UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

COS700 Research Report

Fine-Tuning Low Resource Machine Translation through Data Augmentation: Xitsonga Case Study

Student number: u16153970

Supervisor(s):
Mrs. Seani Rananga
Dr. Abiodun Modupe

Makungu Ndlovu

October 2024

Abstract

This research focuses on enhancing the performance of a translation model for Low-Resource Languages, specifically Xitsonga, a Bantu language spoken in South Africa, through data augmentation techniques. Achieving accurate translations in languages like Xitsonga is challenging due to limited linguistic resources and training data. The scarcity of parallel corpora and linguistic references further complicates the development of robust translation models that capture the unique characteristics of these languages. To address these challenges, this study explored whether retraining existing translation models with augmented data could mitigate the limitations posed by small training sets.

Key objectives included identifying effective data augmentation methods to enrich the Xitsonga dataset and evaluating their impact on model performance. Techniques such as back-translation, random insertion, and word replacement were employed to diversify the dataset and support the fine-tuning of translation models. The findings revealed that data augmentation significantly improved translation quality, with Random Insertion and Word Replacement achieving the highest BLEU and METEOR scores. The study also highlighted the superior performance of transformer-based models, particularly DistilBERT, which outperformed traditional machine learning classifiers.

Cosine similarity analysis was used to ensure that augmented translations maintained semantic consistency with the original sentences. A threshold of 0.84 was determined as optimal for filtering high-quality augmentations, effectively balancing the retention of semantically similar data. This filtering approach contributed to improved model training, helping distinguish between well-preserved and poorly translated sentences and ultimately enhancing overall translation performance.

By bridging the gap between Low-Resource and High-Resource languages, this research not only improved Xitsonga translation accuracy but also provided a framework for enhancing machine translation systems for other under-resourced languages. The outcomes included practical recommendations for improving translation model accuracy, fluency, and robustness, insights into the effectiveness and limitations of various data augmentation strategies, and progress towards more inclusive NLP technologies for underrepresented South African languages.

Keywords:

Low-Resource Languages, Translation Model, Data Augmentation, Fine-Tuning, Neural Machine Translation.

1 Introduction

The past few years have witnessed notable developments in machine translation, particularly with Large Language Models (LLMs). However, this progress has not extended to Low-Resource Languages. Despite these advancements, accurate and reliable translations for languages with few linguistic resources and sparse training data remain challenging[1]. One such language is Xitsonga, a South African Bantu language, which faces several challenges due to limited parallel corpora and linguistic references. This research proposal aims to fill the current gap by investigating data augmentation techniques to fine-tune translation model performance for Low-Resourced language.

The challenges associated with translating Low-Resource Languages like Xitsonga are multifaceted. Firstly, the lack of substantial parallel corpora hampers the ability of translation models to learn accurate mappings between source and target languages. Additionally, limited linguistic references and resources make it difficult to validate and improve translation quality. These issues are compounded by the fact that Low-Resource Languages often have complex grammatical structures and unique cultural contexts that are not easily captured by models trained predominantly on high-resource languages[2]. Addressing these challenges requires innovative approaches that leverage existing data while creating new avenues for generating training material. To this end, incorporating data augmentation can play a crucial role in bridging these gaps.

Literature indicates that certain data augmentation techniques help improve translation quality for LLMs, but little research has been done on Low-Resource African Languages like Xitsonga[3]. Data augmentation methods, which generate synthetic data to increase the training dataset size, have demonstrated potential for improving the performance of machine learning models in various applications[4]. For translation, this can include methods such as back-translation, where a sentence is translated into a pivot language and then back into the original language, thereby creating additional parallel data[5]. This study builds on such insights, hypothesizing that a combination of augmentation techniques can improve translation performance and adaptability.

By systematically applying data augmentation techniques, such as back-translation, paraphrasing, random insertion, and word replacement, this study aims to enrich the training datasets of Low-Resource Languages, improving the robustness and accuracy of translation models. The proposed methodology will include evaluating the augmented data through cosine similarity analysis to ensure semantic alignment between original and augmented sentences. Setting an optimal similarity threshold will help filter high-quality augmentations, preserving data integrity for effective model training.

This study intends to bridge the gap by developing a systematic method for incorporating data augmentation into the process of fine-tuning the Xitsonga translation model. Specifically, the study will explore data augmentation techniques such as back-translation, random insertion, random deletion, word replacement, and synonym replacement to determine which methods most effectively enhance the translation quality for Xitsonga. Additionally, it will assess how these methods impact different types of machine learning and transformer-based models, such as Logistic Regression, Naive Bayes, SVM, and advanced architectures like mBERT and DistilBERT.

The anticipated outcomes of this study are significant for both the field of machine translation and the preservation and accessibility of Low-Resource Languages. Fine-tuning translation models with augmented data is projected to boost translation quality for low-resourced languages[6]. This hypothesis will be tested in the study, potentially offering a new paradigm for supporting linguistic diversity in the digital age. The findings could serve as a foundation for future work on similar languages, demonstrating that data augmentation is a viable strategy for addressing data scarcity in under-resourced linguistic contexts.

2 Problem Statement

This study focuses on how data augmentation techniques can be employed to fine-tune the performance of a translation model for the Low-Resourced Xitsonga language. The lack of linguistic resources and labeled datasets makes it difficult to develop accurate and robust Machine Translation (MT) systems in Low-Resource languages. This problem is significant because it affects the access and inclusivity of NLP technology for users who speak Low-Resource Languages such as Xitsonga, thereby hindering their access to information and good digital communication. Researchers aim to close this gap by addressing this problem while contributing to democratizing NLP technology to promote diverse language use within a society where some minority groups have been marginalized.

Despite its significance, this research problem faces several challenges. First, there is a scarcity of data, which limits the creation of effective MT systems for Low-Resource African Languages. Additionally, limited parallel corpora and annotated datasets are rare, which can negatively affect translation models' performance in terms of accuracy and coverage. Moreover, available data may vary in quality, leading to domain mismatch or sparse data, further degrading Machine-Translation output. Furthermore, fine-tuning MT models with data augmentation introduces additional complexities, such as selecting appropriate augmentation methods and balancing model robustness with computational efficiency. These problems demonstrate why an intensive investigation into improving the translation model quality of Low-Resource Languages remains critical.

