each having 64 units with dense layers following them. They could not perform well in this task. So, it was decided to proceed with transformer based architectures for the Pann classification task. All our models are taken from the HuggingFace transformers library[3]. All these experiments were performed using NVIDIA K80 instances provided by Google Colab. The model parameters are as stated below:

Hyperparameter	Value
Activation Function	Softmax
Max Len	256
Batch Size	64
Optimizer	AdamW
Learning Rate	1e-05
Loss Function	Cross Entropy
Epochs	30

Table 3: Model Hyperparameters

6.2. Results

6.2.1. Pann Classification

Among the mentioned models, LaBSE produced good precision, weighted recall and weighted average F1-score of 0.9151, 0.9127 & 0.9139 respectively on the test dataset. The LaBSE model marginally outclassed mBERT and XLM-RoBERTa models. The results are given in Table 4. The class wise metrics for all the 24 classes is given in Table 7.

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 4: Pann Classification Results

To ensure the validity of these scores, we selected the LaBSE model and performed K-Fold cross validation with K selected as 5. The data was split into 5 equal parts, and in each training phase one fold was kept aside and the model was trained with the remaining 4 folds. The results of the K-Fold cross validation are given in Table 5.

6.2.2. Poem Length - Pann relation

Another noteworthy aspect to notice was that after training, the classifiers were able to correctly identify the Pann of a song from the test dataset by only looking through 40% of the poem in average. We set up a pipeline to partition the poetry into 5 parts, and make the model determine the class label by feeding the data in incremented sections of 20% of the data. The models were able to correctly predict the labels by simply going through maximum one or two divisions of the poem. This is shown in Figure 3.

[3]https://huggingface.co/

Fold No.	Precision	Recall	F1 Score
1	0.90	0.90	0.90
2	0.91	0.91	0.91
3	0.90	0.90	0.90
4	0.91	0.90	0.90
5	0.91	0.91	0.91

Table 5: 5-Fold Validation metrics

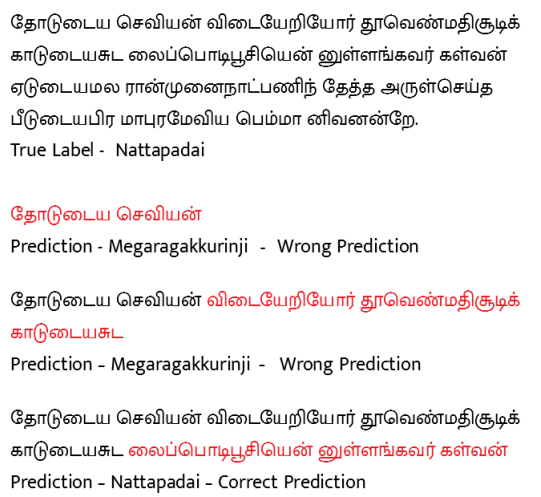

Figure 3: Model results with different lengths of poem.

6.2.3. Author Classification

All the 8416 poems have their corresponding author information available. Out of these Thirungana Sambandhar has written 4438, Thirunavukkarasar 2953 and Sundarar 1025. We perform author classification on these 3 classes using all the 3 transformer models and the baseline BiLSTM model mentioned earlier.

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

Table 6: Author Classification Results

As we can see, the LSTM model performs better in this task, as there is sufficient examples for all the three

No.	Pann	Precision	Recall	F1 Score	Validation No.
1	Indhalam	0.94	0.79	0.86	115
2	Kolli	0.94	0.99	0.96	84
3	Gandharam	0.95	0.91	0.93	84
4	Thakkesi	0.92	0.86	0.89	76
5	Sadhari	0.99	1.00	0.99	75
6	Kurinji	0.91	0.87	0.89	73
7	Natta Ragam	0.81	0.91	0.89	63
8	Thakka Ragam	0.90	0.92	0.91	61
9	Nattapadai	0.93	0.93	0.93	57
10	Pazhampanjuram	0.78	0.94	0.85	52
11	Viyazhak Kurinji	0.72	0.89	0.80	52
12	Gandharapanjamam	0.87	0.89	0.88	48
13	Seekamaram	0.84	0.88	0.86	43
14	Pazhanthakka Ragam	0.90	0.76	0.82	39
15	Koushikam	0.79	0.88	0.84	34
16	Panchamam	0.88	0.97	0.92	32
17	Piyandhai Gandharam	0.97	0.89	0.93	32
18	Sevvazhi	0.84	0.88	0.86	22
19	Kolli Kouvanam	0.95	0.91	0.93	21
20	Puranirmai	0.88	0.94	0.91	18
21	Megaraga Kurinji	0.75	1.00	0.86	15
22	Andhali Kurinji	1.00	1.00	1.00	4
23	Sendhuruthi	1.00	1.00	1.00	4
24	Yazhmuri	1.00	1.00	1.00	4

