African Languages for Emotion Analysis
Enhancing Emotion Classification through Data Augmentation Strategies for Low-Resource African Languages

Mignon Erasmus, Caitlin Simon and Xadrian van Heerdan
University of Pretoria

Abstract

Emotion classification in natural language processing has advanced significantly for high-resource languages, yet African languages remain severely under-represented. This study investigates data augmentation techniques for multilabel emotion classification across three African languages: Afrikaans, Swahili, and Hausa. Using the BRIGHTER dataset and Afro-xlmr-small model, we evaluated Random Insertion, Masked Language Modelling, and Back Translation techniques. Results show Random Insertion achieved the most consistent improvements, with Swahili demonstrating dramatic gains (624% relative improvement in Macro F1-score from 0.0879 to 0.6360), effectively addressing severe class imbalance. The study reveals that balanced class distribution is more critical than dataset size for effective emotion classification. Our findings demonstrate that semantically-aware data augmentation can significantly enhance NLP systems for low-resource African languages, contributing to more inclusive language technologies.

1 Introduction

With the progression of natural language processing (NLP), gaining insight into human sentiment and expression has grown in significance (Ahmad et al.). While advances in machine learning and deep learning have significantly improved performance on emotion classification tasks in high-resource languages like English, these developments have largely excluded low-resource languages, especially those spoken across the African continent (Ronny Mabokela et al., 2025). This disconnect impacts the inclusivity and cultural relevance of NLP systems but also for the communities that are systematically excluded from the benefits of such technologies.

African languages are diverse and rich in linguistic nuance, yet they remain severely under-represented in annotated datasets and NLP research (Ahmad et al.). Emotion expression in these languages often involves culturally specific cues, tonal variations, and structural features that are poorly captured by models trained predominantly on English or other high-resource languages. This lack of representation, compounded by limited training data, poses unique challenges for developing accurate and context-sensitive emotion classification systems.

Recent research in NLP has shown that data augmentation techniques that synthetically increase the size and diversity of training data, can significantly improve model performance, particularly in low-resource or imbalanced settings (Arora and Turcan, 2024) (Koufakou et al., 2023). These techniques have demonstrated success in tasks such as sentiment analysis (Thakkar et al., 2024) and neural machine translation (Gitau and Marivate, 2023). Methods like back translation, contextual embeddings, and paraphrasing enhance model robustness and generalisability. However, while these approaches have been applied in various domains and languages, their effectiveness in the specific context of multilabel emotion classification for low-resource African languages remains largely unexplored.

This study seeks to address this critical gap by investigating the performance and adaptability of data augmentation techniques for emotion classification across Afrikaans, Swahili, and Hausa, three African languages with varying levels of digital resource availability and linguistic structures. By applying and evaluating augmentation strategies across these languages, the research aims to explore both the generalisability and language-specific challenges of emotion classification in low-resource settings.

2 Problem Statement

Emotional classification plays a crucial role in understanding user sentiment, social media as well as the mental health industry. Data augmentation techniques have shown to improve performance of emotional classification on small and imbalanced datasets, but current research has been mostly limited to high-resource languages like English.

There is a critical gap in research and application on applying NLP techniques to low-resource African languages. African languages are both under-represented in annotated datasets (with cultural and linguistic diversity) and NLP development. Although augmentation techniques have been applied in the similar field of sentiment analysis for low-resource languages, the effectiveness and application to multilabel emotion classification remains unexplored.

The research aims to address these challenges by experimenting with context-aware data augmentation techniques to improve multilabel emotion classification in African languages with varying levels of resource availability. Preserving the context and semantic meaning of sentences are especially important in emotion-sensitive tasks.

The above problem statements lead to the following research question: How effective are data augmentation techniques in improving multilabel emotion classification for low-resource African languages?

This question is supported by the following subquestions:

1. How do the proposed augmentation techniques perform across different African languages with different resource levels?