3 Methodology

3.1 Data Collection and Preparation

The dataset utilized in this research consists of Xitsonga-English sentence pairs obtained from *The Vukúzenzele South African Multilingual Corpus*, developed by the Data Science for Social Impact group. The dataset was accessed via Hugging Face and is based on content from the South African government magazine *Vukúzenzele*, produced by the Government Communication and Information System (GCIS). The original PDFs were sourced from the *Vukúzenzele* website and underwent pre-processing to improve translation quality and consistency. The data cleaning process involved removing punctuation, converting text to lowercase, and normalizing whitespace to ensure uniformity throughout the dataset, which is essential for enhancing model performance [7].

Dataset Split	**Number of Rows**
Training set	2,623
Validation set	563
Test set	562

Table 1: Summary of the Xitsonga-English dataset splits.

3.2 Data Augmentation Techniques

To address the limited size of the Xitsonga-English dataset, this study employs and investigates several data augmentation techniques to expand and diversify the dataset. Each technique introduces unique transformations, generating varied sentence pairs that retain the original meaning, but differ in structure or expression. These techniques are as follows:

- **Back-Translation:** This study uses back-translation as a core augmentation technique to introduce linguistic variety into the dataset. The MarianMT model from Hugging Face, specifically *Helsinki-NLP/opus-mt-ts-en*, was chosen for its effectiveness in low-resource language pairs. MarianMT, based on the Transformer architecture, is a robust sequence-to-sequence model known for high-quality translations across many languages, making it suitable for translating between a low-resource language like Xitsonga and a high-resource language like English.

  We use this model to first translate Xitsonga sentences into English as a pivot language and then back into Xitsonga. This two-step translation process creates natural variations in sentence structure and phrasing while retaining the original semantic content [8].

- **Random Deletion:** In this approach, words are randomly deleted from sentences to create more concise variations. This deletion process encourages the model to focus on key content words rather than relying on specific phrasings, which improves its ability to handle missing or omitted information in real-world contexts.

- **Random Insertion:** This technique involves inserting random words into sentences to increase variability. The inserted words are typically selected to fit the context without distorting the meaning. This method helps the model generalize to sentences with added elements or fillers, which can occur in natural language use.

- **Word Replacement:** Words within Xitsonga sentences are replaced at random with other words to introduce alternative phrasing. Unlike synonym replacement, this approach does not strictly adhere to semantic similarity, but instead aims to create diversity. This technique helps the model become less dependent on specific vocabulary choices, promoting flexibility.

- **Synonym Replacement:** Specific words in sentences are replaced with their synonyms, based on semantic similarity[9]. We achieved this by translating Xitsonga words to English, identifying synonyms in English, and then translating these synonyms back to Xitsonga. This process involved several steps:

1. Each word in a Xitsonga sentence was first translated into English using the MarianMT model.
2. Using WordNet, synonyms for the translated English words were identified, enhancing variability in vocabulary while preserving meaning.
3. The selected English synonym was then translated back into Xitsonga, reintroducing the word into the original sentence structure but with an alternative expression.

Each augmentation method contributes to expanding the dataset's linguistic diversity, enabling the model to learn from a richer set of sentence structures and variations, ultimately improving its generalization to unseen data.

3.3 Translation Consistency Detection

A key component of this study's methodology is the translation consistency detection task, designed to evaluate the quality of augmented Xitsonga translations and ensure they maintain semantic fidelity with the original sentences[10]. Cosine similarity was calculated between each augmented sentence and its original counterpart, with a predefined threshold used to classify each augmentation as "well-translated" (label 1) if the similarity score exceeded the threshold, or "poorly translated" (label 0) if it fell below.

This threshold-based classification was applied across all augmentation techniques, generating pseudo-labels that identified strong semantic alignment between augmented and original sentences. The pseudo-labeled data from each augmentation technique were then combined to form a comprehensive set of augmented sentences with associated labels. This filtering step helps maintain data quality by retaining only those augmentations that preserve the core meaning, thereby enhancing the model's training data. By supporting translation consistency detection in this way, the pseudo-labeling approach effectively reinforces the model's ability to distinguish between well-preserved and poorly preserved translations.

3.4 Model Selection and Training

To assess how data augmentation impacts model performance, various models were chosen for tasks related to translation consistency detection and classification. These models included both traditional classifiers and advanced transformer-based models, providing a comprehensive view of how different approaches responded to the augmented Xitsonga-English dataset.

- **Traditional Classifiers:** These models served as baseline classifiers for translation consistency detection and paraphrase detection tasks, providing a foundational comparison for the more advanced transformer-based models[11].

  1. **Logistic Regression:** Logistic Regression is a supervised learning algorithm used for binary classification. It predicts classes by applying a logistic (sigmoid) function to a weighted sum of input features:
     $$f(x) = \frac{M}{1 + e^{-k(x-x_0)}}$$
     where e is Euler's number, x_0 is the midpoint, M is the curve's maximum value, and k controls the steepness. This function maps predictions to probabilities, facilitating class assignments.

     In this study, the Logistic Regression model was used to assess translation consistency detection on the Augmented dataset. The following methodology was employed:

     (a) **Feature Extraction:** TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer was used to transform the textual data into numerical feature vectors. This approach captured the importance of terms across the dataset, with a maximum of 5,000 features set to balance detail and computational efficiency.

     (b) **Data Splitting:** The dataset, including augmented sentences, and their corresponding pseudo-labels, was split into training and testing sets using an 80/20 ratio. This allowed for effective evaluation of model performance.

     (c) **Model Training:** The Logistic Regression model was trained on the TF-IDF-transformed training data. The training process used standard hyperparameters, ensuring that the model learned to classify "well-translated" versus "poorly translated" sentences effectively.

     (d) **Evaluation Metrics:** The trained model was evaluated on the test set, with accuracy and F1 score used as key performance metrics.

  2. **Naive Bayes:** Naive Bayes, a probabilistic model rooted in Bayes' theorem, was employed in this study for detecting translation

consistency due to its straightforward nature and efficiency in handling high-dimensional text data. The model determines whether an augmented sentence preserves translation consistency with the original by computing:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

where $P(C|X)$ represents the probability that the sentence belongs to the "well-translated" category (class C) given the features X, $P(X|C)$ is the likelihood of the features given the class, $P(C)$ is the prior probability of the class, and $P(X)$ is the evidence.

