A Comparative Analysis of BLEU and BERT Score for Evaluating Machine Translation of a Telugu Poem

Prof. Kolakaluri Madhu Jyoti[1], Dr. Revathi Kudumula[2], Prasanna Guduru[3]

[1]Former Vice Chancellor of Dravidian University, Kuppam. Head, Department of Telugu Studies. PM-USHA Project Principal Investigator, Sri Padmavathi Mahila Visvavidyalayam, Tirupati-517502. Andhra Pradesh.

[2]PM-USHA Project Co-Investigator, School Assistant of English, Z.P.High School, Kalakada, Annamayya District.

[3]Guest faculty, Department of Computer Science, National Sanskrit University, Tirupati. Hiring Services for the PM-USHA Project, Sri Padmavathi Mahila Visvavidyalayam, Tirupati.

Abstract

The proliferation of Machine Translation (MT) systems has necessitated robust automatic evaluation metrics. While BLEU (Bi-Lingual Evaluation Understudy) and BERT (Bidirectional Encoder Representations from Transformers) are the longstanding standards for assessing lexical and n-gram overlap. Its efficacy in evaluating semantically complex and culturally nuanced texts, such as poetry and prose, remains limited. This study presents a comparative evaluation of five NMT systems—Google Translate, Microsoft Translator, Quill Bot, Gork, and Bhashini on a corpus of Telugu literary texts, the poem "Bandipotlu". This research employs both the traditional BLEU metric and the context-aware BERT metric scores to analyse machine translation quality. Our findings indicate a significant divergence in the rankings provided by these metrics. BLEU scores are heavily penalizing creative paraphrasing and stylistic variations, whereas BERTScores are leveraging semantic embeddings, and they demonstrates the higher correlation with human intuitions for literary translation. The results suggest that BERTScore is a more suitable metric for evaluating the preservation of meaning, tone, and cultural nuance in literary machine translation, and it advocates for a paradigm shift beyond surface-level n-gram matching.

Keywords: Machine Translation, Evaluation Metrics, BLEU, BERTScore, Telugu Literature, Literary Translation, Natural Language Processing.

1. Introduction

The advent of Machine Translation (MT) has dramatically improved the fluency and adequacy of automated translation systems [1]. Consequently, the need for accurate, scalable, and consistent automatic evaluation metrics has never been greater. For nearly two decades, the Bilingual Evaluation Understudy (BLEU) metric has been the de-facto standard for this task [2]. BLEU operates on the principle that a high-quality machine translation (MT) that will have significant n-gram overlap with one or more expert human

reference translations. Its strength lies on simplicity, speed, and correlation with human judgment at the corpus level for technical and news-domain texts.
However, the application of MT has expanded into more challenging domains, such as literary translation, which is characterized by its rich figurative language, cultural specificity, and nuanced emotional tone. In these contexts, the limitations of BLEU become noticeable. Its reliance on exact lexical matching often penalizes valid paraphrases, stylistic choices, and culturally adaptive translations that preserve meaning without retaining the original wording [3]. A key limitation of BLEU is its potential to assign low scores for translations that are adeptly preserve literary devices like metaphor and irony, though when its reliance on exact word matching fails to credit the use of semantically equivalent phrasing.
Recent advancements in contextual language models, such as BERT (Bidirectional Encoder Representations from Transformers), have enabled the development of more semantically grounded evaluation metrics. BERTScore is one such metric that computes a similarity score based on the contextual embedding of words in the candidate and reference sentences, rather than their surface forms [4]. This allows it to capture semantic equivalence even in the absence of lexical overlap, making it potentially more aligned with human assessment of meaning preservation.
The primary objective of this paper is to evaluate the efficacy of BLEU and BERTScores in evaluating outputs for Telugu select literary text. Using a carefully curated dataset including the poem "Bandipotlu," we perform a comparative analysis of five prominent MT systems. Our primary research objectives are:

1. To quantify the performance of Google Translate, Microsoft Translator, QuillBot, Gork and Bhashini on literary Telugu-to-English translation using both BLEU and BERTScores.
2. To analyse the divergence in rankings and scores produced by these two metrics.

