International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.2 (2025) pp. 2338-2351
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Enhancing Cross Language for English-Telugu pairs through the Modified Transformer Model based Neural Machine Translation

Vaishnavi Sadula[1*], D. Ramesh[2]

[1]Research Scholar, JNTUH, Hyderabad, Telangana, 500085, India
* **Corresponding Author Email:** sadulavaishnavi041@gmail.com - **ORCID:** 0000-0002-5755-101X

[2]Principal, JNTUH University College of Engineering Palair, Khammam, Telangana, 507001, India
Email: dantamr@yahoo.com -**ORCID:** 0000-0002-0751-3824

Article Info:

DOI: 10.22399/ijcesen.1740
Received : 04 February 2025
Accepted : 05 April 2025

Keywords :

Cross Language Translation,
Transformer Networks,
Neural machine translation,
Feed Forward networks,
Multi-scale attention maps.

Abstract:

Cross-Language Translation (CLT) refers to conventional automated systems that generate translations between natural languages without human involvement. As the most of the resources are mostly available in English, multi-lingual translation is badly required for the penetration of essence of the education to the deep roots of society. Neural machine translation (NMT) is one such intelligent technique which usually deployed for an efficient translation process from one source of language to another language. But these NMT techniques substantially requires the large corpus of data to achieve the improved translation process. This bottleneck makes the NMT to apply for the mid-resource language compared to its dominant English counterparts. Although some languages benefit from established NMT systems, creating one for low-resource languages is a challenge due to their intricate morphology and lack of non-parallel data. To overcome this aforementioned problem, this research article proposes the modified transformer architecture for NMT to improve the translation efficiency of the NMT. The proposed NMT framework, consist of Encoder-Decoder architecture which consist of enhanced version of transformer architecture with the multiple fast feed forward networks and multi-headed soft attention networks. The designed architecture extracts word patterns from a parallel corpus during training, forming an English–Telugu vocabulary via Kaggle, and its effectiveness is evaluated using measures like Bilingual Evaluation Understudy (BLEU), character-level F-score (chrF) and Word Error Rate (WER). To prove the excellence of the proposed model, extensive comparison between the proposed and existing architectures is compared and its performance metrics are analysed. Outcomes depict that the proposed architecture has shown the improvised NMT by achieving the BLEU as 0.89 and low WER when compared to the existing models. These experimental results promise the strong hold for further experimentation with the multi-lingual based NMT process.

1. Introduction

India, one of the greatest sub-continents that has deep –rooted civilization and dual-pole language families such as Dravidian and Indo-Aryan [1,2]. Though there are large number of speakers in most Indian languages but still handicapped for language processing operation since the most of the resources are in English [3,4]. Cross-language translation (CLT) is a widely recognized approach and refers to automated systems designed to translate content from one natural language into others [5,6]. With the increasing prevalence of computational applications and the expanding reach of the internet to a diverse, multilingual global audience, the field of CLT is advancing swiftly, driven by ongoing research and innovation [7]. These techniques play an important role in various application in tourism ecosystem [8], global e-commerce [9], educational system [10], language training model [11] and social platforms [12]. Observably, to accurately understand and translate cross –language sentences, Neural Machine Translation (NMT) is widely used that incorporates the back propagation evoked neural networks [13,14]. NMT systems rely heavily on the availability of high-quality corpora to achieve

accurate and reliable translations [15,16]. As a result, NMT performs optimally for languages that are resource-rich, typically those with hundreds of thousands or even millions of parallel sentence pairs. In the Indian context, although Hindi boasts the most extensive collection of parallel datasets, it is still categorized as a medium-resource language compared to many European languages. Other Indic languages, such as Bengali, Telugu, and Tamil, have even scarcer parallel datasets, positioning them as low-resource languages. These low-resource languages are significantly underrepresented in digital content, creating barriers for speakers to leverage advanced technologies, including efficient NMT systems. To mitigate these challenges, researchers have proposed diverse strategies, such as multilingual NMT [17], leveraging transfer learning [18], utilizing related languages, integrating multimodal NMT [19], employing data augmentation techniques [20], curating filtered pseudo-parallel corpora [21] and adopting meta-learning approaches [22].

1.1 English –Telugu NMT techniques

With the rise of deep learning architecture and transformer networks, NMT has changed its own application dimension to handle the resource constraint Indian languages. Telugu one of the rich cultured languages dated back in 1400 years ago [23] and finds it rich heritage in the costal part of Indian sub-continent. The development of machine translation for English–Telugu language pairs plays a pivotal role in showcasing the richness of the Telugu language and its cultural heritage to those unfamiliar with it. Beyond serving as a communication tool, every language embodies the unique essence of its region. From an academic standpoint, Telugu is a part of the curriculum in schools across Telangana and Andhra Pradesh. The significance of Telugu continues to expand over time, yet the language remains underexplored in terms of technological advancements. Potential applications for an English–Telugu neural machine translation (NMT) system include translating governance documents from English to Telugu, enhancing healthcare services through text-to-speech systems with robust MT frameworks, and fostering educational opportunities where students can either learn in their native language or gain deeper understanding through it.

1.2 Motivation and Contribution of the Research

The potential deployment of Monolingual (English to Telugu NMT) systems in tourism, health care systems and education system has gained the more importance to improve the economic development of the Indian state. Specially, this type of NMT systems are used in the education sector to improve the student ‘s calibre in understanding the education in better way. To achieve the higher performance of translation, NMT systems requires the large amounts of annotated data which is not possible with low resource language. Moreover, quality of the translation will also decrease as the amount of data decreases. Motivated by the aforementioned problem, this research article proposes the Enhanced Transformer Model with the Multi-headed attention with multi-fast feed forward networks. To best of our knowledge, the proposed framework is first of its kind in designing the NMT techniques for an efficient English-Telugu Translation process. The main contribution of the paper is as follows

1. The paper proposes the Novel Encoder-Decoder Architecture based on Enhanced version of transformer model to design the NMT techniques for achieving the better English-Telugu Translation.
2. The paper introduces the architecture of multi-headed attention network with the multi fast feed forward networks to achieve the better translation performance.
3. Extensive experimentation is carried out using the performance metrics such as Bilingual Evaluation Understudy (BLEU), character-level F-score (chrF) and Word Error Rate (WER) are evaluated and compared with the other state-of-the-art architectures. Results demonstrate the proposed model has shown the better performance in terms of achieving high rate of translation that can be used for the better education system.

1.3 Structure of the Paper

The rest of the paper is organized as follows: Section-2 presents the related works demonstrated by the different authors. The proposed architecture, data pre-processing, feature extraction process is illustrated in Section-3. The experimental analysis, result demonstration and comparative evaluations are presented in Section-4. Finally, the paper is concluded with the future direction in Section-5.

