An Analysis on Semantic Interpretation of Tamil Literary Texts

Anita Ramalingam* and
Subalalitha Chinnaudayar Navaneethakrishnan

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur-603203, Tamil Nadu, India
E-mail: anitaramalingam17@gmail.com; subalalitha@gmail.com
**Corresponding Author*

Received 23 August 2021; Accepted 13 November 2021;
Publication 11 January 2022

Abstract

The interaction between a computer and a human or natural language is known as Natural Language Processing (NLP). The ultimate goal is to make the natural language text understandable, which in turn, requires its meaning to be captured. Text can be analyzed on several levels, such as lexical, syntax, semantics, discourse, and pragmatics. These NLP tasks deal with text at different levels, such as word, phrase, sentence, paragraph, and document. Discourse analysis is a type of text analysis that goes beyond the sentence level. The discourse analysis is currently performed on expository (essay) type of texts. There are currently no state-of-the-art NLP applications that handle Tamil literary texts at a discourse level. Tamil classical literature is rich with ethical, moral, and philosophical values that should be explored for the benefit of society. This paper proposes an automatic semantic interpretation framework for Tamil literary texts using discourse parsing by giving works on discourse parsing, text classification, discourse-based clustering and information retrieval, and Tamil language and Tamil literatures. This semantic interpretation can be developed as a smart mobile application using

Journal of Mobile Multimedia, Vol. 18_3, 661–682.
doi: 10.13052/jmm1550-4646.1839

multimedia components. This paper also discusses how the Tamil literary text processing differs from the essay type of text.

Keywords: Discourse parsing, tamil literature, text classification, discourse-based clustering, information retrieval, mobile application, multimedia, natural language processing.

1 Introduction

The natural language text can be analyzed on several levels, such as lexical, syntax, semantics, discourse, and pragmatics. The study of how words are generated from morphemes, which are smaller meaning-bearing components, is known as morphology; syntax refers to the rules that define how words combine to form phrases, clauses, and sentences; semantics is concerned with the interpretation of the meaning of words and the sentences they form; discourse is concerned with the coherent sequences of sentences; and pragmatics is concerned with how the interpretation of a text is influenced by the use of words and sentences in various contexts. It can be observed that the morphology deals mostly with words; syntax deals with sentences; semantics deals with both words and sentences, discourse and pragmatics deal texts beyond sentences.

Collection of sentences does not make sense. Sentences must be related to each other for a better interpretation. A discourse comprises of coherent sentences. A discourse structure captures the semantic relations that relate the sentences. There are theories to construct the discourse structure. Rhetorical Structure Theory (RST) introduced by Mann, and Thompson at the University of Southern California (Thompson and Mann [1], Mann and Thompson [2]), is a widely used theory to analyse the text at the discourse level and to form a discourse structure representation called, rhetorical structure. Tamil language and literature have a long and illustrious history, with written Tamil reaching back to 600 BC. Tamil ancient literature is alive with ethical, moral, and philosophical values that should be investigated for the benefit of society. It motivates the paper to propose a semantic interpretation framework for Tamil literary texts using discourse parsing by works on text classification, discourse-based clustering and information retrieval, and Tamil language and Tamil literatures. It can further be developed as a smart mobile application by using multimedia components.

The rest of the paper is organized as follows: Section 2 discusses related works and Section 3 gives the semantic interpretation of Tamil literary texts.

Section 4 discusses the results and discussion, and the conclusions and future work are presented in Section 5.

2 Related Works

The state-of-the-art works on RST based discourse parsing are discussed in this section. Using the shift reduce parsing method and WordNet, Subba and Di [3] discovered discourse relations. The authors have used linguistic cues as features. The document was examined sentence by sentence. Hernault et al. [4] built a discourse parser by employing a Support Vector Machine (SVM) Classifier to build a discourse tree. The document was processed beyond the sentence level, and feature sets were created using a combination of syntactic and lexical features such as lexical heads, words, and POS tags. For discourse relation classification, Hernault et al. [5] utilized a semi-supervised approach called Feature Vector Extension. The technique was centred on detecting co-occurring features in unlabelled data, which were then utilised to extend the classifier's feature vectors. Word pairs, parse tree production rules, and Lexico-Syntactic context at the boundary between two units of text were used as features in the method.

This section also discusses the works that have been done on discourse-based clustering. Alonso et al. [6] used discourse connectives to try hierarchical clustering. The discourse connectives were found in the MACO morphological analyzer's Spanish dictionary. They tested their method with a 16 million word corpus (5.5 million words of balanced Spanish text LEX-ESP and 10.5 million words of newspaper text). Precision was employed as the criterion for evaluation. Miltsakaki et al. [7] proposed employing discourse connectives to annotate discourse structure. The discourse connectives were discovered using WordFreak, an annotation tool. They annotated the PDTB with Penn Tree Bank and Propbank. Ramesh and Yu [8] proposed employing conditional random fields - supervised machine learning classifier to automatically identify discourse connectives in biomedical text. They employed PDTB as a data set and used biomedical articles for cross validation. The evaluation metrics were precision, recall, and F1-score.

The related works on text classification are also discussed in this section. Al-Salemi and Aziz [9] used the simple Naïve Bayes algorithm, multinomial Naïve Bayes algorithm, and multi-variant Bernoulli Naïve Bayes algorithm to automatically categorize Arabic documents into four classes, with the most frequent terms in the documents serving as features. To test their findings,

they used their own Arabic news collections as a dataset, in all on 3172 Arabic documents. Macro average is the average of F1-measure of all categories. Precision, recall, F1-measure and macro-average were used as evaluation metrics and achieved overall Macro-F1, 0.941. Rizzo et al. [10] classified research articles into two classes, relevant and non-relevant papers, using the multinomial Naïve Bayes classifier. The classification assists researchers retrieve relevant papers for their research. Authors' names, names of journals, journal references, abstracts, introductions and conclusions were used as features. Their work was tested on 2215 papers from the benchmark Systematic Literature Review dataset, with recall as the evaluation metric and achieved 95%.

