Evaluating Cross-Lingual Semantic Relatedness in African Languages using Transfer Learning and Explanatory Techniques

Student Number: u20464348
Jaimen Govender
University of Pretoria
Email: u20464348@tuks.co.za

Student Number: u25735056
Kago Motlhabane
University of Pretoria
Email: u25735056@tuks.co.za

Student Number: u20435216
Nevin Thomas
University of Pretoria
Email: u20435216@tuks.co.za

Abstract—**This project investigates the effectiveness of multilingual pre-trained language models for measuring semantic relatedness in low-resource African languages, with a focus on Afrikaans. Using the SemRel datasets from 2022 and 2024, we benchmark traditional statistical approaches and embedding-based baselines against fine-tuned transformer models such as AfroXLM-R and LaBSE. We apply transfer learning, back-translation-based data augmentation, and cross-lingual techniques to address data scarcity. Furthermore, interpretability methods—including LIME—are used to examine model reasoning and reveal linguistic challenges unique to African languages. The study aims to provide insights for building robust and transparent NLP tools for underrepresented languages.**

I. Problem Statement

- Which multilingual pre-trained language models—such as AfroXLM-R, LaBSE—achieve the best performance in identifying semantic relatedness in African language texts when fine-tuned using transfer learning?
- What are the performance limitations and challenges specific to these models in low-resource settings?
- Can interpretability tools such as LIME help explain how these models arrive at their predictions?

II. Introduction

Many African languages are classified as low-resource, meaning they lack the extensive annotated datasets typically required for training state-of-the-art natural language processing (NLP) models [8]. This scarcity hampers the development of robust language technologies for a significant portion of the global population. Transfer learning has emerged as a viable solution by adapting large pre-trained multilingual models to low-resource settings.

This project measures semantic relatedness between sentence pairs in Afrikaans—a low-resource language. By leveraging multilingual transformer models such as AfroXLM-R [8] and LaBSE [4], we evaluate the potential of transfer learning to bridge the resource gap. The SemEval-2024 Task 1 [9] provides a comprehensive framework for evaluating semantic textual relatedness across 14 African and Asian languages, including Afrikaans.

Our key objectives are:

- To benchmark semantic relatedness performance on African languages using multilingual pre-trained models;
- To assess interpretability using LIME;
- To apply back-translation for data augmentation in low-resource settings;
- To identify challenges and ethical considerations in adapting NLP tools to African contexts.

III. Literature Survey

A. Semantic Relatedness in NLP

Semantic relatedness refers to the degree of relation in meaning between two pieces of text. Relation refers to varied common associations in usage contexts, broader than simply similarity in meaning. Early methods relied on lexical resources such as WordNet [6], and statistical measures such as cosine similarity between TF-IDF vectors. With the advent of deep learning, distributed representations like Word2Vec [5], GloVe [10], and later contextual embeddings like BERT [3] significantly improved performance on relatedness tasks.

B. Transfer Learning for Low-Resource Languages

Transfer learning, especially through multilingual transformers, has shown promise for low-resource language processing. Pre-trained models such as mBERT [3], XLM-R [2], and LaBSE [4] are capable of cross-lingual transfer, enabling them to generalize to languages with limited annotated data. AfroXLM-R [8] specifically tailors this approach to African languages by including them in the pre-training corpus, improving performance on downstream tasks.

C. Cross-Lingual Transfer and Data Augmentation

Data scarcity remains a problem. Cross-lingual transfer from high-resource to low-resource languages allows models to reuse shared representations [11]. Back-translation [13], where sentences are translated to a pivot language and back, is a common data augmentation method that generates paraphrases for improved generalization. The SemEval-2024 shared task [9] has demonstrated the effectiveness of machine translation for data augmentation in addressing low-resource challenges.

D. Interpretability in NLP Models

As large language models become more complex, interpretability tools are essential for understanding their decision-making. LIME (Local Interpretable Model-agnostic Explanations) [12] provides local explanations by approximating model behavior around a specific instance. In NLP, LIME highlights which words contribute most to a model's prediction.