2. How does the performance of models trained with the augmented dataset compare to the base model trained on the original unaugmented dataset?

3 Background

3.1 Key Concepts and Terminology

1. **Emotion Classification** - Emotion classification is a natural language processing (NLP) task that involves identifying and categorising emotions expressed in text. Emotion classification provides a granular understanding by detecting multiple, often overlapping emotional states such as joy, anger, sadness, or fear. Multilabel emotion classification allows for the assignment of more than one emotion to a single text instance, reflecting the complexity of human emotional expression (Plaza-del Arco et al., 2024).

2. **Low-Resource Languages** - Low-resource languages are languages that lack extensive digital linguistic resources such as large annotated corpora, pre-trained language models, and computational tools. Many African languages fall into this category due to limited availability of datasets and NLP research focus(Ògúnremí et al., 2023).

3. **Data Augmentation** - Data augmentation refers to techniques used to artificially increase the size and diversity of training datasets by generating synthetic examples. In NLP, common augmentation methods include back translation (translating text to another language and back), synonym replacement, paraphrasing using language models, and contextual word substitutions. These approaches help mitigate problems related to data scarcity and class imbalance, improving model robustness and generalisability (Feng et al., 2021).

3.2 Related Work

Data augmentation has shown to be a promising technique in NLP for improving performance on emotion classification and related tasks in settings with limited annotated data. By generating synthetic training examples, these techniques help address class imbalance and data scarcity which are two persistent challenges in emotion classification in low resource African languages.

In a study done by Arora and Turcan (2024) data augmentation methods were evaluated for multilabel emotion classification using the IESO dataset (877 entries with 15 emotions mapped to Ekman's framework) and a Reddit corpus (224 493 unlabeled posts). The researchers implemented three key techniques: Back Translation for generating diverse paraphrases; BERT/RoBERTa for masked language modeling focused on emotional tokens; Pseudo-labeling using the NTUA-SLP Bi-LSTM model trained on SemEval-2018 data. The methodology revealed Back Translation outperformed autoencoder approaches, improving classification accuracy by 12–18% when generating 3-5 synthetic examples per training instance.

Koufakou et al. (2023) examined data augmentation strategies for emotion classification in low-resource text datasets, focusing on three English-language corpora: COVID-19 survey responses (2,408 records), EmoEvent-EN tweets (7,303 records), and WASSA-21 essays (1,860 records) annotated with Ekman's emotions. Researchers applied four augmentation methods: Easy Data Augmentation (EDA) with synonym replacement/insertion, static embeddings (GloVe), contextual embeddings (BERT), and ProtAugment using BART paraphrasing. While the study exclusively analysed English data, its methodology demonstrates the effectiveness of augmentation for imbalanced emotion distributions which is a critical challenge for low resource African languages. The paper's findings that BART-based paraphrasing improved classification accuracy by 12-18% on small datasets suggest potential applicability to African language contexts.

In both studies, data augmentation significantly improved emotion classification performance on small datasets, showing its effectiveness in low-resource scenarios. However, both studies focused solely on English datasets, neither extended these methods to low-resource African languages highlighting a critical research gap.

Data augmentation techniques have also been applied to the related NLP task of sentiment analysis. Sentiment analysis focuses on positive, negative, or neutral sentiment, while emotion classification provides a more granular understanding of human expression. The methodologies and data augmentation strategies employed in sentiment analysis, particularly for low-resource languages, can be highly relevant and applicable to enhancing emotion classification in similar contexts.

Thakkar et al. (2024) explores data augmentation techniques for sentiment analysis, specifically addressing challenges in low-resource settings, focusing on Croatian and other South Slavic languages. The authors used a dataset comprising Croatian text, employing a custom Masked Language Model (MLM) CLARE augmentor with constraints from the TextAttack library. They filtered out augmented sentences with a cosine similarity below 0.80 to ensure semantic consistency. The study demonstrates the application of transfer learning and contextual embeddings for data augmentation in languages with limited resources, providing a methodology that could be adapted for emotion classification tasks in similarly under-resourced African languages.

