Evaluating Transfer Learning for Multilabel Emotion Classification in Bantu Languages using AfriBERTa, XLM-R and Serengeti

Charlize Hanekom
(u22487222)

Jayson du Toit
(u22571532)

Nonkululeko Ntshele
(u21668452)

Abstract

This study addresses the critical gap in NLP tools for isiZulu, isiXhosa, and Kiswahili by evaluating transfer learning for multilabel emotion classification in these Bantu languages(Acheampong et al., 2020). Using the BRIGHTER dataset, we fine-tuned three multilingual models: AfriBERTa, XLM-Roberta, and Serengeti. Our results demonstrate that transfer learning with strategic fine-tuning significantly enhances model performance in resource-constrained settings. AfriBERTa emerged as the most effective model, achieving the best balance between accuracy and computational efficiency. We evaluated models using ROC AUC, F1 scores (Micro/Macro), and Hamming Loss, with detailed error analysis revealing challenges in detecting minority emotions. This research advances language equity for underrepresented African languages and contributes to culturally relevant NLP applications.

1 Introduction

Emotion recognition and analysis - the task of classifying emotions present in a piece of text - is a crucial aspect of Natural Language Processing (NLP) tools, as it enables machines to understand human communication beyond what is directly said in the text, enhancing human-computer interaction and user experience(Acheampong et al., 2020). While emotion analysis has already been thoroughly researched and implemented in various applications, these works are mostly focused on high-resource languages, neglecting languages with limited digital resources and linguistic data(Muhammad et al., 2025). Bantu languages such as isiZulu, isiXhosa and Kiswahili fall into this category, despite being commonly spoken throughout southern and eastern Africa. These languages were chosen due to their similarities in grammar rules, noun classes, and vocabulary.

The lack of large, high-quality, annotated datasets for these languages presents unique challenges for emotion classification, especially since everyday speech can express multiple nuanced emotions at once(Muhammad et al., 2025). This study aims to overcome this challenge by using transfer learning and various fine-tuning strategies to accomplish improvements in the multi-label emotion classification of low-resource Bantu languages, specifically isiZulu, isiXhosa and Kiswahili. Three pretrained models, AfriBERTa, XLM-Roberta, and Serengeti, were therefore chosen and fine-tuned for these three languages using the BRIGHTER+ dataset. Their performance was then compared to determine which models can successfully be adapted for this task, along with which fine-tuning strategies are most beneficial towards this goal. This study aims to lay the groundwork for further research to utilise transfer learning when implementing NLP tools in low-resource languages so that more advanced applications can include underrepresented languages in their technologies.

2 Background

Multilabel emotion classification is a complex task in Natural Language Processing (NLP), particularly for languages with limited resources. It involves assigning multiple emotional categories—such as anger, joy, or sadness—to a single piece of text, recognizing the multifaceted nature of human emotion(Zhang and Zhou, 2013). While significant progress has been made in this area for high-resource languages like English and Mandarin, African languages, especially Bantu languages such as isiZulu, isiXhosa, and Kiswahili, remain vastly underrepresented in both datasets and NLP tools(Muhammad et al., 2025; Hedderich et al., 2020).

Recent studies have sought to bridge this gap by leveraging transfer learning, particularly with multilingual transformer models like XLM-R (Conneau et al., 2019), which exhibit strong cross-lingual generalization. AfriBERTa, trained specifically on African languages, and Serengeti, trained on 517 African languages and dialects(Adebara et al., 2022), have also emerged as promising candidates for adapting emotion classification to African contexts(Ogueji et al., 2021). However, the majority of prior research has focused on monolingual or binary classification, with few studies exploring multilabel emotion classification in low-resource African languages.

One of the main limitations of past work is the lack of large-scale, balanced, annotated datasets. The BRIGHTER+ dataset(Muhammad et al., 2025) addresses this by offering multilingual emotion-labeled data, including annotations for isiZulu, isiXhosa, and Kiswahili. Yet, challenges persist: many of the emotion categories are imbalanced, and the linguistic diversity of Bantu languages introduces complexities such as rich morphology, agglutination, and tonal variation(Nekoto et al., 2020). These factors contribute to difficulty in training models that generalize well across languages and emotional categories.

