Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

MULTI-SCALE DOMAIN ADAPTIVE TRANSFER BASED SENTIMENT ANALYSIS FOR MORPHOLOGICAL RICH TAMIL TEXT

R. REGAN*

Assistant Professor, Department of Computer Science and Engineering, University College of Engineering, Villupuram, India. *Corresponding Author Email: reganr85@gmail.com

S. SENTHILKUMAR

Assistant Professor, Department of Computer Science and Engineering, University College of Engineering, BIT Campus, Tiruchirappalli, India. Email:senthilucepkt@gmail.com

M. RATHINRAJ

Project Assistant, Department of Computer Science and Engineering, University College of Engineering Villupuram, India. Email:rathin.vimprotech@gmail.com

Abstract

Sentiment analysis is the procedure of extracting information from given text about several functions, human beings, systems and facts. Sentimental analysis utilizes data mining and natural language processing (NLP) techniques to unearth, discover, retrieve and refine the information from the World Wide Web's limitless textual information. The sentiment analyzers for several high-resource languages and certain Indic languages are completely developed. However, Tamil that is considered as a morphological rich language has not experienced these advancements. This paper proposes a novel method called, Multi-Scale Attention and Domain Adaptive Transfer Learning (MSA-DATL) for performing intelligent sentiment analysis as positive, negative or neutral. The proposed method includes three complementary components. They pre-processing, keyword extraction and classification for performing morphological rich sentiment analyses using Tamil text. First, Multilingual Text-To-Text-Transfer-Transformer-based Pre-processing is performed. Following which Multi-Scale Attention Mechanism-based Keyword Extraction using Sliding Window for short range contextual dependencies and Sparse Attention function for long range contextual dependencies are done. Finally, Domain Adaptive Transfer Learning-based Classifier is applied for efficient sentiment analysis. Experimental results on the dataset used in our work show that the proposed MSA-DATL method achieves 29% accuracy, with weighted-average precision, recall of 29%, 28% respectively, significantly outperforming baseline methods. These findings illustrate the robustness and efficiency of our method making a notable contribution to intelligent analysis of morphological rich Tamil text.

Keywords: Morphological Rich Text, Sentiment Analysis, Multi-Scale Attention, Sliding Window, Domain Adaptive, Transfer Learning.

1. INTRODUCTION

The emergence of the internet has triggered an eruption of content generated by user, enabling humans to share their opinions on several topics. Nevertheless, with the increase in the volume of data, the issue of timely and effective relevant information extraction becomes increasingly critical, emphasizing the requirement for sophisticated computational linguistic techniques. Sentiment Analysis within NLP plays a demanding role is ascertaining the sentiments expressed in textual data.A novel method called, CMSA-mBERT was proposed in [1] for multiclass sentiment analysis of low-resource language using multilingual transformers. To perform sentiment analysis followed by conventional NLP techniques, transformer based libraries were exploited for performing

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

tokenization and embedding. Next, Multilingual Bidirectional Encoder Representations from Transformers (mBERT) model was applied with the intent of performing optimization and were trained for multiclass sentiment analysis therefore ensuring notable enhancement in terms of precision, recall and accuracy for sentiment analysis in low resource language. In spite of improvement observed in terms of precision and accuracy however the time consumed in multiclass sentiment analysis of low-resource language was not performed. To focus on this aspect in this work prior to keyword extraction, pre-processing using Multilingual Text-To-Text-Transfer-Transformer is applied that via tokenization and encoding generates results in a computationally efficient manner for further processing. A novel utilization of adapter-based tuning in multilingual transformer models, called, an adapter-based mDeBERT was proposed in [2] for classifying abusive comments in Tamil. Initially to increase the frequency of comments using preprocessing and text augmentation were performed. Followed by which adapter-based model was applied for not only extracting relevant feature but also to fine tune them that in turn minimized the trainable parameter frequency with improved precision and accuracy. Despite improvement in terms of precision and accuracy however the false positive rate involved for classifying abusive comments in Tamil was not conducted. To address on this aspect, Domain Adaptive Transfer Learning-based Classifier is used that based on the adversarial loss and sentiment classification loss in turn reduces the overall falsification of sentiment analysis in an accurate manner.

Conventional methods involved in sentiment analysis combined with static embeddings frequently lack in encapsulating deep contextual relationships within text. In [3], a hybrid deep learning method incorporating RoBERTa embeddings and convolutional neural network (CNN) was presented to analyze movie reviews for real time sentiment prediction. Nevertheless sentiment analysis in low-resource language received less attention because of language complexity. To address this research gap an ensemble method for predicting sentiments in Tamil languages was proposed in [4] with improved accuracy.

The resilience of recognition of character based on handwritten text depends on its potentiality to generalize across different hand writing styles due to their complicated structures and structural complexities. In spite of Tamil's linguistic significance and constrained resources have posed issues for designing efficient effective recognition methods. Multilingual transformer model was applied in [5] to enhance NLP task performance for low-resource languages. However computational efficiency was not focused.Deep learning based method tailored for Tamil script digitization to improve computational efficiency without compromising accuracy was presented in [6]. Yet another method focusing of computationally effectiveness along with accuracy employing fast recurrent neural networks was proposed in [7].

Sentiment Analysis (SA) has attained notable evolution over the recent decades among the advancing applications, to name a few being, social networks and decision-support systems. Moreover of late the mushrooming of data volumes for Dravidian languages has made SA a notable task in generating useful information. A comprehensive review on the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

application of machine learning (ML) and deep learning (DL) techniques for sentiment analysis was investigated in [8]. Moreover a comprehensive discussion of the issues, disadvantages and elements contributing to enhanced accuracy within the domain of sentiment analysis using DL algorithms was presented in [9]. The standard works are concentrated more on the analysis of monolingual language and lack a comprehensive study of the recent trends. A survey on the advancing perspective of SA in morphological low-resource language with a specific concentration on Dravidian languages was investigated in [10].

Morphological synthesis is one of the foremost elements to be considered when both of the source and target languages are morphologically rich. The two languages in India considered to be morphologically rich as well as agglutinative are Malayalam and Tamil. Nonetheless, researchers are progressively focused in navigating the prospective of transformed based methods that have been pre-trained for numerous natural language processing applications, specifically for limited resource languages. The efficiency of four pre-trained transformer models was investigated in [11] for sentence wise sentiment analysis accurately.

A dual branch neural network was presented in [12] contributing to safer online environments with improved accuracy. A plethora of DL algorithms for morphological synthesis in Tamil at the character level was designed in [13] with enhanced accuracy. Three under resourced Dravidian languages were analyzed in [14] using ML and DL methods therefore ensuring minimal false positive rate. A hybrid morphological analyzer contributing to agglutinative and morphological nature was designed in [15].

1.1 Research Gap

In spite of extensive studies on morphological rich Tamil text sentiment analysis involving high quality and specialized learning methods, gaps in research persist. One area with room for exploration is the use of transfer learning methods to process morphological rich Tamil text based on both source and target data. Moreover, the interpretability of transfer learning and sentiment analysis with Tamil text is in the inception stage, with constrained mechanisms for comprehensively understanding keyword significance. Addressing these gaps is crucial for improving sentiment analysis results.

1.2 Contributions of The Work

To address on the above said issues, like reducing false positive rate and measure accurate sentiment analysis for morphological rich Tamil text, a method called, Multi-Scale Attention and Domain Adaptive Transfer Learning (MSA-DATL) is designed. The contributions of the MSA-DATL are listed as given below.

- To design a robust sentiment analysis for morphological rich Tamil text using Multi-Scale Attention and Domain Adaptive Transfer Learning (MSA-DATL) based on pre-processing, keyword extraction and classification.
- To reduce execution time involved in sentiment analysis, Multilingual Text-To-Text-Transfer-Transformer-based Pre-processing is used.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

- To design Multi-Scale Attention Mechanism-based Keyword Extraction algorithm that employing sliding window-based local attention and sparse attention-based global attention extracts both fine-grained keywords and coarse-grained keywords, therefore contributing to improved precision and reducing the overall false positive results.
- To propose Domain Adaptive Transfer Learning-based Classification algorithm for morphological rich Tamil text sentiment analysis where by applying adversarial training loss and sentiment classification loss improving overall sentiment analysis accuracy.
- Finally, comprehensive experimental assessment is carried out with five dissimilar performance metrics, precision, recall, accuracy, execution time and false positive rate to illustrate the proposed MSA-DATL method over traditional methods.

1.3 Organization of The Paper

This paper is organized as follows: In Section 2, we prompted our study by presenting the background and related work regarding sentiment analysis using deep learning and transfer learning methods. In Section 3, we present the details of our proposed Multi-Scale Attention and Domain Adaptive Transfer Learning (MSA-DATL) for sentiment analysis. A comprehensive qualitative analysis and inferences are described in Section 4. Experimental results with elaborate discussion using traditional methods using table and graphical representations are provided in Section 5. Finally, concluding remark is included in Section 6.

2. RELATED WORKS

Low-resource domain area sentiment analysis has observed notable evolutions over the past few years. A review of sentiment analysis of low-resource text using DL algorithms was presented in [16]. The specific objective here remained in identifying suitable methods for sentiment analysis in low-resource domain. Also the role of transfer learning in low-resource language based sentiment analysis was also narrated. Morphological analysis over the informal Tamil content becomes complicated as the language has differing representation for same word.

In [17] Yarowsky algorithm was used for contextual meaning analysis with minimal time complexity. An unsupervised method for analyzing sentiments using artificial intelligence (AI) techniques was proposed in [18] based on polarities of extracted opinions, therefore ensuring accuracy of detection. Nevertheless matter of concern is that it lacks text context semantics. A model for sentiment classification based on manifold dimensions was proposed in [19] to address sentiment analysis in the absence of text context semantics. Correct sentiment interpretation was conducted in [20] using ML and DL methods with improved accuracy.

Research works on NLP are specifically performed in English, with very few navigating languages that are said to be under-resourced, including the Dravidian languages. In [21]

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

a novel method in offensive language detection employing a corpus obtained from YouTube containing comments in Tamil was presented. The work notably strived to make an in-depth comparison between supervised and unsupervised ML algorithms with the intent of detecting offensive patterns present in textual communications achieving an impressive accuracy.