For this study, the Multinomial Naive Bayes classifier was utilized, following these steps:

(a) **Training Process:** The model was trained using a TF-IDF-transformed training dataset, implemented through the `scikit-learn` library's MultinomialNB. This choice was due to its suitability for text-based data, where features are term frequencies.

(b) **Evaluation Metrics:** The model's performance was tested on the validation set, using accuracy and F1 score as key performance indicators. These metrics were vital for gauging how effectively the model categorized sentences as "well-translated" or "poorly translated" based on augmented data.

The Naive Bayes classifier served as a fundamental baseline for comparison with other models, demonstrating its capability in processing TF-IDF features and identifying translation consistency in augmented datasets.

3. **Support Vector Machine (SVM):** A classification model designed to find the optimal hyperplane that separates classes in a high-dimensional feature space, making it highly effective for binary classification problems. The decision boundary it constructs is defined by:

$$f(x) = w \cdot x - b = 0$$

where w denotes the weight vector, x represents input features, and b is the bias term. This hyperplane maximizes the margin between classes, promoting better generalization and performance.

We trained the SVM classifier using the following methodology:

(a) **Training Process:** The model was trained on the training set transformed with TF-IDF features, utilizing the `SVC` implementation provided by `scikit-learn`. The SVM was selected for its strength in managing complex data distributions and its ability to maintain robustness across various classification tasks.

(b) **Evaluation Metrics:** The performance of the trained model was assessed on the test set using accuracy and F1 score as the main metrics. These metrics provided a clear view of how well the SVM classifier could identify and differentiate between "well-translated" and "poorly translated" sentences.

- **Transformer-Based Models:** These advanced models were fine-tuned on the augmented Xitsonga-English dataset, leveraging pre-trained language representations to capture complex linguistic patterns[12].

    1. **mBERT (Multilingual BERT):** A pre-trained multilingual BERT model, fine-tuned for sequence classification, providing robust semantic representation across multiple languages.
    2. **DistilBERT:** A distilled version of BERT, smaller and faster, providing efficient performance on tasks requiring extensive processing of augmented data.
    3. **XLM-R (XLM-RoBERTa):** A cross-lingual model pre-trained on a vast multilingual corpus, optimized for understanding low-resource languages in cross-lingual tasks.
    4. **RoBERTa:** A robustly optimized variant of BERT with improved language representation, allowing an in-depth analysis of data augmentation's impact on translation consistency.

Each model was trained by setting important parameters like learning rate, batch size, and the number of training epochs to get the best performance on the augmented dataset. For transformer models, additional settings were applied to manage memory and improve gradient updates. During training, metrics such as accuracy, F1 score, and validation loss were tracked at each step to see how the models were learning. These metrics helped show how well each model adjusted to the new data and allowed for comparing the effects of data augmentation on translation quality and consistency.

3.5 Evaluation Metrics

To assess the quality of the augmented Xitsonga data and the translation models, we used both text similarity and classification evaluation metrics:

- **Cosine Similarity:** In this study, it is used to assess the semantic similarity between original and augmented sentence pairs by comparing their term vectors:

  $$\text{Sim}(\mathbf{q}, \mathbf{d}) = \frac{\sum_{i=1}^{k} w_{qi} \cdot w_{di}}{\sqrt{\sum_{i=1}^{k} (w_{qi})^2} \cdot \sqrt{\sum_{i=1}^{k} (w_{di})^2}}$$

  where $\mathbf{q}$ and $\mathbf{d}$ represent the query (original sentence) and document (augmented sentence) vectors, respectively, w_{qi} and w_{di} are the term frequencies for term i in $\mathbf{q}$ and $\mathbf{d}$, and k is the total number of terms. This measure evaluates the effectiveness of each augmentation technique by quantifying the alignment in semantic content [13].

- **BLEU Score:** The BLEU score assesses the quality of augmented Xitsonga text by comparing it to the original text based on n-gram precision. It combines the geometric mean of modified n-gram precisions up to length N, multiplied by a brevity penalty (BP):

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

  where p_n is the modified precision for each n-gram between the augmented and original text, and w_n are weights (typically $\frac{1}{N}$) [14].

- **METEOR Score:** The METEOR score evaluates the quality of augmented sentences by emphasizing recall, synonym matching, and stemming compared to the original sentences. It combines a harmonic mean of precision (P) and recall (R) with a penalty for fragmented matches:

  $$\text{METEOR} = (1 - \text{Penalty}) \times \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R}$$

  where α balances precision and recall, and the Penalty reduces the score for non-contiguous matches between the augmented and original sentences [15].

- **Classification Accuracy and F1 Score:** These metrics evaluate model performance in translation consistency detection, with Accuracy measuring overall correct classifications:

  $$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

and F1 Score capturing the harmonic mean of Precision and Recall:

$$\text{F1-Score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

[16].

We then analyzed the data augmentation techniques to assess their effectiveness by comparing the BLEU, METEOR, and cosine similarity scores. Additionally, validation loss and accuracy curves were plotted to visualize the performance of the models across epochs.

4 Background

Natural Language Processing (NLP) has recently witnessed significant breakthroughs due to advanced language model development and deep learning approaches. Nonetheless, despite benefiting highly from abundant curated data and resources on well-represented languages, Low-Resource Languages have remained challenging due to limited linguistic resources and knowledge about them. One such challenge is the scarcity of labeled datasets and parallel corpora, which are essential for training accurate and robust models. Traditional machine-learning approaches struggle to achieve satisfactory performance without enough data, thereby hampering the development and deployment of effective NLP systems for Low-Resource Languages[17]. In this regard, augmentation techniques that can artificially supplement existing datasets have been proposed to enhance the performance of NLP models on Low-Resource Languages.