This project aims to address these gaps by evaluating and fine-tuning AfriBERTa, XLM-R, and Serengeti for multilabel emotion classification in isiZulu, isiXhosa, and Kiswahili. By doing so, it contributes to a growing body of responsible NLP research that centers African languages, acknowledging the historical and structural biases that have led to their exclusion. It also emphasizes transparency by using open-access datasets, while remaining aware of limitations in data balance and model generalization.

3 Methodology

3.1 Data Preparation

We used the isiZulu, isiXhosa and Kiswahili datasets from the BRIGHTER+ dataset collection, which has been published for academic use; these datasets were already split into training and test sets. The following steps were implemented in the data preparation process:

- **Text cleaning:** Text was lowercased, non-alphanumeric characters were removed using regex and trailing white-spaces were trimmed.
- **Label Extraction:** Six emotion labels (`anger`, `disgust`, `fear`, `joy`, `sadness`, `surprise`) were extracted as float vectors.
- **Structured for training:** Data was formatted into dictionaries containing: Cleaned text, emotion label vectors, and language identifiers

3.2 Models

The models were chosen based on their language-specific pertaining, low-resource efficiency and proven success in African NLP tasks. **AfriBERTa** was pre-trained on 17 diverse African languages. This direct exposure to similar languages allows for a better understanding of the inherent linguistic structures of our chosen languages. **XLM-R** is a more generalised model that was trained across more than a hundred languages and has proved cross-lingual transfer capabilities. Finally, **Serengeti** is extremely specialised since it was trained on 517 African languages and dialects, including many Bantu languages.

3.3 Hyperparameter Optimization

To obtain optimal parameters for the models and languages, a grid search fine-tuning method was applied. This means all the possible combinations of the parameter values were tested to find the best performing values. Tested parameters and best configurations:

Parameter	Tested Values
Learning rate	1.00×10^{-5}, 3.00×10^{-5}
Batch size	8, 16
Dropout	0.1, 0.3
Weight decay	0.01, 0.1
Warmup steps	0, 100
Gradient clip	0.5, 1.0

Table 1: Tested hyperparameters

It was found that most of the optimal parameter value combinations tended to be similar.

This leads to the belief that these are the optimal parameters for this type of implantation or that a greater and more aggressive tuning search needs to be done, as the values that could potentially provide the best result were not included here.

Table 2: Optimal hyperparameters per language-model pair

Model	Language	Learning rate	Batch size	Warmup steps	Dropout	Weight decay	Gradient clip
AfriBERTa	Zulu	3.00×10^{-5}	8	100	0.1	0.01	1
AfriBERTa	Xhosa	3.00×10^{-5}	8	100	0.1	0.01	1
AfriBERTa	Swahili	3.00×10^{-5}	8	100	0.1	0.01	1
XLM-R	Zulu	1.00×10^{-5}	16	0	0.1	0.01	1
XLM-R	Xhosa	3.00×10^{-5}	16	0	0.1	0.01	1
XLM-R	Swahili	3.00×10^{-5}	8	0	0.1	0.1	1
Serengeti	Zulu	3.00×10^{-5}	8	100	0.1	0.01	1
Serengeti	Xhosa	3.00×10^{-5}	8	100	0.1	0.01	1
Serengeti	Swahili	3.00×10^{-5}	8	100	0.1	0.01	1

3.4 Training Strategies

- **Layer freezing and gradual unfreezing:** Freezing most of the initial layers while training the new layers, then unfreezing lower layers per epoch. This protects valuable pre-trained knowledge during fine-tuning and allows for a controlled adaptation.

- **Binary cross-entropy loss for multi-label classification:** Using a loss function and treating each emotion label as an independent yes or no prediction. To handle the labelling of multiple emotions in a single text.

- **Early stopping:** Halts training automatically when validation loss stops improving to prevent overfitting.

- **Gradient clipping and learning rate warmup:** To ensure stable training, clipping prevents exploding gradients, and warmup avoids large, destabilising early updates.