2. Related Works

Shinde et al. (2024) [24] propose a multilingual neural machine translation system for Indic languages using an LSTM-based translation model integrated with the pre-trained Google/MT-small model within the transformer architecture. The

system supports bidirectional translations between Indian and foreign languages with features for text, image, and video translation, facilitated by a Flask API for seamless integration. The inclusion of tools like EasyOCR and MoviePy enhances its adaptability across multiple modalities. However, the reliance on pre-trained models may limit performance for underrepresented languages, highlighting the need for extensive language-specific training data. Sudhansu Bala Das et al. (2024) [25] proposed a multilingual neural machine translation (MNMT) framework for Indic-to-Indic languages using the Samanantar and Flores-200 corpora, evaluated with BLEU scores. The study revealed that related language grouping benefited West Indo-Aryan languages but negatively impacted East Indo-Aryan and had inconclusive effects on Dravidian languages. Pivot-based models using English significantly improved translation quality, and transliteration enhanced BLEU scores for lexically rich languages like Malayalam and Tamil. However, the approach showed limited performance improvements for certain language pairs and relied heavily on the availability of high-quality pivot corpora. Lu et al. (2024) [26] proposed a Deep Reinforcement Learning (DRL) based Improved Deep Q-Network (IDQN) approach to enhance English-Chinese Neural Machine Translation (NMT). The study aimed to overcome challenges in translating both short and long sentences, which are common in traditional models utilizing Deep Neural Networks (DNNs) with attention mechanisms. They employed the IWSLT dataset and applied preprocessing for word vector generation. The optimized Convolutional Neural Network (CNN) extracted informative features for translation, achieving 95.2% precision, 95.3% accuracy, 95.3% F1-score, and 42.13% BLEU. However, the study did not explore the scalability of the IDQN approach for more complex languages beyond English and Chinese, which could limit its applicability in broader NMT tasks. Prasanna et al. (2023) [27] proposed a bi-lingual machine translation model for translating text from English to Tamil using a Gated Recurrent Unit (GRU) Long Short-Term Memory (LSTM) network. Their approach incorporates the Repeat Vector function for both the encoder and decoder parts of the network, optimized with the Adam optimizer for faster execution and reduced memory consumption. The dataset used for training the model includes resources from the Technology Development for Indian Languages (TDIL), Linguistic Data Consortium for Indian Languages (LDCIL), Kaggle, and Ishikahooda. The proposed system achieves a BLEU score of 0.9, a Meteor score of 0.98, a TER score of 0.5, and a WER score of 20%, indicating significant improvements over existing systems. Additionally, the system achieves a high accuracy and adequacy rate of 5 (on a 5-point scale). The novelty of their work lies in fine-tuning the space complexity of the model to 256 units of memory and optimizing the number of layers to reduce execution time. Unicode Transformation Format (UTF-8) encoding was used to incorporate Tamil characters, making it suitable for Tamil language translation. However, a limitation of this system is its dependency on large text corpora, which might be unavailable or difficult to access for languages with fewer digital resources, limiting its scalability for lesser-known regional languages. Sudhansu Bala Das et al. (2023) [28] proposed a Multilingual Neural Machine Translation (MNMT) system to address the challenges of translating low-resource Indian languages (ILs) effectively. The model utilizes a shared encoder-decoder architecture supporting 15 language pairs and incorporates techniques like data augmentation, back-translation, and domain adaptation to improve translation quality. High-resource languages from the same linguistic family were leveraged to boost performance for low-resource ILs. Experimental results demonstrated superior BLEU scores compared to baseline models, highlighting the model's efficiency. However, the approach relies heavily on augmentation strategies, which may limit scalability for languages with extremely scarce parallel corpora. Shailashree et al. (2023) [29] provide a comprehensive survey on Neural Machine Translation (NMT) for Indic languages, highlighting the evolution from Statistical Machine Translation (SMT) to NMT. The paper categorizes languages into High Resource Languages (HRLs), Low Resource Languages (LRLs), and Zero Resource Languages (ZRLs) based on corpus availability, addressing challenges in developing NMT models for low-resource languages like Kashmiri and Dogri. While the paper discusses various NMT architectures for languages like Hindi, Tamil, Kannada, Marathi, Sinhala, and Nepali, it acknowledges that translation quality remains low for many Indic languages due to insufficient resources. One key drawback is the lack of automated machine translation models for some less spoken Indic languages, hampering progress in the field. Sharma et al. (2023) [30] provide a comprehensive review of machine translation systems, exploring classical, statistical, and deep learning approaches. They highlight the significant impact of advancements in hardware, as well as the availability of monolingual and bilingual datasets, in driving the success of machine translation. The paper offers a comparative analysis of various models, focusing on hybrid models, neural machine translation (NMT), and statistical machine translation (SMT), and evaluates them using benchmark datasets and metrics

such as BLEU, TER, and METEOR. However, the authors acknowledge that while these models have achieved impressive results, challenges remain in handling low-resource languages and ensuring the preservation of contextual meaning in translations. Despite the advancements, issues such as computational complexity and domain adaptation still pose limitations to their effectiveness. Kandimalla et al. (2022) [31] developed Neural Machine Translation (NMT) systems for English-to-Indian languages using the transformer architecture, addressing the scarcity of parallel data for training. The study highlights the utility of back-translation in enhancing BLEU scores for English-to-Hindi and English-to-Bengali translations, particularly benefiting weaker baseline models. Manual evaluation revealed that the BLEU metric inadequately assesses translation quality, with English-Bengali outputs performing better than BLEU evaluations suggested. Despite improvements, the study notes a limitation in BLEU's reliability for quality analysis and the challenges posed by insufficient parallel datasets. Syed Abdul Basit Andrabi and Abdul Wahid et al. (2022) [32] proposed a deep learning-based machine translation system for English to Urdu languages using a neural network approach. The study utilized a parallel corpus of approximately 30,923 sentences comprising frequently used day-to-day expressions, achieving a BLEU score of 45.83. Automatic evaluation metrics were employed to validate the system's performance, and results were compared with Google Translator. However, the system's reliance on a limited corpus size may restrict its generalizability to broader contexts and complex linguistic structures. Saini and Sahula (2020) [33] explored neural machine translation (NMT) systems for English to six Indian languages, focusing on Hindi due to its extensive datasets and extending the study to Bangla, Tamil, Telugu, Urdu, and Malayalam. They experimented with eight NMT architectures and compared them to statistical machine translation (SMT), highlighting NMT's ability to perform well with smaller datasets. Despite these advantages, the study noted that translation quality for low-resource languages remains limited. This drawback emphasizes the need for further refinement in NMT for languages with less available data. The results demonstrated promising translation performance with minimal training sentences.

3. Proposed Methodology

Figure 1 depicts the steps involved in developing the proposed transformer based NMT model. The raw data which are collected are then pre-processed to make hem suitable for further processing. The pre-processing steps are required for the increasing the accuracy of the translated outputs. Following the pre-processing steps, enhanced transformer architecture is developed and trained using the training datasets. The proposed architecture is tested an evaluated using the English-Telugu corpses of data and its performance was evaluated. The detailed description of the each and every module is as follows

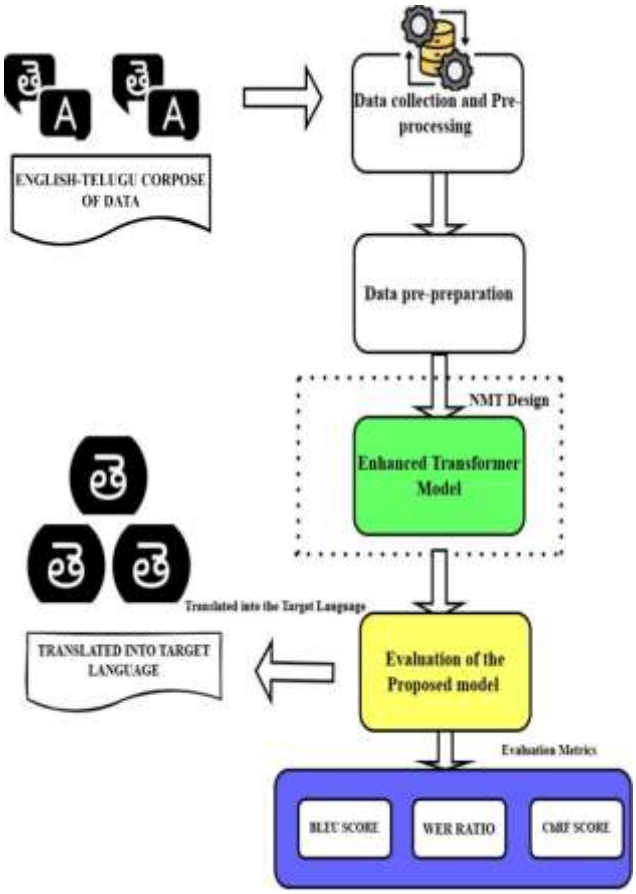

Figure 1. *Architecture for the Recommended Transformer model*

3.1 Materials and Methods

The datasets used for this research is adopted in [34]. The English-Telugu parallel corpora which was mentioned in the databases. The mean words per sentences was 9 for English an 8 for Telugu. The datasets consist of 61,000k English –Telugu sentences in Unicode. The vocabulary is built from 19000 sentences in the training and target languages separately. Figure 2 presents word-length distribution in the corpuses of datasets. Figure 3 shows the word cloud count both in English and Telugu languages. Figure 4 shows the sentence counts in the database corpuses. Here, 20000 sentences in length of 11-20 is found in Telugu corpses of data whereas 30000 sentences with the length of 11-20 is found in English corpuses. Table 1 is the list of curated databases for the training the proposed NMT using transformer model.