Marivate et al. present an insightful study addressing these challenges for Setswana and Sepedi, two underrepresented South African languages[3]. Their work underscores the critical need for clear guidelines on collecting, curating, and preparing datasets to enable various NLP use cases for low-resource languages. Unlike well-resourced languages, which benefit from extensive annotated corpora and research resources, low-resource languages like Setswana and Sepedi faces significant barriers due to limited available data and tools. Marivate and his colleagues focused on creating datasets of news headlines for Setswana and Sepedi and developing a news topic classification task. They utilized data sources from the South African Broadcasting Corporation (SABC) and social media streams to compile these datasets, highlighting the necessity of innovative data collection methods, particularly when direct access to comprehensive news reports is unavailable.

Additionally, Marivate and Catherine addressed the data scarcity issue in Low-Resource African languages through data augmentation techniques, specifically focusing on three popular methods: synonym replacement, random insertion, and contextual augmentation[18]. They compared the performance of these techniques with a baseline Neural Machine Translation (NMT) model using English-Swahili (En-Sw) datasets. Their findings indicate that data augmentation can significantly enhance the performance of translation models, making these techniques valuable for other low-resource languages as well.

Furthermore, Nzama provided a comparison of two data augmentation techniques, namely back-translation data augmentation and multilingual data augmentation, against a baseline system for the Nguni languages of isiZulu and isiXhosa[5]. The study expected both the back-translation system and multilingual system to outperform the baseline system, with the multilingual system providing the best performance. However, the results showed that in both the isiZulu and isiXhosa contexts, the back-translation system provided the best performance. This underscores the variability and context-specific efficacy of different data augmentation techniques, suggesting that tailored approaches are necessary for different languages and datasets.

Data augmentation is crucial for enhancing translation quality because most language pairs involved in Machine Translations (MT) are considered Low-Resourced due to inadequate parallel corpora. For instance, some types of augmentation techniques including back-translation, paraphrasing, and contextual augmentation have been found useful in improving the performance of Neural Machine Translation (NMT) systems for resource-poor languages[19]. Studies have shown that augmenting parallel corpora with synthetic data created from translation and other augmentation methods can significantly improve the quality of translations as measured by different metrics such as BLEU, ChrF, and Meteor[20].

Expanding on these studies, the research conducted by Ortiz Suárez et al. introduces an innovative approach to tackle the data scarcity issue for Guarani, an indigenous language spoken by nearly 10 million people in Paraguay and neighboring regions[21]. Their study explores grammar-based data augmentation to enhance Guarani-Spanish neural machine translation (MT) systems. The proposed method involves automatically generating synthetic Spanish text, which is then syntactically transferred to Guarani. This technique is based on building a grammar-based system that can create high-quality parallel corpora. The authors conducted several experiments to pretrain models using this synthetic text, demonstrating that MT systems pretrained with grammar-based synthetic data perform better than previous baselines.

The study on Data Augmentation for Low-Resource Neural Machine Translation for Sotho-Tswana Languages by Maxwell and Jan further explored the enhancement of NMT models for Sepedi, a low-resource South African language[22]. The authors employed two main data augmentation techniques: backtranslation and word replacement. Backtranslation involved translating monolingual Sepedi text into English to create synthetic parallel corpora, thereby increasing the amount of training data. A reverse NMT model was trained for this purpose, generating synthetic English sentences from Sepedi sentences. This technique has shown improvements in BLEU scores across multiple datasets, including JW300, FLoRes, and Autshumato.

This study is within this stream of literature by fine-tuning the translation model's performance via data augmentation for the Low-Resource Language, Xitsonga. This study will apply back-translation, paraphrase generation, and contextual augmentation techniques. These methods involve augmenting existing parallel corpora by generating additional training samples and refining the NMT model's ability to handle the complexities of Low-resource South African language pairs. The expected outcomes include improved translation accuracy, better preservation of linguistic nuances, and increased robustness in handling diverse language contexts[19]. This study, therefore, aims to overcome challenges associated with inadequate language resources and promote equalization of access to NLP technologies for Low-Resourced languages.

Adding to the body of work on Low-Resource Language dataset creation, Lastrucci and others introduce two multilingual government-themed corpora in various South African languages[23]. The Vuk'uzenzele corpus and the ZA-gov-multilingual corpus were created by collecting South African government newspaper articles and speeches were translated into all 11 official South African languages. These corpora serve multiple downstream NLP tasks, providing researchers with valuable resources to study the language used in government publications and how officials communicate with their constituents. The creation process involved gathering, cleaning, and making the corpora available, resulting in parallel sentence corpora for the Neural Machine Translation (NMT) tasks using Language-Agnostic Sentence Representations (LASER) embeddings. Fine-tuning a massively multilingual pre-trained language model with these aligned sentences has provided NMT benchmarks for nine indigenous languages, further supporting the development of NLP resources for Low-Resourced Languages.

In summary, the literature highlights the importance of innovative data collection and augmentation techniques for enhancing NLP performance for low-resource languages. By addressing the scarcity of datasets and leveraging augmentation methods, researchers are making significant strides in improving translation accuracy and overall NLP capabilities for underrepresented languages. This study aligns with these efforts, focusing on Xitsonga and aiming to promote equitable access to advanced NLP technologies.

5 Related Work

The translation of low-resource languages, such as Xitsonga, presents significant challenges due to limited parallel corpora and linguistic resources. Recent studies have explored various data augmentation techniques to enhance machine translation (MT) performance in these contexts.

5.1 Back-Translation

Back-translation consists of taking monolingual data from the target language and translating it back into the source language to generate synthetic parallel corpora. This method has proven effective in improving translation quality for low-resource languages. Sennrich et al. (2016) demonstrated that back-translation enhances neural machine translation (NMT) systems by providing additional training data, leading to improvements in translation fluency and adequacy [24].

A major advantage of back-translation is its ability to generate data that retains natural linguistic patterns while introducing subtle variations. These variations help models learn to handle diverse sentence structures and phrasings, making them more robust when translating new content. By leveraging high-resource pivot languages, back-translation can enrich low-resource language datasets with meaningful, contextually appropriate translations. This process supports improved generalization and reduces overfitting, as the model is exposed to both standard and varied forms of the target language, enhancing its overall translation capabilities.