3.5 Evaluation Metrics

- **F1-score (Micro/Macro):** Overall classification performance measurement.

- **ROC AUC**: Effective in conveying the performance of multi-label classification models.

- **Precision and Recall:** correctness of positive predictions and coverage of actual positives, respectively.

- **Hamming Loss:** Fraction of incorrectly predicted labels (lower is better).

- **Computational efficiency** (training time, GPU memory) for resource footprint measurement.

4 Experiments and Results

4.1 Performance Metrics

Table 3: Model performance across languages and metrics

Model	Language	F1 Micro	F1 Macro	Hamming Loss	ROC AUC
AfriBERTa	Zulu	0.002	0.001	0.08	0.71
AfriBERTa	Xhosa	0.505	0.182	0.22	0.82
AfriBERTa	Swahili	0.006	0.004	0.10	0.65
XLM-R	Zulu	0.153	0.148	0.92	0.63
XLM-R	Xhosa	0.336	0.241	0.61	0.81
XLM-R	Swahili	0.187	0.155	0.74	0.56
Serengeti	Zulu	0.000	0.000	0.08	0.71
Serengeti	Xhosa	0.492	0.177	0.25	0.83
Serengeti	Swahili	0.000	0.000	0.10	0.64

4.2 Key Findings

- **Language Performance**: In Figure 1 isiXhosa models showed the most significant improvement after tuning, reaching a minimum of 0.8 ROC AUC score with all three models. This might indicate that the isiXhosa dataset is very balanced or even that isiXhosa is very effective at conveying emotion, with specific indicators.

- **Model Comparison**: AfriBERTa outperformed other models across the languages.

- **Error Patterns**: Confusion between "fear" and "sadness" is most common, especially in the isiZulu language. This may indicate an unbalanced dataset with little data on these emotions.

- **Minority Emotions**: Low F1 Macro scores indicate difficulty detecting rare emotions. This can be attributed to similar circumstances as above.

- **Fine Tuning Evaluation**: Finetuning significantly increased the performance of all the models on Xhosa. This might be due to isiXhosa having had the smallest dataset, allowing small parameter changes to have a big impact. isiZulu and Kiswahili stayed relatively the same meaning more testing and tuning is needed to see if the performance boost isiXhosa saw can be replicated.

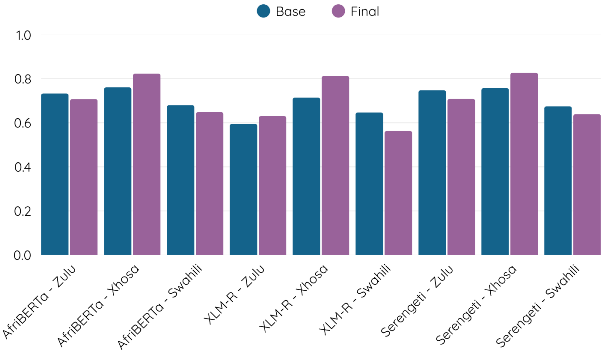

Figure 1: ROC AUC Comparison between base model and Fine-Tuned model

4.3 Computational Efficiency

As can be seen in Figures 2, 3 and 4:

- **Training Speed**: AfriBERTa fastest (3× faster than Serengeti)

- **GPU Memory**: XLM-R required the most memory (up to 10GB)

- **Power Consumption**: Serengeti is the most power-intensive (80W avg)

