International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(3):1252-1271 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i03.04963

# Cross-Lingual Machine Translation: An Integrated Approach with Contextualsequencexl and Goldenhawk Search Optimization Algorithm (GHSO)

Kandula Narasimharao*, Angara SV Jayasri

Department of Computer Science and Engineering, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India. *Corresponding Author's Email: kandulanarasimharao49@gmail.com

## Abstract

Cross-Lingual Machine Translation (CLMT) remains a complicated issue because of linguistic diversity, lack of parallel corpora, and the necessity of effective semantic alignment. ContextXL is a new and complete CLMT framework that combines sophisticated pre-processing, semantic representation, and feature optimization strategies. This method starts with Named Entity Recognition (NER) and Byte Pair Encoding (BPE) to maintain semantic units and to deal with rare words. It then improves language representation with Cross-Lingual Word Embeddings (MUSE) and sub word-sensitive FastText embedding. Bidirectional Encoder Representations from Transformers (BERT) and Embedding from Language Models (ELMo) are used to extract richer features using contextual embedding. An effective feature selection is performed using a new Golden Hawk Search Optimization (GHSO) algorithm which is a combination of Golden Section Search and Chaotic Harris Hawk Optimization. Transformer-XL is the translation engine and models long-range dependencies with segment-level recurrence and memory caching. Experimental analysis reveals, ContextXL has a high translation accuracy of 98.77%, and good results in precision (98.97%), recall (98.57%), and F-score (98.85%). The model also performs better than the state-of-the-art baselines, Neural Machine Translation (NMT), BERT, Transformer, and RoBERTa in various evaluation metrics like Mathews Correlation Coefficient (MCC), sensitivity, and specificity, False Negative Rate (FPR), False Positive Rate (FPR) and Negative Predictive Value (NPV). Its effectiveness is also confirmed by a human evaluation, which gives it high scores in contextual preservation (4.52/5), fluency (4.47/5), and appropriateness (4.38/5). These findings point to the strength of ContextXL, which is appropriate to process low-resource, morphologically rich, even informal or code-switched language data.

**Keywords:** Bidirectional Encoder Representations from Transformers, Byte Pair Encoding, Cross-Lingual Machine Translation, Embedding from Language Models, Golden Hawk Search Optimization, Named Entity Recognition.

## Introduction

Within Natural Language Processing (NLP), English is blessed with an abundance of labeled data, which feeds the appetite of data-hungry deep learning algorithms on tasks such as named entity identification, natural language inference, and part-of-speech tagging. Cross-lingual transfer learning is essential since many languages lack task-specific data (1-4); therefore, this benefit is not universal. This procedure entails using information from languages with a wealth of data to improve performance in languages that have little or no task-specific data. Advances in Neural Machine Translation (NMT) have led to the development of multilingual systems that can translate text between many source languages and numerous target languages all inside the same model. These Multilingual NMT (mNMT) systems demonstrate impressive gains in translation quality, particularly for low-resource languages, which is attributed to the models' capacity to acquire transferable representations between languages (5). Even with the increased scholarly focus on cross-lingual communication research, multilingual textual analysis is still difficult. This phrase emphasizes the necessity of extra procedures and efforts in processing such data and captures the challenge of gathering, disseminating, and evaluating multilingual textual data across national borders (6-8). Due to the increasing demand for multilingual analysis, communication scientists have resorted to two methods to address this problem: either train individual topic models for each language or use Machine Translation (MT) to translate multilingual content into a single language (e.g., English) for analysis. This work addresses the latter strategy, outlining its



(Received 14th April 2025; Accepted 14th July 2025; Published 30th July 2025)

drawbacks and putting forth a word embedding strategy that yields findings that are more valid, repeatable, and consistent with open research. Cross-Lingual Machine Translation becomes an important area of study when considering cross-lingual problems in a larger sense (9-11). With a focus on overcoming the difficulties presented by linguistic variances and structures, this field aims to create techniques and algorithms that make it easier to translate text or voice between languages. Overcoming language barriers and improving comprehension and communication between speakers of various languages is the ultimate objective of cross-lingual machine translation (12, 13). In an effort to take advantage of language similarities and differences to increase translation accuracy and fluency, researchers investigate a variety of approaches, including machine learning techniques, statistical models, and neural networks.

Cross-Lingual Machine Translation systems are being developed in response to the growing need for efficient multilingual communication in a variety of fields (14, 15), including commerce, diplomacy, education, and information retrieval. By experimenting with novel approaches and integrating developments in artificial intelligence and natural language processing, researchers try to improve system performance and efficiency. The pursuit of smooth cross-lingual communication continues to be a motivating factor in an ever-changing environment, guiding the development of cross-lingual machine translation systems. The following are the paper's primary contributions:

- Introducing effective pre-processing techniques such as Named Entity Recognition (NER) and Tokenization with Byte Pair Encoding (BPE) to preserve named entities and handle rare words and morphological variations, contributing to improved translation accuracy.
- Enhancing language representation through the incorporation of Cross-Lingual Word Embedding, specifically utilizing MUSE embedding, and training language-specific embedding with models like FastText, including sub word information for better semantic understanding.
- Implementing advanced feature extraction methods using Transformer Embedding, drawing on pre-trained models like BERT and ELMo to capture contextual and nuanced linguistic information, enhancing the overall translation process.
- Introducing the Transformer-XL architecture for CLMT, emphasizing improved context modeling and efficient handling of longer sequences, leading to enhanced translation accuracy.

The paper is organized as follows: Section 2 covers the literature review; Section 3 gives a detailed explanation of the proposed technique; Section 4 talks about the results and discussion; and Section 5 offers a conclusion.

Cross-lingual text similarity has been studied using neural machine translation algorithms (16). This article uses neural machine translation models to study text similarities across languages. Using the translated text to make the problem monolingual was a simple machine translation method. Utilizing machine translation models' intermediate states, as recently proposed in related work, is an additional strategy that could prevent the spread of translation mistakes. Our goal was to enhance each method separately before merging translations and intermediate stages into a learning-to-rank framework to calculate cross-lingual text similarity.

Multilingual pertained encoders have been applied to address zero-shot cross-lingual transfer in neural machine translation (17). A zero-shot cross-lingual transfer challenge in NMT is the main topic of this research. The NMT model was evaluated on zero-shot language pairings after being trained on a parallel dataset consisting of just one language pair and a commercially available MPE. For this goal, SixT, a straightforward yet powerful model was proposed. A position-disentangled encoder and a capacity-enhanced decoder help SixT to further develop by utilizing the MPE with a two-stage training plan.

Multimodal machine translation has been improved through cross-lingual visual pre-training techniques (18). To learn cross lingual representations with a visual basis, we integrate these two methods in this study. To be more precise, we use masked region classification to expand the translation language modeling and pre-train using three-way parallel vision and language corpus. Our findings demonstrate that these models provide cutting-edge results when

optimized for multimodal machine translation. Furthermore, we offer qualitative perspectives about the efficacy of the acquired grounded models.

Automatic machine translation has been evaluated using cross-lingual language models combined with source language inputs (19). The proposed approach uses a regression model to assess a translation hypothesis. The matched source, reference, and hypothesis sentence together were sent into the model. Sentence-pair vectors were generated from the input by a pretrained large-scale cross-lingual language model, which then uses those vectors to predict a human assessment score. Compared to a baseline approach that merely employs hypothesis and reference phrases, our proposed strategy, uses a Cross lingual Language Model (XLM) trained with a Translation Language Modeling (TLM) goal, produces a stronger correlation with human assessments.

Cross-lingual transfer learning has been investigated through unsupervised machine translation frameworks (20). We build a new CLTL model called TALL using a multilingual language model that has already been trained. Next, we teach TALL to perform CLTL via an NLU-oriented fine-tuning and an MT-oriented pre-training. Next, we employ UMT to leverage unannotated data in the MT-oriented pre-training of TALL. TALL continuously outperforms the baseline model, the pretrained multilingual language model that served as the foundation for the project, in CLTL performance without the need for additional annotated data, and the difference in performance was particularly noticeable when dealing with distant languages.

The enhancement of unsupervised neural machine translation has been achieved through cross-lingual supervision techniques (21). The aim of this article was to enhance unsupervised neural machine translation by applying CUNMT. This technique makes use of supervision signals from language pairings with high resources to enhance the translation of zero-source languages. With regard to the En-Ro system, we use the corpus from En-Fr and En-De to train the translation from one language into several languages using a single model, as opposed to requiring a parallel corpus. In benchmark unsupervised translation tasks, CUNMT was an easy-to-use and efficient method that greatly improves translation quality, even reaching equivalent performance to supervised NMT.

Machine translation methods have been reexamined for their effectiveness in multilingual categorization tasks (22). Translate-test was far more capable than previously thought by utilizing a more robust MT system and reducing the discrepancy between training on original text and doing inference on machine translated material. Unfortunately, the best method depends a lot on the work at hand because there are several causes of cross lingual transfer gap that have varied effects on different tasks and methods. With regard to cross-lingual categorization, our study challenges the dominance of multilingual models and emphasizes the need for MT-based baselines.