Novel sophisticated ML techniques were proposed in [22] for performing sentiment analysis in digital communication systems. Yet another multi stage deep learning architecture was presented in [23] to efficiently analyze sentiments by performing well with low-resource languages. This in turn significantly outperformed baseline methods in terms of accuracy. A survey of ML and DL techniques for performing sentiment analysis and their practical applications were detailed in [24].

A recent review on methods, issues faced while performing sentiment analysis employing ML and DL techniques for low-resource language was investigated in [25]. A comprehensive study investigating the methodological aspects, performing a holistic planning process analysis along with the procedures being implemented and mechanism utilized in ascertaining the data or keywords to be included or exclude in the evaluation framework were detailed in [26]. A sentiment analysis mechanism using DL for text including low-resource and high-resource was proposed in [27] with improved accuracy. An exhaustive review on sentiment analysis for Dravidian language employing pre-trained architecture and DL algorithms was presented in [28].

Overall, prior research reveals a clear advancement from rule-based and statistical methods toward deep learning based architectures across Tamil NLP domains. In spite of notable improvement has been made, the consistent issue of constrained annotated source data for morphological rich Tamil text underscores the requirement for transfer learning to boost classification performance.

3. MATERIALS AND METHODOLOGY

3.1 Data Collection and Dataset Description

The ThamizhiMorph dataset (i.e. source) used in our work extracted from https://github.com/sarves/thamizhi-morph is an open source Tamil morphological analyzer and generator utilized for tokenizing and POS tagging (i.e. nouns and verbs). The dataset contains 80K nouns and 18 verb classes. On the other hand, the MADTRAS dataset (i.e. target) extracted from https://data.mendeley.com/datasets/p59cfx4vx6/2 includes the following comments positive, negative or neutral.

3.2 Multi-Scale Attention and Domain Adaptive Transfer Learning Based Sentiment Analysis for Morphological Rich Tamil Text

The proposed method is a novel mechanism for sentiment analysis that strives to analyze and classify sentiments in a morphological rich target Tamil text language by harnessing knowledge from inflected words a more low-resource source language. The proposed method contains of three associated stages intended to address the issues concerning

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

linguistic complexities and morphological richness of the Tamil language efficiently. Morphological synthesis is the procedure of combining two words in accordance with the Sandhi rules of the morphologically rich Tamil language in India as well as agglutinative. We formulated the task of the morphological generation of nouns and verbs of Tamil as a character to-character sequence tagging problem. The proposed method architecture is illustrated in figure 1.

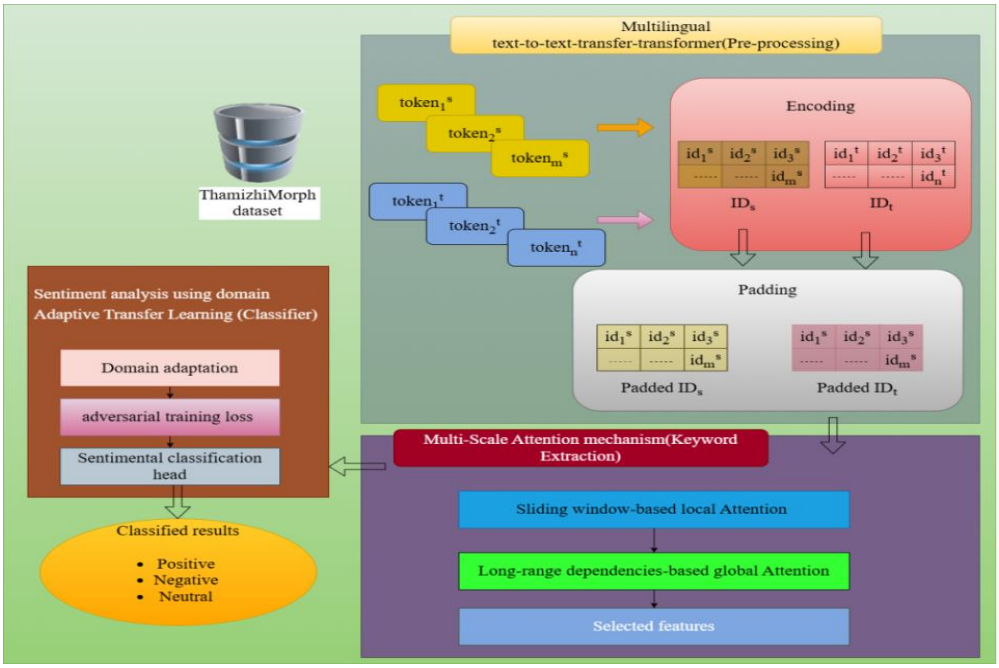


Figure 1: Architecture of MSA-DATL Based Sentiment Analysis

As illustrated in the above figure the overall process of intelligent sentiment analysis of morphological rich Tamil text is split into three sections. They are pre-processing, keyword extraction and classification for intelligent sentiment analysis of morphological rich Tamil text. First, with the samples attained as input, Multilingual Text-To-Text-Transfer-Transformer-based Pre-processing via tokenization, encoding and padding is performed for conducting sentiment analysis of morphological rich Tamil text.

Next, Multi-Scale Attention Mechanism-based Keyword Extraction via Sliding Window-based Local Attention and Sparse Attention-based global function is performed, therefore generating relevant keywords for further processing. Finally, Sentiment Analysis is performed using Domain Adaptive Transfer Learning-based Classifier. The numerous steps in the proposed method are demonstrated in the following subsections. By splitting down inflected Tamil words (i.e. source) into roots or morphological features (i.e. target) address agglutination that is critical for accurate sentiment analysis. The generated results are in positive state (i.e. reviews contain an implicit/explicit clue in the content recommending that the speaker is in a positive state), negative state (i.e. reviews contains an implicit/explicit clue in the content recommending that the speaker is in a negative

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

state or neutral state (i.e. reviews does not contain an implicit/explicit clue of the speaker's emotional state).

3.3 Multilingual Text-To-Text-Transfer-Transformer-Based Pre-Processing

In the input preprocessing stage of the proposed method for intelligence sentiment analysis of morphological rich Tamil text, raw text data from both the source and target models are processed and converted into a pertinent arrangement for upcoming stages. In the proposed method the Multilingual Text-To-Text-Transfer-Transformer tokenizer and encoder are used in tokenizing and encoding the raw sample text data. Figure 2 shows the structure of Multilingual Text-To-Text-Transfer-Transformer-based Pre-processing model.

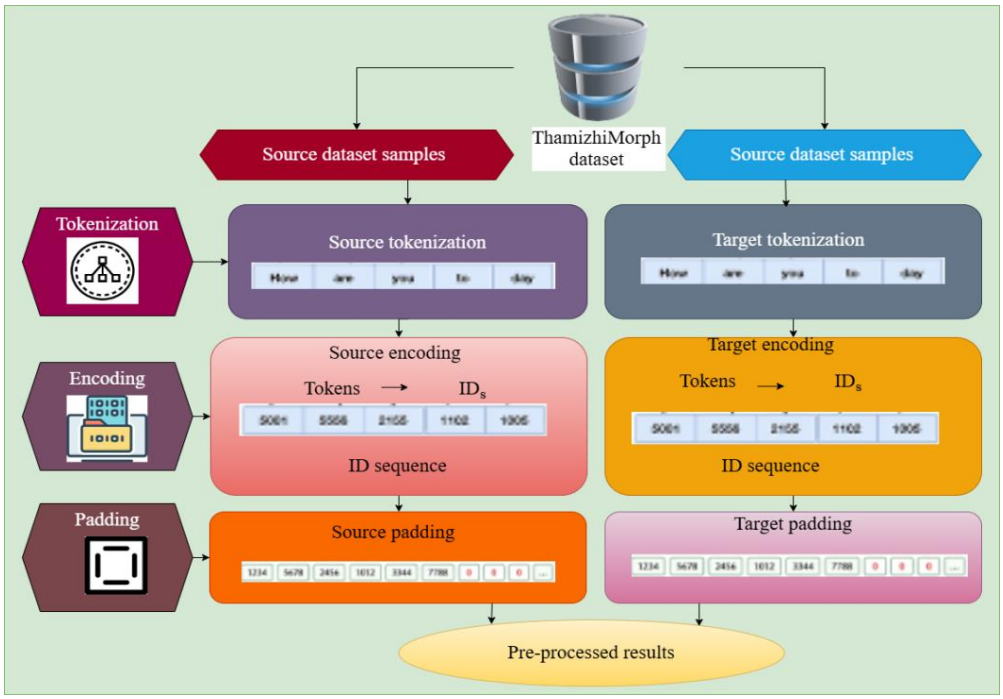


Figure 2: Structure of Multilingual Text-To-Text-Transfer-Transformer-Based Pre-Processing Model

As illustrated in the above figure the overall procedure of pre-processing is split into three parts. They are tokenization, encoding and padding. To start with the raw sample text data obtained from both ThamizhiMorph dataset (i.e. source) and Tamil Sentiment Analysis dataset (i.e. target) are subjected to pre-processing. Here both tokenization and encoding procedures are performed by applying Multilingual Text-To-Text-Transfer-Transformer function. Here, character-level and word-level tokenization is performed for both the source and target models to convert the sample text data into token sequences. Next the process of mapping based on pre-defined vocabulary is performed via encoding. Finally, to attain similar length padding is performed therefore generating pre-processed results.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

Tokenization is the procedure of splitting down the raw sample text data into sequences of tokens including words, sub-words or characters. The Multilingual Text-To-Text-Transfer-Transformer tokenizer by exploiting sub-word tokenization concept on the basis of the SentencePiece library incorporates the advantages of both character-level and word-level tokenization. Let 'P_S' (i.e. Tamizhi Morph dataset), 'P_T' (i.e. Tamil Sentiment Analysis dataset) denote the raw sample text data in the source and target languages. The Multilingual Text-To-Text-Transfer-Transformer tokenizer converts the sample text data into token sequences as given below.