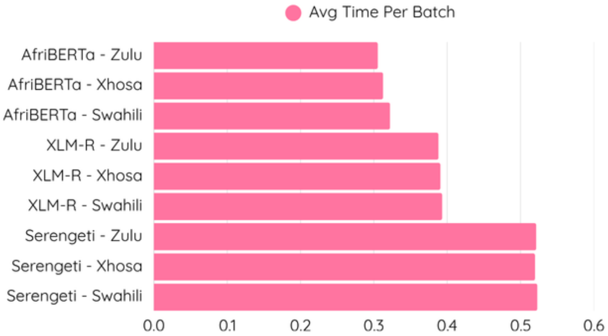

Figure 2: System metrics average time per batch

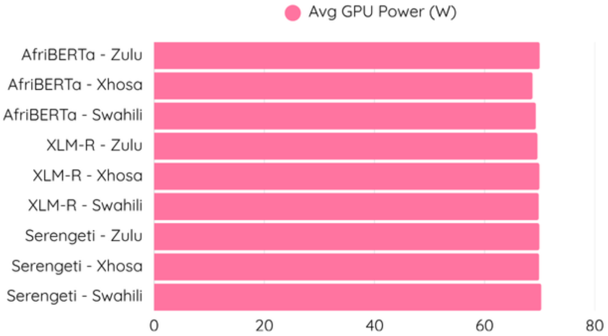

Figure 3: System metrics average GPU power usage per batch

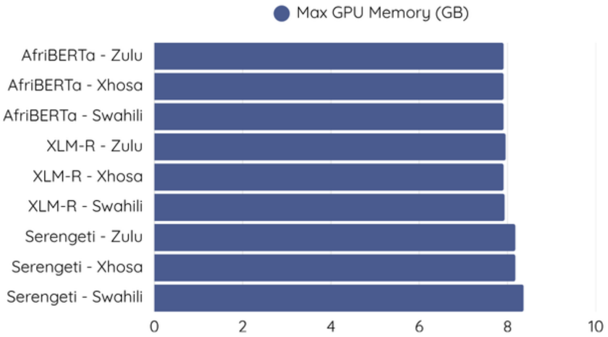

Figure 4: System metrics max GPU memory usage

4.4 Subset Accuracy and ROC AUC Analysis

The evaluation of models across languages using Subset Accuracy and ROC AUC reveals several key trends:

- ROC AUC scores were generally higher than Subset Accuracy across all models and languages, as expected for multilabel classification tasks.
- AfriBERTa-large achieved strong ROC AUC values, particularly in isiXhosa (0.8241) and isiZulu (0.7080), with moderate Subset Accuracy in Zulu (0.5770).
- Serengeti-E250 demonstrated competitive ROC AUC and notably better Subset Accuracy in Swahili and isiZulu compared to AfriBERTa.
- XLM-RoBERTa-base achieved decent ROC AUC in isiXhosa (0.8133) but had zero Subset Accuracy across all languages.
- isiXhosa consistently scored highest in ROC AUC across all models, while Swahili generally lagged.

This suggests that while models can individually detect relevant emotions, exact multilabel prediction remains challenging. The high ROC AUC but low Subset Accuracy gap points to the need for improved thresholding strategies and better modeling of label dependencies.

Table 4: ROC AUC and Subset Accuracy across all models and languages

Model	Variant	Language	ROC AUC	Subset Accuracy
AfriBERTa	large	Zulu	0.7080	0.5770
AfriBERTa	large	Xhosa	0.8241	0.1192
AfriBERTa	large	Swahili	0.6486	0.4432
XLM-R	base	Zulu	0.6308	0.0000
XLM-R	base	Xhosa	0.8133	0.0000
XLM-R	base	Swahili	0.5636	0.0000
Serengeti	E250	Zulu	0.7090	0.5774
Serengeti	E250	Xhosa	0.8276	0.0144
Serengeti	E250	Swahili	0.6397	0.4444

5 Reflections and Discussion

5.1 Performance Analysis

The superior performance of AfriBERTa can be attributed to its specialisation for African languages. isiXhosa's significant improvement in F1 and ROC AUC Score after tuning suggests its smaller dataset responded better to optimisation. This means further tuning with more extreme parameter values might prove beneficial towards increasing the performance of the models for the isiXhosa and Kiswahili. The exceptionally low scores for Zulu across all models indicate fundamental challenges in processing this language, possibly due to morphological complexity. This leads to the belief that a larger training dataset is needed to enable the models to achieve an understanding of the language.