This section also describes the works on Information Retrieval (IR) systems. Fauzi et al. [11] proposed an IR system to retrieve information from Arabic fiqh texts. In all, 13 Arabic fiqh e-books were used as a dataset. The pages of the 13 books were treated as documents, and an inverse book frequency for ranking developed. They tested their work with precision, recall, and F-measure evaluation metrics and achieved 76% precision, 74% recall and 75% F-measure. Zamani et al. [12] proposed a standalone neural ranking model for document retrieval, using an inverted indexing technique to index the documents collected. Their work was tested using newswire and web collections. Precision, mean average precision (MAP), recall, and normalized discounted cumulative gain were used as evaluation metrics. Liu et al. [13] proposed a parallel indexing method to index traditional Chinese medical reports. A multifactor ranking model was applied for ranking, and the medical reports displayed using a template-based visualization method. Precision, MAP, normalized discounted cumulative gain, and average response times were used as evaluation metrics.

This section also examines computational works carried out on Tamil language and literature. Elanchezhiyan et al. [14] proposed the Kuralagam search engine for the Thirukkural, which retrieves couplets based on keywords, concepts and expanded query words. It retrieves couplets that are conceptually relevant to the query. MAP was used as the evaluation metric and achieved the score of 0.83. Madhavan et al. [15] classified Tamil poems into four protocols called "Paa", using a rule-based approach, and context-free grammar to create the rules. Tamil poems have been parsed, an intermediate representation created, and the poems subsequently classified into four categories. They have achieved classification accuracy of 90%. Sridevi and Subashini [16] classified 11th century Tamil handwritten texts using the probabilistic neural network. Line, word, character segmentation

and feature extraction were done before the classification, with structural and syntactic features put to use. Testing involved the use of 500 characters, with accuracy as the evaluation metric. They have achieved the classification accuracy of 80.52%.

It can be observed that the existing works have been done on the essay type of texts. The NLP tools are available for processing the texts and datasets are also available. The existing works have not focussed on the literary type of text. The proposed work focusses on the Tamil literary type of texts. The NLP tools and datasets are not available. For most Indian languages, language tools such as ontology, parts-of-speech tagging, and morphological analyzers have yet to be fully developed. These are the challenges for processing Tamil literary type of texts than essay type of texts. The proposed work builds the semantic interpretation framework for Tamil literary type of texts by discourse processing and it can also be implemented as a smart mobile application by using various multimedia components such as text, audio, and video etc.

3 Semantic Interpretation of Tamil Literary Texts

A semantic interpretation framework for Tamil literary texts using discourse parsing is proposed in this paper by giving works on discourse parsing, text classification, discourse-based clustering and information retrieval, and Tamil language and Tamil literatures.

3.1 Works on Discourse Parsing

For the Penn Discourse Tree Bank (PDTB), Ghosh et al. [17] identified explicit discourse connectives. For token level argument segmentation, they developed shallow discourse parsing. The document was examined sentence by sentence. As features, lexical, syntactic, and semantic features were used. Ghosh et al. [18] used a constraint-based strategy based on conditional random fields to increase the recall of a shallow discourse parser. Ghosh et al. [19] developed a parser that employs both local and global constraints. They used lexico-syntactic features to assess the text at the inter-sentence level. Subalalitha and Ranjani [20] utilized suthras and sangatis, notions from Tamil and Sanskrit literature, as well as modern text processing techniques like RST, Universal Networking Language, to identify semantic indices for Tamil documents. Suthras are used to represent text in a clear and concise manner.

For the Indian languages of Tamil, Malayalam, and Hindi, Sobha et al. [21] proposed automatic detection of connectives and their arguments. They employed a machine learning technique called Conditional Random Fields. As a corpus, they have used 3000 sentences from the health domain. Sobha et al. [22] annotated the discourse relations in three language corpora: Tamil, Malayalam, and Hindi. Lin et al. [23] developed a PDTB-style end-to-end discourse parser. Their parser found the sense of relation between arguments by identifying all discourse and non-discourse relations, and labelling the arguments. At the paragraph level, the document was examined. The features that were used were lexical, syntactic, and semantic. Ji and Eisenstein [24] used a representation learning strategy to convert surface features into latent space, making RST-based discourse parsing easier. They evaluated the document at the sentence level with a shift reduced discourse parser. For RST-based discourse parsing, Sidarenka et al. [25] segmented the German text. They looked over the text sentence by sentence.

Using RST-based discourse parsing and a recursive neural network, Bhatia et al. [26] suggested document level sentiment analysis. They used lexical features as features in their text analysis at the document level. In Tamil documents, Subalalitha and Ranjani [27] discovered 13 RST Relations. Beyond the sentence level, the Naïve Bayes probabilistic classifier was utilised to analyse the Tamil documents. Their discourse parser, which was inherited from Universal Networking Language, employed the high level semantic properties to generate rhetorical structure trees. RST and Segmented Discourse Representation Theory were used to annotate the corpus by Stede et al. [28]. The argumentation annotation was also added to it. The document was examined sentence by sentence. As a feature set, syntactic and semantic features were used. For identifying the discourse relation between adjacent sentences, Ji et al. [29] suggested a latent variable recurrent neural network. They used lexical features to examine the text at the inter-sentence level. Using a recursive neural network and RST, Ji and Smith [30] presented text categorization. The document was examined sentence by sentence.

