
Developing a Mehri–Arabic Neural Machine Translation System Using a Multilingual Transformer Model (mT5)

Khaled Hassan Balhaf

Department of Computer Science, College of Applied and Health Sciences, Mahrah,
Mahrah University, Yemen
Khaledbalhaf2021@gmail.com

Nabil Ali Munassar

Information Technology Department, College of Engineering and Computing, College, University
of Science and Technology, Aden, Yemen
N.munassar@ust.edu

Sulaiman A. Maeli

Department of Computer Science, College of Applied and Health Sciences, Mahrah,
Mahrah University, Yemen
s.maeli@mhru.edu.ye

Lamiaa Mohammed Salem Akoosh

Department of Computer Science and Engineering, School of Engineering Sciences and
Technology, New Delhi, India
lamiaaakoosh@gmail.com

Abstract

Machine translation for low-resource and endangered languages remains a challenging task due to the scarcity of parallel corpora and limited computational resources. Despite its linguistic and cultural significance, the Mehri language has received little attention in natural language processing research. This paper presents a Mehri–Arabic machine translation system based on a multilingual neural model. A custom Mehri–Arabic parallel corpus was constructed and preprocessed using rule-based techniques. The multilingual mT5 model was fine-tuned via transfer learning and evaluated using both automatic metrics and human assessment. Experimental results indicate that satisfactory Arabic translations can be achieved despite limited data availability. This work represents an early computational effort in Mehri–Arabic machine translation and highlights the potential of multilingual pretrained models for supporting under-resourced languages.

Keywords: Mehri, Machine Translation, MT5 model, Arabic Translations, Resourced Languages.

1. Introduction

Machine Translation (MT) has emerged as a critical research area within Natural Language Processing (NLP), largely driven by recent advances in deep learning and neural network architectures. In particular, Neural Machine Translation (NMT) models based on the Transformer architecture have achieved substantial improvements in translation quality by effectively modeling long-range dependencies and capturing rich contextual semantics, outperforming traditional statistical methods [1] [2]. Nevertheless, despite these advancements, NMT systems remain highly dependent on the availability of large-scale parallel corpora, which presents a significant limitation for low-resource and endangered languages.

The Mehri language belongs to the Modern South Arabian branch of the Semitic language family and is primarily spoken in eastern Yemen and southern Oman. Despite its rich linguistic heritage and historical significance, Mehri remains under-resourced and insufficiently documented. It lacks standardized orthography, annotated digital corpora, and dedicated computational tools [3] [4]. Consequently, contemporary language technologies—such as machine translation, automatic speech recognition, and information retrieval systems—provide little to no support for Mehri.

Arabic serves as the primary language of instruction and written communication in regions where Mehri is spoken. Consequently, Mehri–Arabic translation plays a critical role in facilitating effective communication as well as in the documentation and preservation of linguistic heritage. However, the lack of parallel Mehri–Arabic corpora has significantly hindered the development of robust automatic translation systems. Addressing this limitation is essential not only for advancing technological inclusivity but also for supporting the preservation of endangered languages and their associated cultural heritage.

Recent advances in low-resource machine translation indicate that pretrained multilingual Transformer-based models, such as mBART, mT5, and MarianMT, achieve competitive performance even with limited parallel corpora by leveraging transfer learning and fine-tuning strategies [5] [6] [7]. These models employ shared cross-lingual representations, facilitating effective knowledge transfer from high-resource to low-resource languages. This paradigm is particularly advantageous for Mehri, given its typological and lexical proximity to other Semitic languages.

This project aims to develop the first Mehri–Arabic machine translation system by constructing a high-quality parallel corpus comprising approximately 500 phrase pairs and fine-tuning pretrained Transformer-based models. The methodology follows a systematic pipeline, including data collection, manual translation and validation, data preprocessing and cleaning, model training, evaluation, and documentation of results. Translation quality will be assessed using automatic metrics, such as BLEU, alongside evaluations conducted by native speakers. The study contributes both technologically and linguistically by integrating advanced neural machine translation techniques with field-based language data collection. Furthermore, it provides essential resources for Mehri natural language processing (NLP) research and offers empirical insights into the effectiveness of

multilingual Transformer models for extremely low-resource language pairs. The remainder of this paper is organized as follows. Section 2 reviews related work in neural machine translation and low-resource language processing. Section 3 describes the methodology, including dataset construction, preprocessing, and model training. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper and outlines directions for future work.

