ADScAI
Applied Data Science & Artificial Intelligence Symposium

Enhancing Thirukkural Couplet Section Classification: A Meta-Model Learning Approach

Rukshani Puvanendran
Department of ICT
University of Vavuniya
Vavuniya, Sri Lanka
rukupuvan@gmail.com

Saliny Nadarajah
Department of ICT
University of Vavuniya
Vavuniya, Sri Lanka
salinada22@gmail.com

Yalini Parameshwaran
Department of Tamil Language
University of Peradeniya
Kandy, Sri Lanka
yaliniparameswaran2000@gmail.com

Thirukkural, a widely translated classical Tamil literary work, is structured into three sections—Arathuppal (virtue), Porutpal (wealth), and Kamathuppal (love). This study proposes a novel meta-model approach to classify Thirukkural couplets, into these sections. In this study, we propose a meta-model approach that integrates various traditional machine learning (ML) algorithms. By leveraging meta-model learning and stacking methodologies, our approach improves classification accuracy and robustness compared to individual models. The results demonstrate the effectiveness of a meta-model learning strategy with TFIDF features in literary text classification.

***Keywords*—Thirukkural, text classification, TFTDF, BERT, DistilBERT, meta-model, machine learning**

I. Introduction

Thirukkural, written by Thiruvalluvar, is a collection of couplets (kurals) divided into three primary sections, which are further divided into 133 chapters (adhikarams). Each chapter contains 10 couplets [1], making a total of 1,330 as shown in TABLE I. This organized classification ensures that Thirukkural serves as a comprehensive guide to life, covering moral values, governance, and personal relationships.

TABLE I
Structure of Thirukkural

Section (Paal)	Theme	No. of Chapters
Arathuppal (0)	Virtue & Moral Values	38
Porutpal (1)	Wealth & Governance	70
Kamathuppal (2)	Love & Human Emotions	25
Total		**133**

The structured nature of Thirukkural, combined with its translation into multiple languages, ensures the digital preservation and integrity of Thiruvalluvar's intended thematic divisions in both digital and scholarly references. With literature becoming increasingly digitized, automatic categorization improves accessibility and searchability, aiding researchers, scholars, students, and readers in finding couplets on different themes. Further, By automating classification, researchers can compare the structure of Thirukkural with similar works across cultures and languages, fostering cross-cultural literary studies. This study introduces a meta-model ensemble approach, combining multiple classifiers to improve accuracy and robustness in classification.

II. Related Works

Meta-model learning has been widely explored in text classification to improve accuracy and generalizability by leveraging multiple base models [2]. Several studies have implemented meta-model learning approaches using machine learning and deep learning algorithms to enhance classification performance [3]. Text classification has seen significant advancements with the introduction of both traditional feature-based approaches such as Term Frequency-Inverse Document Frequency (TF-IDF) and transformer-based models such as BERT and DistilBERT [3]. Moreover, the integration of Large Language Models (LLMs) has significantly proliferated across NLP tasks, further advancing text classification capabilities [4]. Ensemble techniques such as stacking and meta-model learning frameworks have been proposed to combine the strengths of both feature-based and machine learning methods [3], [5]. Applying data augmentation techniques to Thirukkural couplets presents challenges due to the unique characteristics and the concise poetic nature of the two-line structure [1].

Given these limitations, alternative strategies such as transfer learning with pre-trained models and meta-model learning approaches may be more effective in improving classification performance while preserving textual integrity. These approaches aim to optimize text classification by leveraging their performance. This study builds on these methodologies by integrating a meta-model framework for Thirukkural classification.

III. Methodology

A. Dataset Preparation

We utilize the publicly available "Thirukkural" dataset, which comprises 1,330 Thirukkural couplets translated into English and labeled under the three section collections [6].

https://doi.org/10.31705/ADScAI.2025.41

Each couplet undergoes preprocessing steps of conversion to lowercase, removal of punctuation, and stopwords.

B. Feature Extraction Approaches

The study uses the following feature representations.

- TF-IDF: Convert Thirukkural couplets into numerical vectors using TF-IDF with unigrams and bigrams of max_features as 5000.
- BERT and DistilBERT: Extract contextual embeddings from a pre-trained transformer model with rich textual representation.

C. Individual Classifier Performance

We have chosen three traditional ML algorithms such as Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR). Each classifier is used to train and evaluate separately using TF-IDF, BERT, and DistilBERT features. Given the presence of class imbalance in the dataset, the performance evaluation was conducted using accuracy (Acc.) and F1-score (F1), ensuring a balanced assessment of model effectiveness.