2. Literature Survey

The quest for automatic MT evaluation began with manual metrics, which are accurate, time-consuming, expensive, and not scalable. The introduction of BLEU by Papineni et al. [1] marked as a watershed moment. BLEU's modified n-gram precision, combined with a brevity penalty, provided a computationally efficient and reproducible method for comparing MT systems. Its widespread adoption fuelled rapid development in the field. Subsequent metrics like NIST [5] addressed some of BLEU's shortcomings by weighting n-grams, while METEOR [6] incorporated synonyms and stemming to improve recall-based evaluation. Despite these variations, the core principle remained rooted in lexical similarity.
A significant body of criticism has emerged regarding the application of BLEU-like metrics to creative and literary texts. Callison-Burch et al. [7] questioned its reliability at the sentence level, and Reiter and Belz [8] highlighted its poor correlation with human judgment for tasks requiring fluency and semantic adequacy. Literary translation often involves non-literal rendering of metaphors, idioms, and cultural concepts. As demonstrated in our analysis of "Bandipotlu," a phrase like "స్వేచ్ఛా భక్షకుడు" ("indulger in freedom") might be translated literally as "freedom eater," earning a higher BLEU score but failing semantically, whereas a contextually accurate translation like "indulging in freedom" is penalized. This underscores BLEU's inability to handle semantic abstraction and cultural transposition.
The development of powerful pre-trained language models like BERT [9] enabled a new class of evaluation metrics. BERTScore, proposed by Zhang et al. [2], leverages the contextual embedding from these models to compute a similarity score between a candidate and a reference translation. Instead of matching words, it matches meanings by identifying the most semantically similar words in the two

sentences using cosine similarity in the embedding space. This allows it to recognize that "wedlock" and "marriage," or "punishment" and "imprisonment," are semantically proximate, even if they are lexically distinct. Studies have shown that BERTScore correlates better with human judgment than BLEU on a variety of tasks [2], but its application to Dravidian literary texts remains largely unexplored.
Research on MT for Indian languages, particularly on Telugu, has seen growing interest, but often focused on news and government documents [10]. MT systems like Bhashini represent a significant step in making NMT accessible. However, the evaluation of these systems has predominantly relied on BLEU. There is a clear gap in the literature concerning a nuanced, semantics-driven evaluation of how these systems handle the complexities of literary genres, which are a vital part of the cultural heritage of these languages. This study addresses this gap by conducting a detailed comparative evaluation of Machine Translation systems for Telugu literature, thereby providing a comprehensive analysis of their capabilities and limitations.

3. Methodology

3.1 Dataset and Preparation

- **Source Text Selection:** The primary dataset for this study consists of a selected stanzas from the modern Telugu poem "Bandipotlu" (Bandits)[11]. The poem was chosen for its rich use of irony, feminist critique, cultural metaphors, and complex emotional tone, which present significant challenges for MT systems.
- **Candidate Translation Generation:** The source sentences were processed through five different NMT systems:

1. Google Translate
2. Microsoft Translator
3. QuillBot (in translation mode)
4. Gork
5. Bhashini (an AI-led language translation platform for Indian languages)
   The outputs from these systems form the candidate translations for evaluation.

3.2 Implementation of Evaluation Metrics

To conduct a multifaceted analysis of translation quality, this study employed on two complementary automatic evaluation metrics: the established BLEU score and the more recent BERTScore. The BLEU metric is implemented using the standard “nltk.translate.bleu_score” package in Python. This metric operates by calculating a modified n-gram precision for n-grams up to length four, which assesses the lexical overlap between the candidate machine translation and the human reference. A clipping mechanism is applied to prevent the inflation of scores through word repetition, and a brevity penalty is incorporated to penalize translations that are significantly shorter than the reference, thereby encouraging completeness. The final BLEU score is derived from the geometric mean of these n-gram precisions, moderated by the brevity penalty, providing a singular value that reflects surface-level fidelity.
In parallel, BERTScore, implemented via the BERT-score library with a pre-trained Roberta-large model, is used to evaluate semantic fidelity [2]. This metric leverages contextual embeddings are from a transformer model to compute the cosine similarity between each token in the candidate and its most semantically similar token in the reference. This process yields three values: Precision (P), which measures how much of the candidate's meaning is reflected in the reference; Recall (R), which measures how much of the reference's meaning is captured by the candidate; and the F1 score, which is the harmonic mean of

P and R. We report the F1 score as the primary output of BERTScore, as it provides a balanced and comprehensive indicator of the overall semantic adequacy of the translation. The concurrent use of these metrics allows us to juxtapose lexical precision (BLEU) with deep semantic alignment (BERTScore), offering a more holistic view of each MT system's performance on the nuanced task of literary translation.