Another study by Ronny Mabokela et al. (2025) applied data augmentation to the NLP task of Neural Machine Translation (NMT) on the low resource language pair English-Swahili. The researchers used parallel corpora from JW300 and Tanzil, where they applied three augmentation strategies commonly used in text classification: Word2Vec-based synonym replacement, TF-IDF-based random word insertion, and Masked Language Model (MLM)-based contextual augmentation using RoBERTa. Augmentations were applied only to the source language (English or Swahili, depending on translation direction), and the resulting data was used to train Transformer-based NMT models. Evaluation using BLEU, METEOR, and chrF scores showed that MLM-based augmentation led to the most consistent improvements, particularly due to its ability to retain sentence context, while the combined Word2Vec and TF-IDF method had mixed results.

Despite the promising advancements in data augmentation for other NLP tasks like sentiment analysis and Neural Machine Translation, there remains a significant lack of research and applied work specifically targeting emotion classification in low-resource African languages. Existing research is typically trained on a narrow subset of high-resource languages, which leads to biased performance when applied to African contexts. The structure and use of African languages often differ significantly from English, especially in how emotion is expressed through grammar, tone, and cultural context.

3.3 Datasets

The lack of high-quality, annotated datasets for African languages has significantly hindered progress in natural language processing (NLP) (Ronny Mabokela et al., 2025), particularly in emotion classification. This scarcity limits the development of accurate, culturally aware models, posing challenges to creating inclusive NLP systems that reflect Africa's rich linguistic and cultural diversity.

To address this, initiatives have introduced multilingual and culturally grounded emotion datasets such as BRIGHTER (Muhammad et al., 2025a,b) and EthioEmo (Belay et al., 2025). These resources span 28 languages, with a strong focus on under-represented African languages like isiZulu, Yoruba, Hausa, Amharic, and Somali. Their inclusion of both African and global languages enables

comparative linguistic and cultural emotion studies. Complementary datasets like AfriSenti and MasakhaNEWS further expand coverage for sentiment and topic classification in languages like Sesotho and Shona (Mokhosi et al., 2024; Muhammad et al., 2023b,a; Mabokela and Schlippe, 2022).

like EthioEmo and the BRIGHTER dataset enable more inclusive and representative NLP research. This project contributes not only to technical advancement but also to linguistic and cultural equity. In doing so, it helps move NLP towards being a truly global and inclusive field.

4 Methodology

4.1 Training Data

A language subset of the BRIGHTER Emotion Categories dataset was used for this research (Muhammad et al., 2025a,b) . This dataset is a multilingual corpus designed for multilabel emotion classification for low-resource languages.

The following African languages were selected, representing different levels of resource availability:

1. Swahili (swa) - 7721 examples
2. Hausa (hau) - 5017 examples
3. Afrikaans (afr) - 3548 examples

The African languages were chosen to reflect a variety of resource availability with Swahili representing a relatively high-resource, Hausa as moderately resourced, and Afrikaans as a low-resource language in this context. Additionally, these languages differ significantly in their linguistic structures and properties. This diverse selection enables a more comprehensive evaluation of model performance across different levels of data availability and linguistic complexity.

The dataset includes six emotion categories: anger, disgust, fear, joy, sadness and surprise.

To ensure consistency through training, the following preprocessing steps were applied:

1. Neutral label assignment: Entries with no emotion labels were assigned a new label "neutral".
2. Stratified Splitting: The datasets were split into training (80%), validation (10%) and testing (10%) using stratified sampling by emotion label. This ensures balanced representations of emotion types across all splits.
3. Multi-label encoding: Emotion annotations were converted into multi-hot encoded vectors. This allows the model to support multi-label classification, where multiple emotions can be associated with an instance.