$$P_S = \{Token_1^S, Token_2^S, \ldots\ldots, Token_m^S\} \quad (1)$$

$$P_T = \{Token_1^T, Token_2^T, \ldots\ldots, Token_n^T\} \quad (2)$$

From the above equations (1) and (2) the raw sample text data 'P' from the corresponding source 'S' and target 'T' are split into corresponding 'm' and 'n' samples respectively. Upon successful completion of the tokenization process the tokens required to be mapped to their respective IDs prior to being fed into the deep learning model as input. The Multilingual Text-To-Text-Transfer-Transformer encoder does this process of mapping on the basis of a pre-defined vocabulary. Let '$Enc_S(.)$' and '$Enc_T(.)$' denote the encoding functions for source and target models, then, the token sequences are mapped to ID sequences as given below.

$$ID_i^S = Enc_S\left(Token_i^S\right) \quad (3)$$

$$ID_j^T = Enc_T\left(Token_j^T\right) \quad (4)$$

From the above equations (3) and (4) the resulting sequences of ID for the corresponding source and target models are then mathematically defined as given below.

$$ID_S = \{ID_1^S, ID_2^S, \ldots, ID_n^S\} \quad (5)$$

$$ID_T = \{ID_1^T, ID_2^T, \ldots, ID_n^T\} \quad (6)$$

To speed up batch processing, the ID sequences required to be padded with the intent of possessing similar length. The process of padding is done by adding distinctive special padding tokens to the sequences till they outreach maximum length. Let '$l(S)$' and '$l(T)$' indicate the maximum lengths of the ID sequences in the source and target models respectively. Then the padded ID sequences are mathematically denoted as given below.

$$PadID_S = \left\{ID_1^S, ID_2^S, \ldots, ID_{l(S)}^S\right\} \quad (7)$$

$$PadID_T = \left\{ID_1^T, ID_2^T, \ldots, ID_{l(T)}^T\right\} \quad (8)$$

As given above the input pre-processing of the proposed method makes certain that the raw sample text data is tokenized and encoded appropriately, getting ahead pertinent for further processing. By using the Multilingual Text-To-Text-Transfer-Transformer the proposed method efficiently handles the tokenization and encoding of raw sample text data in both source and target models for performing sentiment analysis of morphological

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

rich Tamil text. The pseudo code representation of Multilingual Text-To-Text-Transfer-Transformer-based Pre-processing is given below.

Algorithm 1: Multilingual Text-To-Text-Transfer-Transformer-Based Pre-Processing

Input: Source Dataset 'DS_S', Target Dataset 'DS_T'
Output: Computationally-efficient pre-processed samples
1: **Initialize**Source Dataset Samples 'M', Target Dataset Samples 'N'
2: **Begin**
3: **For**each Source Dataset 'DS_S', Target Dataset 'DS_T', ource Dataset Samples 'M', Target Dataset Samples 'N'
//Multilingual Text-To-Text-Transfer-Transformer-based Tokenization
5: Perform tokenization separately for source and target according to (1) and (2)
//Multilingual Text-To-Text-Transfer-Transformer-based Encoding
6: Perform encoding separately for source and target according to (3) and (4)
7: Generate sequences of ID for corresponding source and target models according to (5) and (6)
//Padding
8: Generate padded ID sequences according to (7) and (8)
9: **Return** pre-processed sample text data '$PadID_S, PadID_T$'
10: **End for**
11: **End**

As given in the above algorithm initially the source and target dataset are subjected to pre-processing. Owing to the agglutinative nature the morphological rich Tamil words are formed by joining several morphemes. The Multilingual Text-To-Text-Transfer-Transformer employed in our work uses subword-level tokenization that efficiently splits down complicated Tamil words into meaningful root or morpheme, circumventing vocabulary sparsity. In addition by merging words or sandhi using Multilingual Text-To-Text-Transfer-Transformer separately via tokenization and encoding notably enhances tokenization quality, therefore reducing overall execution time.

3.4 Multi-Scale Attention Mechanism-Based Keyword Extraction

The Multi-Scale Attention mechanism to sentiment analysis necessitates producing a top-down approach. Owing to the agglutinative nature of morphological rich Tamil text language the mechanism processes pre-processed sample text data at quite a few granularities (i.e. characters, syllables, words). The mechanism allocates higher weights arbitrarily to syllables that convey strong sentimental value found to be pivotal in morphologically rich languages like Tamil. In the proposed method Multi-Scale Attention

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

Mechanism-based Keyword Extraction is employed. The keyword extraction model employed allows the proposed method to look at fine-grained keywords (i.e. suffices) and coarse-grained keywords (i.e. words or phrases) in a simultaneous fashion for morphologically rich languages like Tamil. Figure 3 shows the structure of Multi-Scale Attention Mechanism-based Keyword Extraction model.

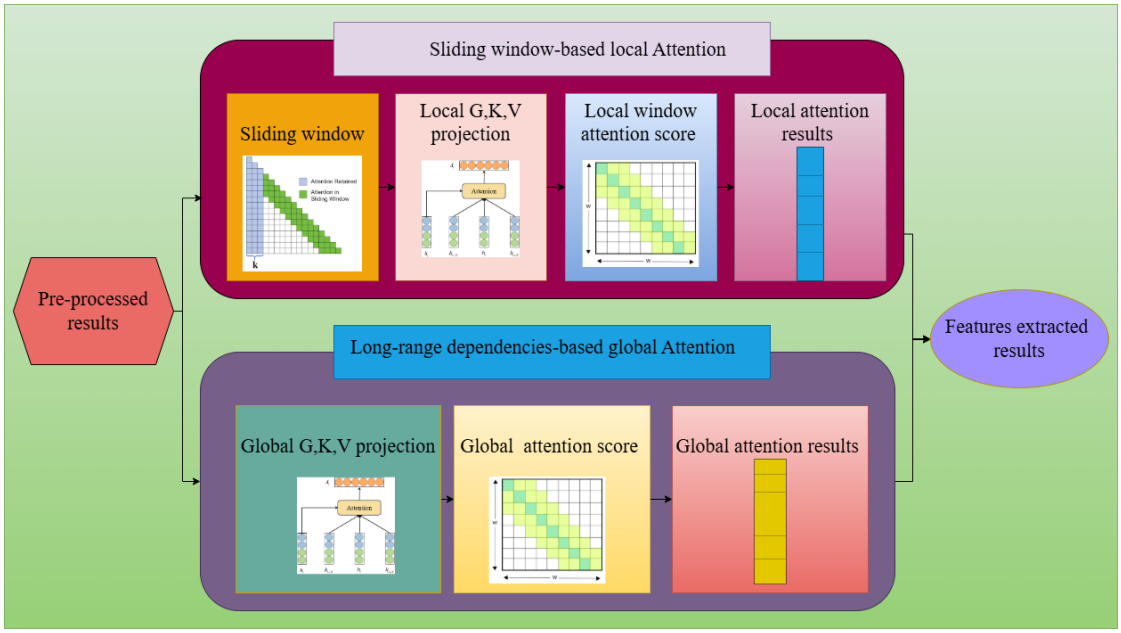


Figure 3: Structure of Multi-Scale Attention Mechanism-Based Keyword Extraction

As illustrated in the above figure the pre-processed sample text data are provided as input to the Multi-Scale Attention Mechanism-based Keyword Extraction model. It enables the keyword extraction process by concentrating on both detailed local features and high level global context information. Here a hierarchical model is employed where the first part encapsulates local word-level features whereas the second part employs global attention to model dependencies between these segments.

Also with the agglutinative structure better management is ensured by splitting down words in morphemes therefore permitting attention to concentrate specifically on the sentiment-handling part instead of their word.

3.4.1 Sliding Window-Based Local Attention

Local attention in our work utilizes a constrained receptive field with 8 attention heads to encapsulate adjacent contextual patterns within a sliding window. For each position in the sequence, local attention function is measured only over adjacent positions. Initially character level tokenization process is involved handle agglutinative word structures. This is mathematically defined as given below.

$$Q_L, K_L, V_L = PadID_S PadID_T W_Q^L, PadID_S PadID_T W_K^L, PadID_S PadID_T W_V^L \quad (9)$$

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

From the above equation (9), 'W_Q^L', 'W_K^L' and 'W_V^L' denotes learnable projection matrices for query, key and value respectively. With the obtained projection value results local window attention scores concentrating on local context via sliding window segmentation is given below.

$$LAtt_{ij} = \begin{cases} \frac{\exp[Q_L(i).K_L(j)]^T/\sqrt{Dim_k}}{\sum_{k\in W_i}\exp[Q_L(i).K_L(k)]^T/\sqrt{Dim_k}}, if\ j \in W_i \\ 0, Otherwise \end{cases} \quad (10)$$

Finally with the above obtained results (10) to encapsulate correlations between adjacent suffices and root words local attention mechanism is formulated as given below.

$$Att_L = LAtt.V_L \quad (11)$$

From the above equation (11) the value vector attention score results are aggregated from equation (10) therefore generating representations that facilitate immediate contextual patterns.

3.4.2 Sparse Attention Long-Range Dependencies-Based Global Attention

Global attention encapsulates long-range discourse patterns by measuring attention over the entire sequence. The reason is that specialized mechanisms are required to handle agglutinative suffices. This is mathematically defined as given below.

$$Q_G, K_G, V_G = PadID_S PadID_T W_Q^G, PadID_S PadID_T W_K^G, PadID_S PadID_T W_V^G \quad (12)$$

From the above equation (12), 'W_Q^G', 'W_K^G' and 'W_V^G' denotes learnable projection matrices for query, key and value respectively for global attention. With the attained projection value results global window attention scores concentrating on global context via long-range dependency mechanism is given below.

$$Att_G = softmax\left(\frac{Q_G K_G^T}{\sqrt{Dim_k}}\right) V_G \quad (13)$$

From the above equation (13) by applying the sparse attention function via '$softmax$' allow sophisticated global tokens to focus on the entire sequence. Finally the optimal fusion strategy between local and global features allowing the method to arbitrarily fine-tune based on the high morphological variance is as given below.

$$KE = \alpha * Att_L + (1-\alpha) * Att_G \quad (14)$$

From the above equation (14) results a learned weight fusion of local and global attention outputs with 'α' controlling the balance at each position and feature dimension by operating at several linguistic levels in a simultaneous fashion.