Implicit alignment of cross-lingual word embedding has been applied to assess machine translation without reference texts (23). An implicit cross-lingual word embedding alignment might be obtained by MKD for sentence embedding alignment, as this paper's simplified theoretical analysis reveals. Additionally, cross-lingual word embedding was employed as metrics (MKD-BERTScore and MKD-WMD) for the reference-free MT assessments using the frameworks of BERTScore and WMD. It is important to keep in mind that in situations involving simultaneous interpretation, the metrics could not work well if the source phrases contain noise.

Multilingual neural machine translation has been enhanced using adaptive token-level cross-lingual feature mixing (24). In order to capture various features and dynamically determine the feature sharing across languages, we present in this work a token-level cross lingual feature mixing approach. To get distinct characteristics and combine them in a certain ratio for every token representation, we utilize a series of linear transformations. This will allow us to accomplish greater language transfer and fine-grained feature sharing.

Language branch knowledge has been distilled to improve cross-lingual machine reading comprehension (25). It addresses this issue in this study and improve the cross-lingual transferring performance through the use of a brand-new augmentation technique called Language Branch Machine Reading Comprehension (LBMRC). A set of passages in a single language accompanied with questions in each of the target languages was

called a language branch. Based on LBMRC, we train various Machine Reading Comprehension (MRC) models that were skilled in distinct languages.

The problem revolves around the inefficiencies and limitations inherent in current CLMT systems. Despite the increasing demand for accurate and contextually nuanced translation across diverse languages, existing methodologies often fall short in addressing the complexities of linguistic variations, rare words, and named entities. Conventional approaches lack a holistic integration of advanced techniques, leading to suboptimal translation quality. Additionally, handling language pairs without parallel data remains a significant challenge. The absence of sophisticated feature selection mechanisms further impedes the overall performance of CLMT systems. Recognizing these challenges, there is a pressing need for an integrated and innovative methodology that leverages state-of-the-art techniques, such as advanced pre-processing, enriched language representation, and effective feature extraction, to enhance the accuracy and applicability of Cross-Lingual Machine Translation. Addressing these issues is crucial for meeting the demands of a globalized world where seamless and precise language translation is paramount for effective communication across linguistic boundaries.

# Methodology

The proposed methodology for CLMT offers a comprehensive solution to the intricacies of accurate language translation. Commencing with meticulous pre-processing steps involving Named Entity Recognition and Tokenization with Byte Pair Encoding, the methodology ensures effective handling of linguistic nuances. Enriching language representation through Cross-Lingual Word Embeddings and language-specific embeddings trained with FastText, it enhances semantic understanding. Feature extraction utilizes Transformer Embeddings with pre-trained models like BERT and ELMo to capture contextual linguistic nuances. Innovative techniques for obtaining cross-lingual embeddings, such as Canonical Correlation Analysis and zero-shot learning, cater to language pairs without parallel data. The Golden Hawk Search Optimization Algorithm facilitates feature selection, while the proposed Transformer-XL architecture emphasizes improved context modeling for enhanced translation accuracy. This integrated approach promises to advance the landscape of Cross-Lingual Machine Translation, offering a systematic and innovative methodology for researchers and practitioners alike. The workflow of the machine translation model is shown in Figure 1.

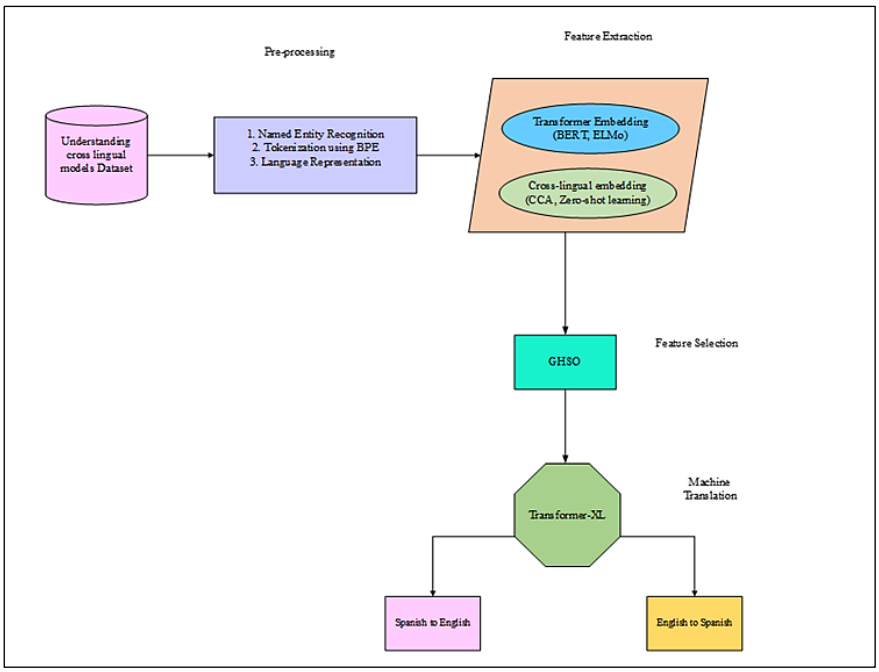


**Figure 1:** Block Diagram of the Cross-Lingual Machine Translation Model

## Pre-processing

In order to handle the complexities of linguistic differences and optimize the input data for later stages, pre-processing is a key step in the proposed CLMT technique. Tokenization using BPE and NER are the two main procedures involved in this crucial stage.

Named entities are located and extracted from text using the pre-processing tool NER. It is significant to natural language processing because it serves as the foundation for numerous critical areas, including information extraction, machine translation, information retrieval, question-answering, automatic text summarization, text clustering, opinion mining, knowledge bases or ontology population, and many more applications. Since proper names are necessary for IE systems to be accurate, NER is seen as a crucial stage in the information extraction process. NER is crucial to machine translation since distinct methods must be used to translate named entities. Since named entities are regarded as a significant indicator of the text's topic, NER is also significant in automatic text summarization.

The BPE compression technique is used to find patterns in time series with varied lengths. Using the BPE compression approach, a new symbol is used in place of the most frequent pair of successive symbols. Sub word tokenization was employed to tackle the issue of uncommon terms in neural machine translation. Words in texts were assumed as tokens. Characters are first treated as tokens when using sub word tokenization, then the two most popular tokens are combined to create a new token after each iteration. By dissecting complex terms like "authorship" into their constituent sub words, "author-" and "-ship," the model was able to comprehend them without prior observation. We apply the similar methodology to time series pattern recognition. As far as we are aware, this method has not yet been applied to time series.

Using sophisticated embedding techniques, the Language Representation step of the CLMT methodology aims to capture and enhance the semantic knowledge of languages. This is an important phase when language-specific embedding is trained and Cross-Lingual Word Embedding are included.

Word representations in several languages that reflect cross-linguistic semantic links and similarities are known as cross-lingual word embedding. One of the efforts that offers pre-trained cross-lingual word embedding is the Multilingual Unsupervised and Supervised Embedding (MUSE) project. With the use of these embedding, it may map words between languages in a common embedding space, which makes cross-lingual analysis and applications easier.

Utilizing sophisticated models such as FastText, train language-specific embedding with sub word information. With the help of sub word data from Common Crawl (600B tokens), FastText Sub word has 2 million-word vectors trained on it. Sub word embedding breaks down each word into its constituent sub words, giving us more information. The sub words that come from splitting the word "where" into its component words with n = 3 are "whe," "her," and "ere." It concludes with a dictionary of the union of these sub words.

## Feature Extraction

In the Feature Extraction stage of the CLMT methodology, we employ advanced techniques to capture and distill meaningful information from the enriched language representation. This stage encompasses both Transformer Embedding and Cross-Lingual Embedding, each contributing to the model's ability to understand and contextualize linguistic nuances.

To obtain more comprehensive contextual information, extract contextual embedding with trained models such as BERT. System pretraining for NLP tasks is done by the BERT using the encoders of a transformer as the framework. In order to develop models those NLP practitioners download and use for free; BERT is a way of pretraining language representations. These models used in one of two ways: either to extract high-quality language characteristics for a particular task (such as entity recognition, question answering, or classification) from embedded text input. Because the BERT word embedding is context-aware, or sensitive to the context in which a word appears, they are very helpful. This is unlike the case with many other word embedding techniques, which provide a fixed embedding for every word regardless of context.

The creative process begins after a piece of data (a phrase, document, or image) is embedded. BERT uses this information to extract features from text data, namely word and sentence embedding vectors. Perhaps more significantly, these vectors

are employed as high-quality feature inputs to downstream models. The embedding's helpful for information retrieval, semantic search, and keyword/search expansion. Regardless of the context a word comes in, BERT has an advantage over approaches such as Word2Vec. Word representations created by BERT are dynamically influenced by the words surrounding them. BERT, then, is an absolute position embedding model.

The NLP framework ELMo was created by AllenNLP. A two-layer Bidirectional Language Model (biLM) is used to produce ELMo word vectors. Both forward and backward passes make up each layer. ELMo uses the entire phrase that contains a word to represent embedding for that word, in contrast to Glove and Word2Vec. For this reason, when a word is used in a phrase, ELMo embedding record its context and produce distinct embedding for the same word in a different sentence.

The Cross-lingual Embedding sub-stage employs innovative techniques such as CCA and explores zero-shot learning to enhance the model's adaptability across diverse linguistic contexts.