IV. Methodology

V. Dataset and Preprocessing

A. Datasets Used

a) SemRel2024 (Afrikaans Subset)::

- Main dataset containing 751 sentence pairs from the SemEval-2024 Task 1 [9].
- Human-annotated gold standard labels for semantic relatedness.

b) SemRel2022 (English)::

- Supplementary dataset with 5499 English sentence pairs.
- Also includes human-annotated gold standard labels.
- Back-translated into Afrikaans using Google Translate following the data augmentation approaches used in recent multilingual semantic relatedness work [9].
- Although back-translation introduces minor accuracy degradation, the quality is acceptable for training.

B. Data Preprocessing

a) Data Format::

- The final dataset is a combination of the SemRel2024 Afrikaans data and the back-translated SemRel2022 data.
- Each record contains:
  - `sentence1`
  - `sentence2`
  - `label` (semantic relatedness score)

C. Dataset Splitting

To ensure fair evaluation and preserve label distribution across splits, stratified sampling was applied:

- 70% training set: 4375 records
- 30% test set: 1875 records

Stratified splitting helps maintain a balanced distribution of semantic similarity scores between the training and test sets.

D. Exploratory Data Analysis

Before building models, it's essential to understand the data. EDA helps uncover data patterns, detect imbalances or noise, and guide model design for more accurate and interpretable results.