This in turn therefore improves the extraction accuracy of relevant features in morphological rich Tamil text. The pseudo code representation of Multi-Scale Attention Mechanism-based Keyword Extraction is given below.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

Algorithm 2: Multi-Scale Attention Mechanism-Based Keyword Extraction

Input: Source Dataset 'DS_S', Target Dataset 'DS_T'
Output: precise keyword extraction
1: **Initialize**pre-processed sample text data '$PadID_S, PadID_T$', sliding window length '$W = 7$', '$\alpha = 0.1$'
2: **Begin**
3: **For**each pre-processed sample text data '$PadID_S, PadID_T$'
//Sliding Window-based Local Attention
//Character-level Tokenization
4: Generate character-level local tokenized results according to (9)
//Sliding window-based Segmentation
5: Generate local window attention score results according to (10)
//Correlated local attention mechanism
6: Generate local attention results according to (11)
7: **End for**
8: **For** each pre-processed sample text data '$PadID_S, PadID_T$'
//Long-range Dependencies-based Global Attention
//Character-level Tokenization
9: Generate character-level global tokenized results according to (12)
//Long range dependency based global attention mechanism
10: Generate global attention results according to (13)
11: Evaluate feature extraction results according to (14)
12: **Return** features extracted 'FE'
13: **End for**
14: **End for**
15: **End**

As given in the above algorithm keyword extraction is performed using Multi-Scale Attention Mechanism via local and global aspects. Given Tamil's rich morphology nature, attention mechanisms are designed in such a manner so as to be morphology-aware. By using this function the sample words (i.e. puththakankal) extracts morphemes (i.e. kal) and lemmas (i.e. puththakam) with the intent of handling suffix variations that is said to be pivotal for ascertaining the true sentiment of words with enhanced precision. This keyword extraction model applies attention at numerous granularities (i.e. morpheme-

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

level and sentence-level) to comprehend contextual associations therefore reducing the overall false positive rate.

3.5 Sentiment Analysis Using Domain Adaptive Transfer Learning-Based Classifier

A transfer learning technique called domain adaptation for morphological rich Tamil text sentiment analysis is first deployed in such a manner that it enables a sentiment classifier trained on a labeled source domain to perform accurate on a target domain. By applying the transfer learning based domain adaptation with the morphological rich Tamil text formed by adding multiple suffices to a root work, the proposed sentiment analysis method understands that 'நன்றாக (good)' and 'நன்றாகயிருந்தது (it was good)' demonstrate similar sentiments irrespective of the context. With this transfer learning based domain adaptation it resolves the issue where sentiment words have different meanings across contexts, improving the overall sentiment analysis accuracy rate. In the proposed method adversarial training is applied with the objective of attaining improved sentiment analysis accuracy. The transfer learning domain adaptation-based classifier is trained with the objective of maximizing the classification accuracy. This is mathematically defined as given below.

$$\mathcal{L}_{Adv} = E_{P_S \sim DD_S(P)}\left[\log DC\left(f(P_S)\right)\right] + E_{P_T \sim DD_T(P)}\left[\log DC\left(f(P_T)\right)\right] \quad (15)$$

From the above equation (15) the adversarial training loss result '$\mathcal{L}_{Adv}$' is arrived at based on 'DD_S' and 'DD_T' denotes the domain classifier for source and target models respectively. Following which sentiment classifier result, a fully connected layer formed using softmax function '$Class(.)$' is included to predict the sentiment analysis results. The classifier is trained on the labeled target model by reducing the sentiment classification loss '$\mathcal{L}_{SC}$' as given below.

$$\mathcal{L}_{SC} = \sum_{i=1}^{N} Res_i \log\left(Class(PadID_T[i])\right) \quad (16)$$

From the above equation (16) 'N', represent the number of pre-processed samples in the Target Dataset 'DS_T' with 'Res_i' denoting the one-hot encoded label for the '$i-th$' pre-processed sample (i.e. positive, negative or neutral) and '$Class(PadID_T[i])$' representing the predicted probability distribution over sentiment classes for '$i-th$' sample respectively. The pseudo code representation of Sentiment Analysis using Domain Adaptive Transfer Learning-based Classifier is given below.

Algorithm 3: Domain Adaptive Transfer Learning-Based Classifier

Input: Source Dataset 'DS_S', Target Dataset 'DS_T'
Output: accurate sentiment classified results
1: **Initialize**pre-processed sample text data '$PadID_S, PadID_T$', features extracted 'FE' 2: **Begin** 3: **For each** pre-processed sample text data '$PadID_S, PadID_T$' and features extracted 'FE' **//Domain adaptation**

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

4: Generate adversarial training loss result according to (15)

//Sentimental analysis Classification

5: Generate sentiment classifier training loss results according to (16)

7: **End for**

8: **End**

As given in the above algorithm with the pre-processed samples and relevant keywords extracted provided as input are subjected to a classifier model with the objective of performing accurate sentiment analysis. As provided in the above algorithm first an adversarial training loss function is evaluated via domain classifier with the intent of reducing the mismatch between source and target data, therefore enhancing overall sentiment analysis classification accuracy. Also by applying Multilingual Text-To-Text-Transfer-Transformer pre-trained models as input are fine-tuned for performing sentiment analysis classification therefore outperforming conventional machine learning methods.

4. CASE STUDY AND INFERENCES

In this section case analysis of morphological rich Tamil text based sentiment analysis is performed by applying a novel method called, Multi-Scale Attention and Domain Adaptive Transfer Learning (MSA-DATL). Though roughly 1300 sentiments are present in the dataset to perform simulation ten samples were analyzed and among them three samples performance analysis are detailed.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

▶ Morphological Analysis

#	WORD (சொல்)	ROOT (வேர்)	SUFFIX (பின்னொட்டு)	POS TAG
1	ஒரு	ஒரு	-	NOUN
2	நல்ல	நல்ல	-	ADJ
3	திரைக்கதை	திரைக்கதை	-	NOUN
4	கொண்டது	கொண்டது	-	VERB
5	வடிவேல்	வடிவேல்	-	NOUN
6	நடிப்பு	நடிப்பு	-	NOUN
7	மிகவும்	மிகவு	ம்	NOUN
8	புதிதாக	புதிதாக	-	NOUN
9	இருந்தது	இருந்தது	இருந்தது	VERB
10	உதயநிதி	உதயநிதி	-	NOUN
11	நடிப்பு	நடிப்பு	-	NOUN
12	மிகவும்	மிகவு	ம்	NOUN
13	புது	பு	து	NOUN
14	மாதிரியாக	மாதிரியாக	-	NOUN
15	இருந்தது	இருந்தது	இருந்தது	VERB
16	ஒரு	ஒரு	-	NOUN
17	நல்ல	நல்ல	-	ADJ
18	படம்	பட	ம்	NOUN
19	இது	இ	து	NOUN

19	14	14	3
TOTAL TOKENS	UNIQUE TOKENS	NOUNS	VERBS

Review #2 — 12 tokens | 9 nouns | 3 verbs

▶ Original Text

பிச்சைக்காரன் 1 மிகவும் யோசிக்கும் வகையில் இருந்தது ஆனால் பிச்சைக்காரன் 2 அதை யோசிக்கும் போது கொஞ்சம் குறைவுதான்

▶ Cleaned Text

பிச்சைக்காரன் மிகவும் யோசிக்கும் வகையில் இருந்தது ஆனால் பிச்சைக்காரன் அதை யோசிக்கும் போது கொஞ்சம் குறைவுதான்

▶ Tokenization (12 tokens)

பிச்சைக்காரன் | மிகவும் | யோசிக்கும் | வகையில் | இருந்தது | ஆனால் | பிச்சைக்காரன் | அதை | யோசிக்கும் | போது | கொஞ்சம் | குறைவுதான்