2. Related work

Natural Language Processing (NLP) systems have advanced significantly in recent years, extending their applicability beyond high-resource languages. This progress has increased the importance of investigating machine translation for low-resource languages. Early machine translation approaches relied on rule-based and statistical methods, which required extensive multilingual dictionaries and large aligned corpora [8]. However, the scarcity of such resources limited their effectiveness for low-resource languages [9]. The emergence of Neural Machine Translation (NMT) marked a significant breakthrough, enabling the development of models that achieve improved fluency and adequacy in translation [1].

The introduction of self-attention mechanisms significantly improved Neural Machine Translation (NMT) by enabling models to capture global dependencies while eliminating the need for recurrent and convolutional architectures. Despite these advancements, such models typically require large-scale training data to achieve optimal performance. Consequently, recent research has focused on multilingual learning strategies to address data scarcity. Multilingual NMT trains a single model on multiple language pairs, allowing it to learn shared representations and transfer knowledge across languages, thereby improving performance, particularly for low-resource scenarios [10].

In the context of multilingual neural machine translation, mBART, proposed by Lewis et al., represents a significant advancement. This model is pre-trained as a multilingual sequence-to-sequence denoising autoencoder on large-scale monolingual corpora and subsequently fine-tuned using parallel data. It has demonstrated strong performance, particularly in low-resource translation scenarios [5].

Similarly, mT5 extends the T5 architecture to support a broader range of languages, thereby enhancing multilingual text understanding and generation capabilities [11]. This extension facilitates cross-lingual transfer learning and improves performance across diverse linguistic tasks [6]. Furthermore, MarianMT employs efficient Transformer based architectures optimized for high-speed translation. These models offer competitive performance and can be further enhanced through domain-specific fine-tuning, making them suitable for practical deployment in customized translation systems [7].

Several studies have demonstrated that even modest parallel corpora can significantly improve translation performance, particularly when combined with pre-trained bilingual models. Rico Sennrich and Zhang showed that leveraging transferable knowledge from linguistically related

languages substantially enhances translation quality in low resource settings [12]. In a similar direction, Chronopoulou et al. introduced language-family adapters to improve the effectiveness of multilingual neural machine translation (NMT) systems for under-resourced languages [13].

Furthermore, research on minority and endangered languages highlights the importance of transparent data collection practices and active community involvement. Steven Bird emphasized that language technologies should support diverse linguistic usage without exacerbating digital inequality [14]. Additionally, Joshi et al. analyzed global NLP coverage and found that many minority languages remain significantly under-represented, advocating for targeted efforts in dataset creation and resource development [15].

Most of the computational studies within the Semitic language family have focused on Arabic, Hebrew, and Amharic [4][3]. In contrast, modern South Arabian languages have received far less attention. Although Mehri has been described in linguistic literature, these works are primarily descriptive rather than computational. Furthermore, there is a lack of digital resources that can be utilized for NLP tasks [16].

There are currently no published studies that develop or evaluate machine translation systems for Mehri. This research enhances existing low-resource neural machine translation techniques by exploring a previously unexamined language pair. It broadens the scope of neural machine translation across multiple languages and provides additional empirical evidence that machine translation can be effective for low resource languages.

3. Methodology

This study adopts an experimental, quantitative methodology to develop and evaluate a neural machine translation (NMT) system for the Mehri–Arabic language pair. Given the low-resource nature of Mehri and the limited availability of computational tools, the approach emphasizes data acquisition from native speakers, rule-based preprocessing, and the fine-tuning of pretrained NMT models. The overall workflow comprises five principal stages: data collection, translation and validation, data preprocessing, model training, and system evaluation. The proposed framework is illustrated in Figure (1).



Figure (1). Our proposed framework

a) Dataset Collection:

For this study, a parallel Mehri–Arabic dataset was constructed. Data collection was conducted through informal interviews with five native speakers proficient in both Mehri and Arabic. These

participants engaged in natural, everyday conversations, from which approximately 500 Mehri sentences were elicited. The collected data encompass a diverse range of linguistic forms, including commonly used expressions, daily conversational utterances, brief descriptive statements, and poetic excerpts, ensuring adequate lexical and syntactic variability.