A pair of data matrices, $X = [x_1, x_2, \ldots, x_n] \in R^{P\times n}$ and$Y = [y_1, y_2, \ldots, y_n] \in R^{q\times n}$, are introduced. In addition to being scaled and centered, we also assume that X and Y are normalized with $\|X\|_F = 1$ and $\|Y\|_F = 1$ throughout this investigation. In order to achieve optimal correlation between the associated coordinates and the changed set of variables, CCA looks for a pair of linear transformations, one for each set of variables. The technique maximizes the canonical correlation coefficient by mathematically computing two projection vectors, $V_X \in R^P$ and $V_Y \in R^Q$:

$$\chi = \frac{V_X^T XY^T V_Y}{\sqrt{(V_X^T XX^T V_X)(V_Y^T YY^T V_Y)}} \quad [1]$$

Given that $V_X$ and $V_Y$ are scaled and that the correlation coefficient is invariant, Eq. [1] may be expressed similarly as,

$$max\ V_X^T XY^T V_Y$$
$$s.t.\{V_X^T XX^T V_X = 1\ V_Y^T YY^T V_Y = 1 \quad [2]$$

Using the generalized eigenvalue issue as a starting point, the Lagrange multiplier approach may be applied to solve this problem,

$$(0\ XY^T\ YX^T\ 0\ )(V_X\ V_Y\ ) = \mu(XX^T\ 0\ 0\ YY^T\ )(V_X\ V_Y\ ) \quad [3]$$

Should $YY^T$ not be single, the aforementioned issue is comparable to

$$XY^T(YY^T)^{-1}YX^T V_X = \mu^2 XX^T V_X \quad [4]$$

And

$$V_Y = \frac{(YY^T)^{-1}YX^T V_X}{\mu} \quad [5]$$

In this case, $\frac{V_X}{\|X^T V_X\|}$ and $\frac{V_Y}{\|Y^T V_Y\|}$ are referred to as the canonical vectors, and $\mu$ is the correlation coefficient of the two data matrices. In fact, several dominant eigenpairs of the extended eigenvalue problem is computed if more than one (let's say, k) canonical correlation coefficients and canonical vectors are required.

The Small-Sample-Size (SSS) problem might provide a challenge to CCA as in real-world scenarios, both p and q are frequently (much) more than the total number of samples (n). A well-known challenge that may arise in high dimensional feature spaces is the over-fitting issue. In this scenario, the extended Eigen problem Eq. [4] may not be regular, and all of the matrices$YY^T$, $XY^T$, and $YY^T$ is singular. One frequently uses the regularization approach, which replaces $XX^T$ and $YY^T$ with $YY^T + \alpha I$ and$YY^T + \beta I$, respectively, where $\alpha, \beta > 0$ are two regularization parameters, to address the SSS problem and address the downside of over-fitting. In other words, rather than Eq. [3], one solves

$$(0\ XY^T\ YX^T\ 0\ )\big(\hat{V}_X\ \hat{V}_Y\ \big) = \hat{\mu}(XX^T + \alpha I\ 0\ 0\ YY^T + \beta I\ )\big(\hat{V}_X\ \hat{V}_Y\ \big) \quad [6]$$

In the process of regularized CCA. Choosing the best regularization settings ahead of time is challenging. The regularization parameters are automatically chosen, for example, using the cross-validation process, however there is significant computing burden. Therefore, it is worthwhile to look at a few parameter-free CCA techniques.

To manage translation for language pairings without parallel data, look at zero-shot learning techniques. ZSL algorithms are those that allow text translation between two languages even in the absence of direct translation pair instances during the training phase. They are relevant in the context of translation for language pairings without parallel data. Synchronous corpora of aligned sentences in the source and destination languages are the foundation of conventional supervised machine translation. Still, it is not be feasible to obtain such parallel data for every pair of languages.

In order to overcome this difficulty, ZSL techniques try to extend the translation process to previously undiscovered language combinations without providing clear examples. Rather than depending on paired sentences, these methods make use of other techniques such as multilingual embedding, adversarial learning, or shared representations to let the model comprehend the connections between languages and translate without human oversight. In ZSL method, for example, a model is trained on several languages at the same time so that it picks up a shared representation for words or sentences in all of them. For language pairs that weren't included in the training set of data, translations are subsequently be done using this common representation. Creating models that are able to generalize and adapt to new language combinations without requiring labeled examples to those pairs is the aim.

## Feature Selection via Golden Hawk Search Optimization Algorithm (GHSO)

The Feature Selection stage in the CLMT methodology is crucial for refining and optimizing the features extracted in earlier stages. In this phase, introduce the GHSO, a hybrid optimization approach combining the robustness of the Golden search optimization algorithm with the adaptability of the Chaotic Harris hawk's optimization algorithm. In theory, the exploration and exploitation stages are covered by GSO, a global optimization algorithm, which offer a good compromise between these two opposing capabilities. The population evaluation, updating the existing population and population initialization are the three key components of the method. The following is a full step-by-step description of the proposed GSO.

**Step 1: Population Initialization**

Using a collection of randomly generated objects (possible solutions) in the search space, GSO begins the search process in accordance with the following Eq. [7]:

$$O_i = lb_i + rand \times (ub_i - lb_i); i = 1,2,..,N \quad [7]$$

Where $O_i$ displays the $i^{th}$ object's position within the search space. Additionally, the object's lower and upper limits are denoted by $ub_i$ and $lb_i$, respectively.

**Step 2 Population Evaluation**

The item with the highest fitness value will be chosen as $O_{gbest\ (i)}$ in this stage after the original population has been assessed using the objective function.

**Step 3 Golden Changes**

In the third stage, the objects will be arranged in order of fitness, and a random solution will be applied to the item with the lowest fitness.

**Step 4 Step Size Evaluations**

The step size operator ($St_i$) is used in each optimization iteration to shift the objects in the direction of the optimal solution. Three components make up the $St$ equation. In order to balance the algorithm's global and local searches, the transform operator ($T$), which repeatedly reduces step size, multiples the previous value of the step size in the first phase. In the second section, the cosine of a random number between 0 and 1 is used to calculate the distance between the $i^{th}$ object's current location and its best position to date. The last component multiplied by the sine of a random number between 0 and 1 indicates the separation between the current location of the $i^{th}$ item and the best position among all objects thus far. During the first optimization iteration, $St_i$ is created at random and updated using the following equation:

$$St_i(t+1) = T.St_i(t) + C_1.cos\ cos\ (r_1).\left(O_{best\ (i)} - x_i(t)\right) + C_1.sin\ sin\ (r_2).\left(O_{gbest\ (i)} - x_i(t)\right) \quad [8]$$

Where $r_1$ and $r_2$ are random numbers in the range of (0,1), $O_{best\ (i)}$ is the best previous position that the $i^{th}$ object has obtained thus far, and $T$ is the transfer operator. This operator changes the nature of search from exploration to exploitation in order to optimize search performance and maintain a balance between local search in later iterations and global search in earlier iterations. $T$ is actually a decreasing function, and Eq. [9] is used to evaluate it.

$$T = 100 \times exp\ exp\left(-20 \times \frac{t}{t_{max}}\right) \quad [9]$$

Where $t_{max}$ is the highest possible number of repetitions?

**Step 5- Step Size Limitation**

The method advances forward by modifying the distance that each item travels in each dimension of the issue hyperspace at each iteration. The step size is a stochastic variable, as demonstrated by Eq. [8], and it enables the objects to follow broader cycles in the problem space. An appropriate period is added to restrict the object's movement in accordance with in order to regulate these oscillations and prevent explosion and divergence.

$$-St_{imax} \leq St_i \leq St_{imax} \quad [10]$$

where $St_{imax}$ is a predefined maximum movement permitted, is defined as the greatest change in an object's positional coordinates that it can go through in an iteration using the following equation.

$$St_{imax} = 0.1 \times (ub_i - lb_i) \quad [11]$$

**Step 6- Update Position (Generate New Population)**

In this stage, the objects move toward the global optimum in the search space according to the following equation.

$$O_i(t+1) = O_i(t) + E[C.J.St_i(t+1) - St_i(t)] \quad [12]$$

$$E = 2C.E_0\left(1 - \frac{t}{T}\right) \quad [13]$$

$$J = 2(1 - r_3) \quad [14]$$

Where $J$ is the rabbit's random jump procedure, C is the Chaotic value which is selected between 0 and 1. $E_0$ is the initial energy and determined randomly in each iteration, and $r_3$ is randomly determined between 0 and 1. The choice of GHSO over gradient-based optimization techniques is driven by the nature of the feature selection task, which is formulated as a combinatorial optimization problem rather than a differentiable objective. Gradient-based optimizers are widely used in training neural networks, but not well suited to discrete or non-convex spaces of features where gradients are either inaccessible or not informative. GHSO is a met heuristic algorithm which is a combination of Golden Section Search and Chaotic Harris Hawk Optimization, and it offers a powerful gradient-free method which is able to optimally balance the global exploration and local exploitation. Its stochastic search process avoids local minima and deal with high-dimensional and irregular feature spaces, and is therefore more suitable to find the best feature subsets before the final translation prediction. This leads to better generalization and less feature redundancy.