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

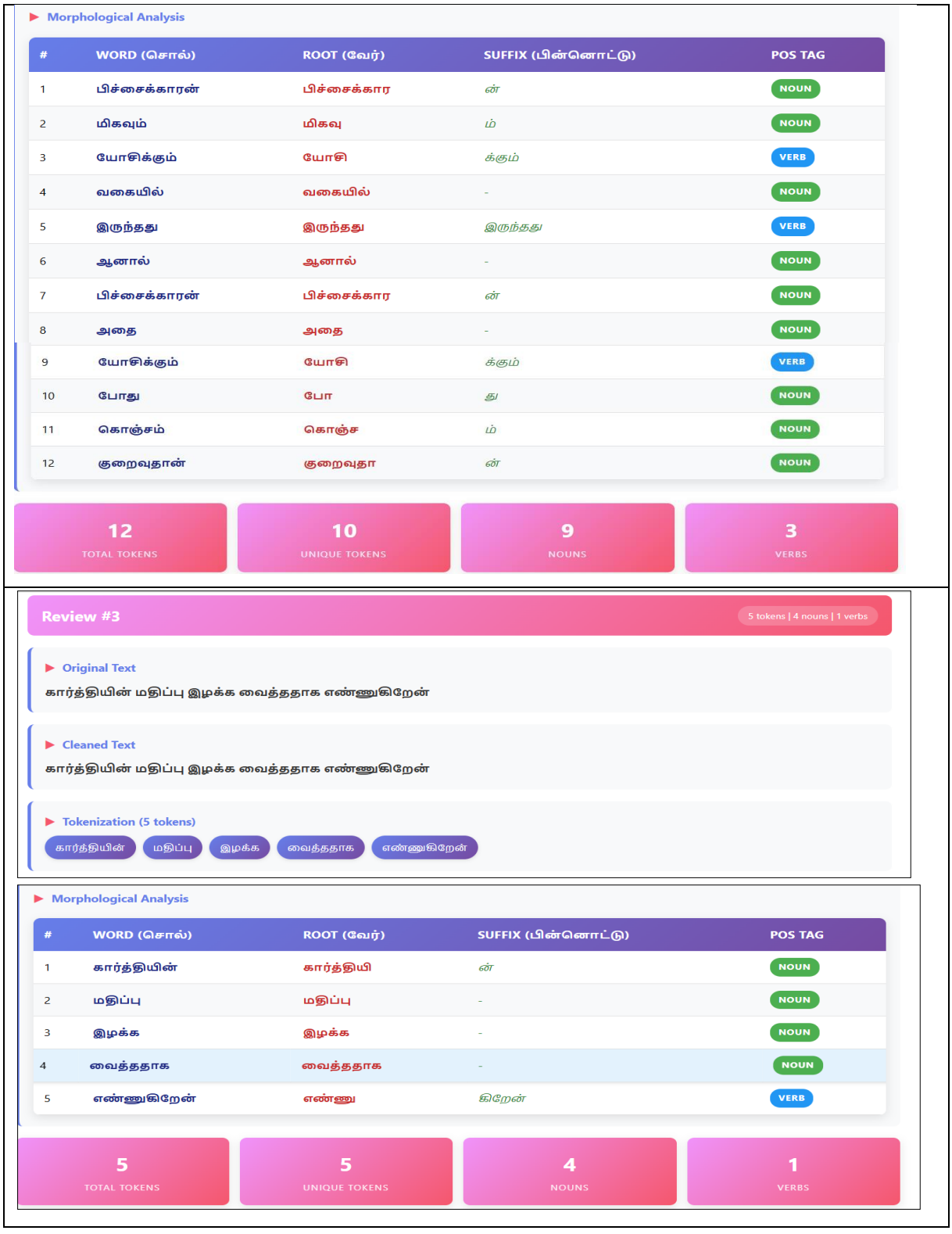
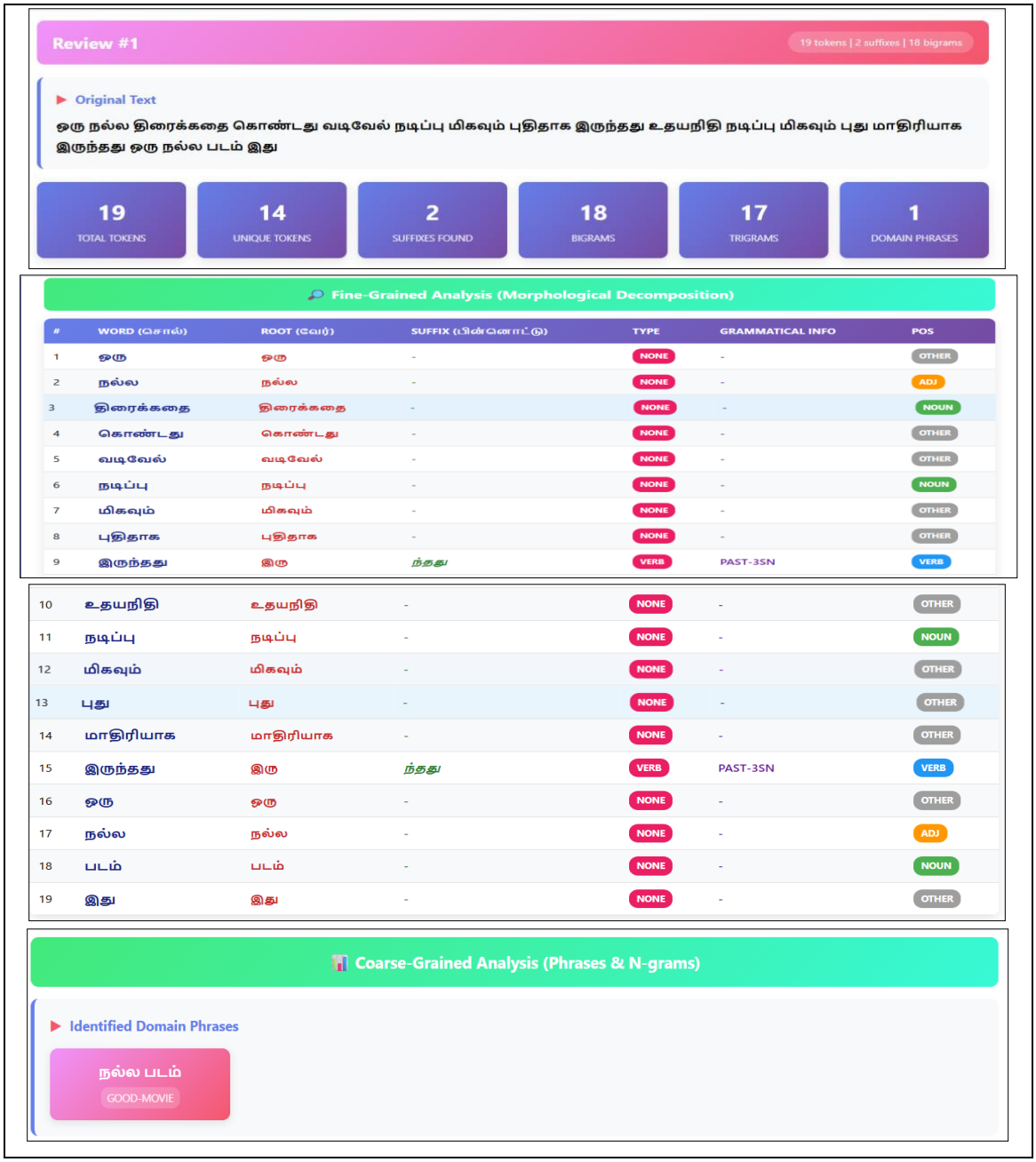


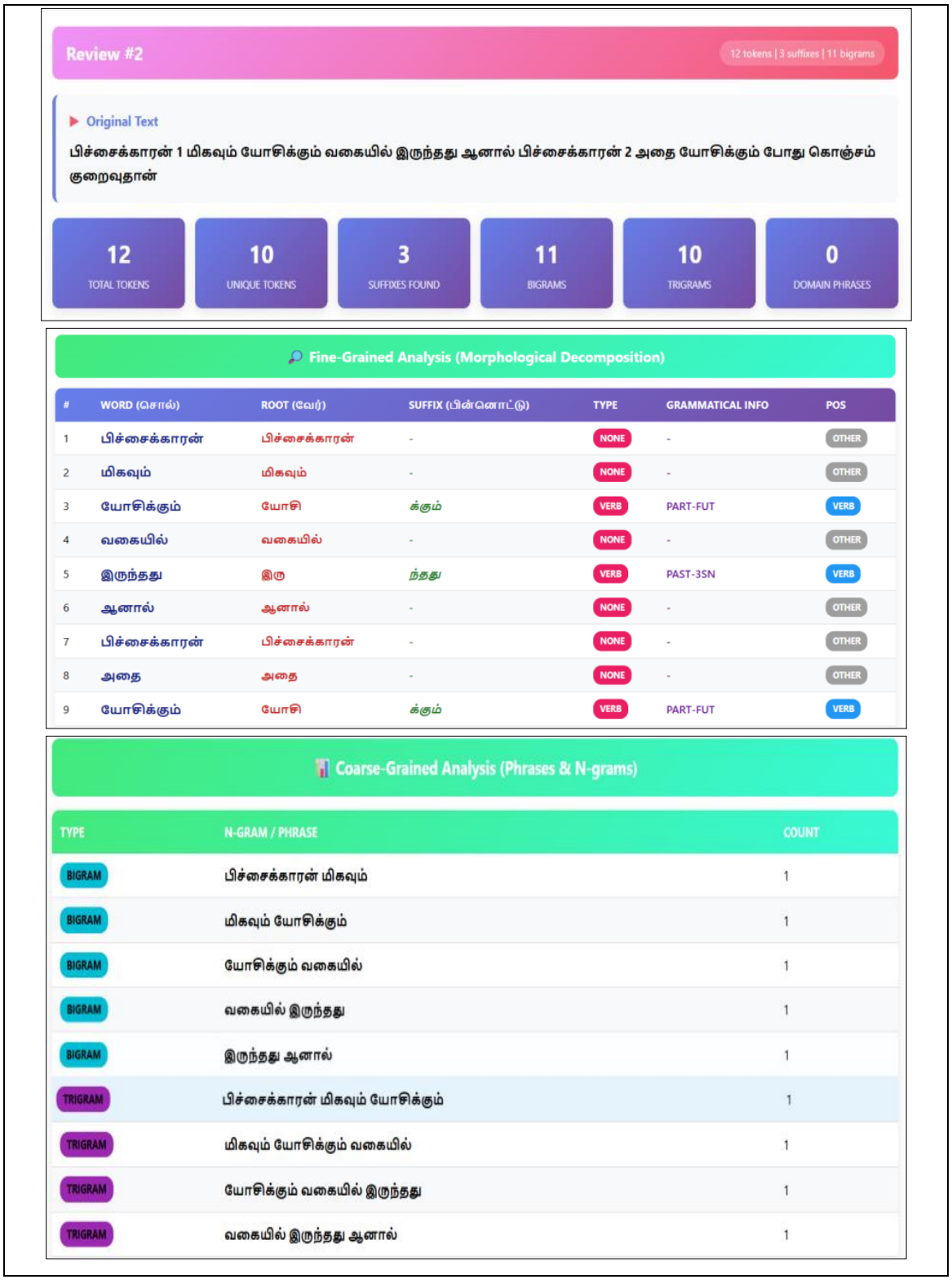
Figure 4: Samples with Pre-Processed Results

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

As shown in the above figure three samples were attained for performing pre-processing using Multilingual Text-To-Text-Transfer-Transformer model. First, samples from target were cleaned and tokenized as shown in the left hand side for further processing. Following which the right hand side results shows the morphological analysis performed for the source data. By performing tokenization along with encoding and padding for the target data and morphological analysis for the source data the execution time involved in the overall process is said to be reduced. Next, the keywords extracted results are provided as given below in figure 5.



Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339



Review #2

12 tokens | 3 suffixes | 11 bigrams

Original Text

பிச்சைக்காரன் 1 மிகவும் யோசிக்கும் வகையில் இருந்தது ஆனால் பிச்சைக்காரன் 2 அதை யோசிக்கும் போது கொஞ்சம் குறைவுதான்

12	10	3	11	10	0
TOTAL TOKENS	UNIQUE TOKENS	SUFFIXES FOUND	BIGRAMS	TRIGRAMS	DOMAIN PHRASES

Fine-Grained Analysis (Morphological Decomposition)

#	WORD (சொல்)	ROOT (வேர்)	SUFFIX (பின்னொட்டு)	TYPE	GRAMMATICAL INFO	POS
1	பிச்சைக்காரன்	பிச்சைக்காரன்	-	NONE	-	OTHER
2	மிகவும்	மிகவும்	-	NONE	-	OTHER
3	யோசிக்கும்	யோசி	க்கும்	VERB	PART-FUT	VERB
4	வகையில்	வகையில்	-	NONE	-	OTHER
5	இருந்தது	இரு	ந்தது	VERB	PAST-3SN	VERB
6	ஆனால்	ஆனால்	-	NONE	-	OTHER
7	பிச்சைக்காரன்	பிச்சைக்காரன்	-	NONE	-	OTHER
8	அதை	அதை	-	NONE	-	OTHER
9	யோசிக்கும்	யோசி	க்கும்	VERB	PART-FUT	VERB

Coarse-Grained Analysis (Phrases & N-grams)

TYPE	N-GRAM / PHRASE	COUNT
BIGRAM	பிச்சைக்காரன் மிகவும்	1
BIGRAM	மிகவும் யோசிக்கும்	1
BIGRAM	யோசிக்கும் வகையில்	1
BIGRAM	வகையில் இருந்தது	1
BIGRAM	இருந்தது ஆனால்	1
TRIGRAM	பிச்சைக்காரன் மிகவும் யோசிக்கும்	1
TRIGRAM	மிகவும் யோசிக்கும் வகையில்	1
TRIGRAM	யோசிக்கும் வகையில் இருந்தது	1
TRIGRAM	வகையில் இருந்தது ஆனால்	1

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

Figure 5: Keyword Extracted Results

As shown in the above results, with three samples provided as input, fine-grained and coarse-grained results are separately obtained. With the obtained results, the keywords are extracted by applying Multi-Scale Attention Mechanism-based Keyword Extraction algorithm. By applying this algorithm attention mechanism at numerous granularities (i.e. morpheme-level and sentence-level) comprehend contextual associations therefore reducing the overall false positive rate with improved accuracy.

Finally with the pre-processed sample results and keyword extracted results as input are applied for the classification process. Figure 6 shows the classified results of ten samples for morphological rich Tamil text-based sentiment analysis.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

#	Review (தமிழ்)	Sentiment Label	Confidence
1	ஒரு நல்ல திரைக்கதை கொண்டது வடிவேல் நடிப்பு மிகவும் புதிதாக இருந்தது உதயநிதி நடிப்பு மிகவும் புது மாதிரியாக இருந்தது ஒரு நல்ல படம் இது	Positive	1.00
2	பிச்சைக்காரன் 1 மிகவும் யோசிக்கும் வகையில் இருந்தது ஆனால் பிச்சைக்காரன் 2 அதை யோசிக்கும் போது கொஞ்சம் குறைவுதான்	Neutral	0.92
3	கார்த்தியின் மதிப்பு இழக்க வைத்ததாக எண்ணுகிறேன்	Negative	0.97
4	சிம்புவின் நடிப்பு மிகவும் அற்புதமான கதாபாத்திரம். இந்தப் படத்தில் பாடல் மிகவும் நன்றாக இருந்தது. சண்டை காட்சி	Positive	0.97
5	ஒரு நல்ல படம் இது சூரி நடிப்பு மிகவும் அற்புதமாக இருந்தது அவர் முதல் முறையாக நடித்த படம் இது அதில் சிறந்த நடிப்பை	Positive	0.99
6	மிகவும் சாதாரணமான கதைத்தளம். பெரிய முன்னணி கதாநாயகன் என்பதால் தேவையற்ற காட்சிகளும் இருந்தது. திரைக்கதை	Negative	0.94
7	விஜய் நடித்ததால் பரவாயில்லை என்று கூறலாம்.	Neutral	0.91
8	சுருக்கமாகச் சொன்னால்....அருமையான கதை சிறந்த இயக்கம் மற்றும் உயிரோட்டமான உச்சநிலை நம் குடும்பங்களிலும்	Positive	0.97
9	பழங்குடியினர் வாழ்க்கை வைத்து சில அரசியல் வாதிகள் எவ்வாறு பயன்படுத்தி வருகின்றனர் என்பதனை காட்டியுள்ளது	Neutral	0.89
10	திரைக்கதையில் திரைக்கதை இல்லை ஊக்கமளிக்கும் திரைப்படத் தொகுப்பில் உள்நோக்கம் இல்லை	Negative	0.96

Figure 6: Classified Results

As provided above Domain Adaptive Transfer Learning-based Classifier was applied to perform sentiment analysis results of positive, negative or neutral results. Here by applying the transfer learning process both adversarial training loss and sentiment classifier loss function results were applied separately for classification processing. This in turn improved overall classification accuracy results for sentiment analysis.

5. EXPERIMENTAL SETUP AND DISCUSSION

The results of the simulations used to validate the method called Multi-Scale Attention and Domain Adaptive Transfer Learning (MSA-DATL) for performing intelligent sentiment analysis for morphological rich Tamil text are described in detail in this section. The experiment uses the ThamizhiMorph dataset (i.e. source) from https://github.com/sarves/thamizhi-morph and MADTRAS dataset (i.e. target) from https://data.mendeley.com/datasets/p59cfx4vx6/2. The proposed MSA-DATL method has been implemented in Python language. The simulation results for the proposed MSA-DATLmethod and existing methods, CMSA-mBERT [1] and adapter-based mDeBERT [2] are detailed below in terms of different performance parameters, execution time, precision, recall, accuracy and false positive rate. To ensure fair comparison between proposed MSA-DATLand existing methods, [1] and [2] similar samples is applied for an average of 10 different simulation runs.

5.1 Performance analysis of Execution time

Evaluation of execution time when conducting sentiment analysis on morphological rich Tamil text is important because increasing computational complexity straightforwardly

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

influences the overall performance. Due to the agglutinative nature of morphological rich Tamil text intensive pre-processing is extensively required for sentiment analysis.

This is due to the inclusion of multiple suffices being included in the root word. While performing this pre-processing a significant amount of time is said to be consumed and this is referred as execution time. The execution time is measured as given below.

$$ET = \sum_{i=1}^{N} S_i * Time(SA) \quad (17)$$

From the above equation (17) the execution time 'ET' is measured based on the samples considered for simulation 'S_i' and the actual time consumed in performing sentiment analysis '$Time(SA)$'. This is measured in terms of seconds (sec).

Table 1 lists the results of execution time using the proposed MSA-DATL and two existing methods CMSA-mBERT [1] and adapter-based mDeBERT [2] respectively.

Table 1: Results of Execution Time

Samples	Execution time (sec)		
	MSA-DATL	CMSA-mBERT [1]	adapter-based mDeBERT [2]
130	1.95	2.21	2.34
260	2.25	2.75	2.9
390	2.55	3.05	3.2
520	2.85	3.25	3.4
650	3.15	3.55	3.7
780	3.35	3.75	3.9
910	3.65	4.05	4.2
1040	3.95	4.35	4.5
1170	4.15	4.55	4.7
1300	4.35	4.75	4.9

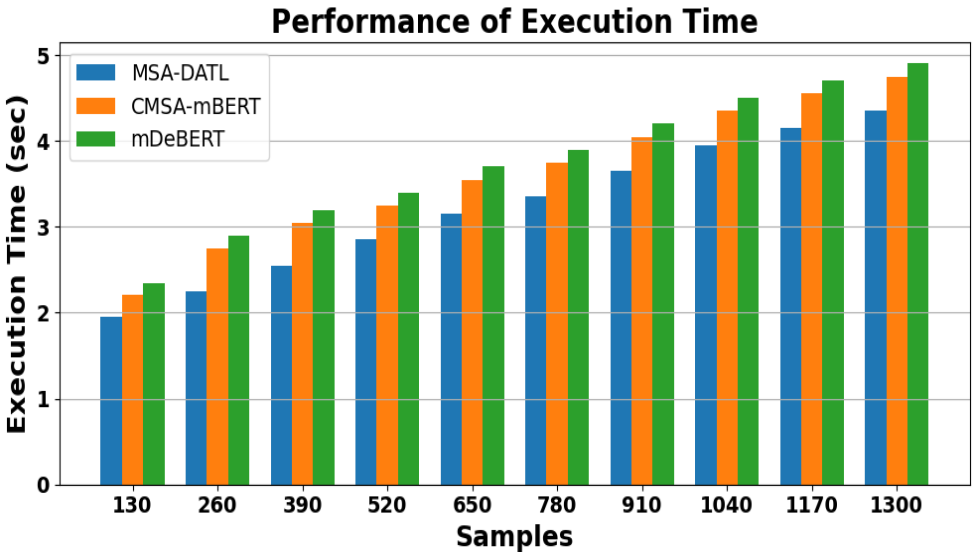


Figure 7: Samples versus Execution Time

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

Figure 7 illustrates the graphical representation of execution time using the proposed MSA-DATL and two existing methods, CMSA-mBERT [1] and adapter-based mDeBERT [2]. To perform fair comparison between proposed MSA-DATL and two existing methods [1], [2], similar set of 1300 reviews are provided as input and accordingly the execution time is measured. The blue color bar denotes the execution time results of MSA-DATL method whereas the orange and green color bar represents the results of existing methods [1] and [2] respectively. From the above figure it is inferred that an increase in the sample size causes a subsequent increase in the execution time also. Simulation performed with 130 samples observed 1.95 sec of execution time when applied with proposed MSA-DATL method whereas 2.21 sec and 2.34 sec using [1] and [2]. From this result the execution time consumed in analyzing sentiments using proposed MSA-DATL method was observed to be better upon comparison to [1] and [2]. Conventional natural language processing for agglutinative languages like Tamil necessitates considerable morphological analysis to handle inflectional variations. By applying the Multilingual Text-To-Text-Transfer-Transformer-based Pre-processing model frequently bypasses this procedure due to the reason that the tokenizer encapsulates subword-level information, notably reducing pre-processing and the overall execution time involved in sentiment analysis using MSA-DATL method by 13% compared to [1] and 18% compared to [2] respectively.