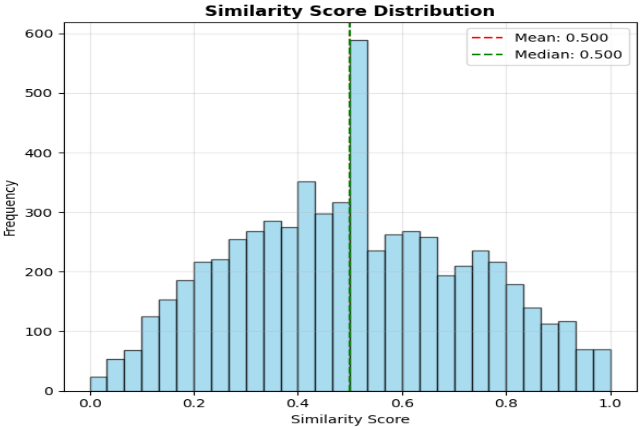

Fig. 1. Similarity score distribution

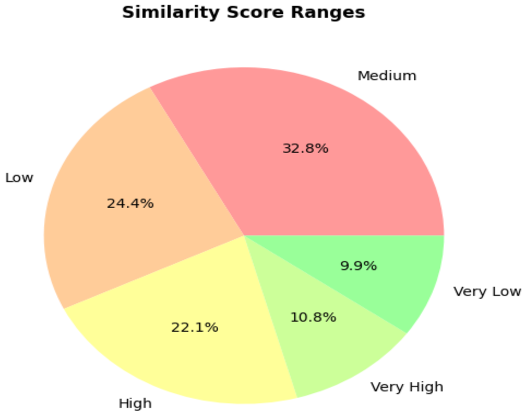

Fig. 2. Similarity Score Ranges

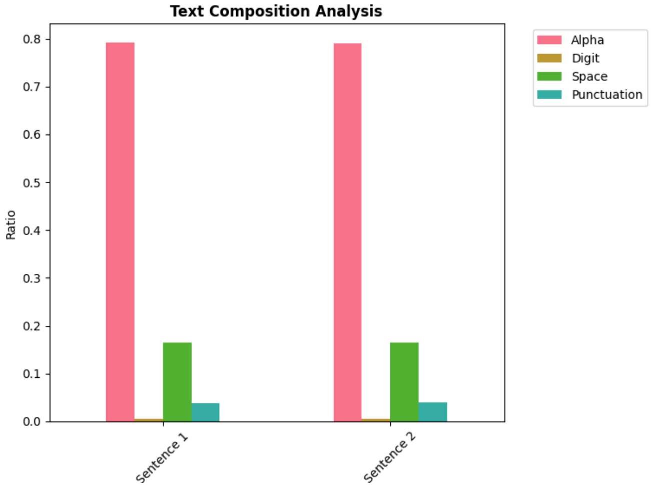

Fig. 3. Text Composition

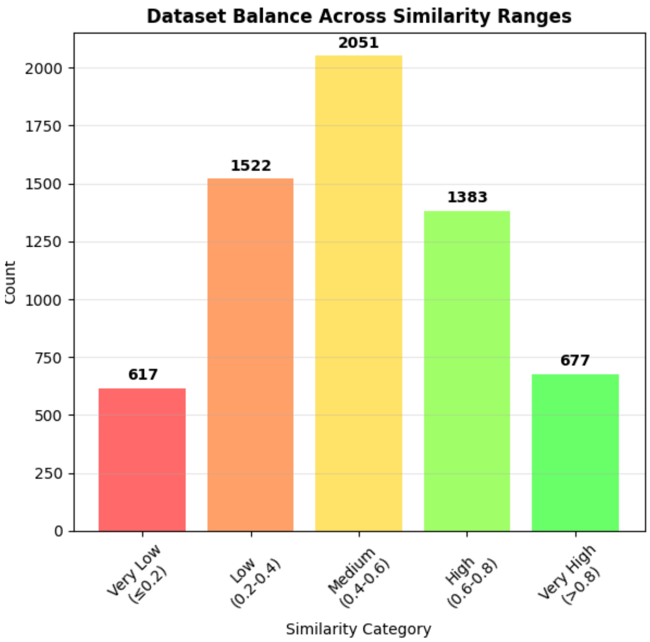

Fig. 4. Dataset Balance Across Similarity Ranges

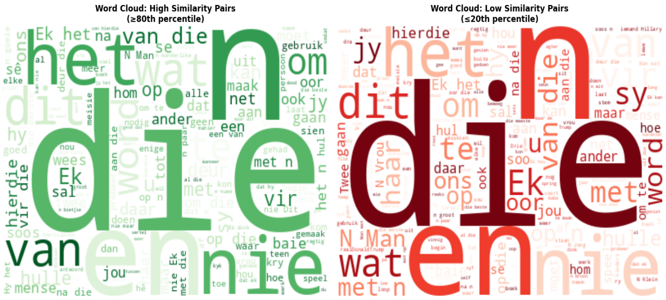

Fig. 5. Word Clouds Showing How Words Affect Relatedness

E. Baseline Models

A baseline model is used to compare the performance of more sophisticated models by providing a simple foundation. Baseline models include:

- TF-IDF to linear regression pipeline
- Freeze AfroXLMR embeddings to linear regression pipeline

F. AfroXLM-R Fine-Tuned Model Architecture

1) Transfer Learning Strategy:

- **Base Model:** `AfroXLM-R Large` — an adaptation of the multilingual XLM-RoBERTa transformer specifically designed for African languages [8].
- **Pre-trained Foundation:** XLM-R is pre-trained on massive multilingual corpora [2].
- AfroXLM-R is further fine-tuned specifically for African languages, enhancing performance on underrepresented linguistic features.

2) Fine-Tuning Process:

- Load the pre-trained AfroXLM-R model and fine-tune it using the custom Afrikaans dataset.
- Predict semantic relatedness scores between 0 and 1 using a regression head.
- Add a regression head using attention layers for the similarity prediction task.
- The model updates its embeddings through backpropagation, adjusting weights per layer via gradient descent to minimize error.

3) Training Parameters:

- Seed: 42
- Batch size: 32
- Epochs: 4
- Learning rate: 2×10^{-5}

4) Model Evaluation Metrics:

- **Loss Function:** Mean Squared Error (MSE)
- Pearson Correlation
- Spearman Correlation
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

G. LaBSE Fine-Tuned Model Architecture

1) Transfer Learning Strategy:

- **Base Model:** `LaBSE (Language-Agnostic BERT Sentence Embeddings)` — a multilingual model for sentence embeddings supporting 109 languages [4].
- **Pre-trained Foundation:** LaBSE is trained on large-scale multilingual corpora including 17 billion monolingual sentences and 6 billion bilingual sentence pairs [4].

2) Fine-Tuning Process:

- Load the pre-trained LaBSE model and fine-tune it using the custom Afrikaans dataset.
- Predict relatedness scores on a continuous scale from 0 to 1.
- Add a regression head using attention layers for semantic similarity estimation.
- Modify internal embeddings through self-attention and gradient-based optimization.
- Backpropagate to minimize error by adjusting weights layer-by-layer.