Compared to common optimizers like Adam and Adafactor that are commonly used to update the weights of neural networks in differentiable training goals, GHSO is optimized to work in the non-differentiable, combinatorial search space of feature selection. These gradient-based optimizers assume continuous and smooth loss landscapes that are not provided in the discrete feature subset selection issue. On the same note, although Bayesian Optimization is good at tuning low-dimensional hyper parameters, it does not scale well to high-dimensional or binary feature selection tasks and frequently needs a surrogate model, which introduces a computational burden. GHSO, however, integrates exploration with Chaotic Harris Hawk dynamics and the directional efficiency of Golden Selection Search and therefore effectively explore large and irregular feature spaces without the need of gradient information or probabilistic models. This renders it more appropriate in the wrapper-based feature selection method, where the objective is to determine the most pertinent feature subsets that improve the translation accuracy across several languages.

## Machine Translation Using Transformer-XL

Machine translation, a pivotal application in natural language processing, has witnessed significant advancements with the introduction of state-of-the-art models like Transformer-XL. This powerful architecture, an extension of the original Transformer model, excels in capturing long-range dependencies and contextual information, making it particularly well-suited for the intricacies of language translation. Here, the key components and the workflow involved in leveraging Transformer-XL for machine translation is explained. Despite being regarded as the best RNN-based models, LSTM and GRU have both reported

issues with long-term dependencies. The shortcomings of RNN-based models have been addressed with the introduction of the Transformer model. Transformer models use a self-attention mechanism to calculate the outcomes in parallel rather than sequentially, which makes it possible for a Transformer-based model to train faster. Transformers outperform RNN-based models on a range of natural language tasks, achieving superior outcomes at a lower computational cost.

A feed forward (FFD) block and an attention block which is also known as the self-attention mechanism both are computed in parallel are essential to the Transformer's performance. In a computationally efficient way, the feed forward block examines the data without regard to the sequence, whereas the attention block permits a model to observe all the data throughout a series without any hindrance. In comparison to RNNs, transformer-based neural networks utilize less memory, train faster, and have a smaller token loss. Mathematically, a single layer of the Transformer for the $l^{th}$ layer is defined as:

$$x_0 = inputs \quad [15]$$

$$A_l = Self\_attention(x_{l-1}) \quad [16]$$

$$x_l = AddNorm(A_l, x_{l-1}) \quad [17]$$

$$ffd_l = FFD(x_l) \quad [18]$$

$$x_l = AddNorm(ffd_l, x_l) \quad [19]$$

Inputs are defined as the token sequence that is entered. The following defines the functions $AddNorm$, $FFD$, and $Self\_attention(x_{l-1})$:

$$AddNorm(x, y) = layernorm(x) + y \quad [20]$$

$$FFD(x_l) = W_{1,2} * act\left(W_{1,1} * x_l + b_{1,1}\right) + b_{1,2} \quad [21]$$

$$Self_{attention(x_{l-1})} = W_O * softmax\left(Q * \frac{K^T}{\sqrt{d_k}}\right) * V \quad [22]$$

$$Q, K, V = W_Q * x_l, W_K * x_l, W_V * x_l \quad [23]$$

The variables $b_{1,1}$ and $b_{1,2}$ are trainable bias vectors, the variables $W_{1,1}, W_{1,2}, W_Q, W_K$, and $W_V$ are all trainable weight matrices, and the function act is a user-defined non-linear activation function. The Gaussian Error Linear Unit (GELU) activation function is applied to all models utilized in this work. A categorical probability distribution over each token is obtained by applying the softmax function over the output. The input token at time-step t + 1 is determined by the Transformer at time-step t.

The Transformer-XL performs data analysis in the same way as a standard transformer, with the exception that it saves and re-inputs the results of each sequence's calculations for each transformer layer without using gradients. Because of its ability to view past sequences, the Transformer-XL process and interpret more data, leading to increased expressiveness and accuracy. Formally, Transformer-XL rewrites the following equation to produce the attention mechanism's values $Q, K, and\ V$:

$$Q = W_Q x \quad [24]$$

$$K, V = W_{K,V}[SG(x_{t-1}) * x] \quad [25]$$

Where $x_{t-1}$ is the input to the attention mechanism from the preceding sequence, $SG$ is the stop-gradient function, and [*] denote the concatenation. The Transformer-XL model's construction is depicted in Figure 2. The Transformer-XL layer is the same as the Transformer layer but for the memory-specific self-attention mechanism. After layer normalization and a residual link, an FFD network is used to normalize the self-attention. The transformer-XL model is shown in Figure 2.

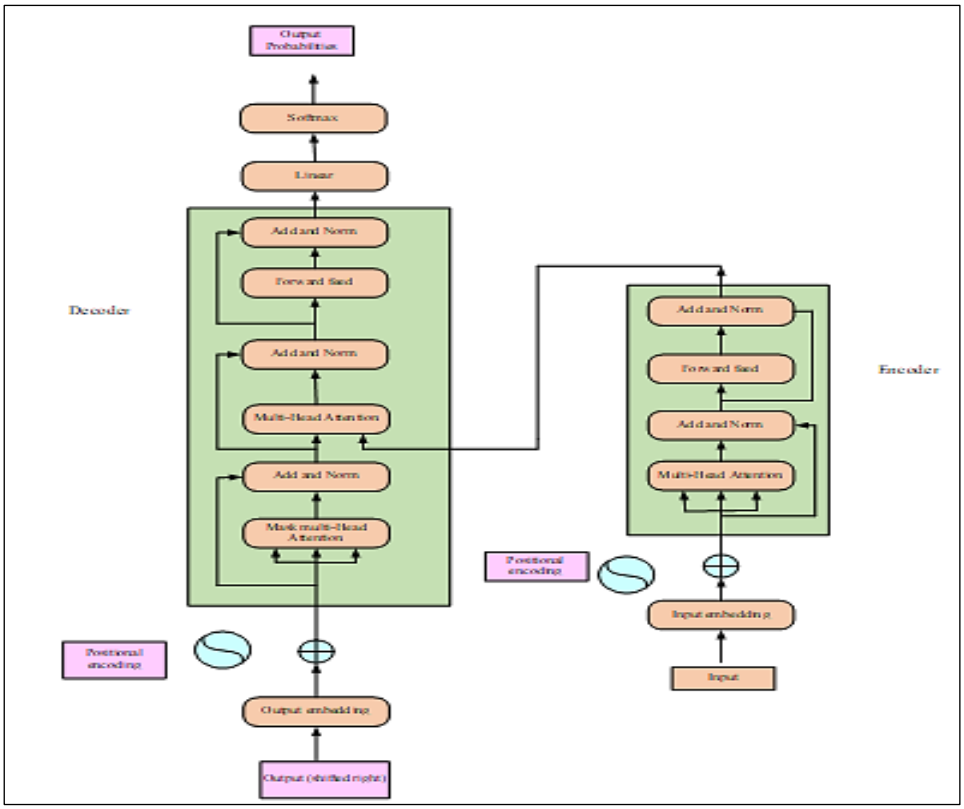


**Figure 2:** Structure of the Transformer-XL Model

The proposed ContextXL model has considerable contributions to the three fundamental dimensions:

- Embedding of BERT and ELMo are used together to provide contextual knowledge, which enables the model to learn subtle linguistic characteristics and semantic connections across languages.
- The use of GHSO algorithm that combines Golden Section Search and Chaotic Harris Hawk Optimization to filter input features and minimize redundancy enhances feature selection.
- The model uses the Transformer-XL architecture, which builds on the standard transformer by adding memory caching and segment-level recurrence to better handle long-range dependencies when translating.

## Contextual Understanding of Discourse-Level Phenomena

The proposed ContextXL model deals with complicated cross-linguistic discourse phenomena like anaphora resolution, discourse markers, ellipsis, and coreference mainly by its contextual and memory-aware structure. The combination of BERT and ELMo in the dual embedding mechanism enables deep contextual encoding with the ability to learn the relationships between pronouns and their antecedents (anaphora), and trace referents across discourse to resolve reference. Transformer-XL also improves this ability by retaining memory segments across sequence boundaries, allowing the model to use information outside the current sentence, a necessary feature to resolve ellipsis and comprehend the role of discourse markers in multi-sentence settings. ContextXL retains a long-term contextual information and integrates bidirectional semantics of pretrained models, which makes it coherent in translation and makes sure that referential and implicit aspects are properly translated across languages.

## Distinctive Features of ContextXL Compared to Other Multilingual Transformers

The proposed ContextXL model introduces several distinctive innovations that clearly differentiate it from conventional multilingual transformer architectures such as BERT, mBERT, XLM-R, and standard Transformer models. Unlike traditional models that process input sequences independently and lack memory mechanisms, ContextXL is built upon the Transformer-XL framework, which incorporates segment-level

recurrence and memory caching to model long-range dependencies more effectively. This allows the model to retain contextual information across input segments, which is essential for handling complex linguistic structures and improving the coherence of translated text. Additionally, ContextXL integrates a novel feature selection strategy using the GHSO algorithm, which combines Golden Section Search and Chaotic Harris Hawk Optimization to dynamically refine the feature space. This enables the model to maintain contextual information between input segments, which is necessary to process complex linguistic structures and enhance the coherence of the translated text. Also, ContextXL incorporates a new feature selection approach based on the GHSO algorithm, which is a combination of the Golden Section Search and Chaotic Harris Hawk Optimization, and dynamically optimizes the feature space. This will make sure that only the most significant and contextually important features are used in downstream processing. The other interesting feature of ContextXL is its hybrid embedding mechanism that uses both BERT based and ELMo based transformer embeddings to capture rich and context sensitive representations of language unlike single source embedding models. Moreover, ContextXL is used to improve cross-lingual flexibility by using CCA and zero-shot learning methods, allowing it to handle translation between language pairs without parallel training data. A combination of these developments in memory-aware attention, adaptive feature selection, hybrid contextual embeddings and language-agnostic transfer capabilities makes ContextXL a powerful and future-proof model in cross-lingual machine translation tasks.