5.2 Paraphrasing and Diversified Rephrasing

Paraphrasing is a valuable data augmentation technique that creates multiple expressions of the same content, enhancing the linguistic variety within training datasets. Gao et al. (2023) proposed a method that leverages high-quality translation models trained on well-resourced languages to generate paraphrased sentences, thereby improving the performance of low-resource NMT systems by introducing syntactic diversity while preserving semantic meaning [25]

Diversified rephrasing extends the concept of paraphrasing by introducing different word choices, grammatical forms, and sentence structures, thereby creating a richer training corpus. Integrating this technique with other augmentation methods, such as synonym replacement or contextual data augmentation, can produce even more varied training data. This enriched training set enables models to learn from a broader spectrum of linguistic expressions, improving their generalization capabilities and robustness in translating complex or unexpected inputs [25].

5.3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) have emerged as a promising technique for generating synthetic data, particularly useful in scenarios where training data is limited, such as the Xitsonga corpas. The application of GANs to data augmentation for low-resource language translation has shown potential in enriching training datasets and enhancing model performance. Zeng (2024) demonstrated that GANs could generate monolingual data that closely mimic real linguistic structures, which, when incorporated into neural machine translation (NMT) systems, results in improved translation accuracy and robustness [26].

One of the significant advantages of using GANs in data augmentation is their ability to model complex distributions and generate diverse outputs. This diversity is critical for exposing translation models to a variety of linguistic forms, including colloquial expressions and syntactically diverse structures. Training GANs, however, can be demanding in terms of computational resources and often requires precise tuning to avoid mode collapse, a situation where the generator outputs a narrow range of results. Despite these challenges, GAN-based augmentation remains a valuable tool for enriching datasets and enhancing the robustness of translation models for low-resource languages [27].

5.4 Morphologically-Aware Data Augmentation

Morphologically-aware data augmentation techniques have been developed to address the challenges posed by languages with rich and complex morphological structures. Alam et al. (2024) introduced a dictionary-based augmentation method that incorporates morphological information to generate synthetic data that better reflects the inherent linguistic features of such languages. By considering the inflectional and derivational aspects of words, this approach ensures that augmented data remains semantically coherent and contextually relevant. This method allows translation models to better generalize by exposing them to a wider range of morphological forms during training.

The effectiveness of morphologically-aware augmentation lies in its ability to create realistic variations that maintain the original sentence's meaning while introducing morphological diversity. For low-resource languages with significant morphological complexity, such techniques can significantly improve the robustness of machine translation models. By training on data that captures subtle grammatical variations, models can handle complex language structures more accurately, enhancing their performance in real-world applications. This method provides a strategic advantage, allowing models to learn linguistic nuances that are often underrepresented in traditional data augmentation approaches [28].

5.5 Application to African Languages

Research focusing on African languages has shown promising results. Gitau and Marivate (2023) explored textual augmentation techniques for Swahili, employing methods such as synonym replacement, random insertion and contextual data augmentation. Their work indicated potential improvements in NMT systems for low-resource African languages, demonstrating the viability of such techniques in real-world applications [18].

These studies collectively highlight the importance of data augmentation in advancing MT systems for low-resource languages. Expanding training datasets through diverse augmentation techniques can significantly improve the performance and reliability of translation models, contributing to the preservation and accessibility of under-represented languages.

6 Discussion

This section offers an in-depth analysis of the results obtained from applying data augmentation techniques to enhance the performance of a Xitsonga-English translation model designed for low-resource settings. The study followed a systematic approach, involving data augmentation, cosine similarity analysis to assess translation quality, threshold setting for labeling translations, and finally, training various machine learning and transformer models on the augmented dataset. The objective was to enhance model performance through careful augmentation and selection of high-quality translated data.

6.1 Data Augmentation Techniques

Augmentation Technique	**BLEU Score**	**METEOR Score**
Back-Translation	54.98	0.6227
Random Deletion	66.59	0.8621
Random Insertion	80.61	0.9189
Word Replacement	89.45	0.9361
Synonym Replacement	40.28	0.4566

Table 2: BLEU and METEOR Scores for Augmentation Techniques

To address the limited data in Xitsonga-English translation, five augmentation techniques were applied: Back-Translation, Random Deletion, Random Insertion, Word Replacement, and Synonym Replacement. These techniques were chosen to introduce controlled variations in the dataset, allowing the model to learn from diverse sentence structures.

The effectiveness of each technique was evaluated using BLEU and METEOR scores, metrics that reflect how well the augmented translations align with reference translations. Higher scores imply closer semantic alignment, indicating better-quality augmentations. The results are summarized in Table 6.1 above.

6.2 Visualization of BLEU and METEOR Scores

Figures 1 and 2 illustrate the BLEU and METEOR scores across different augmentation techniques.

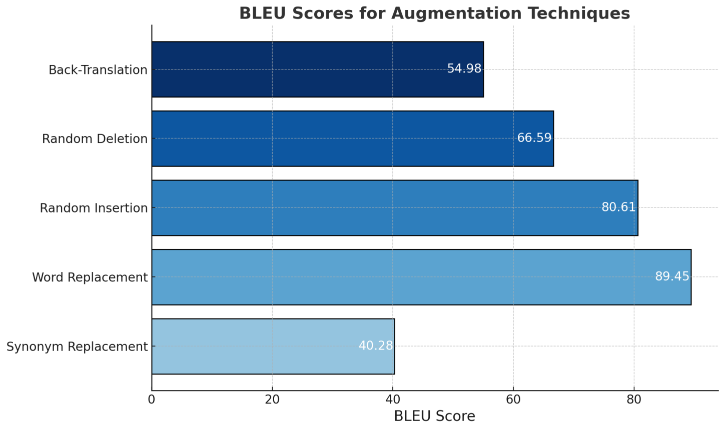

Figure 1: BLEU Scores for Augmentation Techniques

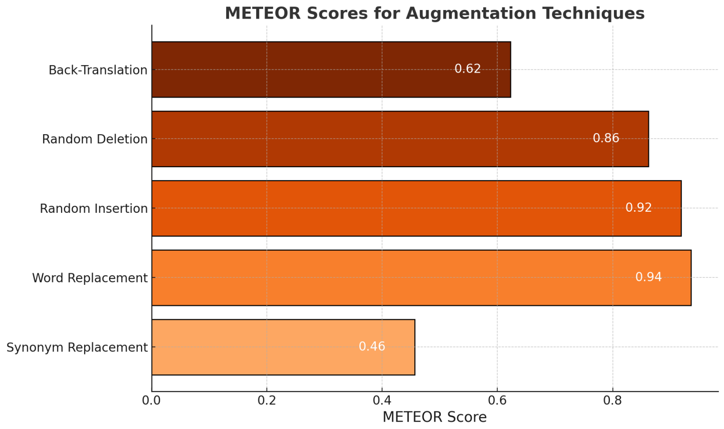

Figure 2: METEOR Scores for Augmentation Techniques

6.3 Cosine Similarity for Quality Assessment

Following augmentation, cosine similarity scores were computed between each augmented sentence and its original sentence to assess semantic consistency. Higher similarity scores indicate that the augmented sentence maintains the meaning of the original, making it suitable for model training. Figures 4 and 3 show the distribution and comparison of cosine similarity scores for each augmentation technique.