5.2 Performance Analysis of Precision, Recall and Accuracy

Precision concentrates on the accuracy of positive predictions (i.e. positive sentiment classified as positive) and minimizing false positives (i.e. falsely making a negative sentiment as positive sentiment). Recall on the other hand evaluates the methods potentiality to identify all actual positive samples (i.e. positive sentiment) crucial for minimizing false negatives. Finally accuracy evaluates the overall ratio of correct prediction (i.e. positive reviews as positive, negative reviews as negative and neutral reviews as neutral) out of the total sample reviews involved. Precision, recall and accuracy is evaluated as given below.

$$Pre = \frac{TP}{TP+FP} \tag{18}$$

$$Rec = \frac{TP}{TP+FN} \tag{19}$$

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \tag{20}$$

From the above equations (18), (19) and (20), precision 'Pre', recall 'Rec' and accuracy 'Acc' results are arrived at based on the true positive rate 'TP' (i.e. positive reviews classified as positive), false positive rate 'FP' (i.e. positive reviews classified as negative), true negative rate 'TN' (i.e. negative reviews classified as negative) and false negative rate 'FN' (i.e. negative reviews classified as positive) respectively. Table 2 lists the results of precision, recall and accuracy using the proposed MSA-DATL and two existing methods CMSA-mBERT [1] and adapter-based mDeBERT [2] respectively.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

Table 2: Results of Precision, Recall and Accuracy

Samples	Precision			Recall			Accuracy		
	MSA-DATL	CMSA-mBERT [1]	adapter-based mDeBERT [2]	MSA-DATL	CMSA-mBERT [1]	adapter-based mDeBERT [2]	MSA-DATL	CMSA-mBERT [1]	adapter-based mDeBERT [2]
130	0.9	0.85	0.8	0.94	0.91	0.86	0.88	0.82	0.75
260	0.88	0.78	0.73	0.93	0.81	0.77	0.87	0.77	0.72
390	0.85	0.75	0.7	0.9	0.78	0.74	0.85	0.75	0.7
520	0.82	0.72	0.67	0.87	0.75	0.71	0.83	0.73	0.68
650	0.8	0.7	0.65	0.85	0.73	0.69	0.81	0.71	0.66
780	0.78	0.68	0.63	0.83	0.71	0.67	0.79	0.69	0.64
910	0.75	0.65	0.6	0.86	0.74	0.7	0.81	0.71	0.66
1040	0.79	0.69	0.64	0.89	0.77	0.73	0.83	0.73	0.68
1170	0.82	0.72	0.67	0.93	0.85	0.81	0.85	0.75	0.7
1300	0.85	0.75	0.7	0.95	0.88	0.84	0.87	0.77	0.72

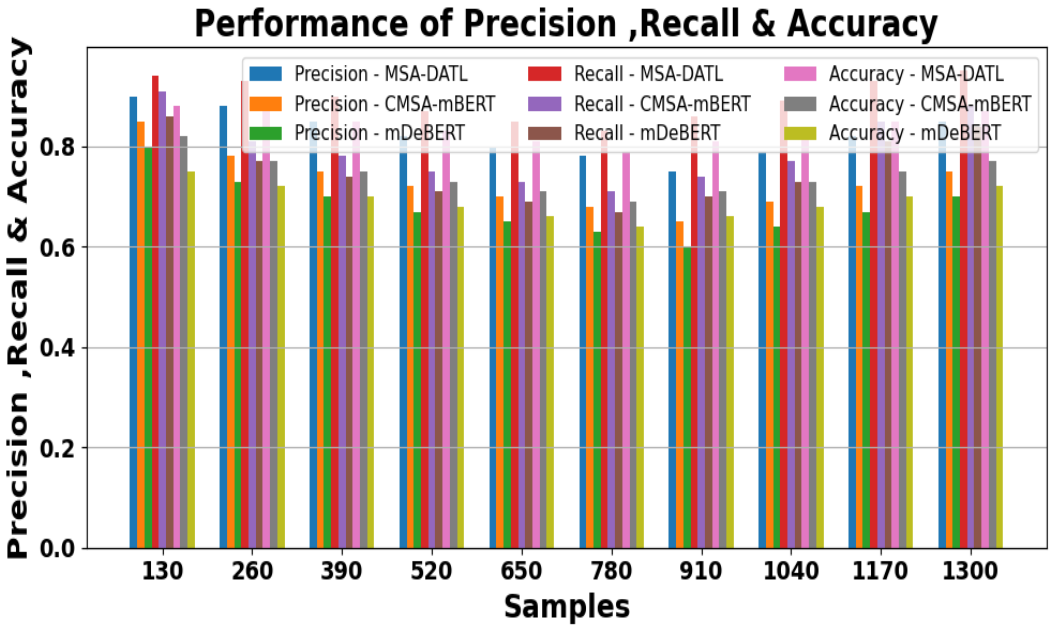


Figure 8: Samples versus Precision, Recall and Accuracy

Figure 8 illustrates the graphical representations of precision, recall and accuracy using the proposed MSA-DATL and two existing methods CMSA-mBERT [1] and adapter-based mDeBERT [2]. An overall ten simulation runs were performed to analyze the performance metrics precision, recall and accuracy with respect to samples ranging between 130 and 1300. From the above figure for the first seven iterations a dip in the resultant precision, recall and accuracy values were observed for all the three methods. However for the rest of three iterations the results were found to be increased when applied with the three methods, therefore increasing the samples does not affect the three performance metric results. Also comparative analysis show improvement in terms of three performance metrics precision, recall and accuracy when applied with proposed MSA-DATL method. The multi-scale attention mechanism applied as keyword extractor

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

permits the system to capture both fine-grained features (i.e. local details) and coarse-grained features (i.e. global context). By incorporating these fine-grained and coarse-grained features the proposed method better understands the context.

This in turn identifies the most relevant keywords and therefore ensures precise sentiment analyzed results using the proposed MSA-DATL method by 12% compared to [1] and 18% compared to [2]. Next by using attention based sliding windows of different sizes (multi-scale), the multi-scale attention mechanism encapsulates both short and long keywords in an efficient manner. This in turn ensures that relevant keywords are not missed, thereby increasing recall of the proposed MSA-DATL method by 12% compared to [1] and 16% compared to [2].Finally by applying the Domain Adaptive Transfer Learning-based Classification algorithm initially an adversarial training loss function is measured via domain classifier that in turn reduce the mismatch between source and target data, therefore improving overall sentiment analysis classification accuracy.

Moreover by using Multilingual Text-To-Text-Transfer-Transformer pre-trained models as input are fine-tuned for performing sentiment analysis classification therefore outperforming conventional machine learning methods of accuracy rate using proposed MSA-DATL method by 11% compared to [1] and 18% compared to [2].

5.3 Performance Analysis of False Positive Rate

Finally in this section false positive rate involved in sentiment analysis is detailed. It refers to the probability that a negative case is incorrectly identified as positive or neutral. The false positive rate is measured as given below.

$$FPR = \frac{FP}{FP+TN} \tag{21}$$

From the above equation (21) false positive rate 'FPR' is measured using the false positive value 'FP' and true negative value 'TN' respectively.

Table 3 lists the results of false positive rate using the proposed MSA-DATL and two existing methods CMSA-mBERT [1] and adapter-based mDeBERT [2] respectively.

Table 3: Results of False Positive Rate

Samples	False positive rate		
	MSA-DATL	CMSA-mBERT [1]	adapter-based mDeBERT [2]
130	0.28	0.4	0.52
260	0.35	0.47	0.55
390	0.38	0.5	0.56
520	0.41	0.53	0.57
650	0.45	0.57	0.58
780	0.52	0.64	0.65
910	0.45	0.67	0.69
1040	0.4	0.6	0.62
1170	0.35	0.5	0.52
1300	0.31	0.56	0.57

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

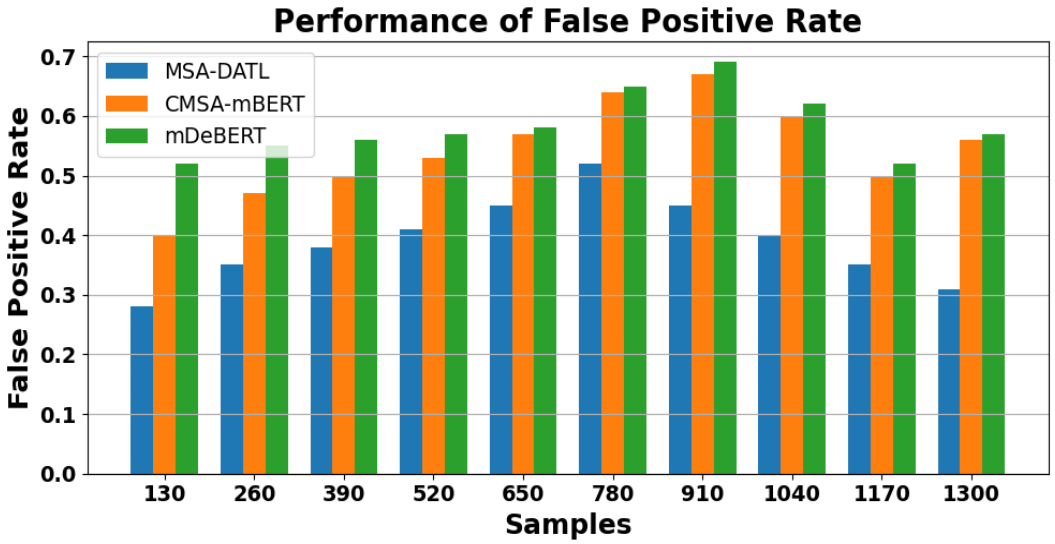


Figure 9: Samples versus False Positive Rate

Figure 9illustrates the graphical representation of false positive rate analyzed using three methods, proposed MSA-DATL, CMSA-mBERT [1] and adapter-based mDeBERT [2] for samples ranging between 130 and 1300. Here the blue color graph denotes the false positive results when applied with MSA-DATL method whereas the orange and green color graph represents the false positive result of [1] and [2] respectively. Mix of increase and decrease in false positive rate results are obtained using all the three methods. Simulations performed with respect to ten iterations observed minimal false positive rate when applied with proposed MSA-DATL method upon comparison to [1] and [2]. The reason was conventional attention mechanisms frequently introduce noise by concentrating on irrelevant and distant words. On the other hand by applying multi-scale local attention mechanisms, using different sliding window sizes suppress this noise (i.e. irrelevant words) notably reducing the false positives of proposed MSA-DATL method by 41% compared to [1] and 52% compared to [2]