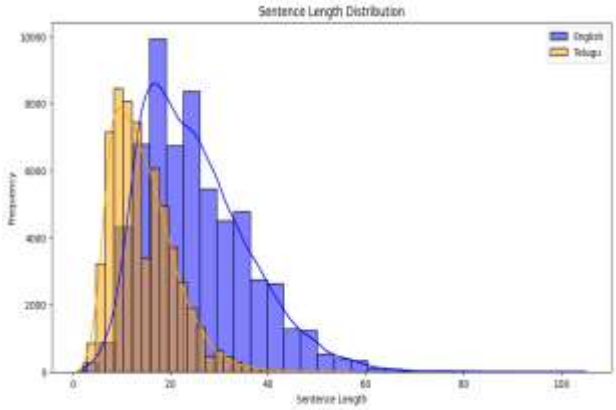

Figure 2. *Sample Word length Distribution in the Total Datasets*

Figure 3. *English –Word Count Mechanism in the Datasets*

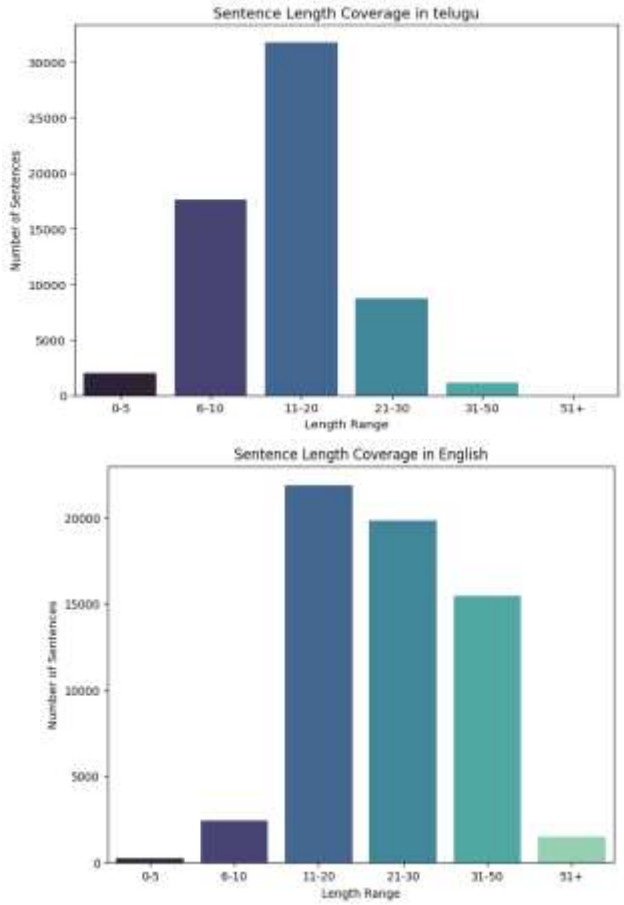

Figure 4. *Word Count Length in a) English b) Telugu Language*

Table 1. *List of Curated Databases for the Training the Proposed NMT using Transformer Model*

Description of the Datasets	Number of Datasets
NLLB_v1	46,376,267
Samanantar_v0.2	4,946,036
Anuvaad_v1	1,578,731
CCAligned_v1	581,652
Joshua-IPC_v1	414,162
XLEnt_v1.2	146,917
WikiMatrix_v1	91,911
Wikimedia_v20230407	69,800
Bible-Uedin_v1	62,191

3.2 Data Pre-processing

Before training the proposed model with the datasets, described in Section 3.1, pre-processing technique is adopted for the parallel datasets by adopting the following steps: (i) Converting all the texts into lower case (ii) Removing the special characters from texts except the apostrophe symbols (iii) tokenizing he source and target parallel sentences into sub word tokens using Keras Libraries [35]. (iv) generating the sub word embeddings as input to the transformer enabled block training. Furthermore, sentences are divided into training, validation and test sets, without sentence overlap. During the training phase, a small validation dataset will help to tune the model and create better translations and the test set is used for the model's performance evaluation. Figure 5 shows correlation between English and Telugu Sentences used for the model training. From the Figure 5, non-overlapping relationship between the two languages.

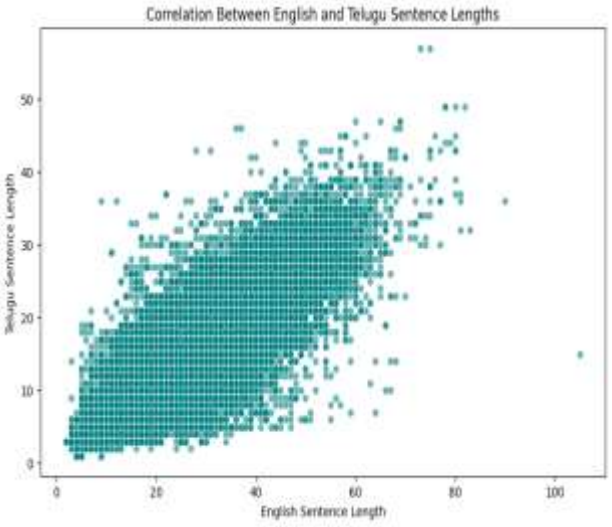

Figure 5. *Correlation Relationship Between the English and Telugu Languages*

3.3 Enhanced Transformer Model

This section discusses about the overview of the transformer based encoder-decoder design and enhanced transformer model