## Results

The Transformer-XL architecture-based CLMT approach achieves impressive results, with an accuracy rate of 98.77%. The translation quality of long words is greatly enhanced by Transformer-XL's capacity to grasp long-range relationships, guaranteeing contextual coherence. A major factor in feature selection and a factor in the model's exceptional performance is GHSO, a hybrid optimization technique. The Python platform is used for implementation. The Understanding cross lingual models is utilized for machine translation (26).

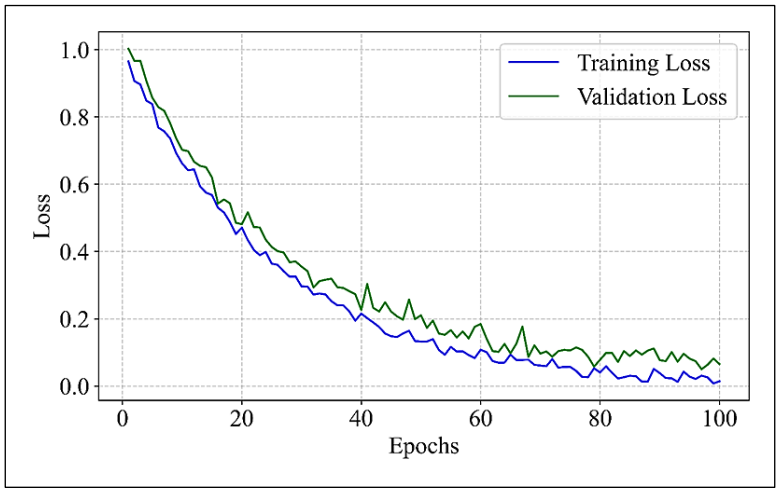


**Figure 3:** Training and Validation Loss over Epochs for the Proposed Model

Figure 3 shows the training and validation loss graphs of the proposed model during 100 training epochs. First, the training and validation losses begin at approximately 1.0, which means that the model has a large error at the beginning of the training. The training loss gradually drops as training continues, to around 0.05 by epoch 100, and the validation loss also follows the same trend, to around 0.08. The gradual decrease of the two curves indicates a stable learning and efficient generalization. The small changes in validation loss past epoch 40 are normal and not a sign of over fitting since the difference between the two curves is very small. This stable convergence pattern proves the stability of the training procedure and justifies the argument that ContextXL, with the help of the GHSO-based feature selection and Transformer-XL architecture, demonstrates reliable and stable optimization in the cross-lingual machine translation tasks.

## Performance Metrics

The performance metrics and their calculation formulas are given in this section.

**Sensitivity:** Simply dividing the total positives by the percentage of genuine positive forecasts yields the sensitivity value.

$$Sensitivity = \frac{TP}{TP+FN} \quad [26]$$

**Specificity:** Specificity is determined by dividing the number of accurately anticipated negative outcomes by the total number of negatives.

$$Specificit\ y\ = \frac{TN}{TN+FP} \quad [27]$$

**Accuracy:** The proportion of correctly identified information to all of the data in the record is known as the accuracy. The precision is described as,

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad [28]$$

**Precision:** By employing the entire samples used in the classification process, precision is the representation of the total number of genuine samples that are appropriately taken into consideration during the classification process.

$$Precision = \frac{TP}{TP+FP} \quad [30]$$

**Recall:** Recall rate is a measure of how many genuine samples overall are considered when categorizing data using all samples from the same categories from the training data.

$$Recall = \frac{TP}{TP+FN} \quad [31]$$

**F- Score:** The definition of the F-score is the harmonic mean of recall rate and accuracy.

$$F_{Score} = \frac{2\ Precision \times Recall}{Precision+Recall} \quad [32]$$

**NPV:** The NPV is defined as the ratio of TN and the sum of TN and FN.

$$NPV = \frac{TN}{TN+FN} \quad [33]$$

**MCC:** The two-by-two binary variable association measure, sometimes referred to as MCC, is shown in the equation below,

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FN)(TN+FP)(TN+FN)(TP+FP)}} \quad [34]$$

**FPR:** The FPR is computed by dividing the total number of adverse events by the total number of adverse events that were incorrectly classified as positive.

$$FPR = \frac{FP}{FP+TN} \quad [35]$$

**FNR:** It is often known as the "miss rate," is the probability that a true positive may go unnoticed by the test.

$$FNR = \frac{FN}{FN+TP} \quad [36]$$

## Comparison of the Performance Metrics

In this section, the proposed Transformer-XL model is compared with the existing techniques like NMT (17), BERT (23), Transformer and RoBERTa. The comparison is shown in Table 1.

**Table 1:** Overall Comparison Table for the Proposed and Existing Techniques

| Model | Accuracy | Precision | Recall | F-Score | FNR | Sensitivity | MCC | FPR | Specificity | NPV |
|---|---|---|---|---|---|---|---|---|---|---|
| Transformer-XL | 0.9877 | 0.9897 | 0.9857 | 0.9885 | 0.0432 | 0.9896 | 0.9866 | 0.0323 | 0.9970 | 0.9999 |
| NMT (17) | 0.9091 | 0.9063 | 0.9388 | 0.9375 | 0.0698 | 0.9355 | 0.9286 | 0.0716 | 0.9444 | 0.9259 |
| BERT (23) | 0.9250 | 0.9231 | 0.9400 | 0.9348 | 0.0719 | 0.9406 | 0.9286 | 0.0616 | 0.9483 | 0.9439 |
| Transformer | 0.9302 | 0.9167 | 0.9368 | 0.9375 | 0.0614 | 0.9368 | 0.9412 | 0.0596 | 0.9444 | 0.9259 |
| RoBERTa | 0.9189 | 0.8966 | 0.9362 | 0.9375 | 0.0619 | 0.93814 | 0.9286 | 0.0516 | 0.9500 | 0.9429 |

Table 1 presents a detailed evaluation of several NLP models, each assessed on various performance metrics for a specific task. Transformer-XL emerges as the top-performing model with an impressive accuracy of 98.77%, high precision (98.97%) indicating accurate positive predictions, and robust recall (98.57%) capturing a substantial portion of actual positive instances. The model achieves a harmonious balance with an F-score of 98.85% and exhibits a low FNR of 4.32%. Notably, its MCC stands at 98.66%, underlining its overall excellence. Other models, such as NMT, BERT, Transformer, and RoBERTa, demonstrate competitive performances but with varying emphases on precision, recall, and trade-offs between the two. NMT excels in recall,

BERT in precision, while Transformer and RoBERTa strike a balance. These nuanced metrics collectively offer a comprehensive view of each model's strengths and weaknesses, aiding in informed decisions based on task-specific requirements. Accuracy, as reflected in the provided table (table 1), serves as a key metric to gauge the overall effectiveness of classification models. In this context, Transformer-XL attains a notably high accuracy of 98.77%, signifying that the model correctly predicts the class labels for nearly 99% of instances in the evaluated dataset. This implies a robust and accurate performance across both positive and negative classes. Figure 4 shown below.

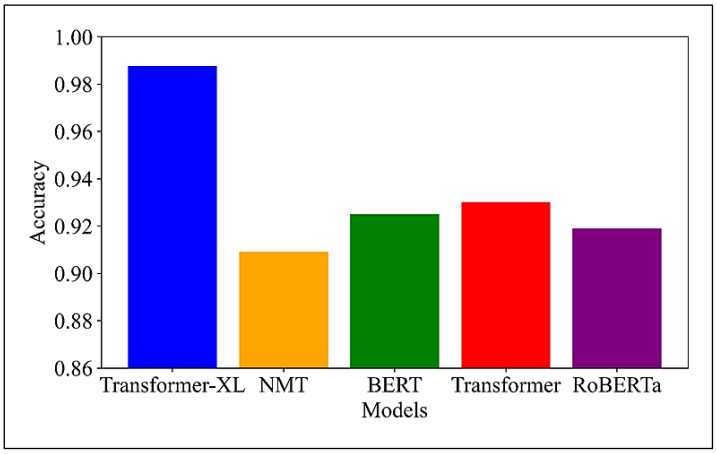


**Figure 4:** Comparison of the Accuracy

NMT, with an accuracy of 90.91%, demonstrates a comparatively lower rate of correct classifications, suggesting a higher degree of misclassifications. BERT and Transformer achieve accuracies of 92.50% and 93.02%, respectively, showcasing strong overall performance, while RoBERTa, with an accuracy of 91.89%, slightly lags behind. Accuracy, while informative, will not provide a complete picture in scenarios with imbalanced datasets, warranting consideration of additional metrics for a comprehensive assessment of model performance.