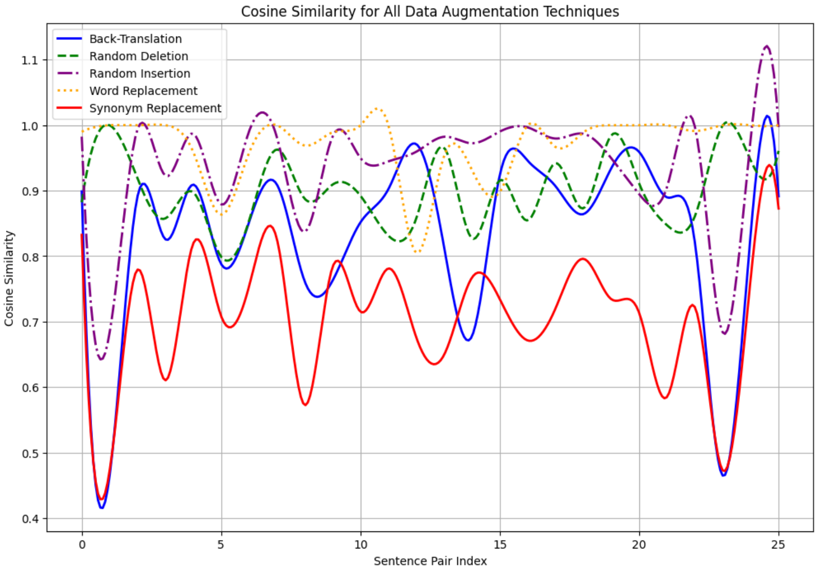

Figure 3: Cosine Similarity Scores for All Data Augmentation Techniques

6.4 Threshold Determination for Labeling

An essential step in this study was selecting an appropriate threshold for cosine similarity to classify augmented translations as "well-translated" or "poorly translated." The purpose of this threshold was to identify augmentations that maintained a high degree of semantic similarity with the original sentences, thereby enhancing the data quality for the translation model.

To determine the optimal threshold, we analyzed the distribution of cosine similarity scores across different augmentation techniques (see Figure 4). Initially, we observed that augmentations with cosine similarity scores above 0.8 typically retained high semantic alignment with the original sentences. To refine this further, we iteratively tested threshold values between 0.8 and 0.9, aiming to achieve a balanced distribution of labels across the augmented dataset.

Ultimately, a threshold of 0.84 was selected as it provided the best balance, labeling approximately half of the augmented sentences as "well-translated." Augmented sentences with a cosine similarity score above 0.84 were labeled as "well-translated," while those below this threshold were labeled as "poorly translated." This threshold, represented by the red dashed line in Figure 4, allowed us to retain high-quality augmentations while filtering out augmentations that deviated too much from the original meaning.

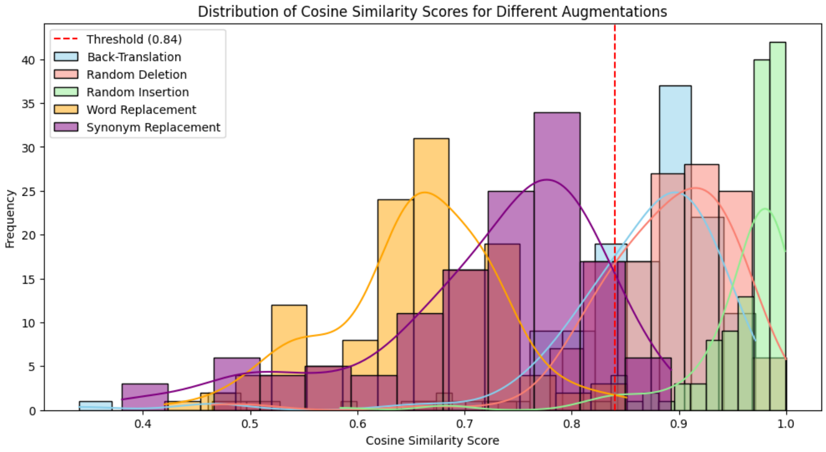

Figure 4: The distribution of Cosine Similarity scores for different Augmentation techniques

6.5 Comparison of Traditional and Transformer-Based Models

After threshold-based filtering, we trained and evaluated both traditional machine learning models (Logistic Regression, Naive Bayes, and SVM) and transformer-based models (mBERT, DistilBERT, XLM-R, and RoBERTa) on the augmented dataset. This comparison aimed to assess the performance impact of data augmentation across different modeling approaches.

6.5.1 Traditional Models

The traditional models included Logistic Regression, Naive Bayes, and SVM. Table 6.5.2 shows their classification metrics on the augmented dataset. These models achieved reasonable accuracy and F1 scores, demonstrating that they could extract useful information from the augmented data. However, their performance was generally lower than that of the transformer-based models, likely due to their limited ability to capture complex language structures.

6.5.2 Transformer-Based Models

Transformer models, especially DistilBERT, XLM-R, and RoBERTa, outperformed traditional models. Table 6.5.2 provides the classification results for each model. The F1 score of 0.8714, along with an accuracy of 0.8626, highlighted DistilBERT's effectiveness in processing diverse and augmented data.Transformer-based models like mBERT, RoBERTa, and XLM-R also performed well, showcasing their strength in managing complex linguistic features introduced by data augmentation. These results emphasize the advantage of transformer architectures in low-resource settings where data quality and variability are crucial.