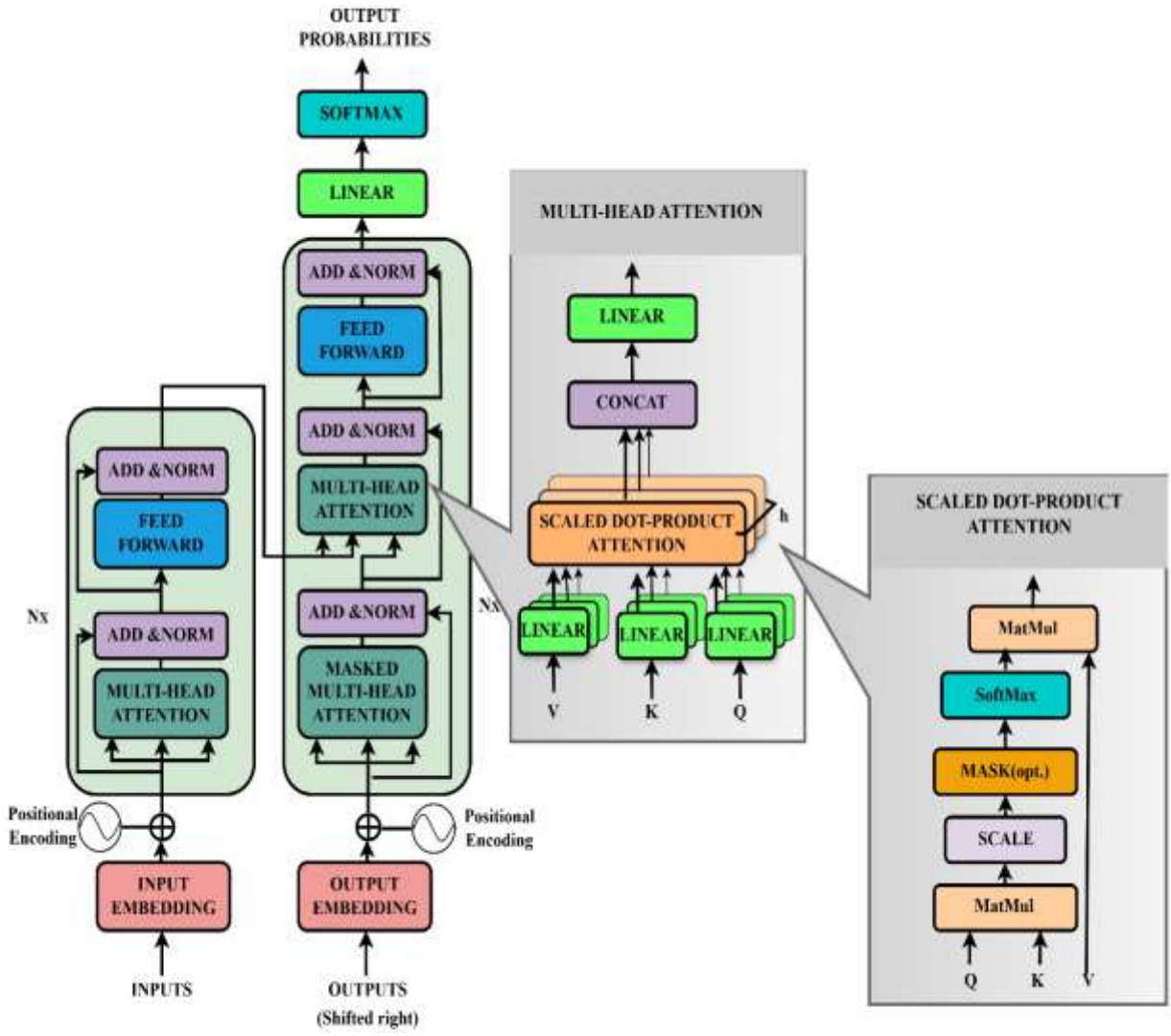

Figure 6. *General Transformer Architecture for the NMT Cross Language Translation*

Transformer with Encoder-Decoder Design-An Overview

The Transformer model comprises two key segments: an encoder and a decoder, each with a stack of six identical layers as shown in Figure 6. The encoder layers are each made up of two sub-layers: a multi-head self-attention mechanism and a position-wise feedforward network, both augmented by residual connections and normalized on a layer basis. The decoder, mirroring the encoder's structure, includes an additional third sub-layer in each of its layers, which performs multi-head attention over the encoder's output. This feature allows the decoder to focus on relevant parts of the input sequence, thereby facilitating more accurate and contextually informed predictions. A key innovation in the Transformer is its use of masked multi-head self-attention in the decoder, which ensures predictions for a given position are dependent only on the known outputs at previous positions, making it particularly suited for sequence generation tasks.

Self –Attention Layers in Transformer Models

The attention map was introduced in 2014 to define the proper words in sequence-to-sequence architecture. Most of the recent works have been carrying adding the attention layers to imitate redundant features that can aid for the precise classification mechanism. The Self attention mechanism also known as the intra-attention mechanism which is done by creating the three vectors Q, K and V for each input sequence. Thus the input sequences from each layers are transformed to the output sequences. In other words, it is technique which maps the Query with the set of key pairs by using scaled dot functions. Mathematically, dot product for self –attention is computed as follows

$$F(K, Q) = ((K), Q^{T}))/(V_K)\text{^}0.5 \quad (1)$$

3.4 Multi-Headed Self Attention Maps (MHSA) – An Overview

MHSA are developed from self-attention maps which maps input sequences and set of key value pairs to a weighted output. where the weights assigned are computed by a compatibility function using the input sequence and corresponding key. As depicted in Figure 6, for each frame in given input image, self-attention computes the queries, keys, values of dimension (V_K) with linear projection. By combining these dimensions, A, B, C, the self-attention output is calculated using

$$F(A, B, C) = (\text{Softmax}(AB^T)C)/(V_K)\text{^}0.5) \quad (2)$$

To exploit the information from the noisy images it is necessary to necessary network depth of self-attention maps. Hence this research article proposes multi-head self -attention maps that uses multiple attention maps that iterates for 'n' times to generate queries, keys, values matrices to overcome the noises from the images. Figure 6 Shows the Multi-Headed Self Attention (MSA) structure deployed in the transformer blocks. Mathematically MHSA is given as follows

$$\text{MSA}(A, B, C) = \text{Concat}(SA1, SA2, SA3 \ldots . SAn)W^o \quad (3)$$

$$\text{SA} = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}) \quad (4)$$

$$\text{Self} - \text{Attention maps calculated using Equation}(4)$$

3.5 Fast Feed Forward Layers – Its Principle of Working

The fast feed forward networks are designed based on the principle of extreme learning mechanism. The proposed model uses the principle of extreme learning machines suggested by G.B.Huang for the high speed and high accurate classification of different grades. This kind of neural network utilizes the single hidden layers in which the hidden layers doesn't require the tuning mandatorily. ELM uses the kernel function to yield good accuracy for the better performance. The major advantages of the ELM are minimal training error and better approximation. Since ELM uses the auto-tuning of the weight biases and non-zero activation functions, ELM finds its applications in classification and classification values. The detailed working mechanism of the ELM. In this sort of system, the' L' neurons in the hidden layer are required to work with an activation function that is vastly differentiable (for instance, the sigmoid function), though that of the output layer is straight. in ELM, hidden layers do not need to be tunes mandatorily. In ELM, the hidden layer compulsorily need not be tuned. The loads of the hidden layer are arbitrarily appointed (counting the bias loads). It isn't the situation that hidden nodes are irrelevant, however they need not be tuned and the hidden neurons parameters can be haphazardly produced even in advance.

That is, before taking care of the training set data. For a single-hidden layer ELM, the system yield is given by eqn (5)

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = h(x)\beta \quad (5)$$

Where x → input features from encoder-decoder

β → output weight vector and it is given as follows as

$$\beta = [\beta_1, \beta_2, \ldots \ldots \ldots \ldots \beta_L]^T \quad (6)$$

h(x)→ output hidden layer which is given by the following equation

$$h(x) = [h_1(x), h_2(x), \ldots \ldots \ldots \ldots h_L(x) \quad (7)$$

To determine Output Vector O which is called as the target vector, the hidden layers are represented by eqn (8)

$$H = \begin{bmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_N) \end{bmatrix} \quad (8)$$

The basic implementation of the ELM uses the minimal non -linear least square methods which are represented in eqn (9)

$$\beta' = H^*O = H^T(HH^T)^{-1}O \quad (9)$$

Where H∗ → inverse of H known as Moore–Penrose generalized inverse.

Above eqn can also be given as follows

$$\beta' = H^T(\frac{1}{C}HH^T)^{-1}O \quad (10)$$

Hence the output function can be find by using the above equation

$$f_L(x) = h(x)\beta = h(x)\ H^T(\frac{1}{C}HH^T)^{-1}O \quad (11)$$

Where h(x) is input feature maps, β is temporal matrix which is solved by Moore–Penrose generalized inverse theorem denoted by H^T, C is constant, B and O are weights and bias factors of the network.