3.2 Works on Discourse-based Clustering

For literary text analysis, Luyckx et al. [31] used K-means clustering. The authors have used parts of speech tags as features. For their experiments, they employed the Wall Street Journal as a dataset. In terms of the text's discourse relation, Rysova and Rysova [32] focused at the discourse connectives. In discourse connectives, they suggested constraints and preferences. The Prague

Dependency Treebank was utilized to test their findings. Rutherford and Xue [33] proposed utilizing Maximum Entropy classifiers to improvise implicit discourse relations by classifying explicit discourse connectives. They tested their method using two datasets: English Gigaword Corpus Version 3, and PDTB 2.0. The evaluation metrics were precision, recall, and F1-score.

Braud and Denis [34] suggested a semi-supervised approach for strengthening implicit discourse relations. A binary classifier was used to identify the discourse connectives. They tested their method using PDTB and newspaper articles from the Wall Street Journal datasets. As evaluation measures, F1-score and accuracy were utilized. Malmi et al. [35] proposed that discourse connectives in shorter sentences and passages be automatically predicted. For this assignment, they employed web-based papers. A decomposable attention model was utilized to predict discourse connectives automatically, and the results were compared to human ratters. The F1-score was utilized as a criterion for evaluation. The primary and secondary discourse connectives have been analyzed by Rysova and Rysova [36]. Primary connectives, such as *therefore*, have only one word and are grammaticalized, but secondary connectives, such as *for this reason*, have multiple words. To test their technique, they employed the Prague Discourse Treebank and Czech newspaper texts as datasets.

3.3 Works on Text Classification

Al-Badarneh et al. [37] investigated different indexing approaches for Arabic text classification using the multinomial Naïve Bayes classifier and identified five categories. Word frequencies were used as features in their 1000 normalized Arabic documents dataset. Micro average accuracy was used as evaluation metric. Xu [38] classified a text into 20 categories, comparing algorithms such as the classical Naïve Bayes classifier as well as the Bernoulli and Gaussian event models. Word frequencies of the document were used as features. Their work was tested on 23020 English documents from the 20 Newsgroups and WebKB datasets. As evaluation measures, precision, recall, and F-measure were used. Bahgat et al. [39] classified the email into two categories using semantic methods. The Principal Component Analysis and Correlation Feature Selection techniques were used for feature selection. The benchmark Enron dataset was used for evaluating their work. The comparative study was performed with different machine learning techniques.

The real time classification of social media data for disaster response and recovery was proposed by Ragini, et al. [40]. The SVM algorithm was used for text classification. The authors have used the data collected from Twitter. They have used precision, recall and F1-score as evaluation metrics. Text classification of Arabic documents was proposed by Elnagar et al. [41]. The authors have used deep learning models for the text classification task. They have used SANAD and NADiA Arabic datasets for classification. Accuracy was used as evaluation metric and obtained 96.94 percent in SANAD dataset and 88.68 percent in NADiA dataset. Meta-feature representation, sparsification and selective sampling were proposed in the pre-processing step of the text classification by Cunha et al. [42]. The authors have used kNN algorithm for selecting the meta-features. They have used four datasets namely, WebKB, 20NewsGroup, a subset of ACM digital library, and Reuters. Micro Averaged F1 and Macro Averaged F1 evaluation metrics were used to evaluate their work.

The fine tuning algorithm for Naïve Bayesian classifier is proposed by El et al. [43]. They have compared their algorithm with the existing Naïve Bayes algorithm. The performance of their work is tested using 47 UCI datasets. The comparison is performed with different machine learning algorithms using 18 UCI text classification datasets. An interaction representation was proposed by Zheng et al. [44] for text classification, which gave interaction among words. They have used five datasets, namely, IMDB dataset, Yelp dataset, open domain fact based questions dataset, AG news corpus and DBpedia dataset. Natured inspired algorithm and ensemble classifier were used by Khurana et al. [45] for optimal text classification. They have used 10 datasets taken from UCI repository and one real-time airline dataset. They tested their work using precision, recall, F-score and accuracy.

The hierarchical multi label Arabic text classification was proposed by Aljedani et al. [46]. The authors have used various machine learning algorithms for the classification. They have used the in house dataset. They have evaluated their work using micro-averaged precision, recall and F-measure. The discharge summary classification based on sentimental analysis was proposed by Waheeb et al. [47]. They have used various machine learning algorithms. They have used 1237 discharge summaries as the dataset. They have evaluated their work using F1-score. The combination of different preprocessing steps were used to improve the text classification by HaCohen-Kerner et al. [48]. They have used three machine learning algorithms, namely, Bayes network, a variant of SVM, and Random Forest. The authors have used

four datasets namely, R8, WebKB, sentiment labelled sentences, and SMS spam collection. They have evaluated their work using accuracy.

Automatic medical protocol classification was proposed by López-Úbeda et al. [49]. The authors have used random forest, and SVM algorithms for their text classification task. They have used HT-medica dataset which composed of CT and MRI examinations. They have evaluated their work using precision, recall, F1-score, and accuracy. The reasoning mechanism was proposed in multi label text classification by Wang et al. [50]. The authors have used AAPD dataset for testing their work. They have used recall, precision, and F1-score for evaluating their work. Luo [51] used SVM, Naïve Bayes, and Logistic Regression machine learning algorithms to classify English documents. Word frequency, question mark, full stop, initial word and final word of the documents were used as features. The author has tested the work with 1033 English documents.