Table 7: Class Wise Weighted Metrics

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

7. Conclusion

We have collected a large dataset of poems from Thirumurai. We have detailed and examined various text processing approaches and transformer based models to detect the Pann employed in Thirumurai poetry, and to classify the author of the poem. For classification, we applied and fine-tuned transformer-based models such as LaBSE, mBERT and XLM-RoBERTa, which achieves an astonishing rise in accuracy when compared to RNN-based approaches like BiLSTM, where the precision score increased from 0.7836 to 0.9151. Our work also illustrates the superiority of Transformer architectures, notably the LaBSE model, for low resource languages like Tamil. Our best model yields a Precision score of 0.9151, weighted Recall score of 0.9127 and weighted F1 Score of 0.9139, which is comparable to human-level classification ability. We also illustrate how the model is able to identify the Pann type by receiving sections of the input data alone.

Dealing with the class imbalance found in the dataset will improve the performance of the models considerably. We can explore the possibilities of achieving good performance with less labeled data through augmentation. 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 Thirunavukkarasar, Sundarar, Thirugnana Sambandhar and Manikkavasagar can also be analyzed through this dataset. We leave all these possibilities to interested researchers to work in the future.

8. Bibliographical References

Balambegai, R. (2007). Carnatic music and its ragas, with special reference to five notes ragas.

Can, E. F., Can, F., Duygulu, P., and Kalpakli, M. (2011). Automatic categorization of ottoman literary texts by poet and time period. In *Computer and Information Sciences II*, pages 51–57. Springer.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.

Deshpande, M. M. (1995). Ancient indian phonetics. In *Concise History of the Language Sciences*, pages 72–77. Elsevier.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dyczkowski, M. S. (1987). *The doctrine of vibration: An analysis of the doctrines and practices associated with Kashmir Shaivism*. SUNY Press.

Fedus, W., Zoph, B., and Shazeer, N. (2021). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Hoisington, H. R. (1853). Brief notes on the tamizh language. *Journal of the American Oriental Society*, pages 387–397.

Kanakaraj, S. (2016). The tamil classic tolkappiyam: Its antiquity and universal appeal. *Indian Literature*, 60(1 (291):184–191.

Karunakaran, G. (2020). Thinai based antics and music-adopting silapathikaram. *International Research Journal of Tamil*, 2(2):31–43.

Kodeeswari, R. (2021). The three tevaram theories and principles of god. *International Research Journal of Tamil*, 3(S-2):122–125.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Mahmudi, A. and Veisi, H. (2021a). Automated grapheme-to-phoneme conversion for central kurdish based on optimality theory. *Computer Speech & Language*, 70:101222.

Mahmudi, A. and Veisi, H. (2021b). Automatic meter classification of kurdish poems. *arXiv preprint arXiv:2102.12109*.

Rajam, V. and Rajam, V. (1992). *A reference grammar of classical Tamil poetry: 150 BC-pre-fifth/sixth century AD*, volume 199. American philosophical society.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad.

Rathakrishnan, S. L. (2019). The history of tamil religion: A general view. *Journal of Indian Studies*, 12:73–81.

Sahin, D. O., Kural, O. E., Kilic, E., and Karabina, A. (2018). A text classification application: Poet detection from poetry. *arXiv preprint arXiv:1810.11414*.

Subalalitha, C. N. (2019). Information extraction framework for kurunthogai.

Thavareesan, S. and Mahesan, S. (2019). Sentiment analysis in tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Venkataraman. (2012). C subramaniya bharathiyar a biographical study 1882 1921.

Yousef, W. A., Ibrahime, O. M., Madbouly, T. M., and Mahmoud, M. A. (2019). Learning meters of arabic and english poems with recurrent neural networks: a step forward for language understanding and synthesis. *arXiv preprint arXiv:1905.05700*.

9. Language Resource References

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran and Ruba Priyadharshini and John P. McCrae. (2020). *Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text*.

Miller, George A. (1995). *WordNet: a lexical database for English*. ACM New York, NY, USA.