3.6 Enhanced Transformer Model with MHMSBA network

Figure 6 shows the enhanced transformer model with the encoder-decoder design model with the proposed multi-scale multiple-headed attention layers (MSMHA)and fast feed forward networks. Each point is subjected to an independent fast feed-forward after MSMHA layers. The proposed MSMHA model was designed. The MSMHA model was constructed using two layers such Channel attention (CA) a Spatial attention (SA). The effective combination of these attention networks aids for an effective extraction of word sequences. In this context, channel attention module is used for compressing the feature maps within the spatial dimension to generate one-dimensional vector pre-operations. In the channel attention modules, spatial information is compressed using Global Average pooling layer to engender the two feature maps. On the other hand, spatial attention module compresses the channel features using Softmax and Global Average pooling layers. In this contrary, convolutional dot operations are employed and concatenated for the generation of weighted feature maps. These two attention module repeats its computations multiple times and finally concatenated to form the final attention maps. Equations (12), (13) and (14) provide the mathematical operation performed in the PSA and CA networks.

$$Y(x) = H(x) + J(x) \tag{12}$$

$$H(x) = S(x) = \text{Softmax}(\text{Transpose}(A(x)) + B(x) \tag{13}$$

$$Y(x) = S(x) = \text{Softmax}\left(\text{Transpose}(A(x)) + B(x)\right) + J(x) \tag{14}$$

$$J(x) = T(x) + U(x) \tag{15}$$

$$T(x) = \text{Softmax}\left(\text{Transpose}(B(x)).A(x)\right) \tag{16}$$

$$J(x) = \text{Softmax}\left(\text{Transpose}(B(x)).A(x)\right)) + U(x) \tag{17}$$

Finally

$$\text{MHDA}(Y(x)) = S(x) = \left(\text{Softmax}\left(\text{Transpose}(A(x)) + B(x)\right) + \text{Softmax}\left(\text{Transpose}(B(x)).A(x)\right)\right) + U(x) \tag{18}$$

Where Y(x) is the Bi-layered attentions' feature maps, x): The feature maps derived from position self-attention are represented as H(x). while J(x)) corresponds to those obtained through channel attention. The outputs of three distinct parallel convolutional operations are denoted as A(x), B(x), and C(x). Additionally, J(x) encapsulates the channel attention feature maps, whereas T(x) signifies the result of the element-wise product between softmax((D(x)) and the input feature maps. Similarly, U(x) refers to the feature maps produced by reshaping, transposing, and reshaping the input feature maps.

The designed MHBA mapping technique evaluates the relevance of each word concerning others within the input sequence while effectively capturing word interdependencies simultaneously. Positional encodings are incorporated into the input embedding to determine token positions. The encoder comprises multiple MHBA layers, each playing a vital role in maintaining contextual relationships within the sequence. Following MHBA, each token undergoes processing through a standalone fast feedforward neural network. The inclusion of attention layers empowers the model to evaluate the relative importance of words in the sequence, fostering a deeper understanding of their connections. Layer normalization is applied after a residual connection encases each sub-layer within the encoder, enhancing training stability and mitigating issues like the vanishing gradient. Similarly, the sentence in the target language is embedded with positional encodings, mirroring the encoder's approach. During the training process, the decoder employs masked self-attention, ensuring that each position in the decoder sequence can only consider preceding positions. This mechanism prevents the model from accessing future tokens during training, preserving the sequential generation characteristic essential for auto-regressive modelling. Additionally, the decoder leverages multi-head attention (MHA) over the encoder's outputs, enabling it to emphasize various sections of the input sequence while producing each token of the output. Following the attention mechanism, a rapid feedforward neural network processes the decoder's output. Like the encoder, the decoder incorporates residual connections and layer normalization in every sub-layer. This refined transformer framework learns vocabulary from the parallel corpus it is trained on, effectively developing a comprehensive understanding of the language.

4. Results and Discussions

This section details about the performance metrics, validation analysis and comparative analysis.

4.1 Performance Metrics Analysis

The evaluation of NMT models commonly utilizes metrics such as BLEU score, chrF, and WER. The BLEU

score is determined by combining the Brevity Penalty (BP) with the geometric mean of the modified precision values, pnp_n, across n-gram levels up to a specified size, N. The BP serves to reduce scores for excessively brief translations in comparison to the reference text. Here, r represents the length of the reference text, c denotes the length of the translated text, and N specifies the n-gram size, typically set at 4. The parameter w_n signifies the weight associated with each n-gram, ensuring the sum equals 1. Equations (19) and (20) define the mathematical formulas for computing BP and BLEU scores.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \quad (19)$$

$$BLEU = BP.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (20)$$

$$chrF = (1+\beta^2)\frac{chrP.chrR}{\beta^2.chrP+chrR} \quad (21)$$

The average chrP evaluates the proportion of n-grams in a translated sentence that match those in the reference sentence. Similarly, the average chrR measures the percentage of n-grams from the reference sentence that are

found in the translation. The parameter β is used to assign greater significance to recall compared to precision. WER for a sentence is determined using Equation (22).

Table 2. *Translation Result Sample Outputs from the Proposed Model*

Input Sentences	Output Translated results	Remarks
thou shalt prepare thee a way and divide the coasts of thy land which the lord thy god giveth thee to inherit into three parts that every slayer may flee thither	ప్రతి నరహంతకుడు పారిపోవునట్లుగా నీవు త్రోవను ఏర్పరచుకొని నీవు స్వాధీనపరచుకొనునట్లు నీ దేవుడైన యెహోవా నీకిచ్చుచున్న దేశముయొక్క సరి హద్దులలోగా ఉన్న పురములను మూడు భాగములు చేయవలెను	Accurate with Exact Translation
blessed is the man whom you discipline yah and teach out of your law	యెహోవా నీవు శిక్షించువాడు నీ ధర్మశాస్త్రమును బట్టి నీవు బోధించువాడు ధన్యుడు	
and the king said unto her what aileth thee and she answered i am indeed a widow woman and mine husband is dead	రాజునీకేమి కష్టము వచ్చెనని అడిగెను అందుకు ఆమెనేను నిజముగా విధవరాలను నా పెనిమిటి చనిపోయెను	
grace to you and peace from god our father and the lord jesus christ	ద్రాక్షారసము త్రాగుటలో ప్రఖ్యాతినొందిన వారికిని మద్యము కలుపుటలో తెగువగలవారికిని శ్రమ	
they said we cant until all the flocks are gathered together and they roll the stone from the wells mouth then we water the sheep	వారుమంద లన్నియు పోగుకాకమునుపు అది మావలన కాదు తరువాత బావిమీదనుండి రాయి పొర్లించుదురు అప్పుడే మేము గొఱ్ఱలకు నీళ్లు పెట్టుదుమనిరి	
yahweh of armies blessed is the man who trusts in you	సైన్యములకధిపతివగు యెహోవా నీయందు నమ్మికయుంచువారు ధన్యులు	
: wherefore the anger of the lord was kindled against amaziah and he sent unto him a prophet which said unto him why hast thou sought after the gods of the people which could not deliver their own people out of thine hand	అందుకొరకు యెహోవా కోపము అమజ్యా మీద రగులుకొనెను ఆయన అతనియొద్దకు ప్రవక్తను ఒకని పంపగా అతడునీ చేతిలోనుండి తమ జనులను విడిపింప శక్తిలేని దేవతలయొద్ద నీవెందుకు విచారణ చేయుదువని అమజ్యాతో ననెను	

$$WER = (T + N + E)/N \quad (22)$$

In this context, T, N, and E correspond to the respective counts of substitutions, insertions, and deletions. N depicts the entire word count.

4.2 Implementation Details

The complete model was developed using Python 3.19 programming and libraries such as matplotlib, pandas, numpy, tensorflow –keras are used for evaluating the proposed model. The experimentation was carried out in the PC workstation with i7 CPU, NVIDIA Tesla GPU,16GB RAM and 3.2 GHZ operating frequency.

4.3 Results Discussion

After training the proposed model with 15000 steps, accuracy by utilizing a test set with 5000 sentences in English. Table 2 presents the translation results and it test accuracy has been checked with the Google translators and reference datasets.