Model	**Accuracy**	**F1 Score**
Logistic Regression	0.7939	0.8302
Naive Bayes	0.7939	0.8364
SVM	0.8244	0.8553
mBERT	0.8244	0.8369
DistilBERT	0.8626	0.8714
XLM-R	0.8244	0.8369
RoBERTa	0.8473	0.8571

Table 3: Classification Metrics for Different Models

6.6 Training and Validation Metrics for Transformer Models

The validation accuracy and validation loss for mBERT, DistilBERT, XLM-R, and RoBERTa over 10 epochs are depicted in Figures 5 and 6. These figures provide a comparative analysis of each model's training progression on the augmented dataset.

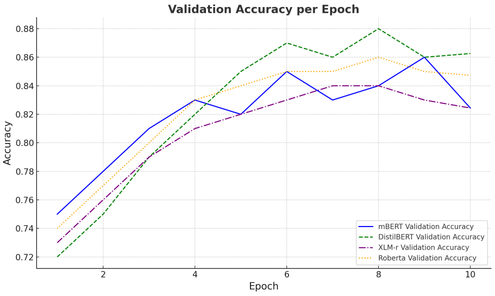

Figure 5: Validation Accuracy per Epoch

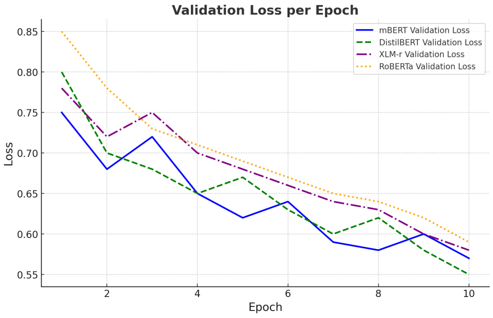

Figure 6: Validation Loss per Epoch

The validation accuracy plot indicates that DistilBERT achieved the highest overall performance, displaying consistent improvements across the epochs and reaching stable accuracy by the 10th epoch. RoBERTa followed with a similarly positive trend, indicating strong learning capabilities.

Both mBERT and XLM-R showed a steady but slightly less pronounced increase in accuracy, with some variability observed throughout the epochs. This suggests that while they effectively learned from the augmented data, their convergence was not as consistent as DistilBERT's and RoBERTa's.

The validation loss plot complements these findings, where DistilBERT demonstrated the most stable decline, highlighting its robust adaptation to the training data. RoBERTa also showed a clear downward trend, signifying effective learning. mBERT and XLM-R experienced more fluctuations in loss, pointing to occasional challenges in model optimization during training.

Overall, the figures emphasize that DistilBERT and RoBERTa outperformed the other models in terms of training stability and final accuracy, showcasing the benefits of using advanced transformer architectures on augmented datasets.

Below are the validation accuracy and validation loss over 30 epochs for the best-performing model, DistilBERT, which are shown in Figures 7 and 8. These trends illustrate successful model convergence, indicating that DistilBERT was well-suited to learning from the augmented dataset. The stable reduction in validation loss and improvement in accuracy further support the effectiveness of using data augmentation in low-resource machine translation.

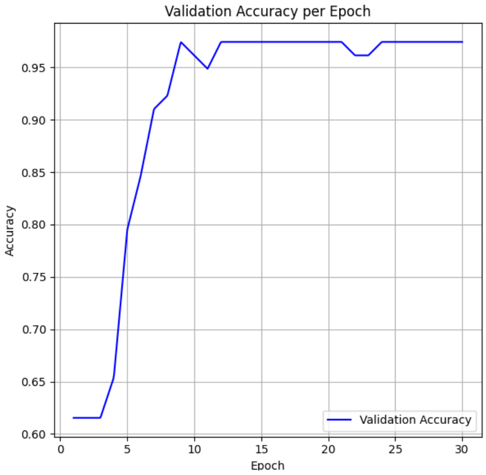

Figure 7: Validation Accuracy per Epoch

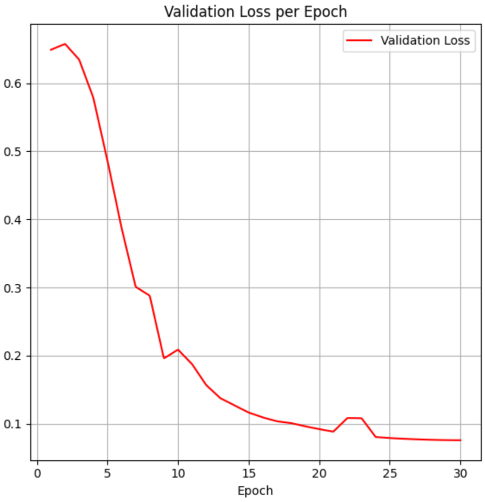

Figure 8: Validation Loss per Epoch

6.7 Summary of Findings

The results indicate that data augmentation significantly improves translation quality for low-resource languages like Xitsonga. Techniques such as Random Insertion and Word Replacement generated the highest-quality augmentations, as reflected in their BLEU and METEOR scores. The threshold of 0.84 for cosine similarity effectively filtered well-translated sentences, maintaining the dataset's semantic integrity.

Among the models tested, DistilBERT achieved the highest classification performance, demonstrating the advantage of transformer models in low-resource settings. While traditional models like Logistic Regression and SVM provided useful baselines, transformer models proved more capable of handling the variability introduced by data augmentation, making them preferable for complex linguistic tasks.

7 Conclusion

This study shows that data augmentation can improve the performance of translation models for low-resource languages like Xitsonga. By applying and evaluating multiple data augmentation techniques, including back-translation, random deletion, random insertion, word replacement, and synonym replacement, we expanded the dataset and introduced variability that improved the model's ability to generalize. The analysis revealed that techniques like Random Insertion and Word Replacement yielded the highest BLEU and METEOR scores, indicating their effectiveness in generating high-quality augmentations.

A key methodological component was the use of cosine similarity to classify augmented sentences based on their semantic consistency with the original sentences. Setting an optimal threshold of 0.84 ensured that only well-aligned augmentations were included for model training, preserving the quality of the dataset and enhancing translation consistency detection. Future work may explore hybrid augmentation techniques or fine-tuning threshold values to further enhance the translation quality for underrepresented languages like Xitsonga.