To maintain consistency and facilitate accurate representation, all Mehri sentences were transcribed using a dedicated Mehri keyboard that supports both Mehri and Arabic scripts. The dataset was initially organized in a Microsoft Excel file, with separate columns for Mehri sentences and their corresponding Arabic translations. Table (1) summarizes the statistical characteristics of the constructed dataset.

table (1). dataset characteristics.

Term	Mehri	Arabic
Number of sentences	500	500
Number of words	1387	1620
Long sentence	28	38
Short sentence	6	7

The parallel corpus utilized in this study comprises 500 sentence pairs, reflecting the low-resource nature and endangered status of the Mehri language. Data collection in such contexts is inherently challenging due to the absence of standardized written materials, necessitating reliance on native speakers for linguistic data acquisition. Prior research has shown that multilingual pre-trained models can achieve meaningful performance even with limited parallel data through transfer learning techniques. Accordingly, the selected dataset size is considered sufficient for a feasibility study, providing a foundational benchmark for subsequent expansion and future research efforts.

b) Translation and Validation:

All Mehri sentences were manually translated into colloquial Arabic by bilingual speakers proficient in both languages. To ensure translation quality and accuracy, a structured validation process was implemented. Each translated sentence underwent systematic proofreading by a multilingual linguist, who reviewed and corrected inconsistencies or linguistic inaccuracies as necessary.

Furthermore, to maintain a high-quality dataset, ambiguous sentence pairs or those exhibiting multiple interpretations were excluded from the corpus. This rigorous manual translation and validation procedure ensures the reliability and linguistic integrity of the dataset, making it suitable for training and evaluating neural machine translation models.

c) Data Preprocessing:

Data preprocessing was conducted using Python due to its flexibility and suitability for customized natural language processing (NLP) tasks. In the absence of standardized preprocessing tools or dedicated linguistic analyzers for the Mehri language, a rule-based preprocessing approach was adopted. The preprocessing pipeline comprised the following steps:

- Text Refinement: Extraneous whitespace, non-Mehri characters, punctuation, and numerical elements were systematically removed during preprocessing. Additionally, character normalization was applied where necessary to ensure consistency and compatibility with the Mehri keyboard layout.
- Getting Rid of Common terms: A custom list of high-frequency, semantically insignificant terms in the Mehri language was manually constructed. This list primarily comprises common function words, including prepositions and other frequently occurring tokens with limited semantic contribution. These terms were subsequently removed from the corpus to enhance the model's ability to focus on informative and contextually relevant lexical items, thereby improving overall translation performance.
- Light Stemming: Due to the absence of a dedicated Mehri lemmatizer, a rule-based light stemming approach was adopted. This method involved the removal of frequent affixes, including common prefixes (like "ال", "ب", and "ل") and suffixes (like "هم", "ه", and "ك"), where applicable, while preserving the semantic integrity of the root word [17].

The preprocessing procedures were applied exclusively to the Mehri text, and the resulting dataset was stored in a newly created Excel file for subsequent model training. This approach aligns with established best practices in low-resource language processing, emphasizing efficient and systematic data preparation to ensure clarity and model readiness.

d) Model Training and Evaluation:

Following the completion of the data preprocessing stage, the Mehri–Arabic parallel corpus was utilized to train a neural machine translation (NMT) model based on the mT5 architecture. Owing to the limited size of the dataset and the low-resource nature of the Mehri language, a transfer learning approach was adopted. Specifically, a pretrained mT5 model was fine-tuned on the prepared corpus, rather than training a model from scratch. This strategy leverages prior multilingual knowledge to enhance translation performance under low resource conditions.

The mT5 (Multilingual Text-to-Text Transfer Transformer) model is adopted in this study as the core architecture for multilingual neural machine translation. mT5 extends the Text-to-Text Transfer Transformer framework to support over one hundred languages within a unified text-to-text paradigm, in which all natural language processing tasks, including machine translation, are formulated as sequence generation problems. This unified formulation enhances model flexibility and facilitates seamless adaptation across diverse linguistic tasks and domains.