3) Training Parameters:

- Seed: 42, 1042, 2042, 3042, 4042, 5042, 6042, 7042, 8042, 9042
- Batch size: 32
- Epochs: 4
- Learning rate: 2×10^{-5}

4) Model Evaluation:

- **Loss Function:** Cosine similarity loss

5) Evaluation Metrics:

- Pearson Correlation
- Spearman Correlation
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

VI. Experiments and Results

The entire experiment is looped 10 times, each with a different random seed. This ensures that model performance is not the result of a particular initialization. This ensures robust results following best practices in multilingual NLP evaluation [9].

A. Evaluation Metrics

- **Pearson Correlation:** Measures the linear relationship between the predicted and true similarity scores.
- **Spearman Correlation:** Measures the rank-order correlation to evaluate monotonic relationships.
- **Mean Squared Error (MSE):** Captures the average squared differences between predicted and actual values.
- **Mean Absolute Error (MAE):** Reflects the average magnitude of the absolute prediction errors.

B. LIME Visualization

Local Interpretable Model-agnostic Explanations (LIME) [12] is used to interpret the predictions of complex models by approximating them locally with an interpretable model. In this project, LIME was applied to analyze the feature importance for semantic similarity predictions. Specifically, it highlights which parts of the input sentences contributed most to the model's output, providing insights into the model's decision-making process.

The visualization includes:

- Feature importance scores for each sentence individually.
- Combined feature importance to compare contributions from both sentences.
- Highlighted text segments indicating influential words or phrases.

This approach helps validate the model's behavior and ensures transparency in semantic similarity assessment.