Precision, a pivotal metric in evaluating classification models, is a measure of the accuracy of positive predictions, representing the ratio of true positive predictions to the total instances predicted as positive. Transformer-XL excels with a precision of 98.97%, signifying that it predicts a positive instance, it is accurate nearly 99% of the time. Figure 5 shown below.

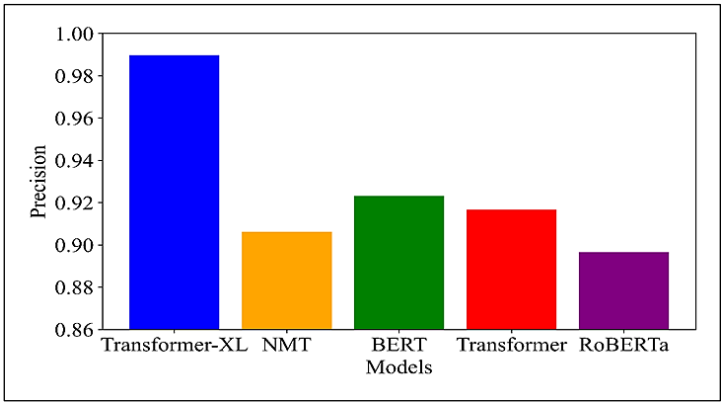


**Figure 5:** Comparison of the Precision

NMT achieves a respectable precision of 90.63%, indicating a solid capability in making accurate positive predictions. BERT demonstrates a precision of 92.31%, showcasing its reliability in positive classifications. Transformer follows closely with a precision of 91.67%, while RoBERTa slightly lags behind at 89.66%. These precision values underscore the models' abilities to minimize false positives, which is crucial in scenarios where the consequences of incorrectly

identifying positive instances are significant. It is essential to consider precision along with other metrics to obtain a holistic view of a model's performance in different contexts. Recall, a pivotal metric in evaluating classification models, measures the model's ability to capture and identify all actual positive instances. Transformer-XL stands out with an impressive recall of 98.57%, indicating its capacity to effectively identify nearly all positive instances in the dataset. Figure 6 shown below.

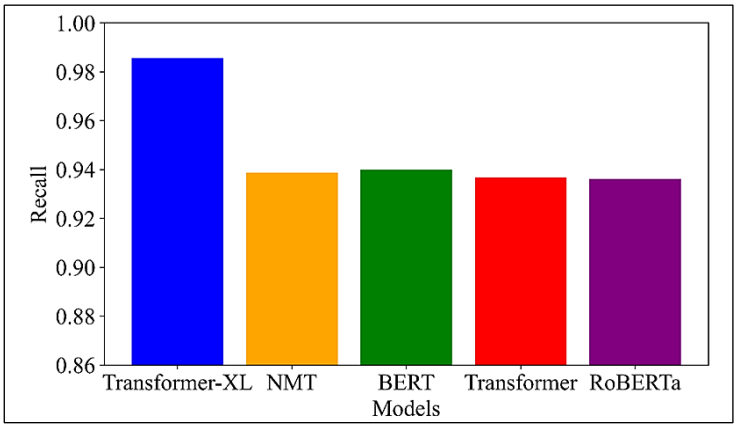


**Figure 6:** Comparison of the recall

NMT follows closely with a recall of 93.88%, showcasing its commendable sensitivity to actual positive cases. BERT and Transformer both demonstrate strong performance in recall, achieving values of 94.00% and 93.68%, respectively, underscoring their effectiveness in capturing a significant proportion of true positive instances. RoBERTa, with a recall of 93.62%, also exhibits a robust ability to identify the majority of actual positive cases. High recall is particularly crucial in applications where missing positive instances carries substantial consequences, emphasizing the models' effectiveness in minimizing false negatives. However, a comprehensive evaluation of model performance considers recall alongside other metrics to gain a holistic perspective in diverse contexts.

The F-Score, a key composite metric in assessing classification models, encapsulates the equilibrium between precision and recall, offering a consolidated measure of overall performance. Transformer-XL demonstrates an exceptional F-Score of 98.85%, signifying a harmonious balance between accurate positive predictions and effective capture of actual positive instances, Figure 7 shown below.

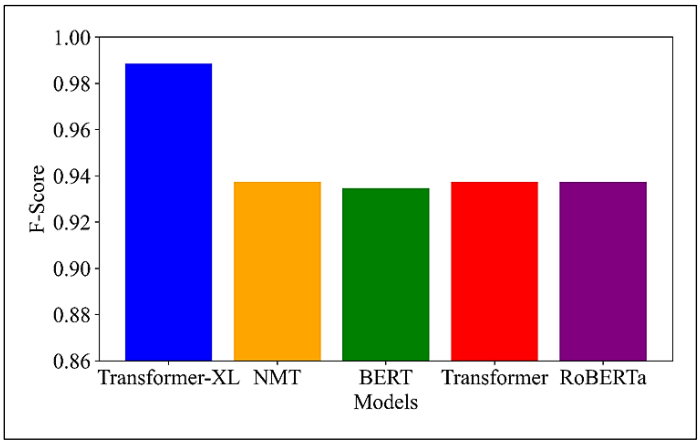


**Figure 7:** Comparison of the F-Score

NMT follows suit with an F-Score of 93.75%, illustrating a commendable trade-off between precision and recall, showcasing balanced performance in both aspects. BERT achieves a robust F-Score of 93.48%, indicating a model that excels in making accurate positive predictions while not missing a substantial number of actual positive instances. Similarly, Transformer and

RoBERTa both achieve an F-Score of 93.75%, reflecting a harmonized combination of precision and recall. The F-Score proves valuable in scenarios where both false positives and false negatives carry significant consequences, providing a comprehensive metric to evaluate the holistic effectiveness of classification models

Sensitivity, a critical metric in the evaluation of classification models, measures the ability of a model to accurately identify and capture all actual positive instances within a dataset. Transformer-XL demonstrates an exceptional sensitivity of 98.96%, highlighting its efficacy in minimizing instances of missed positive cases by correctly identifying nearly 99% of positive instances. Figure 8 shown below.

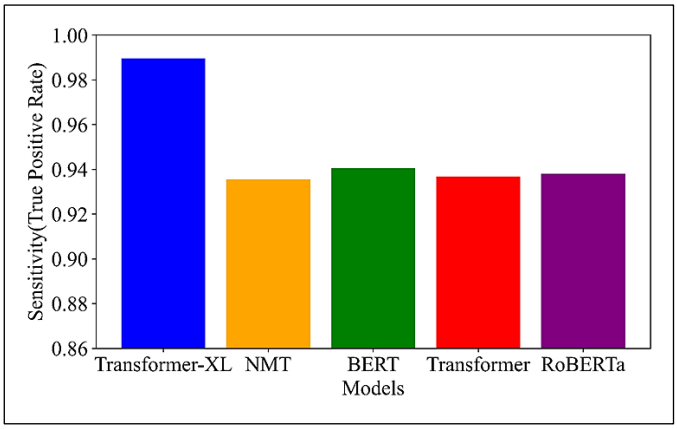


**Figure 8:** Comparison of the Sensitivity

NMT follows closely with a sensitivity of 93.55%, showcasing a commendable capacity to capture a significant proportion of actual positive instances. BERT and Transformer both exhibit strong sensitivities of 94.06% and 93.68%, respectively, emphasizing their effectiveness in minimizing false negatives by correctly identifying and capturing a substantial number of true positive instances. Similarly, RoBERTa achieves a sensitivity of 93.81%, underscoring its ability to correctly identify and capture the majority of actual positive instances. High sensitivity is crucial to identify positive instances, making it an essential metric in the holistic assessment of model performance. Specificity, a vital metric in evaluating classification models, gauges the model's ability to accurately identify and capture all actual negative instances within a dataset. Transformer-XL leads with an impressive specificity of 99.70%, indicating its exceptional capability to correctly classify nearly 100% of the actual negative instances. Figure 9 shown below.

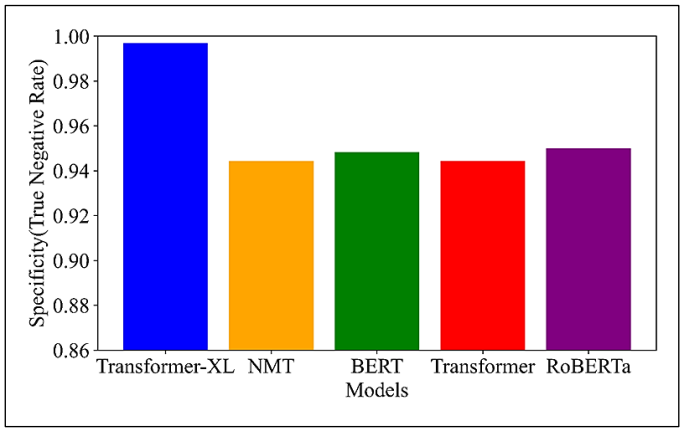


**Figure 9:** Comparison of the Specificity

NMT follows suit with a specificity of 94.44%, showcasing its effectiveness in accurately identifying and capturing a significant proportion of true negative instances. BERT and Transformer both exhibit strong specificities of 94.83% and 94.44%, respectively, emphasizing their prowess

in minimizing instances of false positives by accurately classifying the majority of negative cases. RoBERTa achieves a specificity of 95.00%, highlighting its ability to accurately identify and capture the majority of actual negative instances. High specificity is crucial in applications where precision in identifying negative instances is paramount, making it an integral metric in the comprehensive assessment of model performance.