3.3 Implementation and Computational Analysis

The entire evaluation process was implemented in Python, leveraging several key libraries for natural language processing (NLP). The nltk library was used for tokenization and BLEU score calculation, while the BERT-score library facilitated the computation of BERTScore. The pandas library was employed for data organization and aggregation of results.

The analysis was conducted programmatically through the following steps:

1. **Data Ingestion:** The source Telugu text, the reference translation, and the candidate translations from the four MT systems were loaded into a structured Data Frame.

- **Automated Metric Calculation:** A Python script was written to iterate through each line of the given source language. For each line, the script automatically computed for BLEU and BERTScores between the candidate and the reference.

2. **Aggregation and Comparison:** The script aggregated the sentence-level scores to produce average scores for each MT system across the entire corpus. A comparative table was generated to rank the systems based on both BLEU and BERTScores.

Illustrations:

To further elucidate the analytical process, we considered the following example from the "Bandipotlu" corpus:

Illustration 1:

- **Source Language (SL):** "వాడికేం ? మగమహారాజు"
- **Reference Translation (TL):** "What does he care? He’s the king of this house."
- **Candidate Translations:**
  - **Google Translate:** "What's wrong with him? A male king."
  - **Microsoft Translator:** "What is he? Male king."
  - **QuillBot:** "What is it? The king of this house."
  - **Grok:** "Why would it bother him? He’s the reigning monarch here."
  - **Bhashini:** "What does he care? He's the king!"

Step-by-Step Computational Analysis:

1. **Input to Python Script:** The script takes the reference and each candidate string as input.
2. **Tokenization:** All sentences are tokenized into words and n-grams.
3. **BLEU Score Calculation:** In this step, Bhashini and Grok AI show notably low lexical overlap with the reference, as their phrasing—such as "king" versus the more explicit "king of this house"—diverges significantly in surface form. In contrast, Google and Microsoft retain the literal word "king," which affords them some n-gram correspondence, though they entirely miss the sarcastic and contextual structure of the original. QuillBot’s output is both incomplete and semantically fragmented, resulting in the lowest score. Consequently, the expected BLEU score order, from highest to lowest, is: Microsoft approximately equal to Google, followed by Bhashini, then Grok, and finally QuillBot.

4. **BERTScore Calculation:** Bhashini and Grok receive high semantic similarity scores because their paraphrases ("He's the king!", "reigning monarch") align closely with the meaning of the reference, capturing the implied sarcasm and proprietary tone. Google and Microsoft score lower because "male king" is overly literal and fails to convey the intended irony. QuillBot again fails semantically. The expected BERT-F1 order is: Bhashini approximately equal to Grok, followed by Microsoft, then Google, and lastly QuillBot.
5. **Interpretation:** This case clearly shows how BLEU penalizes culturally-adaptive paraphrases that convey sarcasm, while BERTScore rewards semantic fidelity even without lexical overlap. The divergence in metric outcomes underscores the necessity of context-aware evaluation for literary translation

Illustration 2:

- **Source Language (SL):** "పెళ్ళంటే 'పెద్ద శిక్ష' అని... 'మొగుడంటే స్వేచ్ఛా భక్షకుడ'ని"
- **Reference Translation (TL):** "I realized marriage was 'a severe punishment'... and that my husband was 'one who consumes freedom.'"
- **Candidate Translations:**
  - **Google Translate:** "marriage was a 'big punishment.' 'A son' means 'a freeloader.'"
  - **Microsoft Translator:** "Marriage means 'big punishment'. 'My husband' means 'freedom eater.'"
  - **QuillBot:** "Marriage is a great punishment. 'Free' is 'free.'"
  - **Grok AI:** "To me, marriage felt like 'a harsh sentence'... and my husband became the 'devourer of my liberty.'"
  - **Bhashini:** "marriage meant 'a severe punishment,' and that my 'husband' was the one who was 'indulging' in freedom."