The model is pretrained on large-scale multilingual corpora, enabling it to learn shared semantic and syntactic representations across languages. Such cross-lingual representation learning supports effective knowledge transfer from high-resource languages to low-resource counterparts. This characteristic is particularly beneficial for Mehri language, given its linguistic proximity to other

Semitic languages. The overall workflow and implementation steps of the mT5-based approach employed in this study are illustrated in Figure (2).

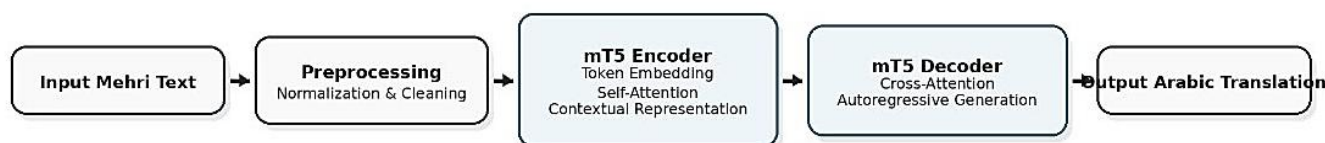


Figure (2). Workflow of the mT5-based Mehri–Arabic machine translation model.

The dataset was partitioned into three subsets: 80% for training, 10% for validation, and 10% for testing. Model finetuning was done using Python in conjunction with established deep learning frameworks. During training, Mehri text served as the source input, while the corresponding Arabic translations were treated as target outputs.

To evaluate the performance of the fine-tuned mT5 model, both automatic and human evaluation methods were employed. The Bilingual Evaluation Understudy (BLEU) score was used to quantify the similarity between machine generated translations and reference Arabic translations, providing an objective measure of translation quality. In addition, human evaluation was conducted by native speakers to assess translation fluency and adequacy [18]. This step was particularly critical due to the absence of standardized evaluation benchmarks for the Mehri language. The integration of automated metrics and human judgment enabled a comprehensive and reliable assessment of the proposed translation system.

4. Results and analysis

This section presents and analyzes the outcomes of optimizing the mT5 model on the Mehri–Arabic parallel corpus. The evaluation considers both quantitative and qualitative perspectives to provide a comprehensive assessment of the translation system's performance. Following the data preparation phase, the final dataset was organized into a structured Excel file, where Mehri sentences were aligned with their corresponding Arabic translations. As outlined in the methodology, the dataset was partitioned into training, validation, and testing subsets. Model performance was evaluated on the test set using the BLEU metric, which measures the similarity between machine-generated translations and human reference translations.

The results indicate that the fine-tuned mT5 model was capable of generating Arabic sentences that are both grammatically coherent and semantically relevant. Despite the limited size of the dataset, the model demonstrated a reasonable level of translation quality, highlighting its effectiveness in low-resource language settings.

The results indicate that the proposed method performs more effectively on short, frequently occurring Mehri phrases, particularly those related to everyday topics and high frequency vocabulary. In contrast, the model demonstrates reduced performance when handling longer sentences, as well as

those containing rare terminology or complex syntactic structures. This behavior is consistent with known limitations of neural machine translation systems in low-resource settings Neural Machine Translation [19]. Figure (3) illustrates the corresponding evaluation results.

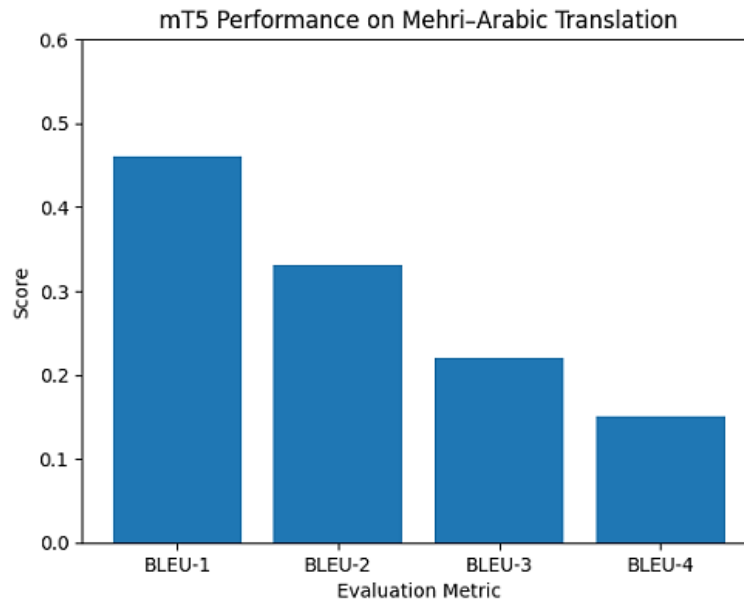


Figure (3). BLEU scores achieved by mT5 Performance on Mehri-Arabic Translation.