6. CONCLUSION

This paper presented a novel Multi-Scale Attention and Domain Adaptive Transfer Learning (MSA-DATL) method for performing intelligent sentiment analysis using morphological rich Tamil text. The proposed MSA-DATL method performs the tasks of pre-processing, keyword extraction and classification for performing sentiment analysis. With the collected source and target data, Multilingual Text-To-Text-Transfer-Transformer-based Pre-processingwas applied to convert into a pertinent arrangement for upcoming stages and minimize overall execution time. Then, the most prominent keywords were learnt by applyingMulti-Scale Attention Mechanism-based Keyword Extraction model. Finally the actual classification process using pre-trained source along with the pre-processed and relevant keywords was performed by applying Domain Adaptive Transfer Learning-based Classifier with enhanced precision and accuracy. The

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

algorithm is outlined to optimize five performance metrics, improving precision, recall, accuracy, execution time and false positive rate. The proposed MSA-DATL method is assessed on different numbers of simulation runs. The performance of the proposed MSA-DATL method is compared to the popular sentiment analysis methods such as CMSA-mBERT and adapter-based mDeBERT using different performance metrics. Experimental results prove that the proposed MSA-DATL method efficiently analyzes and detects positive, negative and neutral sentiment for morphological rich Tamil text in a significant manner.

Acknowledgement

The authors gratefully acknowledge the financial support provided by the Chief Minister's Research Grant (CMRG), Department of Higher Education, Government of Tamil Nadu, under Proceedings No. CMRG/3005/H3/2024/050, dated 16 July 2025, for Project ID CMRG2400678, titled "An Intelligent Sentimental Analysis Morphological Rich Tamil Ttext". The authors also thank Anna University, Chennai, and the University College of Engineering, BIT Campus, Tiruchirappalli, for providing the necessary facilities and institutional support to carry out this research.

References

1) Muhammad Kashif Nazir, Haseeb Ahmad, CM Nadeem Faisal, Muhammad Asif Habib, "Leveraging Multilingual Transformer for Multiclass Sentiment Analysis in Code-Mixed Data of Low-Resource Languages", IEEE Access, Vol. 13, Jan 2025 [CMSA-mBERT]

2) Malliga Subramanian, Sathishkumar Veerappampalayam Easwaramoorthy, Nandhini Subbarayan, Usha Moorthy, "Empowering sentiment analysis in social media: a comprehensive approach to enhance the classification of abusive Tamil comments using transformer models", Journal of Big Data, Springer Nature Link, Vol. 12, Aug 2025 [adapter-based mDeBERT]

3) Biplov Paneru, Bipul Thapa, Bishwash Paneru, "Sentiment analysis of movie reviews: A flask application using CNN with RoBERTa embeddings", Systems and Soft Computing, Elsevier, Vol. 7, Dec 2025

4) Pradeep Kumarroy, "Deep Ensemble Network for Sentiment Analysis in Bi-lingual Low-resource Languages", ACM Transactions on Asian and Low-Resource Language", Vol. 23, Jan 2024

5) Anton Vijeevaraj, Sinthusha, Eugene Y. A. Charles, Ruvan Weerasinghe, "Machine Reading Comprehension for the Tamil Language With Translated SQuAD", IEEE Access, Vol. 13, Jan 2025

6) Srinithi Jayachandran, Yakesh Selvakumar, S. Anubha Pearline, "Multi-Attention Based Convolutional Neural Network for Tamil Handwritten Character Recognition", IEEE Access, Vol. 13, Oct 2025

7) Vinotheni C., S. Lakshmana Pandian, "Fast Recurrent Neural Network with Bi-LSTM for Handwritten Tamil Text Segmentation in NLP", ACM Transactions on Asian and Low-Resource Language, Vol. 23, May 2024

8) Yonatan Mamani-Coaquira, Edwin Villanueva, "A Review on Text Sentiment Analysis with Machine Learning and Deep Learning Techniques", IEEE Access, Vol. 12, Dec 2024

9) Md. Shofiqul Islam, Muhammad Nomani Kabir, Ngahzaifa Ab Ghani, Kamal Zuhairi Zamli, Nor Saradatul Akmar Zulkifli, Md. Mustafizur Rahman, Mohammad Ali Moni, "Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach", Artificial Intelligence Review, Springer Nature Link, Vol. 57, Mar 2024

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

10) Namitha Shambhu Bhat, Kuldeep Sambrekar, Venkatesh Bhandage, Prashant Y. Niranjan, “An exhaustive review of the recent advancements in sentiment analysis on Dravidian languages”, Discover Applied Sciences, Springer Nature Link, Vol. 8, Feb 2026

11) Kannaiah Chattu, K. Adi Narayana Reddy, Sai babu veesam, Pardha Saradhi Chirumamilla, Vunnava Dinesh Babu, Krishna Prakash, Mohammad Rashed Iqbal Faruque, Shonak Bansal, K. S. Al-mugren, “Sentiment classification for telugu using transformed based approaches on a multi-domain dataset”, Scientific Reports, Nature Portfolio, Oct 2025

12) Muhammad Shoibamin Ahmada, Alzahrani, Summaira Jabeen, Abdol Raheem Khader, Ali Ahmed, Sadaf Hussain, “Dual-Branch Neural Network for Bridging Semantic Gap in Harmful Meme Detection”, IEEE Access, Vol. 13, Jul 2025

13) Premjith B., Soman K. P, “Deep Learning Approach for the Morphological Synthesis in Malayalam and Tamil at the Character Level”, ACM Transactions on Asian and Low-Resource Language, Vol. 20, Aug 2021

14) Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, John P. McCrae, “DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text”, Language Resources and Evaluation, Springer Nature Link, Vol. 56, Feb 2022

15) Musica Supriya, Dinesh Acharya Udupi, Ashalatha Nayak, ArjunaSrirangapatna Raghavendra, “Developing a Hybrid Morphological Analyzer for Low-Resource Languages”, Applied Sciences, MDPI, Oct 2025

16) Yusuf Aliyu, Aliza Sarlan, Kamaluddeen Usman Danyaro, Abdullahi Sani B. A., Rahman, Mujaheed Abdullahi, “Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources”, IEEE Access, Vol. 12, May 2024

17) Thivaharan. S, Srivatsun.G, “Morphological and Contextual Meaning Analysis in Colloquial Tamil Using Yarowsky Influenced Modified Apriori Algorithm”, International Journal of Intelligent Systems and Applications in Engineering, Jun 2024

18) Toqira Rana, Kiran Shahzadi, Tauseef Rana, Ahsanarshad, Mohammad Tubishat, “AnUnsupervised Approach for Sentiment Analysis on Social Media Short Text Classification in Roman Urdu Sentiment analysis on short text classification in RomanUrdu”, ACM Transactions on Asian and Low-Resource Language, Vol. 21, Nov 2021

19) Roop Ranjan, Dilkeshwar Pandey, Ashok Kumar Rai, Pawan Singh, Ankit Vidyarthi, Puranam Revanth Kumar, SachiNandanMohanty, Deepak Gupta, “AManifold-Level Hybrid Deep Learning Approach for Sentiment Classification Using an Autoregressive Model”, Applied Sciences, MDPI, Oct 2023

20) Tajinder Kaur, Gagninder Kaur, “Sentiment Analysis for Low Resource Language Using Machine Learning and Deep Learning”, International Conference on Artificial Intelligence, Machine Learning and Cybersecurity, May 2025

21) Vimala Balakrishnan, Vithyatheri Govindan, Kumanan N. Govaichelvan, “Tamil Offensive Language Detection: Supervised versus Unsupervised Learning Approaches”, ACM Transactions on Asian and Low-Resource Language, Vol. 22, Mar 2023

22) Rayees Ahamad, Kamta Nath Mishra, “Exploring sentiment analysis in handwritten and E-text documents using advanced machine learning techniques: a novel approach”, Journal of Big Data, Springer, Oct 2025

23) Jothi Prakash, Arul Antranvijay S, “Cross-lingual Sentiment Analysis of Tamil Language Using a Multi-stage Deep Learning Architecture”, ACM Transactions on Asian and Low-Resource Language, Vol. 22, Dec 2025

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 07:2026
DOI: 10.5281/zenodo.21719339

24) Nikhil Sanjay Suryawanshi, “Sentiment analysis with machine learning and deep learning: A survey of techniques and applications”, International Journal of Science and Research Archive, Jun 2024

25) Nor Zakiah Lamin, Azwa Abdul Aziz, “Cross-Lingual Sentiment Analysis in Low-Resource Languages: A Recent Review on Tasks, Methods and Challenges”, International Journal of Advanced Computer Science and Applications, Vol. 16, Nov 2025

26) Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, Bader Alshemaimri, “BERT applications in natural language processing: a review”, Artificial Intelligence Review, Springer Nature Link, Vol. 58, Mar 2025

27) Kogilavani Shanmugavadivel, V. E. Sathishkumar, Sandhiya Raja, T. Bheema Lingaiah, S. Neelakandan, Malliga Subramanian, “Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data”, Scientific Reports, Jul 2022

28) Namitha Shambhu Bhat, Kuldeep Sambrekar, Venkatesh Bhandage, Prashant Y. Niranjan, “An exhaustive review of the recent advancements in sentiment analysis on Dravidian languages”, Discover Applied Sciences, Springer Nature Link, Vol. 8, Feb 2026

29) https://github.com/sarves/thamizhi-morph

30) https://data.mendeley.com/datasets/p59cfx4vx6/2