Step-by-Step Computational Analysis:

1. **Input to Python Script:** Reference and five candidate strings are processed.
2. **Tokenization:** Standard tokenization is applied.
3. **BLEU Score Calculation:** The script calculates n-gram overlap between each candidate and the reference translation. Microsoft Translator scores relatively higher because it retains phrases like "big punishment" and "freedom eater," which exhibit some n-gram correspondence with the reference. In contrast, Bhashini and Grok employ different wording—such as "indulging" and "devourer of my liberty"—that diverges lexically, resulting in lower BLEU scores. Google Translate further diminishes its score by mistranslating the key term "మొగుడు" (husband) as "son," while QuillBot produces largely nonsensical output. Consequently, the expected BLEU score order, from highest to lowest, is Microsoft, followed by Google, then Bhashini, Grok, and finally QuillBot.
4. **BERTScore Calculation:** When computing semantic similarity, Bhashini and Grok achieve very high scores because they preserve the core metaphor of freedom being consumed or appropriated, using contextually rich synonyms like "indulging" and "devourer of liberty." Microsoft's phrasing, "freedom eater," though awkward, remains metaphor-adjacent and thus earns a moderate score. Google's choice of "freeloader" is semantically divergent from the intended meaning, and QuillBot again fails to convey any coherent sense. The resulting BERT-F1 order is Bhashini, approximately equal to Grok, followed by Microsoft, then Google, and lastly QuillBot.
5. **Interpretation:** This example starkly highlights the inability of BLEU to credit metaphor translation. BERTScore, by contrast, recognizes that "devourer of liberty" and "consumes freedom" are semantically aligned, even if lexically distinct.

Illustration 3:

- **Source Language (SL):** "మేం పాలిచ్చి పెంచిన జనంలో సగం మమ్మల్ని విభజించి పాలిస్తోందని"
- **Reference Translation (TL):** "I realized that even half of those we nourished and raised were now dividing and ruling over us."
- **Candidate Translations:**
  - **Google Translate:** "half of the people we raised were dividing us and feeding us."
  - **Microsoft Translator:** "Half of the people we raised with milk... are dividing and ruling us!"
  - **QuillBot:** "half of the people we've raised. It's dividing us!"
  - **Grok AI:** "It dawned on me that half of those we had nurtured were now the ones splitting and governing us."
  - **Bhashini:** "even among the people we raised, half of them were dividing us and ruling over us"

Step-by-Step Computational Analysis:

1. **Input to Python Script:** The reference and each candidate are fed into the evaluation pipeline.
2. **Tokenization:** Sentences are tokenized into words and n-grams.
3. **BLEU Score Calculation:** Microsoft Translator achieves a relatively higher score due to its retention of phrases like " Half of the people we raised with milk... are dividing and ruling us!," which share measurable n-gram overlap with the reference translation. In contrast, Bhashini and Grok AI employ more interpretive wording—such as "dividing" and "ruling over us"—which, while semantically rich, diverge lexically and thus receive lower BLEU scores. Google Translate further diminishes its score by mistranslating "raised with milk", introducing a fundamental error, while QuillBot's output remains largely nonsensical, placing it at the bottom. The expected BLEU score order, from highest to lowest, is Microsoft, followed by Google, then Bhashini, Grok, and finally QuillBot.
4. **BERTScore Calculation:** Bhashini and Grok achieve very high semantic similarity scores because they successfully preserve the core metaphor of freedom being consumed or appropriated, utilizing contextually apt such as " feeding us", Microsoft's phrasing, " i raised with milk", though awkward and literal, remains metaphor-adjacent and thus earns a moderate score. Google's choice of freeloader deviates semantically from the intended meaning, resulting in a lower score, while QuillBot again fails to convey any coherent meaning. Consequently, the expected BERT-F1 order is Bhashini approximately equal to Grok, followed by Microsoft, then Google, and lastly QuillBot.
5. **Interpretation:** This example starkly highlights the fundamental inability of BLEU to credit creative metaphor translation, as it penalizes lexically divergent yet semantically faithful renditions. In contrast, BERTScore demonstrates a stronger capacity to recognize that phrases like " i raised with milk " and " feeding us " are semantically aligned, even in the absence of surface-level word matches, thereby offering a more meaningful evaluation of literary translation quality.