The in-depth review of the derived outcome in Table 2 is examined further. Based on the table 2, it is apparent that the translation provided by the recommended model is accurate and matches the expected Google translation. Therefore, one can infer that sentences of any length tend to offer better translation performance. By the outcomes depicted in the table 2, recommended transformer architecture captures the exact meaning of the English word and generates an accurate Telugu translation which gives the different shade of correlation context as it really correlates for the understandable system for the education. In the above examples, though the translations obtained the transformer model with MHHBSA has a better

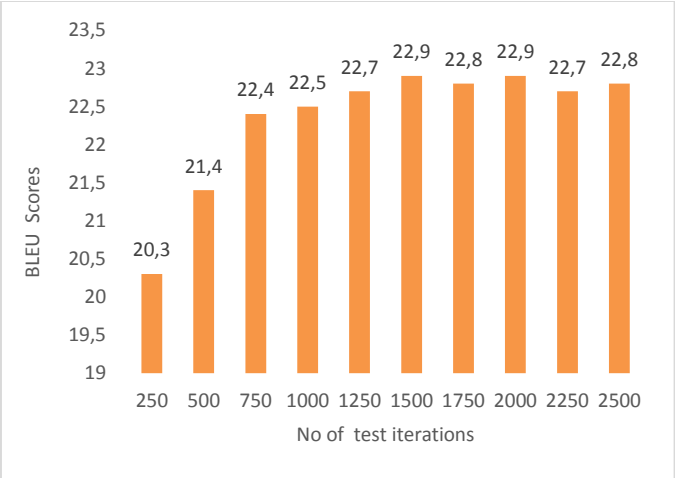

Figure 7. *BLEU Scores for the Proposed Model for the increasing word lengths*

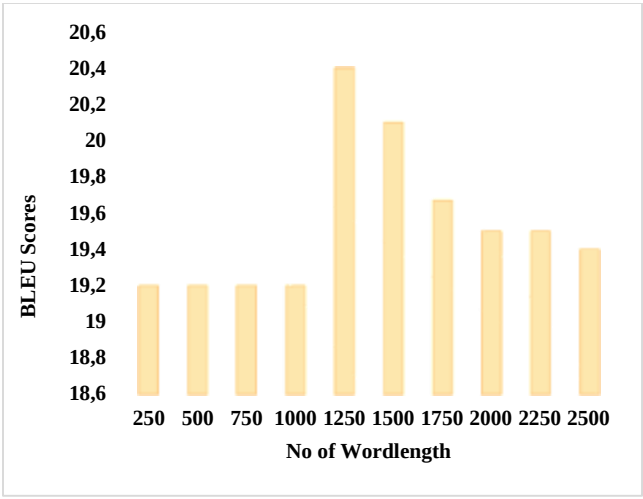

Figure 8. *BLEU Scores for the Traditional Model for the increasing word lengths.*

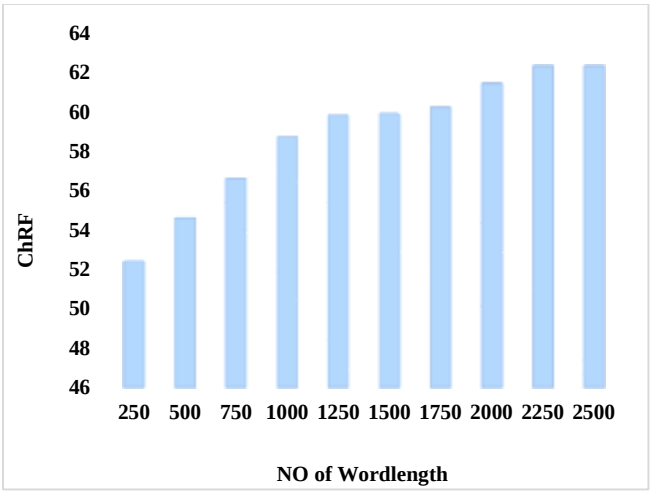

Figure 9. *ChRF Scores for the Proposed Model for the increasing word lengths*

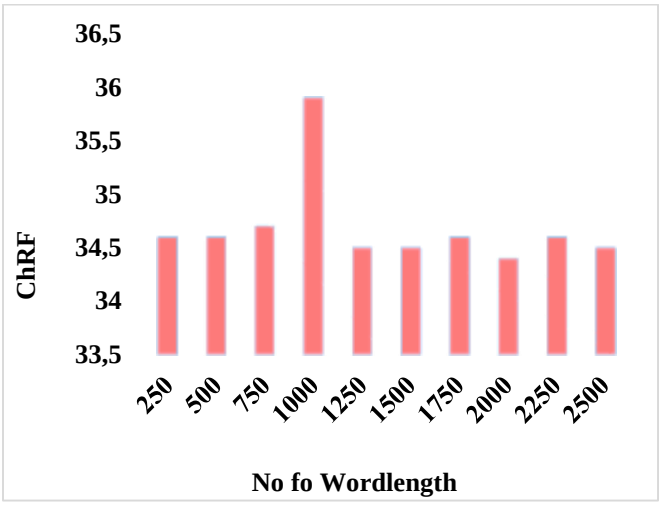

Figure 10. *BLEU Scores for the Proposed Model for the increasing word lengths*

translation. From the analysis, it is apparent that the recommended model attains optimal translation outcomes. The validation of the model relies on translation scores, employing performance metrics such as BLEU score, chrF, and average chrF. Figure 7 shows the evaluation metrics for every 500 steps and it is evident that highest BLEU, Chrf and average Chrf metrics are produced by the proposed model. Figure 7-10 shows the BLeU scores for the proposed transformer model and traditional transformer model used in the NMT design for cross language translation English-Telugu. From the figure 7, it is evident that the BLeU scores of the recommended model has excelled than the traditional model. The addition of the MHMHBSA model has provided the significant role in the improvisation in the translation efficiency in the model. Figure 9 shows the ChRF scores for the traditional and proposed transformer model. The similar fashion is witnessed as in Figure 9 in which the recommended model has produced the average ChRF as 62.3 and the traditional transformer has produced ChRF as 36.2.

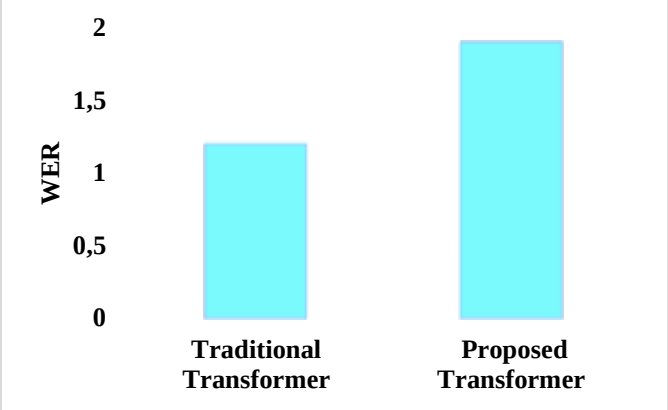

Figure 11. *WER Scores for the Proposed Model and Traditional Transformer Models in the NMT*

Figure 11 shows the WER Scores for the traditional and proposed transformer used in the cross-language translation mechanism. WER scores are inversely proportional to the quality of translation. The score is derived by evaluating the insertions, deletions, and substitutions necessary to align the output with the reference sentence. As shown in Figure 11, WER is higher for the proposed transformer model than the traditional transformers used in the NMT designs.

4.4 Comparative Analysis

To compare the superiority of the recommended transformer model, residing models like LSTM, Traditional transformer, Transformer +ATG, Transformer+POS and Hybrid RNN-LSTM Models. Every model is trained with the curated data used in the proposed research and 2500 steps are used for the validation and comparison for the different models.