3.4 Works on Information Retrieval Systems

Meng et al. [52] proposed a scheme for indexing and retrieving social media data. Experiments were undertaken on two image datasets and an e-commerce dataset. Precision, recall, MAP, and response time were used as evaluation metrics. Tekli et al. [53] proposed an approach for semantic indexing to retrieve information from structured, unstructured and semi-structured data. An algorithm developed to facilitate searching was tested on the IMBD movie dataset. Their work was evaluated using precision, recall, F-score and MAP metrics.

Samia and Khaled [54] proposed an Arabic plagiarism detection system. Semantic indexing was carried out using Arabic ontology and parts-of-speech tagging. The AraPlagDet corpus was used to evaluate their work using precision, recall and F-score metrics. Agosti et al. [55] proposed an unsupervised neural framework for IR. Semantic indexing, synonymy, and polysemy were used to eliminate semantic gaps, indicating a mismatch between document terms and queries. The TREC CDS and OHSUMED collections were used to test their work, with precision and recall as the evaluation metrics.

3.5 Computational Works on Tamil Language and Tamil Literature

Prasath et al. [56] proposed a cross-language IR approach for a given user query in another language. A corpus-driven query suggestion approach for

re-ranking was used on Tamil and English news collections of the FIRE corpora, with precision as the evaluation metric. Subalalitha and Anita [57] proposed a page ranking algorithm based on discourse relations, to retrieve web pages. The rhetorical structure theory was applied to ascertain semantic relations between web pages, as well as hyperlinks in the web pages. In all, 500 Tamil and 50 English tourism web pages were tested, using precision as the evaluation metric. Giridharan et al. [58] proposed a scheme to retrieve information from ancient Tamil texts inscribed in temples, and after that transform the epigraphy into current Tamil digital texts, along with their meaning. The Brahmi database was used as a dataset, and accuracy of 84.57% obtained.

Sankaralingam et al. [59] put forward a methodology for IR for Tamil texts, using ontology to convert ontological structures into visual representations that aid retrieval. Lexical and semantic relations such as homonymy, synonymy, antonymy, and meronymy were used on their 50000-word general domain dataset. Thenmozhi and Aravindan [60] suggested an ontology-based cross-lingual IR system. Tamil queries were translated into English and the relevant documents were retrieved in English. Ambiguity was eliminated from Tamil and English queries with word-sense disambiguation and ontology, respectively. A Tamil-English bilingual dictionary, a Tamil morphological analyzer, and a named entity database were used to translate Tamil queries into English. Their methodology was evaluated for the agricultural domain, with precision as the evaluation metric. Subalalitha and Poovammal [61] constructed an automatic bilingual dictionary for the Thirukkural using the Naïve Bayes machine learning algorithm. They used G. U. Pope's English translation and commentary, as well as Dr. M. Varadharajan and Dr. Solomon Pappaiya's Tamil explanations. Precision was used as an evaluation metric.

Subalalitha [62] proposed an information extraction scheme for the Tamil literary work, Kurunthogai. Food, flora, fauna, vessels, waterbodies, noun unigrams, verb unigrams, adjective-noun bigrams, and adverb-verb bigrams were all extracted for information. N-grams were extracted using a Tamil morphological analyzer tool. Precision was used as the evaluation metric. Clustering of Thirukkural couplets based on discourse connectives was proposed by Anita and Subalalitha [63]. The machine learning algorithm used was K-means clustering. Cluster purity, Rand index, precision, recall, and F-score were used as evaluation metrics, resulting in 79 percent purity, 92 percent overall Rand index, 79 percent precision, 80 percent recall, and a 79 percent F-score. A rule-based method for creating a discourse parser

for the Thirukkural was proposed by Anita and Subalalitha [64]. Discourse connectives have been used as the features. Precision and recall were used as evaluation metrics and achieved 81.5% precision and 81.86% recall.

Saravanan [65] proposed a cluster-based Tamil document retrieval system using semantic indexing. The K-means algorithm was used on a dataset taken from the Tamil Language Consortium Repository. F-score was used as an evaluation metric. Tamil handwritten character recognition was proposed by Vinotheni et al. [66]. The authors have used modified convolution neural network model for this work. They have collected the dataset from various schools. They have evaluated their work with accuracy. A discourse-based information retrieval system for the Tamil literary texts such as, Thirukkural and Naladiyar was proposed by Anita and Subalalitha [67]. The discourse structure was constructed and discourse-based inverted indexing, searching and ranking were performed for information retrieval. The work is compared with the Google search and keyword-based search. Mean Average Precision (MAP) was used to evaluate their work and achieved 89 percent.

4 Results and Discussion

Existing discourse approaches analyzed text in English documents and expository type Tamil documents, as can be shown. This paper aids to propose a discourse technique that employs RST to identify semantic/discourse relations in a Tamil literature text that lacks a consistent pattern for semantic analysis. Unlike English, which uses a fixed Subject Verb Object (SVO) sentence pattern, Tamil expository texts use either SVO or Subject Object Verb (SOV). Tamil literature, on the other hand, does not follow the SOV or SVO pattern. Tamil literatures contain a diverse range of morphological variants (Anand et al. [68]). The usage of words in expository documents is different from literature-type text.

The existing works have targeted building IR system for English, Chinese and Arabic languages. These IR systems are capable of semantically interpreting the query, whereas, Indian language based IR systems are still at the keyword-based search level. This is due to the fact that language tools like ontology, parts-of-speech tagging, and morphological analyzers are yet to be comprehensively developed for most Indian languages. It is observed that much of the existing research carried out has restricted itself to the Tamil expository text, with very few studies undertaken on Tamil literary

texts. These challenges has to be met to achieve high accuracy and better performance in the discourse-based interpretation framework construction for Tamil literary texts and it can be developed as a smart mobile application by using multimedia components.