Although mT5 does not explicitly support the Mehri language, the model demonstrated satisfactory translation performance, particularly for short and structurally simple sentences. The results suggest that multilingual Transformer based architectures can exhibit partial generalization to previously unseen Semitic languages when sufficient parallel data is available. However, the relatively lower BLEU-3 and BLEU-4 scores highlight persistent challenges, including morphological complexity and variability in word order. Additionally, the limited size of the training dataset constrains the model's ability to capture broader contextual dependencies. Overall, the obtained performance is consistent with findings reported in prior studies on low-resource neural machine translation, reinforcing the inherent difficulties associated with such settings.

The human evaluation was conducted by two native Mehri speakers who were also fluent in Arabic. Each evaluator independently assessed a randomly selected subset of the translated sentences. The evaluation focused on two criteria: fluency and adequacy. A three-point rating scale was employed, where 1 indicates poor quality, 2 indicates acceptable quality, and 3 indicates high quality. Discrepancies between evaluators were resolved through discussion until a consensus was reached.

The human evaluation scores, presented in Figure (4), range between 1.5 and 2. These results indicate that the majority of the translations are of acceptable quality, though not yet optimal. The findings highlight the challenges associated with limited linguistic resources and the relatively small size of the training corpus for the Mehri language. Nevertheless, the model demonstrates the ability to preserve the core meaning of the source sentences, suggesting that the proposed system provides a viable foundation for future improvements and further development.

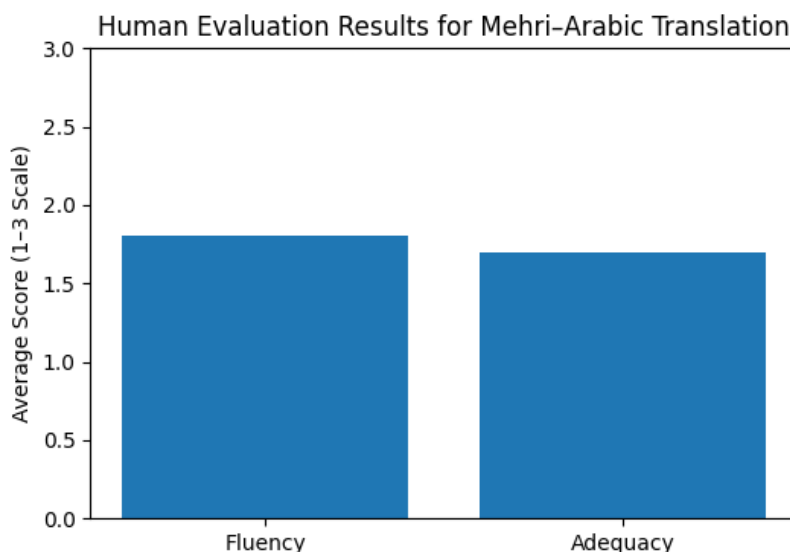


Figure (4). Human assessment outcomes of the Mehri–Arabic machine translation system, evaluated for fluency and sufficiency on a three-point scale.

5. Conclusion and future work

This study investigated the feasibility of developing a Mehri–Arabic machine translation system using the mT5 framework. Despite the low-resource nature of the Mehri language and the limited size of the available parallel corpus, the experimental results demonstrate that it is possible to generate meaningful and reasonably accurate Arabic translations. These findings confirm that pretrained multilingual models, when leveraged through transfer learning, provide an effective and scalable approach for supporting under-resourced and endangered languages. The analysis further indicates that the model performs more effectively on short and structurally simple sentences, while longer and syntactically complex inputs remain challenging. This limitation is primarily attributed to the constrained dataset size and limited linguistic diversity within the training corpus. Moreover, while the system succeeds in preserving the core semantic content of the source text, there is still room for improvement in fluency and adequacy.