The results also highlighted the superior performance of transformer-based models over traditional machine-learning classifiers. DistilBERT, in particular, achieved the highest accuracy and F1 score, showcasing the potential of advanced architectures in managing the complexities introduced by data augmentation. While traditional models like Logistic Regression and SVM provided valuable baselines, transformer models proved to be more capable of handling the semantic richness and structural variability of the augmented data.

In conclusion, this research underscores the effectiveness of data augmentation as a strategy to overcome the data scarcity challenges in low-resource language translation. The findings contribute to the broader goal of improving translation accuracy and linguistic inclusivity for underrepresented languages. Future research could explore hybrid augmentation techniques and further fine-tune selection thresholds to optimize translation quality and robustness.

References

[1] P. Chiguvare and C. W. Cleghorn, "Improving transformer model translation for low resource south african languages using bert," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8, IEEE, 2021.

[2] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech communication*, vol. 56, pp. 85–100, 2014.

[3] V. Marivate, T. Sefara, V. Chabalala, K. Makhaya, T. Mokgonyane, R. Mokoena, and A. Modupe, “Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi,” *arXiv preprint arXiv:2003.04986*, 2020.

[4] A. Ziyaden, A. Yelenov, F. Hajiyev, S. Rustamov, and A. Pak, “Text data augmentation and pre-trained language model for enhancing text classification of low-resource languages,” *PeerJ Computer Science*, vol. 10, p. e1974, 2024.

[5] F. Nzama, “A comparison of data augmentation techniques for nguni languages statistical and neural machine translation models,” 2021.

[6] M. Fadaee, A. Bisazza, and C. Monz, “Data augmentation for low-resource neural machine translation,” *arXiv preprint arXiv:1705.00440*, 2017.

[7] V. Marivate, D. Njini, A. Madodonga, R. Lastrucci, and J. Dzingirai, Isheanesu Rajab, “The vuk’uzenzele south african multilingual corpus,” Feb. 2023.

[8] A. Sugiyama and N. Yoshinaga, “Data augmentation using back-translation for context-aware neural machine translation,” in *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)* (A. Popescu-Belis, S. Loáiciga, C. Hardmeier, and D. Xiong, eds.), (Hong Kong, China), pp. 35–44, Association for Computational Linguistics, Nov. 2019.

[9] L. F. A. O. Pellicer, T. M. Ferreira, and A. H. R. Costa, “Data augmentation techniques in natural language processing,” *Applied Soft Computing*, vol. 132, p. 109803, 2023.

[10] D. Kurokawa, C. Goutte, and P. Isabelle, “Automatic detection of translated text and its impact on machine translation,” in *Proceedings of Machine Translation Summit XII: Papers*, 2009.

[11] C. N. Kamath, S. S. Bukhari, and A. Dengel, “Comparative study between traditional machine learning and deep learning approaches for text classification,” in *Proceedings of the ACM Symposium on Document Engineering 2018*, pp. 1–11, 2018.

[12] R. Kora and A. Mohammed, “A comprehensive review on transformers models for text classification,” in *2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pp. 1–7, 2023.

[13] F. Rahutomo, T. Kitasuka, M. Aritsugi, *et al.*, "Semantic cosine similarity," in *The 7th international student conference on advanced science and technology ICAST*, vol. 4, p. 1, University of Seoul South Korea, 2012.

[14] 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 of the Association for Computational Linguistics* (P. Isabelle, E. Charniak, and D. Lin, eds.), (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.

[15] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, eds.), (Ann Arbor, Michigan), pp. 65–72, Association for Computational Linguistics, June 2005.

[16] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint arXiv:2008.05756*, 2020.

[17] Y. Aliyu, A. Sarlan, K. Usman Danyaro, A. S. B. A. Rahman, and M. Abdullahi, "Sentiment analysis in low-resource settings: A comprehensive review of approaches, languages, and data sources," *IEEE Access*, vol. 12, pp. 66883–66909, 2024.

[18] C. Gitau and V. Marivate, "Textual augmentation techniques applied to low resource machine translation: Case of swahili," *arXiv preprint arXiv:2306.07414*, 2023.

[19] W.-H. Her and U. Kruschwitz, "Investigating neural machine translation for low-resource languages: Using bavarian as a case study," *arXiv preprint arXiv:2404.08259*, 2024.

[20] W. Yang and G. Nicolai, "Neural machine translation data generation and augmentation using chatgpt," *arXiv preprint arXiv:2307.05779*, 2023.

[21] A. Lucas, A. Baladón, V. Pardiñas, M. Agüero-Torales, S. Góngora, and L. Chiruzzo, "Grammar-based data augmentation for low-resource languages: The case of Guarani-Spanish neural machine translation," in *Proceedings of the 2024 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (K. Duh, H. Gomez, and S. Bethard, eds.), (Mexico City, Mexico), pp. 6385–6397, Association for Computational Linguistics, June 2024.

[22] M. Mojapelo and J. Buys, "Data augmentation for low resource neural machine translation for sotho-tswana languages," 2023.

[23] R. Lastrucci, I. Dzingirai, J. Rajab, A. Madodonga, M. Shingange, D. Njini, and V. Marivate, "Preparing the vuk'uzenzele and za-gov-multilingual south african multilingual corpora," *arXiv preprint arXiv:2303.03750*, 2023.

[24] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (K. Erk and N. A. Smith, eds.), (Berlin, Germany), pp. 86–96, Association for Computational Linguistics, Aug. 2016.

[25] Y. Gao, F. Hou, H. Jahnke, and R. Wang, "Data augmentation with diversified rephrasing for low-resource neural machine translation," in *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pp. 35–47, 2023.

[26] L. Zeng, "Generative-adversarial networks for low-resource language data augmentation in machine translation," in *2024 6th International Conference on Natural Language Processing (ICNLP)*, pp. 11–18, IEEE, 2024.

[27] L. Zeng, "A generative-adversarial approach to low-resource language translation via data augmentation," *Journal of Student Research*, vol. 12, no. 4, 2023.

[28] M. M. I. Alam, S. Ahmadi, and A. Anastasopoulos, "A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages," *arXiv preprint arXiv:2402.01939*, 2024.