The proposed work is evaluated using precision, recall, F-score and accuracy as the evaluation metrics using Equations (1)–(3) (Ravi et al. [69]) and (4) (Makaju et al. [70]). The fraction of relevant outcomes among the retrieved results is known as precision. The percentage of relevant results that are successfully retrieved is known as recall. The harmonic mean of precision and recall is the F-score. The percentage of right decisions is calculated by accuracy.

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F-score} = \frac{2\,\text{P}\,\text{R}}{\text{P} + \text{R}} \tag{3}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

Where true positive, true negative, false positive, and false negative are represented by the letters TP, TN, FP, and FN. The number of similar elements in the same class is true positive; the number of dissimilar elements in separate classes is true negative; the number of dissimilar elements in the same class is false positive, and the number of similar elements in different classes is false negative.

Figures 1–4 show the performance comparison of the existing works using various machine learning algorithms and datasets. The algorithms are Logistic Regression (LR) algorithm, Support Vector Machine (SVM) algorithm, Random Forest (RF) algorithm, and Multinomial Naïve Bayes (MNB) algorithm. Figure 1 shows the performance comparison of the algorithms using precision. It shows that MNB algorithm works well (85.78%) than other algorithm.

Figure 2 shows the performance comparison of the algorithms using recall. It shows that MNB algorithm has good recall value (86.42%) than others.

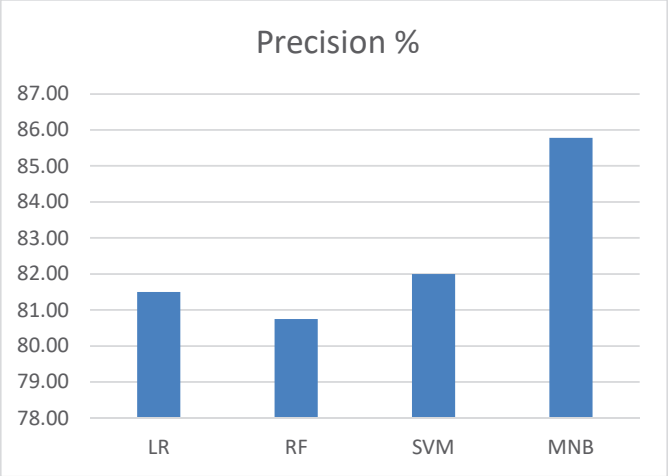

Figure 1 Performance comparison of precision.

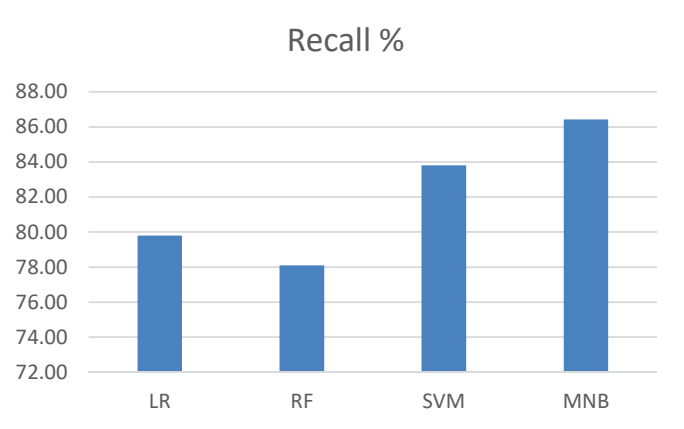

Figure 2 Performance comparison of recall.

Figure 3 shows the performance comparison of the algorithms using F-Score. It shows that MNB algorithm works well (86.1%) than other algorithm.

Figure 4 shows the performance comparison of the algorithms using accuracy. It can be observed that the accuracy of 87.98 percent is achieved in MNB, which shows the MNB algorithm outperforms the other algorithms.

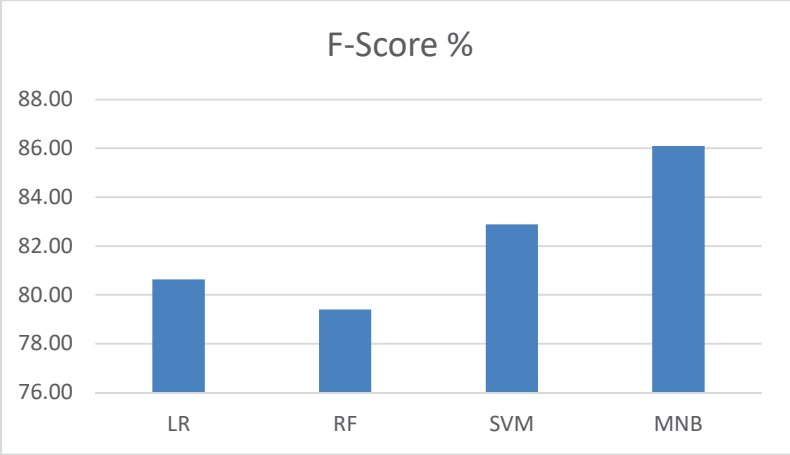

Figure 3 Performance comparison of F-score.