The study also highlights several important implications. First, multilingual architectures such as mT5 offer a strong baseline for Mehri and other low-resource Semitic languages. Second, data availability

remains a critical factor, as increasing corpus size is expected to significantly improve evaluation metrics such as BLEU score. Third, incorporating linguistic features or leveraging transfer learning from closely related Semitic languages may further enhance model performance.

However, this work is subject to several limitations. The dataset is relatively small, only a single automatic evaluation metric (BLEU) was employed, and no comparative analysis with baseline or alternative models was conducted. These constraints limit the comprehensiveness of the evaluation and the generalizability of the findings. Future work should focus on expanding the Mehri–Arabic parallel corpus, improving preprocessing and normalization techniques, and exploring data augmentation strategies to mitigate data scarcity. Additionally, integrating linguistic knowledge and experimenting with alternative neural architectures or hybrid approaches may yield further improvements. Extending the system to support additional modalities, such as speech translation, and promoting open access to linguistic resources will contribute to both technological advancement and the preservation of the Mehri language [20].

Acknowledgment

The authors like to extend their heartfelt thanks to AlMahrah University for its academic assistance and support during this research. We would also want to thank the Mehri Language Centre for Studies and Research for their helpfulness, language knowledge, and ongoing support in supplying resources connected to the Mehri language, which was a big part of the success of our study

References

- [1] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in ICLR, 2016.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," in 31st Conference on Neural Information Processing Systems, CA, USA., 2017.
- [3] A. Rubin, The Mehri Language of Oman, Brill, 2010.
- [4] J. C. Watson, The structure of Mehri, Otto Harrassowitz, 2012.
- [5] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," Transactions of the Association for Computational Linguistics, pp. 726-742, 8 2020.
- [6] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua and C. Raffel, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies, 2021.

- [7] M. Junczys-Dowmunt, R. Grundkiewicz, . T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji and N. Bogoychev, "Marian: Fast Neural Machine Translation in C++," 2018.
- [8] P. Koehn, Statistical machine translation, Cambridge University Press, 2009.
- [9] P. Singh and S. Kumar, "AdiBhashaa: A Community-Curated Benchmark for Machine Translation into Indian Tribal Languages," arXiv preprint arXiv, 2025.
- [10] M. Johnson and et al, "Google's multilingual neural machine translation system: Enabling zero-shot translation," Transactions of the Association for Computational Linguistics, pp. 339-351, 5 2017.
- [11] S. M. M. Billah, , A. A. Subarna, S. N. Sarna, . A. S. Wasit, A. Fariha, A. Sushmit and A. Y. Sadeque, "Towards santali linguistic inclusion: building the first Santali-to-English translation model using mT5 transformer and data augmentation," arXiv preprint arXiv, 2024.
- [12] R. Sennrich and. B. Zhang, "Revisiting low-resource neural machine translation: A case study," in 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.
- [13] A. Chronopoulou, D. Stojanovski and A. Fraser, "Language-family adapters for low-resource multilingual neural machine translation," in Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023), 2023.
- [14] S. Bird, "Decolonising speech and language technology," in Proceedings of the 28th international conference on computational linguistics, 2020.
- [15] P. Joshi, S. Santy, A. Budhiraja, K. Bali and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," in ACL, 2020.
- [16] . T. M. Johnstone, Mehri lexicon, Routledge, 2012.
- [17] K. Balhaf, O. Darwish, E. Rawashdeh, M. A. Awad, D. Darweesh, Y. Tashtoush and S. Rawashdeh, "Classifying Arabian Gulf Tweets to Detect People's Trends: A case study," in 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2022.
- [18] H. K. eklehaymanot, G. Gidey and W. Nejd, "Low-Resource EnglishTigrinya MT: Leveraging Multilingual Models, Custom Tokenizers, and Clean Evaluation Benchmark," 2025.
- [19] V. Mujadia and D. M. Sharma, "BhashaVerse: Translation Ecosystem for Indian Subcontinent Languages," arXiv preprint arXiv, 2024.
- [20] I. Sel and D. Hanbay, "Efficient Adaptation: Enhancing Multilingual Models for Low-Resource Language Translation," Mathematics, 19 12 2024.