The BRIGHTER dataset may also only reflect the cultural and linguistic characteristics of the sources from which the data was drawn. Biases in emotion expression may vary across different cultures and languages and it could also impact the model's ability to generalise across different languages and contexts. The BRIGHTER dataset is licensed under the CC-BY 4.0, which allows for the use and modification of the data as long as recognition is given to the original source .

4.2 Model Description

The model used was the AfroXLMR-small, which is a multilingual transformer model trained on 17 African languages, and 3 high-resource languages. It is built upon the XLM-R-base architecture and is compact with 140M params, which is ideal for low-resource settings and for fine-tuning small datasets. The three chosen languages (Afrikaans, Hausa and Swahili) were included in the model's training data which contributed to our choice of model.

The AfroXLMR-small was fine-tuned on the chosen three African languages for emotion classification. The model was configured for multilabel classification as instances may contain more than one emotion. The datasets were tokenised using the model's tokeniser.
The training arguments are as follows:

1. Batch size = 16
2. Epochs = 5
3. The seed is set globally to ensure reproducibility

These arguments were chosen as a result of parameter tuning using a grid search method. To obtain the probability of each emotion being present, a sigmoid function was applied to the model's raw output (logits) individually. These probabilities are then compared against a predefined threshold (0.3 in this case) to determine whether each emotion is predicted for a given instance.

Table 1 shows the evaluation results on the MasakhanNER dataset, demonstrating the Afro-xmlr-small model's strong performance in Named

Language	XML-R-miniLM	XML-R-base	Afro-xmlr-small
hau	74.5	89.5	91.4
swa	86.3	87.4	88.7

Table 1: NER F-score results on MasakhaNER; Afro-xmlr-small shows strong performance.

Entity Recognition (NER) tasks for African languages. Afro-xmlr-small balances performance and efficiency since it is specifically trained on African language data, hence outperforming generic multilingual transformer models like mBERT in this domain.

4.3 Data Augmentation Techniques

The techniques were deliberately chosen to ensure diversity in approach and complexity. They were also selected to introduce linguistic and structural variation into the dataset while preserving the original emotional content. The selected techniques were Random Insertion, Masking Language Model based contextual augmentation, and back translation.

All of the above augmentation techniques were to increase the size of the datasets and to balance the distribution of emotion labels.

4.3.1 Random Insertion Technique

This augmentation technique applies random insertion guided by pre-trained word embeddings. Specifically, fastText word vectors trained on Common Crawl and Wikipedia for Afrikaans, Swahili, and Hausa, as provided by Grave et al. (2018) were used. These embeddings enable us to compute semantic similarity between words based on their vector representations.

The augmentation process begins by randomly selecting a word from the input sentence. Using cosine similarity, semantically similar words were retrieved from the fastText embedding space. A synonym was then selected using the cosine similarity values as weights. The selected synonym was inserted at a random position in the sentence. This method is repeated 5 times to get 5 augmentations for the instance.

To ensure quality, augmented sentences are compared with the original using fastText embeddings. Cosine similarity was used to filter out candidates that deviated too much from the original meaning. If the most similar augmented sentence exceeds a threshold of 0.75 in cosine similarity and is not identical to the original, it is retained. Otherwise, it is discarded. The corresponding emotion labels remain unchanged. This method increases training data diversity while preserving the semantic and emotional integrity of the original text.

4.3.2 Masking Language Model based Contextual Augmentation Technique

This technique generates new training instances by predicting masked words in a sentence. This helps the model to learn more robust and generalised emotion representations. A instance is augmented by removing random words and then replacing them with predictions from the Masked Language Model (MLM).

Each instance is first tokenised using regular expressions to separate words from punctuation. Then, approximately 15% of the valid tokens were randomly masked and replaced with the <mask> token. The Afro-xmlr-small model was used to predict suitable replacements for these masked tokens. This ensures substitutions are informed by the surrounding context. To maintain data quality and semantic consistency, augmented sentences inherit the original emotion labels.