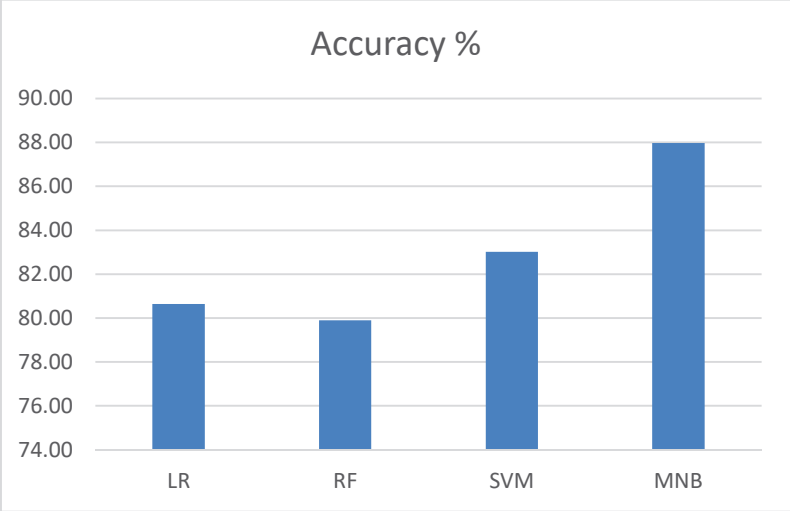

Figure 4 Performance comparison of accuracy.

5 Conclusions and Future Work

Tamil literary classics, dating prior to 300 BCE, are a treasure trove of invaluable information. Computationally analyzing them to make them accessible to today's generation is inevitable. A discourse/semantic analysis on the Tamil literary texts must be performed. In this paper, various methodologies

with different datasets for text classification, clustering, and IR has been discussed. This paper has observed the gaps related to the works done on essay type of texts which has to be performed to achieve high accuracy and better performance in Tamil literary texts. This paper has proposed automatic semantic interpretation of Tamil literary texts by giving works on discourse parsing, text classification, discourse-based clustering and information retrieval systems. The semantic interpretation framework can be developed as a mobile application by using multimedia components. This mobile application can be used as a search engine for retrieving relevant information from the Tamil literary texts for the given user query.

References

[1] Thompson, S. A. and Mann, W. C. (1987). Rhetorical structure theory: A theory of text organization. The structure of discourse, Norwood NJ, Ablex.

[2] Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text, 8(3), 243–281.

[3] Subba, R. and Di Eugenio, B. (2009, June). An effective discourse parser that uses rich linguistic information. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 566–574).

[4] Hernault, H., Prendinger, H., du Verle, D. A. and Ishizuka, M. (2010). HILDA: A discourse parser using support vector machine classification. Dialogue & Discourse, 1(3), 1–33.

[5] Hernault, H., Bollegala, D. and Ishizuka, M. (2010, October). A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 399–409).

[6] Alonso, L., Castellón, I., Gibert, K. and Padró, L. (2002, October). An empirical approach to discourse markers by clustering. In Catalonian Conference on Artificial Intelligence (pp. 173–183). Springer, Berlin, Heidelberg.

[7] Miltsakaki, E., Joshi, A., Prasad, R. and Webber, B. (2004). Annotating discourse connectives and their arguments. In Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004 (pp. 9–16).

[8] Ramesh, B. P. and Yu, H. (2010). Identifying discourse connectives in biomedical text. In AMIA Annual Symposium Proceedings (Vol. 2010, p. 657). American Medical Informatics Association.

[9] Al-Salemi, B. and Aziz, M. J. A. (2011). Statistical Bayesian learning for automatic Arabic text categorization. Journal of computer Science, 7(1), 39.

[10] Rizzo, G., Tomassetti, F., Vetro, A., Ardito, L., Torchiano, M., Morisio, M. and Troncy, R. (2017). Semantic enrichment for recommendation of primary studies in a systematic literature review. Digital Scholarship in the Humanities, 32(1), 195–208.

[11] Fauzi, M. A., Arifin, A. Z. and Yuniarti, A. (2017). Arabic book retrieval using class and book index based term weighting. International Journal of Electrical & Computer Engineering, 7(6), 2088–8708.

[12] Zamani, H., Dehghani, M., Croft, W. B., Learned-Miller, E. and Kamps, J. (2018, October). From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In Proceedings of the 27th ACM international conference on information and knowledge management, (pp. 497–506).

[13] Liu, L., Liu, L., Fu, X., Huang, Q., Zhang, X. and Zhang, Y. (2018). A cloud-based framework for large-scale traditional Chinese medical record retrieval. Journal of biomedical informatics, 77, 21–33.

[14] Elanchezhiyan, K., Geetha, T. V., Ranjani, P. and Karky, M. (2011). Kuralagam - Concept Relation based Search Engine for Thirukkural. In Tamil Internet Conference, University of Pennsylvania, Philadelphia, USA, 19–23.

[15] Madhavan, K. V., Nagarajan, S. and Sridhar, R. (2012). Rule based classification of tamil poems. International Journal of Information and Education Technology, 2(2), 156.

[16] Sridevi, N. and Subashini, P. (2013). Optimized Framework for Classification of 11th Century Handwritten Ancient Tamil Scripts using Computational Intelligence. International Journal of Computer Science. 2 (2), 14–23.

[17] Ghosh, S., Johansson, R., Riccardi, G. and Tonelli, S. (2011, November). Shallow discourse parsing with conditional random fields. In Proceedings of 5th International Joint Conference on Natural Language Processing (pp. 1071–1079).

[18] Ghosh, S., Johansson, R., Riccardi, G. and Tonelli, S. (2012, May). Improving the Recall of a Discourse Parser by Constraint-based Post-processing. In LREC (pp. 2791–2794).

[19] Ghosh, S., Riccardi, G. and Johansson, R. (2012, July). Global features for shallow discourse parsing. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 150–159).