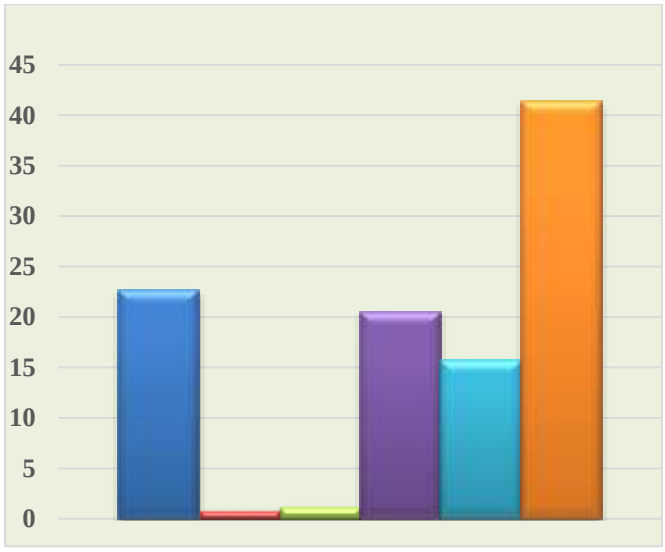

Figure 12. *Comparative Analysis of the BLeU Scores obtained from the different models*

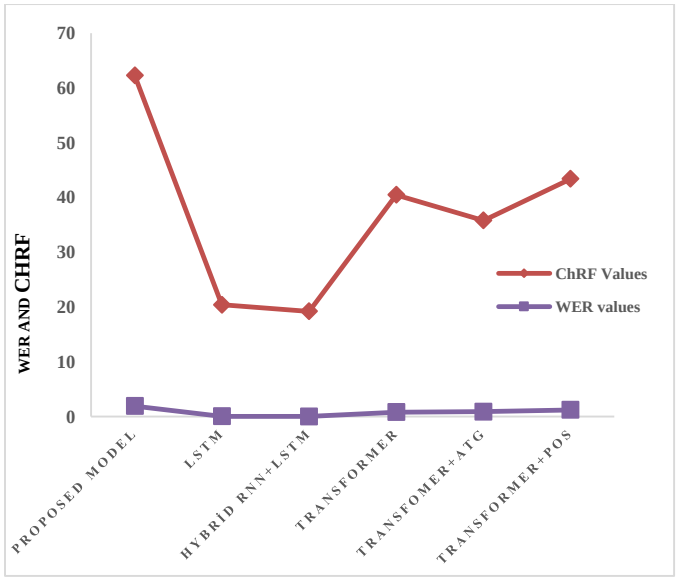

Figure 13. *Comparative Analysis of the WER and ChRF Scores obtained from the different models*

Figure 13 shows the comparative analysis of the varied approaches by calculating the BLeU scores for the neural machine translation. From the Figure 12, LSTM and hybrid model has produced the least BLeU scores. Transformer+POS model has produced the better BLeU scores but lesser than the recommended approach. The

recommended approach has produced the better performance in detecting the corrected words while translating from the English to Telugu language corpuses. It is inferred from the Figure 10-11, WER and ChrF values in the recommended approach has been examined with the varied residing models. The integration of MHMSBHA model in the transformer network has improvised the quality of translation from English to Telugu language that can be suitable for the education system. Neural machine has been applied in different fields and reported [36-51].

5. Conclusion and Future Enhancement

The recommended cross-language translation system integrates linguistic features into the source side of machine translation for under-resourced language pairs. Translating between low- and extremely low-resource languages presents challenges due to the absence of parallel datasets. To overcome this aforementioned problem, enhanced transformer architecture is recommended with the low –resource Dravidian languages such as Telugu language corpuses. The novel MSMHBA maps and fast feed forward networks are introduced in the proposed transformer model. The extensive experimentation is carried out using the 61000 datasets and performance metrics such as BLeU, WER and ChRf are measured and analysed. Furthermore, performance of the recommended approach is compared with the varied residing models. Experimental outcomes demonstrated that the recommended transformer model has produced the better translation efficiency than the existing models. It has proved that integration of MSMHBA and fast feed forward networks has played the significant role in improving the translation performance that play an important role in the education system so that students /kids can enhance their knowledge by their own mother tongue. As the future direction, proposed model needs for the multi-lingual translation with the video and audio inputs. This could even increase the literacy rate of any country and making the future generation to gain the disruptive knowledge in their own mother tongue.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Chen, Z., Jiang, C., & Tu, K. (2023). Using interpretation methods for model enhancement. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language* (pp. XX–XX). Singapore, December 6–10.

[2] Yin, K., & Neubig, G. (2022). Interpreting language models with contrastive explanations. *arXiv Preprint*, arXiv:2202.10419.

[3] Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2018). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *arXiv Preprint*, arXiv:1801.07772.

[4] Ekin, A., Dale, S., Jacob, A., Tengyu, M., & Denny, Z. (2023). What learning algorithm is in-context learning? Investigations with linear models. In *Proceedings of the International Conference on Learning Representations (ICLR)* (pp. XX–XX). Kigali, Rwanda, May 1–5.

[5] He, S., Tu, Z., & Wang, X. (2019). Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. XX–XX). Hong Kong, China, November 3–7.

[6] Qiang, J., Liu, K., Li, Y., Zhu, Y., Yuan, Y. H., Hu, X., & Ouyang, X. (2023). Chinese lexical substitution: Dataset and method. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language* (pp. XX–XX). Singapore, December 6–10.

[7] Kalpana, P., Almusawi, M., Chanti, Y., Sunil Kumar, V., & Varaprasad Rao, M. (2024). A deep reinforcement learning-based task offloading framework for edge-cloud computing. In *Proceedings of the 2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)* (pp. 1–5). Raichur, India. https://doi.org/10.1109/ICICACS60521.2024.10498232

[8] Tan, S., Shen, Y., Chen, Z., Courville, A., & Gan, C. (2023). Sparse universal transformer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language* (pp. XX–XX). Singapore, December 6–10.

[9] Müller, M., Jiang, Z., Moryossef, A., Rios, A., & Ebling, S. (2023). Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of*

the 61st Annual Meeting of the Association for Computational Linguistics (pp. XX–XX). Toronto, Canada, July 9–14.

[10] Kalpana, P., Narayana, P. L., Madhavi, S., Dasari, K., Smerat, A., & Akram, M. (2025). Health-Fots—A latency-aware fog-based IoT environment and efficient monitoring of body's vital parameters in smart healthcare environment. *Journal of Intelligent Systems and Internet of Things, 15*(1), 144–156. https://doi.org/10.54216/JISIoT.150112

[11] Kai, V., & Frank, K. (2024). Cluster-centered visualization techniques for fuzzy clustering results to judge single clusters. *Applied Sciences, 14*(1102).

[12] Woosik, L., & Juhwan, L. (2024). Tree-based modeling for large-scale management in agriculture: Explaining organic matter content in soil. *Applied Sciences, 14*(1811).

[13] Kalpana, P., Malleboina, K., Nikhitha, M., Saikiran, P., & Kumar, S. N. (2024). Predicting cyberbullying on social media in the big data era using machine learning algorithm. In *Proceedings of the 2024 International Conference on Data Science and Network Security (ICDSNS)* (pp. 1–7). Tiptur, India. https://doi.org/10.1109/ICDSNS62112.2024.10691297

[14] Rikters, M., Pinnis, M., & Krišlauks, R. (2018). Training and adapting multilingual NMT for less-resourced and morphologically rich languages.

[15] Aly, R., Caines, A., & Buttery, P. (2021). Efficient unsupervised NMT for related languages with cross-lingual language models and fidelity objectives. In *Workshop on NLP for Similar Languages, Varieties, and Dialects*

[16] Kalpana, P., Kodati, S. Smitha, L., Sreekanth, D., Smerat, N., & Akram, M. (2025). Explainable AI-driven gait analysis using wearable internet of things (WIoT) and human activity recognition. *Journal of Intelligent Systems and Internet of Things, 15*(2), 55–75. https://doi.org/10.54216/JISIoT.150205

[17] Prasanna, R. K., Sudharson, S., Reddy, A. A., Reddy, B. S. J. N., & Anvitha, V. (2023). A comparative analysis of transformers for multilingual neural machine translation. In *Proceedings of the 2023 IEEE 7th Conference on Information and Communication Technology (CICT)* (pp. 1–6). Jabalpur, India. https://doi.org/10.1109/CICT59886.2023.10455324

[18] Ji, B., Zhang, Z., Duan, X., Zhang, M., Chen, B., & Luo, W. (2020). Cross-lingual pre-training based transfer for zero-shot neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(01), 115–122. https://doi.org/10.1609/aaai.v34i01.5341

[19] Dabre, R., Chu, C., & Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. *ACM Computing Surveys, 53*(5), Article 99. https://doi.org/10.1145/3406095

[20] Kudugunta, S. R., Bapna, A., Caswell, I., Arivazhagan, N., & Firat, O. (2019). Investigating multilingual NMT representations at scale. *arXiv Preprint*, arXiv:1909.02197.