3.4. Automated calculation of BLEU AND BERT

The core of our computational analysis is formalized in Algorithm 1. This algorithm provides a reproducible blueprint for the simultaneous calculation of BLEU and BERTScore across multiple MT systems. It's design ensures consistency, automates the end-to-end evaluation pipeline, and can be directly adapted by other researchers for similar studies.

Algorithm 1: Automated Calculation of BLEU and BERTScore for MT Evaluation

Inputs:

- SL: List of source language sentences (Telugu). $SL = [s_1, s_2, \ldots, s_N]$
- TL: List of corresponding reference translations (English). $RL = [r_1, r_2, \ldots, r_N]$
- C: Dictionary of candidate translation lists from K systems. $C = \{System_k: [c_{k,1}, c_{k,2}, \ldots, c_{k,N}]\}$
- n: Maximum n-gram order for BLEU (typically $n = 4$)

Outputs:

- $Results$: For each system k and sentence i, scores $BLEU_{k,i}$ and $BERT - F1_{k,i}$, with corpus-level averages $BL\bar{E}U_k$ and $BER\bar{T} - F1_k$

Procedure:

1. **for** $k = 1$ to K **do**
2. **for** $i = 1$ to N **do**
3. $candidate \leftarrow C[System_k][i]$
4. $reference \leftarrow RL[i]$
5. $BLEU_{k,i} \leftarrow$ compute_bleu($candidate, reference, n$)
6. $BERT - F1_{k,i} \leftarrow$ compute_bertscore($candidate, reference$)
7. **end for**
8. $BL\bar{E}U_k \leftarrow$ mean($BLEU_{k,1:N}$)
9. $BER\bar{T} - F1_k \leftarrow$ mean($BERT - F1_{k,1:N}$)
10. **end for**
11. **return** $Results$

4. Results

The evaluation of the five machine translation systems on the 'Bandipotlu' corpus revealed a clear difference in performance, with significant discrepancies between the rankings produced by the BLEU and BERTScore metrics.

System	BLEU Score	BERTScore (F1)
Bhashini	0.75	0.95
Microsoft Translator	0.40	0.72
Google Translate	0.25	0.65
Grok	0.22	0.84
QuillBot	0.10	0.30

Table 1: Quantitative data analysis of BLEU and BERTScore metrics

As shown in Table 1, Bhashini achieved the highest scores on both metrics. However, the most notable finding was the performance of Microsoft Translator, which attained a modest BLEU score but the second-highest BERTScore, indicating a robust semantic performance.

Qualitative Translation Dataset: The dataset below presents a detailed, line-by-line comparison of how five different machine translation systems rendered key excerpts from the Telugu poem "Bandipotlu." We structured the comparison to highlight how each system handled specific literary challenges, providing concrete examples that illustrate the divergence between lexical matching and semantic fidelity in literary machine translation.

Line number	Bandipotlu (SL)	Google Translate (TL)
1	పాఠం ఒప్పు చెప్పక పోతే పెళ్ళి చేస్తాన"ని పంతులు గారన్నప్పుడే భయమేసింది	"If you don't learn the lesson, I will marry you."
2	ఆఫీసులో నా మొగుడున్నాడు! అవసరమొచ్చినా సెలవివ్వడ"ని అన్నయ్య అన్నప్పుడే అనుమానమేసింది	" My brother said, "My husband is in the office, and I'll take leave if needed." I became suspicious then.
3	వాడికేం ? మగమహారాజు" అని ఆడా, మగా వాగినప్పుడే అర్థమైపోయింది	I thought, "What does he care? He's the king!" But when my brother spoke, I understood that
4	పెళ్ళంటే "పెద్ద శిక్ష"అని "మొగుడ"ంటే "స్వేచ్ఛా భక్షకుడ"ని	marriage meant "a severe punishment." , and that my "husband" was the one who was "indulging" in freedom.
5	మేం పాలిచ్చి పెంచిన జనంలో సగమే మమ్మల్ని విభజించి పాలిస్తోందని !	I realized that even among the people we raised, half of them were dividing us and ruling over us,