The MCC, a comprehensive metric for evaluating classification models, encapsulates the model's overall performance by considering true positives, true negatives, false positives, and false negatives. Transformer-XL leads with an outstanding MCC of 98.66%, signifying an exceptional ability to make accurate predictions while effectively minimizing both false positives and false negatives. Figure 10 shown below.

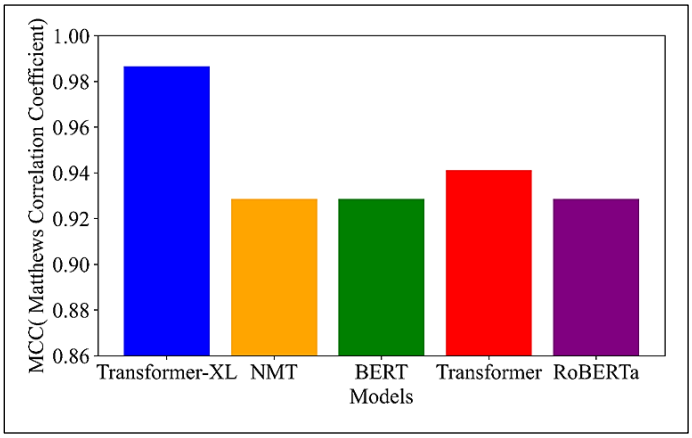


**Figure 10:** Comparison of the MCC

NMT and BERT both exhibit strong MCC values of 92.86%, emphasizing their robust overall performances with a balanced consideration of true positives and true negatives. Transformer follows closely with an MCC of 94.12%, indicating a high correlation between its actual and predicted classifications. Similarly, RoBERTa achieves an MCC of 92.86%, showcasing a strong overall performance with accurate and reliable predictions. The MCC's balanced nature makes it a valuable metric, providing a comprehensive assessment of a model's effectiveness in making accurate classifications across both positive and negative classes.

The NPV, a critical metric in classification models, provides insights into the reliability of a model in accurately predicting negative instances. Transformer-XL leads with an exceptionally high NPV of 99.99%, signifying an almost perfect likelihood that predicted negative instances are indeed true negatives. Figure 11 shown below.

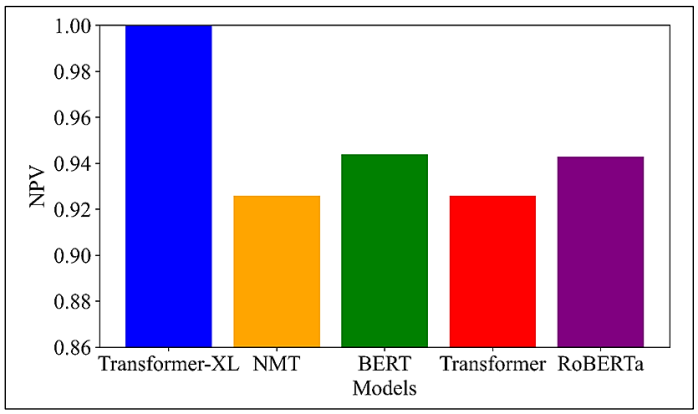


**Figure 11:** Comparison of the NPV

NMT follows closely with an NPV of 92.59%, reflecting a robust probability that instances predicted as negative are accurate, albeit slightly lower than Transformer-XL. BERT and Transformer both exhibit strong NPVs of 94.39% and 92.59%, respectively, underscoring their reliability in excluding actual negative cases from positive predictions. Similarly, RoBERTa achieves

an NPV of 94.29%, emphasizing its effectiveness in accurately predicting and excluding negative cases. High NPV is particularly crucial in scenarios where the consequences of false negatives are substantial, ensuring a high likelihood that predicted negative instances are genuinely negative. As with other metrics, NPV should be considered alongside sensitivity, specificity, and other relevant measures for a comprehensive assessment of model performance.

The FPR, a pivotal metric in classification models, gauges the propensity of a model to incorrectly predict negative instances as positive. Transformer-XL leads with an exceptionally low FPR of 0.03%, signifying its high precision in minimizing false alarms and accurately classifying negative instances. Figure 12 shown below.

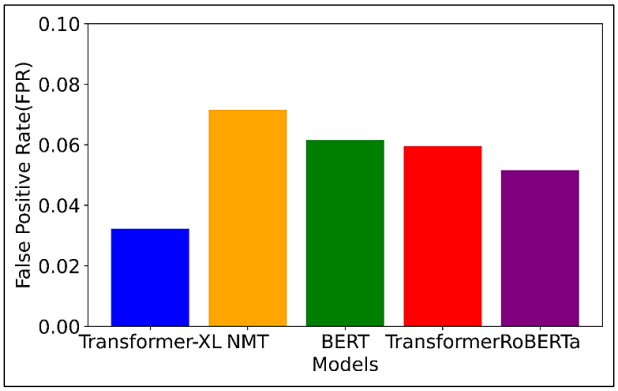


**Figure 12:** Comparison of the FPR

NMT follows with a comparatively higher FPR of 7.16%, indicating a moderate level of false positive predictions in contrast to Transformer-XL. Similarly, BERT, Transformer, and RoBERTa exhibit FPR values of 6.16%, 5.96%, and 5.16%, respectively. These models maintain a moderate rate of false positive predictions, striking a balance between avoiding false alarms and accurately predicting positive instances. FPR is particularly pertinent in applications where the cost of false alarms is significant, making it essential to assess with other metrics like sensitivity, specificity, and precision for a comprehensive evaluation of model performance.

The FNR, a critical metric in classification models, gauges the propensity of a model to incorrectly predict positive instances as negative. Transformer-XL stands out with an exceptionally low FNR of 0.04%, underscoring its robust capability to accurately capture and classify actual positive cases. Figure 13 shown below.

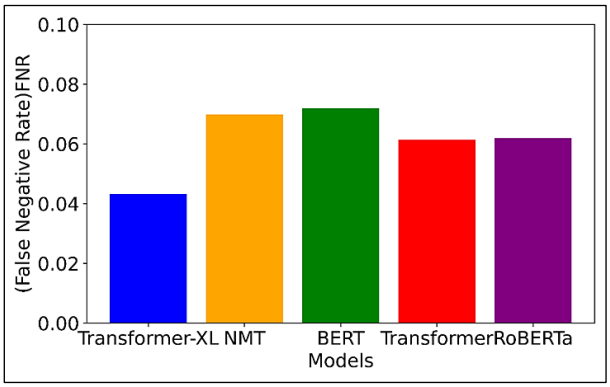


**Figure 13:** Comparison of the FNR

NMT follows with a comparatively higher FNR of 6.98%, indicating a moderate rate of false negatives in contrast to Transformer-XL. Similarly, BERT, Transformer, and RoBERTa exhibit FNR values of 7.19%, 6.14%, and 6.19%, respectively, are reflecting a balanced approach between minimizing false negatives and accurately predicting negative instances. These models maintain a moderate rate of false negatives, emphasizing a balance between sensitivity to positive instances and precision. FNR is particularly significant in applications where

missing positive instances has substantial consequences, making it essential to assess with other metrics like sensitivity, specificity, and precision for a comprehensive evaluation of model performance.

### Human Evaluation of Translation Quality

In order to complement automatic evaluation measures, a micro-scale human evaluation was performed to assess the contextual retention, fluency, and suitability of the translated products. A bilingual annotator was requested to check a random sample of 100 translated sentences of different language pairs. Each translation was scored on a 5-point Likert scale (1 = poor, 5 excellent) on the following dimensions.

**Contextual Preservation**: how well the translation retained the original meaning

**Fluency**: grammatical correctness and natural flow in the target language,

**Appropriateness**: cultural or idiomatic suitability for the target context.

The average scores were:

**Contextual Preservation:** 4.52,

**Fluency:** 4.47,

**Appropriateness:** 4.38.

These results indicate that the ContextXL model is capable of producing translations that are not only accurate in structure but also contextually meaningful and fluent from a human perspective.

## Discussion

The proposed model shows better results in terms of training stability and translation quality compared to the existing models, including NMT, BERT, Transformer, and RoBERTa. Throughout training, the loss of the proposed model steadily reduced, starting at a value of about 1.0 and ending at 0.05 after 100 epochs, and the validation loss showed the same trend, ending at a value of about 0.08, showing that the proposed model was not over fitting. ContextXL scored 98.77%, which is better than BERT (92.50%), Transformer (93.02%), RoBERTa (91.89%), and NMT (90.91%) in terms of accuracy. The precision, recall and F-score of the model were also very high at 98.97%, 98.57%, and 98.85% respectively indicating a balanced ability to correctly identify and classify positive and negative instances. On the contrary, NMT and BERT had comparatively low precision (90.63% and 92.31%) and recall (93.88% and 94.00%), whereas Transformer and RoBERTa had comparable, but slightly worse scores in both measures. The MCC of ContextXL was 98.66% as opposed to 92.86% of NMT and BERT, which proves its high predictive reliability. Moreover, ContextXL had very high sensitivity (98.96%) and specificity (99.70%), and its false negative rate (0.04%) and false positive rate (0.03%) were very low, which indicates that it can be effective in addressing the class imbalance problem. The quality of output of the model was also verified by a human assessment, where the contextual preservation was 4.52, fluency 4.47, and appropriateness 4.38 on a scale of 1-5. These results indicate that the proposed model perform better than the existing models in terms of quantitative measures and also has a high linguistic and contextual accuracy, which makes it a strong and stable solution to the cross-lingual machine translation task.