To maintain semantic consistency between the original and augmented text, LaBSE (Language-agnostic BERT Sentence Embedding)(Feng et al., 2020) is used to measure the cosine similarity between sentence embeddings. LaBSE is a multilingual sentence encoder capable of capturing semantic similarity across over 100 languages (including Afrikaans, Hausa and Swahili), making it particularly effective for low-resource or morphologically rich languages. If the augmented sentence is either identical to the original or fails to meet the defined semantic similarity threshold (0.75), it is discarded. This ensures that the original emotional intent of the sentence is preserved and no duplicates are added to the dataset. LaBSE was selected over FastText because it generates sentence-level embeddings that capture the overall meaning and context of a sentence.

4.3.3 Back Translation Technique

This augmentation technique applies back-translation to generate semantically diverse paraphrases for emotion classification across Afrikaans, Swahili, and Hausa. Specifically,

pre-trained translation models were tailored to each language pair: Helsinki-NLP's Opus-MT models for Afrikaans and Hausa, and Meta's NLLB-200 for Swahili. These models enable the translation of input sentences to English and then back to the original language, producing alternate phrasings while preserving the original meaning.

The augmentation process begins by identifying under-represented emotion labels in the training set. For each such label, the number of additional samples required were determined to approach a balanced label distribution. Samples associated with these labels are then selected and passed through the back translation pipeline. Each sentence is translated to English and subsequently back to the source language using the appropriate translation model. To maintain data quality and semantic consistency, augmented sentences inherit the original emotion labels.

4.4 Augmentation Workflow

To evaluate the impact of data augmentation on emotion classification performance, a two-phase experimental setup was employed:

1. **Baseline Model Training:** The Afro-xlmr-small model was first fine-tuned on the original, unaugmented training data for each language to establish baseline performance.

2. **Augmentation-Specific Training**: The model was then separately fine-tuned on datasets augmented using each technique: Random Insertion, MLM-based contextual augmentation, and Back Translation. For each augmentation method, only one technique was applied at a time, and the model was retrained to isolate its individual impact.

This sequential setup enables a clear comparison between the baseline and each augmentation method, allowing us to quantify how much each technique improves or impacts classification performance. The validation and test sets remained unchanged to ensure a fair and consistent evaluation across all experiments.

The augmentation process began by identifying under-represented emotion labels in the training set. For each such label, the number of additional samples required were determined by the current label distribution, where instances with under-represented labels were augmented multiple times while moderately-represented labels where augmented once and over-represented labels were not augmented. This aimed to both increase the overall size of the dataset while attempting to balance the emotion label distribution in the dataset.

4.5 Evaluation Strategy

To evaluate model performance across different languages and augmentation strategies, multi-metric evaluation was used. Given the multilabel nature of the task, traditional single-label metrics like accuracy are not sufficient. Therefore, the following evaluation metrics were used:

1. **Micro F1 Score**: Aggregates the contributions of all classes to compute the average F1 score. This metric is sensitive to class imbalance and is appropriate for multi-label problems.

2. **Macro F1 Score**: Computes the F1 score independently for each emotion label and then takes the average. This ensures that performance is assessed fairly across all emotion categories, including those with fewer samples.

3. **Hamming Loss**: Measures the fraction of incorrect labels to the total number of labels. A lower value indicates better performance, especially relevant for multilabel tasks.

4. **Subset Accuracy (Exact Match Ratio)**: The proportion of examples where the predicted set of labels exactly matches the true set. Although strict, it provides a sense of how well the model captures the complete emotional content.

All metrics were computed on the validation and test sets after each training phase. This allowed for:

- Comparing baseline model performance to that of models trained with data augmentation.

- Evaluating the generalisation ability of the model across languages with varying resource availability.

Additional evaluation methods, such as confusion matrices for each emotion label across different languages, were employed to gain deeper insights into model performance. ROC curves and associated AUC values were also used to assess the models' ability to distinguish between emotion classes.