[20] Subalalitha, C. N. and Ranjani, P. (2014). A unique indexing technique for discourse structures. Journal of Intelligent Systems, 23(3), 231–243.

[21] Sobha Lalitha Devi, Lakshmi S and Sindhuja Gopalan (2014). "Discourse Tagging for Indian Languages", In A. Gelbukh (ed), Computational Linguistics and Intelligent Text Processing, Springer LNCS Vol 8403, pp. 469–480.

[22] Sobha Lalitha Devi, Sindhuja Gopalan, Lakshmi S (2014). Automatic Identification of Discourse Relations in Indian Languages. In proceedings of 2nd Workshop on Indian Language Data: Resources and Evaluation, Organized under LREC2014, Reykjavik, Iceland.

[23] Lin, Z., Ng, H. T. and Kan, M. Y. (2014). A PDTB-styled end-to-end discourse parser. Natural Language Engineering, 20(2), 151–184.

[24] Ji, Y. and Eisenstein, J. (2014, June). Representation learning for text-level discourse parsing. In Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 13–24).

[25] Sidarenka, U., Peldszus, A. and Stede, M. (2015). Discourse Segmentation of German Texts. J. Lang. Technol. Comput. Linguistics, 30(1), 71–98.

[26] Bhatia, P., Ji, Y. and Eisenstein, J. (2015). Better document-level sentiment analysis from RST discourse parsing. arXiv preprint arXiv:1509.01599.

[27] Subalalitha, C. N. and Ranjani, P. (2015). Building a Language-Independent Discourse Parser using Universal Networking Language. Computational Intelligence, 31(4), 593–618.

[28] Stede, M., Afantenos, S., Peldzsus, A., Asher, N. and Perret, J. (2016, May). Parallel discourse annotations on a corpus of short texts. In 10th International Conference on Language Resources and Evaluation (LREC 2016) (pp. 1051–1058).

[29] Ji, Y., Haffari, G. and Eisenstein, J. (2016). A latent variable recurrent neural network for discourse relation language models. arXiv preprint arXiv:1603.01913.

[30] Ji, Y. and Smith, N. (2017). Neural discourse structure for text categorization. arXiv preprint arXiv:1702.01829.

[31] Luyckx, K., Daelemans, W. and Vanhoutte, E. (2006). Stylogenetics: Clustering-based stylistic analysis of literary corpora. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy.

[32] Rysová, M. and Rysová, K. (2014, December). The centre and periphery of discourse connectives. In Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing (pp. 452–459).

[33] Rutherford, A. and Xue, N. (2015). Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 799–808).

[34] Braud, C. and Denis, P. (2016, November). Learning connective-based word representations for implicit discourse relation identification. In Empirical Methods on Natural Language Processing.

[35] Malmi, E., Pighin, D., Krause, S. and Kozhevnikov, M. (2017). Automatic prediction of discourse connectives. arXiv preprint arXiv:1702.00992.

[36] Rysová, M. and Rysová, K. (2018). Primary and secondary discourse connectives: Constraints and preferences. Journal of Pragmatics, 130, 16–32.

[37] Al-Badarneh, A., Al-Shawakfa, E., Bani-Ismail, B., Al-Rababah, K. and Shatnawi, S. (2017). The impact of indexing approaches on Arabic text classification. Journal of Information Science, 43(2), 159–173.

[38] Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science, 44(1), 48–59.

[39] Bahgat, E. M., Rady, S., Gad, W. and Moawad, I. F. (2018). Efficient email classification approach based on semantic methods. Ain Shams Engineering Journal, 9(4), 3259–3269.

[40] Ragini, J. R., Anand, P. R. and Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. International Journal of Information Management, 42, 13–24.

[41] Elnagar, A., Al-Debsi, R. and Einea, O. (2020). Arabic text classification using deep learning models. Information Processing & Management, 57(1), 102121.

[42] Cunha, W., Canuto, S., Viegas, F., Salles, T., Gomes, C., Mangaravite, V. Resende, E., Rosa, T., Goncalves M. A. and Rocha, L. (2020). Extended pre-processing pipeline for text classification: On the role

of meta-feature representations, sparsification and selective sampling. Information Processing & Management, 57(4), 102263.

[43] El Hindi, K. M., Aljulaidan, R. R. and AlSalman, H. (2020). Lazy fine-tuning algorithms for naïve Bayesian text classification. Applied Soft Computing, 96, 106652.

[44] Zheng, J., Cai, F., Chen, H. and de Rijke, M. (2020). Pre-train, Interact, Fine-tune: a novel interaction representation for text classification. Information Processing & Management, 57(6), 102215.

[45] Khurana, A. and Verma, O. P. (2020). Novel approach with nature-inspired and ensemble techniques for optimal text classification. Multimedia Tools and Applications, 79(33), 23821–23848.

[46] Aljedani, N., Alotaibi, R. and Taileb, M. (2020). HMATC: Hierarchical multi-label Arabic text classification model using machine learning. Egyptian Informatics Journal.

[47] Waheeb, S. A., Khan, N. A., Chen, B. and Shang, X. (2020). Machine learning based sentiment text classification for evaluating treatment quality of discharge summary. Information, 11(5), 281.

[48] HaCohen-Kerner, Y., Miller, D. and Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. PloS one, 15(5), e0232525.

[49] López-Úbeda, P., Díaz-Galiano, M. C., Martín-Noguerol, T., Luna, A., Ureña-López, L. A. and Martín-Valdivia, M. T. (2021). Automatic medical protocol classification using machine learning approaches. Computer Methods and Programs in Biomedicine, 200, 105939.