5.2 Error Analysis

Confusion matrices revealed:

- Strongest signals for Sadness and Joy in isiXhosa
- isiZulu models failed to detect most emotions
- Kiswahili marginally better than isiZulu for Joy and Surprise

Confusion Matrix Analysis: AfriBERTa Xhosa sadness classification example

- TP = 25 (correctly predicted sadness)
- FP = 3 (wrongly predicted sadness)
- FN = 7 (missed actual sadness)

Inference Time Analysis: We evaluated how quickly each model makes predictions by logging average inference time at several training steps with Weights & Biases. Figure 5 shows AfriBERTa sitting at the front, consistently clocking between 0.05 and 0.07 seconds per call. XLM-R trails just behind, with slightly higher but still stable times. Serengeti, by contrast, not only ran slower overall, peaking dramatically around step 3.

Together, these findings confirm AfriBERTa as the best choice for tight hardware budgets.

5.3 Limitations

- GPU memory constraints limited batch size options
- Low F1 scores for minority emotions (e.g., ”surprise”)

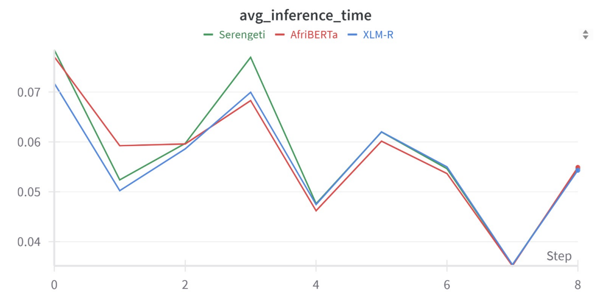

Figure 5: Inference Time Analysis

- Model confusion between semantically close emotions
- Limited dataset size, especially for Zulu

5.4 Responsible NLP Considerations

The project tackled several angles of responsible AI. First, it pushed for language equity by working with a set of African languages that are often ignored.

Second, all data collected by BRIGHTER is openly shared and gathered in an ethical way. No personally identifiable information was included. Researchers have noted that the dataset is still unbalanced, especially for emotions spoken by smaller groups, and that shortcoming probably hurt how well the model performs. They also tried to use the same methods for every language, yet noticeable differences in behavior still exist.

6 Conclusion

This study demonstrates that transfer learning with strategic fine-tuning significantly enhances emotion classification for Bantu languages. AfriBERTa emerged as the most effective model, achieving the best balance between accuracy and computational efficiency. Key findings include:

- Hyperparameter optimisation is crucial for low-resource languages, but more tuning with a wider variety of values is needed to ensure the models are given the chance of being improved.
- Language-specific tuning needed, as each language has its own nuances and challenges. As seen, the different performance of the models on the different languages.
- Minority emotions remain challenging to detect; thus, greater datasets focused on emotion balance are needed.
- Computational efficiency is critical for real-world deployment, thus using models like AfriBERTa is a good choice as it was computationally more effective than the other models.

Future Work:

- Test greater amounts and extreme hyperparameters on larger datasets
- Incorporate the EthioEmo dataset for richer emotion labels and greater training amount.
- Explore linguistic similarity effects (genealogical families), to improve the combined detection of languages.
- Develop lightweight models for mobile deployment so that high processing power is not needed.
- Address GPU limitations through cloud optimization

Dataset Availability Limitation: Even though the original plan featured both the BRIGHTER and EthioEmo datasets, the work reported here draws solely on BRIGHTER. Access issues kept EthioEmo off the table at the implementation stage. Later studies can bring it in, broadening coverage and letting researchers compare emotional patterns more fully across Bantu languages.

This research advances language equity for underrepresented African languages and contributes to culturally relevant NLP applications while highlighting key implementation challenges.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, and 1 others. 2025. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. 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*, pages 116–126.

Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

A Confusion Matrix Example (AfriBERTa Xhosa)

Table 5: Emotion-specific performance across languages

Metric	isiZulu	isiXhosa	Swahili
Sadness F1	0.33	0.58	0.00
Joy F1	0.00	0.47	0.27
Surprise F1	0.00	0.08	0.21
Hamming Loss	0.10	0.25	0.11
Macro F1	0.056	0.19	0.08