D. Stacking Classifier Approach

In Meta-model (M-M) approach, we have used "StackingClassifier" from sklearn.ensemble to combine the individual classifiers as base models, where the optimal weightings are automatically learned. We have chosen LR, RF, and SVM as the meta-model to learn from the predictions of base models.

E. Comparison and Analysis

In an imbalanced dataset, where accuracy may be misleading due to the dominance of the majority class, F1-score serves as a more reliable evaluation metric as it balances precision and recall, ensuring that both false positives and false negatives are accounted for effectively. Finally, the best combination of feature extraction and classifier is identified for Thirukkural section classification with the optimal F1-score.

IV. Results and Discussions

The TABLE II presents the accuracy and F1-score for different classification models— RF, SVM and LR —using three feature extraction methods: TF-IDF, BERT, and DistilBERT. Additionally, it includes the performance of a meta-model (M-M) based on these classifiers. TF-IDF consistently provides the best performance across individual and meta-models, while DistilBERT shows competitive performance, often surpassing BERT.

TABLE II
Performance Metric with Different Feature Representations

Model	TF-IDF (%)		BERT (%)		DistilBERT (%)	
	Acc.	F1	Acc.	F1	Acc.	F1
RF	64.21	66.91	56.91	61.65	61.16	63.90
SVM	68.37	69.62	65.23	65.03	68.03	68.04
LR	65.02	67.66	69.62	69.92	68.03	72.18
M-M (RF)	68.62	69.17	68.98	69.54	65.99	66.54
M-M (SVM)	**71.93**	**73.68**	70.69	71.80	72.02	72.55
M-M (LR)	71.25	72.18	70.54	71.80	71.41	72.55

A. Performance of Individual Base Models

TF-IDF achieves the highest F1-score with SVM (69.62) and performs the worst with RF (66.91). This suggests that SVM is more effective in handling text features extracted through TF-IDF.

B. Stacked Meta-Model Performance

Meta-Models outperform individual classifiers in most cases, confirming that an ensemble approach benefits classification. M-M (SVM) achieves the highest F1-score (73.68) with TF-IDF, outperforming individual classifiers, as evidenced by the confusion matrix in TABLE III. The model (MM-SVM) struggles to differentiate between these two classes (Class 0 and Class 1), as evidenced by the high misclassification rate between them, as they have some overlapping or similar textual patterns. The model classifies Class 2 more accurately, exhibiting only a few misclassifications.

TABLE III
The confusion matrix for Meta-model SVM (M-M SVM) for TF-IDF feature representation

Actual \Predicted	**Class 0**	**Class 1**	**Class 2**
Class 0	29	44	2
Class 1	15	121	5
Class 2	1	3	46

V. Conclusion

This study presents an initiative to use AI techniques in literary text classification using meta-model approach for classifying Thirukkural couplets in English translation into their respective section collections. The results indicate that the meta-model approach significantly improves classification accuracy, with Meta-model SVM achieving the highest performance in terms of F1-score and generalizability with TF-IDF feature representations. However, the class imbalance further influences the mismatch in the classification. For future work, we aim to explore advanced meta-model learning techniques with additional transformer-based architectures to further enhance classification performance.

References

[1] "Thirukkural," Wikipedia, The Free Encyclopedia. Available: https://ta.wikipedia.org/wiki/. Accessed: Mar. 1 2025.

[2] Hu, Z., Dychka, I., Potapova, K., Meliukh, V. (2024). Augmenting sentiment analysis prediction in binary text classification through advanced natural language processing models and classifiers. Int. J. Inf. Technol. Comput. Sci, 16, 16-31.

[3] V. Kusumaniswar, M. Chinta, and S. N. Shareff, "Bug classification using stacking classifier," in 2024 2nd International Conference on Disruptive Technologies (ICDT), pp. 1259-1265, IEEE, 2024.

[4] F. Wei et al., "Empirical Study of LLM Fine-Tuning for Text Classification in Legal Document Review," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 2786-2792.

[5] M. Parmar and A. Tiwari, "Enhancing text classification performance using stacking ensemble method with TF-IDF feature extraction," in 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), pp. 166-174, IEEE, 2024.

[6] R. V. K. S., "Thirukkural Dataset," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/rahulvks/thirukkural/data. [Accessed: 11-Feb-2025].