[21] Ha, T. L., Niehues, J., & Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. *arXiv Preprint*, arXiv:1611.04798.

[22] Kumari, D., Ekbal, A., Haque, R., Bhattacharyya, P., & Way, A. (2021). Reinforced NMT for sentiment and content preservation in low-resource scenarios. *ACM Transactions on Asian and Low-Resource Language Information Processing, 20*(4), Article 70. https://doi.org/10.1145/3450970

[23] Chan, K. H., Ke, W., & Im, S. K. (2020). CARU: A content-adaptive recurrent unit for the transition of hidden state in NLP. In H. Yang, K. Pasupa, A. C. S. Leung, J. T. Kwok, J. H. Chan, & I. King (Eds.), *Neural information processing (Lecture Notes in Computer Science, Vol. 12532, pp. 123–135)*. Springer.

[24] Shinde, V. R., et al. (2024). Multilingual neural machine translation system for Indic languages. *International Journal of Engineering Research & Technology (IJERT), 10*(1).

[25] Das, S. B., Panda, D., Mishra, T. K., Patra, B. K., & Ekbal, A. (2024). Multilingual neural machine translation for Indic to Indic languages. *ACM Transactions on Asian and Low-Resource Language Information Processing, 23*(5), Article 65. https://doi.org/10.1145/3652026

[26] Lu, Y. (2024). Research on English-Chinese neural machine translation based on improved deep Q-network approach. *Second International Conference on Data Science and Information System (ICDSIS)*, Hassan, India. https://doi.org/10.1109/ICDSIS61070.2024.10594422

[27] Prasanna ASD, Latha CBC. (2023) Bi-Lingual Machine Translation Approach using Long Short–Term Memory Model for Asian Languages. Indian Journal of Science and Technology. 16(18):1357-1364. https://doi.org/10.17485/IJST/v16i18.176.

[28] Das, S. B., Biradar, A., Mishra, T. K., & Patra, B. K. (2023). Improving multilingual neural machine translation system for Indic languages. *ACM Transactions on Asian and Low-Resource Language Information Processing, 22*(6), Article 169. https://doi.org/10.1145/3587932

[29] Shailashree, K. S., Gupta, D., & Costa-Jussà, M. R. (2023). A voyage on neural machine translation for Indic languages. Procedia Computer Science, 218, 2694–2712. https://doi.org/10.1016/j.procs.2023.01.242.

[30] Sharma, S., Diwakar, M., Singh, P., Singh, V., Kadry, S., & Kim, J. (2023). Machine translation systems based on classical-statistical-deep-learning approaches. *Electronics, 12*(1716). https://doi.org/10.3390/electronics12071716

[31] Kandimalla, Akshara & Lohar, Pintu & Maji, Kumar & Way, Andy. (2022). Improving English-to-Indian Language Neural Machine Translation Systems. Information. 13. 245. 10.3390/info13050245.

[32] Andrabi, S. B., & Wahid, A., et al. (2022). Machine translation system using deep learning for English to Urdu. *[Journal Name]*. https://doi.org/10.1155/2022/7873012

[33] Saini, S., & Sahula, V. (2020). Setting up a neural machine translation system for English to Indian languages. In G. R. Sinha & J. S. Suri (Eds.), Cognitive informatics, computer modelling, and cognitive science (pp. 195–212). Academic Press. https://doi.org/10.1016/B978-0-12-819443-0.00011-8.

[34] https://opus.nlpl.eu/results/en&te/corpus-result-table

[35] https://huggingface.co/datasets/allenai/nllb

[36] LAVUDIYA, N. S., & C.V.P.R Prasad. (2024). Enhancing Ophthalmological Diagnoses: An Adaptive Ensemble Learning Approach Using Fundus and OCT Imaging. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.678

[37] P. Padma, & G. Siva Nageswara Rao. (2024). CBDC-Net: Recurrent Bidirectional LSTM Neural Networks Based Cyberbullying Detection with Synonym-Level N-Gram and TSR-SCSOFeatures. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.623

[38] Rajani Kumari Inapagolla, & K . Kalyan Babu. (2025). Audio Fingerprinting to Achieve Greater Accuracy and Maximum Speed with Multi-Model CNN-RNN-LSTM in Speaker Identification: Speed with Multi-Model CNN-RNN-LSTM in Speaker Identification. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1138

[39] Mekala, B., Neelamadhab Padhy, & Kiran Kumar Reddy Penubaka. (2025). Brain Tumor Segmentation and Detection Utilizing Deep Learning Convolutional Neural Networks: Enhanced Medical Image for Precise Tumor Localization and Classification. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1051

[40] Rajitha Kotoju, B.N.V. Uma Shankar, Ravinder Reddy Baireddy, M. Aruna, Mohammed Abdullah Mohammed Alnaser, & Imad Hammood Sharqi. (2025). A Deep auto encoder based Framework for efficient weather forecasting. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.429

[41] M. Kannan, & K.R. Ananthapadmanaban. (2025). Students Performance prediction by EDA analysis and Hybrid Deep Learning Algorithms. *International Journal of Computational and Experimental Science and Engineering*, 11(2). https://doi.org/10.22399/ijcesen.1524

[42] Kumar, N., & T. Christopher. (2025). Enhanced hybrid classification model algorithm for medical dataset analysis. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.611

[43] Olola, T. M., & Olatunde, T. I. (2025). Artificial Intelligence in Financial and Supply Chain Optimization: Predictive Analytics for Business Growth and Market Stability in The USA. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.18

[44] A, V., & J Avanija. (2025). AI-Driven Heart Disease Prediction Using Machine Learning and Deep Learning Techniques. *International Journal of Computational and Experimental Science and Engineering*, 11(2). https://doi.org/10.22399/ijcesen.1669

[45] Ibeh, C. V., & Adegbola, A. (2025). AI and Machine Learning for Sustainable Energy: Predictive Modelling, Optimization and Socioeconomic Impact In The USA. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.19

[46] Polatoglu, A. (2024). Observation of the Long-Term Relationship Between Cosmic Rays and Solar Activity Parameters and Analysis of Cosmic Ray Data with Machine Learning. *International Journal of Computational and Experimental Science and Engineering*, 10(2). https://doi.org/10.22399/ijcesen.324

[47] Fowowe, O. O., & Agboluaje, R. (2025). Leveraging Predictive Analytics for Customer Churn: A Cross-Industry Approach in the US Market. *International Journal of Applied Sciences and Radiation Research* , 2(1). https://doi.org/10.22399/ijasrar.20

[48] Kumar, A., & Beniwal, S. (2025). Depression Sentiment Analysis using Machine Learning Techniques:A Review. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.851

[49] Hafez, I. Y., & El-Mageed, A. A. A. (2025). Enhancing Digital Finance Security: AI-Based Approaches for Credit Card and Cryptocurrency Fraud Detection. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.21

[50] S, P., & A, P. (2024). Secured Fog-Body-Torrent : A Hybrid Symmetric Cryptography with Multi-layer Feed Forward Networks Tuned Chaotic Maps for Physiological Data Transmission in Fog-BAN Environment. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.490

[51] García, R., Carlos Garzon, & Juan Estrella. (2025). Generative Artificial Intelligence to Optimize Lifting Lugs: Weight Reduction and Sustainability in AISI 304 Steel. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.22