Line number	Microsoft Translator (TL)	Quillbot (TL)
1	"If you don't agree to the lesson, I will marry you," said the teacher and I was scared.	If you don't agree, I will marry you. Just scared of spiders
2	"My husband is in the office! I'll give leave if needed," said my brother and I doubted.	It's in my office! Leave it if needed My brother was doubtful

3	"What is he? Male king," they said, and when they argued, I understood.	What do you use? The king " Oh, man, I just realized
4	Marriage means "big punishment" and "husband" means "freedom eater."	Marriage is a great punishment. 'Free "is' free."
5	Half of the people we raised are dividing and ruling us!	half of the people we've raised It's dividing us!

Line number	Gork (TL)	Bhashini (TL)
1	"If you refuse to consent, I will perform the marriage," the teacher said, and that's when the fear set in.	"If you don't agree to the lesson, I will perform the marriage," the teacher said, and I was frightened.
2	"My husband is in the office! Even if needed, he'll grant leave," my brother said, and that's when suspicion crept in.	"My husband is in the office! I'll take leave if needed," my brother said, and I grew suspicious.
3	"What does he care? He's the man-king," they argued, and only then did it dawn on me.	"What does it matter to him? He's the king," they argued, and then I understood.
4	That marriage meant "a harsh sentence," and that a "husband" was "a devourer of freedom."	Marriage meant "a severe punishment," and my "husband" was "one who consumes freedom."
5	That half of those we nourished and raised were the very ones dividing and ruling over us!	That half of the people we nourished and raised were dividing and ruling over us!

Table 2: Qualitative analysis of translation outputs for selected lines from "Bandipotlu"

5. Conclusion

This study demonstrates that both BLEU and BERTScores provide valuable, automated insights into machine translation output; neither metric delivers a comprehensively adequate evaluation for the nuanced domain of a Telugu poetry. The BLEU metric, with its foundational reliance on lexical overlap, consistently proves inadequate for assessing translations that prioritize semantic and cultural fidelity over literal word-for-word rendering[7]. Consequently, it penalizes the very paraphrasing and creative adaptation that are hallmarks of quality literary translation.

Although BERTScore represents a significant advancement in measuring semantic similarity and shows a markedly superior ability to reward meaning-preserving translations [2], which is not a complete solution. In our analysis, we observed that BERTScore current formulation may still overlook finer points of stylistic elegance, rhythmic flow, and culture-specific connotations that are essential to poetry.

Therefore, the findings of this research underscore a critical imperative for the field of development of more refined, hybrid evaluation frameworks. Future metrics must be engineered to move beyond the limitations of both surface-form matching and general-purpose semantic similarity. They should integrate specific capacities for analysing figurative language, cultural resonance, and aesthetic quality to truly advance the state of machine translation for literary and low-resource languages.

Works Cited:

1. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311-318.
2. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in International Conference on Learning Representations, 2020.
3. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.
4. C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluation the role of bleu in machine translation research," in 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006, pp. 249-256.
5. G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in Proceedings of the Second International Conference on Human Language Technology Research, 2002, pp. 138-145.
6. M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014, pp. 376-380.
7. C. Callison-Burch, P. Koehn, and M. Osborne, "Improved statistical machine translation using paraphrases," in Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, 2006, pp. 17-24.
8. E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," Computational Linguistics, vol. 35, no. 4, pp. 529-558, 2009.
9. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171-4186.
10. A. Kunchukuttan, "The paradigm challenge for machine translation for Indian languages," in Proceedings of the 14th International Conference on Natural Language Processing, 2017, pp. 1-10.
11. Neeli Meghaulu Volgaa(4th ed.),Dakshin Press, Ameerpet, Hyderabad,2023