5 Results and Discussion

5.1 Baseline Model

The baseline results reveal significant performance variations across the three African languages, highlighting the challenges of emotion classification in low-resource settings. Looking at Table 2. Hausa demonstrated the strongest baseline performance with a Micro F1-score of 0.7690 and Macro F1-score of 0.7624, indicating balanced performance across all emotion categories. Afrikaans achieved competitive results with a Micro F1-score of 0.7609, though its Macro F1-score of 0.6169 suggests some class imbalance issues. Referring to Figure 5, the ROC curves for both Afrikaans and Hausa demonstrate that the base model is able to reliably distinguish between the different labels. The AUC (Area Under Curve) values are provided for each emotion, for Afrikaans and Hausa these AUC values do not fall below 0.8, indicating high model performance in the models' capabilities to distinguish between positive and negative instances across the seven emotion classes.

Swahili exhibited the worst baseline performance with a Micro F1-score of 0.4353 and a particularly low Macro F1-score of 0.0879 as shown in Table 2, indicating severe difficulties in emotion classification for this language despite having the largest dataset (7721 instances). This is further shown by the poor hamming loss value of 0.1647 which indicates frequent mislabelling.

Referring to the confusion matrices for each emotion in Swahili (Figure 1), it is evident that all test instances were labelled as neutral, with no instances assigned to any of the other emotion labels. This combined with the poor hamming loss and macro F1 results in Table 2 can be attributed to the extreme class imbalance in the Swahili dataset as shown in Figure 9. Upon first look at the dataset it is one of the highest resourced African languages in the BRIGHTER dataset but once looking at the distribution it can be seen that around 45% of the dataset is made up of only neutral text examples.

The AUC values for Swahili, ranging from 0.41 to 0.56 (Figure 5), suggest performance close to random guessing. Since there were no True Positives, the True Positive Rate remains at 0, resulting in diagonal lines on the ROC curves. An AUC value of less than 0.5 means that the base model has no predictive power.

The Subset Accuracy (exact match ratio) results further emphasise the varying performance levels across languages. Afrikaans achieved the highest subset accuracy of 0.6901, meaning that approximately 69% of test instances had their complete emotion label sets predicted correctly. However, this relatively strong subset accuracy performance appears inflated due to class imbalance issues within the Although Afrikaans achieved the highest subset accuracy (69%), this metric is likely inflated due to class imbalance, with the model favouring dominant emotions like joy and neutral (Figure 9). The lower Macro F1-score (0.6169) suggests the model's performance was biased toward frequent label combinations, highlighting how subset accuracy can be misleading in imbalanced multi-label settings.

Hausa demonstrated a moderate subset accuracy of 0.6175, indicating that while the model performed well on individual emotion predictions (as evidenced by strong F1-scores), achieving perfect multi-label matches remained challenging for about 38% of cases. This suggests that while the model could identify emotions present in the text, it occasionally included false positive predictions or missed secondary emotions.

Swahili's subset accuracy of 0.4443 appears deceptively reasonable given its poor F1-scores, but this result is, again, misleading. Since the model predicted only neutral labels for all instances, the subset accuracy merely reflects the proportion of test instances that were actually neutral. The remaining 56% of instances that contained non-neutral emotions were completely misclassified, highlighting the model's complete failure to detect emotional content beyond the dominant neutral class. This reinforces the severity of the class imbalance problem in the Swahili dataset.

5.2 Data Augmentation Techniques

5.2.1 Random Insertion Technique

The Random Insertion technique demonstrated the most consistent and substantial improvements across all three African languages, with particularly notable results for the Swahili dataset. Looking at Table 3, the technique showed distinct performance patterns aligned with each language's baseline challenges.