[50] Wang, R., Ridley, R., Qu, W. and Dai, X. (2021). A novel reasoning mechanism for multi-label text classification. Information Processing & Management, 58(2), 102441.

[51] Luo, X. (2021). Efficient English text classification using selected machine learning techniques. Alexandria Engineering Journal, 60(3), 3401–3409.

[52] Meng, L., Tan, A. H. and Wunsch II, D. C. (2019). Online multimodal co-indexing and retrieval of social media data. In Adaptive resonance theory in social media data clustering, (pp. 155–174). Springer, Cham.

[53] Tekli, J., Chbeir, R., Traina, A. J. and Traina Jr, C. (2019). SemIndex+: A semantic indexing scheme for structured, unstructured, and partly structured data. Knowledge-Based Systems, 164, 378–403.

[54] Samia, Z. and Khaled, R. (2020). Multi-agents indexing system (MAIS) for plagiarism detection. Journal of King Saud University-Computer and Information Sciences.

[55] Agosti, M., Marchesin, S. and Silvello, G. (2020). Learning unsupervised knowledge-enhanced representations to reduce the semantic gap in information retrieval. ACM Transactions on Information Systems (TOIS), 38(4), 1–48.

[56] Prasath, R., Sarkar, S. and O'Reilly, P. (2015, April). Improving cross language information retrieval using corpus based query suggestion approach. In International Conference on Intelligent Text Processing and Computational Linguistics, (pp. 448–457). Springer, Cham.

[57] Subalalitha, C. N. and Anita, R. (2016). An approach to page ranking based on discourse structures. Journal of Communications Software and Systems, 12(4), 195–200.

[58] Giridharan, R., Vellingiriraj, E. K. and Balasubramanie, P. (2016, April). Identification of Tamil ancient characters and information retrieval from temple epigraphy using image zoning. In 2016 International conference on recent trends in information technology (ICRTIT), (pp. 1–7). IEEE.

[59] Sankaralingam, C., Rajendran, S., Kavirajan, B., Kumar, M. A. and Soman, K. P. (2017, September). Onto-thesaurus for Tamil language: Ontology based intelligent system for information retrieval. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), (pp. 2396–2396). IEEE.

[60] Thenmozhi, D. and Aravindan, C. (2018). Ontology-based Tamil–English cross-lingual information retrieval system. Sādhanā, 43(10), 1–14.

[61] Subalalitha, C. N. and Poovammal, E. (2018). Automatic bilingual dictionary construction for Tirukural. Applied Artificial Intelligence, 32(6), 558–567.

[62] Subalalitha, C. N. (2019). Information extraction framework for Kurunthogai. Sādhanā, 44(7), 1–6.

[63] Anita, R. and Subalalitha, C. N. (2019, July). An Approach to Cluster Tamil Literatures Using Discourse Connectives. In 2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP) (pp. 1–4). IEEE.

[64] Anita, R. and Subalalitha, C. N. (2019, December). Building Discourse Parser for Thirukkural. In Proceedings of the 16th International Conference on Natural Language Processing (pp. 18–25).

[65] Saravanan, M. S. (2020). Semantic document clustering based indexing for Tamil language information retrieval system. Journal of Critical Reviews, 7(14), 2999–3007.

[66] Vinotheni, C., Pandian, S. L. and Lakshmi, G. (2021). Modified convolutional neural network of Tamil character recognition. In Advances in Distributed Computing and Machine Learning, 127 (pp. 469–480). Springer, Singapore.
[67] Anita, R. and Subalalitha, C. N. (2021). A discourse-based information retrieval for Tamil literary texts. Journal of Information and Communication Technology, 20(3), 353–389.
[68] Anand Kumar, M., Dhanalakshmi, V., Soman, K. P. and Rajendran, S. (2010). A sequence labeling approach to morphological analyzer for tamil language. International Journal on Computer Science and Engineering, 2(06), 1944–1951.
[69] Ravi, L., Subramaniyaswamy, V., Vijayakumar, V., Chen, S., Karmel, A. and Devarajan, M. (2019). Hybrid location-based recommender system for mobility and travel planning. Mobile Networks and Applications, 24(4), 1226–1239.
[70] Makaju, S., Prasad, P. W. C., Alsadoon, A., Singh, A. K. and Elchouemi, A. (2018). Lung cancer detection using CT scan images. Procedia Computer Science, 125, 107–114.

Biographies

Anita Ramalingam received the B.E. degree in Computer Science and Engineering from Manonmaniam Sundaranar University in 2004, the M.E. degree in Computer Science and Engineering from Sathyabama University in 2012, and she is pursing PhD in Computer Science and Engineering in SRM Institute of Science and Technology, Chennai. She is currently working as an Assistant Professor in Department of Computer Science and Engineering, S.R.M Institute of Science and Technology, Kattankulathur, Chennai, India. She has over fifteen years of experience in Teaching. She has published her

research papers in various refereed International Journals, National Journals and International Conferences. Her research areas include Natural Language Processing and regional language computing.

Subalalitha Chinnaudayar Navaneethakrishnan has finished her PhD in the Natural Language Processing domain in the year 2014 in College of Engineering Guindy (CEG), Anna University, India. She was working as a Junior Research Fellow for the Indian Government funded project titled, "Cross Lingual Information Access" for about 3.5 years in CEG. She has published her research papers in various refereed International Journals, National Journals and International Conferences. Her research interests inclines towards the domains namely, Natural Language Processing and Computational Linguistics. She is currently working as Assistant Professor in SRM Institute of Science and Technology, Tamil Nadu, India.