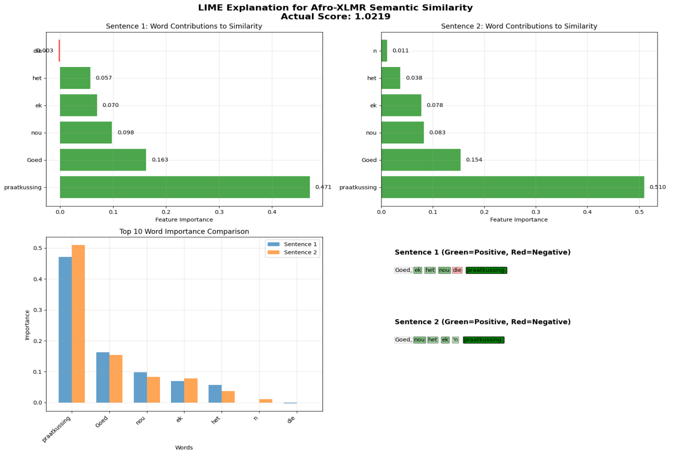

Fig. 6. High Similarity Case AFRO-XLMR

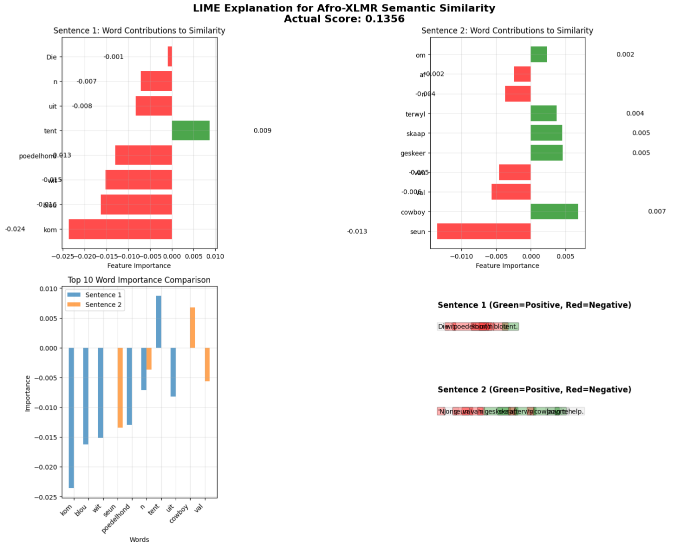

Fig. 7. Low Similarity Case AFRO-XLMR

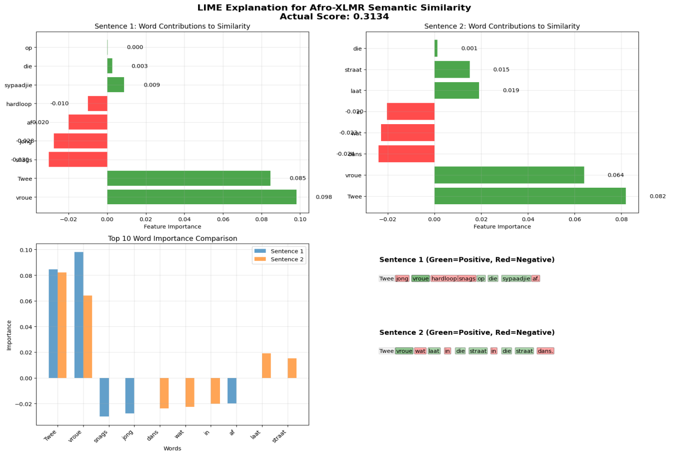

Fig. 8. Largest Prediction Error Case AFRO-XLMR

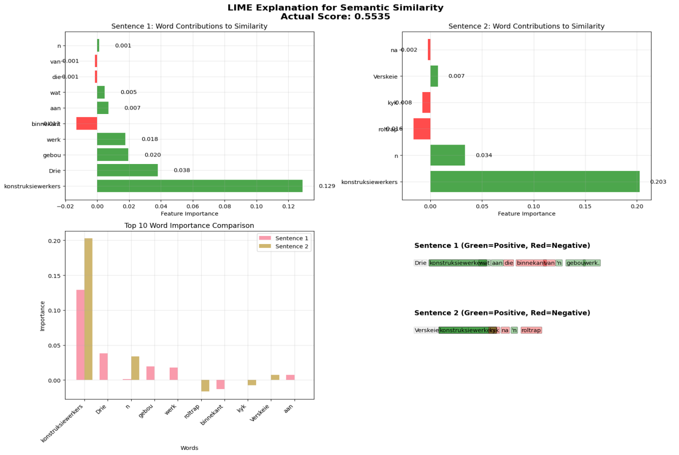

Fig. 9. High Similarity Case LaBSE

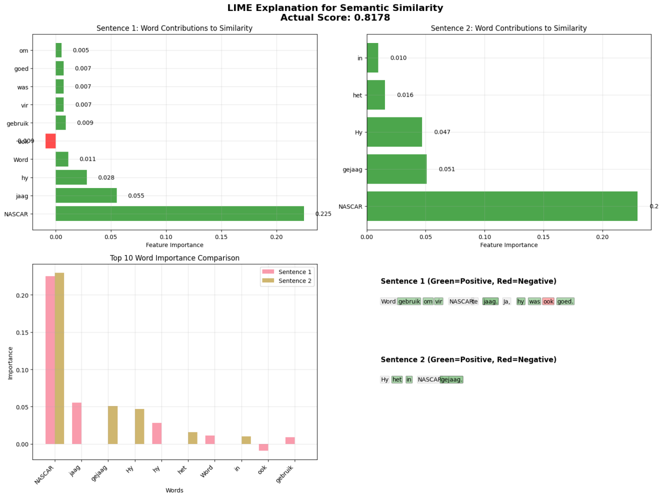

Fig. 10. Low Similarity Case LaBSE

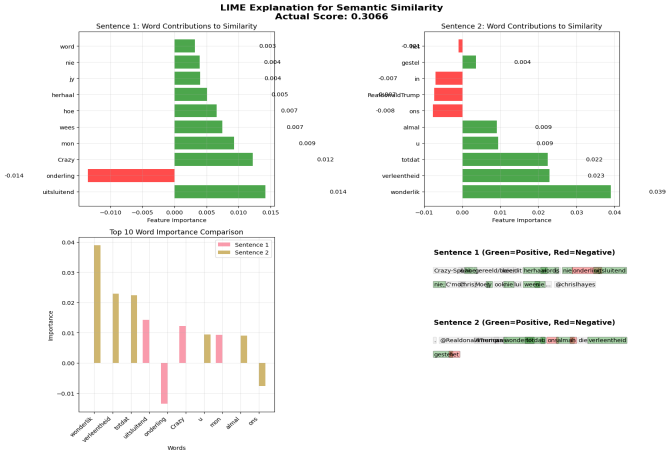

Fig. 11. Largest Prediction Error Case LaBSE

C. Results

Average results over 10 runs:

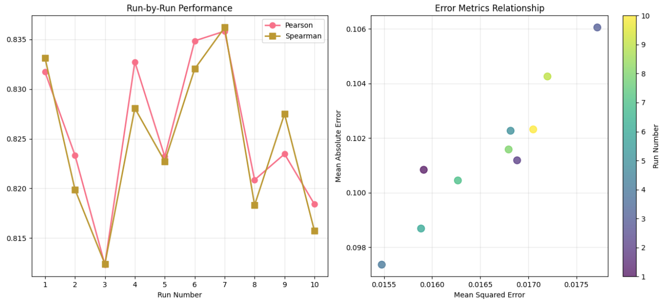

Fig. 12. Fine tuned LaBSE with Transfer Learning Results over 10 runs

VII. Conclusion

This study evaluated various models for semantic relatedness prediction in Afrikaans using the SemRel datasets [9]. The results demonstrate that transformer-based models significantly outperform traditional baseline methods. Among the models tested, LaBSE [4] achieved the highest performance across all evaluation metrics, with an average Pearson correlation of 0.8253 and the lowest MSE and MAE, indicating strong alignment with human-annotated similarity scores. AfroXLM-R [8] also performed well but was slightly less effective than LaBSE. In contrast, traditional methods such as TF-IDF and CLS Embeddings with Logistic Regression showed poor correlation and higher error rates. These findings highlight the effectiveness of multilingual pre-trained models for semantic tasks in low-resource languages and align with recent findings from the SemEval-2024 shared task on multilingual semantic relatedness [9].

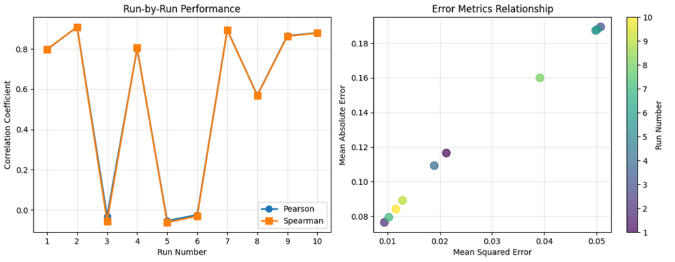

Fig. 13. Fine Tune AFRO-XLMR with Fine Tuning Results over 10 runs

Model	Avg Pearson	Avg Spearman	Avg MSE	Avg MAE
AfroXLM-R	0.6508	0.6532	0.0324	0.1282
LaBSE	0.8253	0.8252	0.0166	0.1015
TF-IDF	0.1559	0.1691	0.0841	0.2303
CLS	0.1982	0.2014	0.0668	0.2062

TABLE I
Performance comparison of multilingual models

VIII. Project Repository

The source code and additional resources for this project are available on GitHub: https://github.com/kaglet/afrikaans_sem_rel

References

[1] A. T. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow, ”AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages,” arXiv preprint arXiv:2211.03263, 2022.

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, ”Unsupervised Cross-lingual Representation Learning at Scale,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440-8451.

[3] 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.

[4] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, ”Language-agnostic BERT Sentence Embedding,” in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 878-891.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, ”Distributed Representations of Words and Phrases and their Compositionality,” in Advances in Neural Information Processing Systems, 2013, pp. 3111-3119.

[6] G. A. Miller, ”WordNet: A Lexical Database for English,” Communications of the ACM, vol. 38, no. 11, pp. 39-41, 1995.

[7] W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Fagbohungbe, S. O. Akinola, S. Muhammad, S. Kabongo, S. Osei, F. Freshia, et al., ”Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2144-2160.

[8] K. Ogueji, Y. Zhu, and J. Lin, ”Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages,” in Proceedings of the 1st Workshop on Multilingual Representation Learning, 2021, pp. 116-126.

[9] N. Ousidhoum, S. Mohammad, M. Abdalla, S. Abdalla, A. Ahmad, R. Alam, B. Alphonso, et al., ”SemEval-2024 Task 1: Semantic Textual Relatedness for African and Asian Languages,” in Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), 2024, pp. 4138-4157.

[10] J. Pennington, R. Socher, and C. Manning, ”GloVe: Global Vectors for Word Representation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532-1543.

[11] T. Pires, E. Schlinger, and D. Garrette, ”How Multilingual is Multilingual BERT?” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4996-5001.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ”Why Should I Trust You?: Explaining the Predictions of Any Classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

[13] 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, 2016, pp. 86-96.