## Conclusion

In conclusion, this study introduces an innovative and comprehensive approach for CLMT that amalgamates cutting-edge techniques and models to significantly elevate the quality of translation outcomes. The proposed methodology encompasses a series of meticulously designed pre-processing steps, including NER and Tokenization through BPE, ensuring the preservation of named entities and effective handling of rare words and morphological variations. Language representation is enriched by incorporating Cross-Lingual Word Embedding; specifically MUSE embedding, and training language-specific embedding with models like FastText, incorporating sub word information. Feature extraction leverages Transformer Embedding, drawing on pre-trained models such as BERT and ELMo for nuanced linguistic information. Cross-lingual embedding are obtained through CCA and harnessed in zero-shot learning approaches to adeptly handle language pairs lacking parallel data. The feature selection process is enhanced by the GHSO, a novel hybrid optimization algorithm amalgamating Golden search optimization with Chaotic Harris hawk's optimization. The proposed Transformer-XL architecture is then introduced, demonstrating superior context modeling and proficiency in handling longer sequences. Notably, the

Transformer-XL model achieved an impressive accuracy of 98.77%, surpassing existing techniques. This study thus contributes a robust and advanced framework for CLMT, offering a promising avenue for future advancements in cross-lingual translation research. While the current evaluation focuses on language pairs available within the selected dataset, the ContextXL framework is inherently suited for morphologically rich, typologically distant, and low-resource languages. This is enabled through zero-shot learning, sub word-based embedding, and CCA-based alignment. Future work will include empirical validation on language pairs such as English–Tamil and English–Amharic to demonstrate the adaptability and robustness of the model across diverse linguistic structures.

Even though the present assessment did not explicitly involve code-switched, noisy, or informal text, the ContextXL model is naturally built to accommodate such variation to some degree because it uses sub word tokenization (Byte Pair Encoding), contextual embedding (BERT and ELMo), and the Transformer-XL memory mechanism. BPE is useful to decompose rare and hybrid words that are common in informal or code-switched texts, and contextual embedding are useful to capture meaning based on the language context around it, even when grammar or word choice is not standard. Moreover, FastText subword-based embedding give resilience to misspellings, slang, and morphological variations. All these elements add to the generalizing ability of the model over syntactically irregular or stylistically informal language. Future research includes expanding the dataset to have code-switched and user-generated content in order to empirically evaluate and optimize performance under these conditions in the real world environment.

## Abbreviation

None.

## Acknowledgment

The authors would like to thank the Deanship of GITAM (Deemed to be University), Visakhapatnam for supporting this work.

## Author Contributions

Kandula Narasimharao: methodology, Conceptualization, Data collection, writing the study, Angara S. V. Jayasri: Analysis of overall concept, writing, editing.

## Conflict of Interest

The authors declare that we have no conflict of interest.

## Ethics Approval

No ethics approval is required.

## Funding

On Behalf of all authors the corresponding author states that they did not receive any funds for this project.

# References

1. Gupta A, Rallabandi SK, Black A. Task-specific pre-training and cross lingual transfer for code-switched data. 2021 Feb 24. https://arxiv.org/pdf/2102.12407
2. Ansell A, Ponti EM, Korhonen A, Vulić I. Composable sparse fine-tuning for cross-lingual transfer. 2021 Oct 14. https://arxiv.org/pdf/2110.07560
3. Ahmad WU, Li H, Chang KW, Mehdad Y. Syntax-augmented multilingual BERT for cross-lingual transfer. 2021 Jun 3. https://arxiv.org/pdf/2106.02134
4. Muller B, Elazar Y, Sagot B, Seddah D. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. 2021 Jan 26. https://arxiv.org/pdf/2101.11109
5. Adelani DI, Ruiter D, Alabi JO, Adebonojo D, Ayeni A, Adeyemi M, Awokoya A, España-Bonet C. The Effect of Domain and Diacritics in Yor\ub\'a-English Neural Machine Translation. 2021 Mar 15. https://arxiv.org/pdf/2103.08647
6. Maier D, Baden C, Stoltenberg D, De Vries-Kedem M, Waldherr A. Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections. Communication methods and measures. 2022 Jan 2;16(1):19-38.
7. Licht H. Cross-lingual classification of political texts using multilingual sentence embeddings. Political Analysis. 2023 Jul;31(3):366-79.
8. Takeshita S, Green T, Friedrich N, Eckert K, Ponzetto SP. Cross-lingual extreme summarization of scholarly documents. International journal on digital libraries. 2024 Jun;25(2):249-71.
9. Lee C, Yang K, Whang T, Park C, Matteson A, Lim H. Exploring the data efficiency of cross-lingual post-training in pretrained language models. Applied Sciences. 2021 Feb 24;11(5):1974.
10. Muneer I, Nawab RM. Cross-lingual text reuse detection using translation plus monolingual analysis for English-Urdu language pair. Transactions on Asian and Low-Resource Language Information Processing. 2021 Oct 31;21(2):1-8.
11. Chi HV, Anh DL, Thanh NL, Dinh D. English-Vietnamese cross-lingual paraphrase identification Using MT-DNN. Engineering, Technology & Applied Science Research. 2021 Oct 12;11(5):7598-604.

12. Lim H, Cosley D, Fussell SR. Understanding cross-lingual pragmatic misunderstandings in email communication. Proceedings of the ACM on Human-Computer Interaction. 2022 Apr 7;6(CSCW1):1-32.
13. Horbach A, Pehlke J, Laarmann-Quante R, Ding Y. Crosslingual content scoring in five languages using machine-translation and multilingual transformer models. International Journal of Artificial Intelligence in Education. 2024 Dec;34(4):1294-320.
14. Nosek TV, Suzić SB, Pekar DJ, Obradović RJ, Sečujski MS, Delić VD. Cross-lingual neural network speech synthesis based on multiple embeddings. Engineering Applications of Artificial Intelligence.. 2021;7(2):110-120. https://doi.org/10.1016/j.engappai.2021.104781
15. Li X, Fang L, Zhang L, Cao P. An interactive framework of cross-lingual NLU for in-vehicle dialogue. Sensors. 2023 Oct 16;23(20):8501.
16. Seki K. Cross-lingual text similarity exploiting neural machine translation models. Journal of Information Science. 2021 Jun;47(3):404-18.
17. Ma S, Yang J, Huang H, Chi Z, Dong L, Zhang D, Awadalla HH, Muzio A, Eriguchi A, Singhal S, Song X. Xlm-t: Scaling up multilingual machine translation with pretrained cross-lingual transformer encoders. 2020 Dec 31. https://arxiv.org/pdf/2012.15547
18. Caglayan O, Kuyu M, Amac MS, Madhyastha P, Erdem E, Erdem A, Specia L. Cross-lingual visual pre-training for multimodal machine translation. 2021 Jan 25. https://arxiv.org/pdf/2101.10044
19. Takahashi K, Sudoh K, Nakamura S. Automatic machine translation evaluation using source language inputs and cross-lingual language model. InProceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Jul:3553-3558. https://aclanthology.org/2020.acl-main.327.pdf
20. Wang C, Gaspers J, Do TN, Jiang H. Exploring cross-lingual transfer learning with unsupervised machine translation. InFindings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 2021 Aug:2011-2020. https://aclanthology.org/2021.findings-acl.177.pdf
21. Wang M, Bai H, Zhao H, Li L. Cross-lingual supervision improves unsupervised neural machine translation. 2020 Apr 7. https://arxiv.org/pdf/2004.03137
22. Artetxe M, Goswami V, Bhosale S, Fan A, Zettlemoyer L. Revisiting machine translation for cross-lingual classification. 2023 May 23. https://arxiv.org/pdf/2305.14240
23. Zhang M, Yang H, Zhao Y, Qiao X, Tao S, Peng S, Qin Y, Jiang Y. Implicit cross-lingual word embedding alignment for reference-free machine translation evaluation. IEEE Access. 2023 Mar 30;11:32241-51.
24. Liu J, Huang K, Li J, Liu H, Su J, Huang D. Adaptive token-level cross-lingual feature mixing for multilingual neural machine translation. InProceedings of the 2022 Conference on Empirical Methods in Natural Language Processing 2022 Dec:10097-10113.
25. Liu J, Shou L, Pei J, Gong M, Yang M, Jiang D. Cross-lingual machine reading comprehension with language branch knowledge distillation. arXiv preprint 2020 Oct 27. https://arxiv.org/pdf/2010.14271
26. Dataset.https://www.kaggle.com/code/mobassir/understanding-cross-lingual-models/input