Swahili showed the most dramatic transformation under Random Insertion augmentation. The Micro F1-score increased substantially from 0.4353 to 0.6883 (Δ +0.25), representing a 58% relative improvement. More significantly, the

Macro F1-score improved from 0.0879 to 0.6360 (Δ +0.55), a 625% relative improvement that indicates the technique successfully addressed the severe class imbalance issues. The Hamming Loss decreased from 0.1647 to 0.0959 (Δ -0.069), demonstrating a substantial reduction in mislabelling frequency. The Subset Accuracy improvement from 0.4443 to 0.6477 (Δ +0.20) represents a 46% relative improvement, indicating that the model now correctly predicts complete emotion label sets for nearly 65% of instances. This is particularly significant given that the baseline's subset accuracy was artificially inflated by neutral class dominance. The augmented model now achieves genuine multilabel accuracy across all emotion categories. Referring to Figure 2, the confusion matrices reveal that the model now successfully predicts non-neutral emotions, with clear improvements in emotion detection across all categories. The ROC curves in Figure 6 show dramatically improved AUC values, with most emotions now achieving AUC scores above 0.7, compared to the baseline's random guessing performance of 0.41-0.56.

Afrikaans showed moderate but consistent improvements across all metrics. The Micro F1-score increased from 0.7609 to 0.7937 (Δ +0.03), while the Macro F1-score improved from 0.6169 to 0.6655 (Δ +0.05). The Hamming Loss decreased from 0.0816 to 0.0684 (Δ -0.013), indicating more precise label assignments. The Subset Accuracy improved from 0.6901 to 0.7380 (Δ +0.05), achieving a 74% exact match rate for complete emotion label sets. This 7% relative improvement demonstrates that Random Insertion not only enhanced individual emotion predictions but also improved the model's ability to capture the full complexity of multilabel emotional expressions, which is particularly noteworthy given that subset accuracy is the most stringent evaluation metric. Given Afrikaans' already strong baseline performance, these improvements represent meaningful enhancements to an already well-performing model. The ROC analysis shows maintained high AUC values above 0.8 for most emotions, with slight improvements in distinguishing power across emotion categories.

Hausa demonstrated the most modest improvements, which aligns with its strong baseline performance. The Micro F1-score increased marginally from 0.7690 to 0.7779 (Δ +0.01), and the Macro F1-score improved slightly from 0.7624 to 0.7689 (Δ +0.01). The Hamming Loss showed minimal improvement from 0.0865 to 0.0833 (Δ -0.003). The Subset Accuracy increased from 0.6175 to 0.6494 (Δ +0.03), representing a 5% relative improvement that brings the exact match rate to nearly 65%. This improvement indicates that Random Insertion helped refine the model's precision in complete multilabel predictions, reducing instances where secondary emotions were either missed or incorrectly added.

5.2.2 Masking Language Model based Contextual Augmentation Technique

The Masking technique generally had a positive impact on the emotion classification for the three languages. Its aim was to generate more robust and generalised emotion representations by augmenting data with contextually informed word replacements. The masking technique appears most effective for languages with poor baseline performance. Swahili had the worst baseline results but showed the most significant improvement, while Hausa (which had decent baseline performance) showed minimal to no gains.

Swahili had the most noticeable improvements. The Macro F1 score improved from 0.0879 to 0.4878 (Table 4), representing an 459% improvement that reflects the model's ability to handle severely imbalanced classes. While still substantial, the Micro F1 and accuracy improvements were more moderate at 43% and 20% respectively, with Micro F1 rising from 0.4353 to 0.6221 and Subset Accuracy improving from 0.4443 to 0.5330 as shown in Table 4. The disparity between the macro and micro metric improvements highlights how the masking technique particularly excelled at addressing the severe class imbalance that had rendered the baseline model essentially unusable for multiclass emotion detection.

The confusion matrix (Figure 3) further illustrates that the augmented model began correctly identifying emotional categories that the baseline model completely missed. This suggests that the augmented data introduced critical diversity and contextual richness that the original training set lacked, enabling the model to learn more meaningful emotional representations. Supporting this, the AUC values (Figure 7) improved significantly, from a baseline range of 0.41 to 0.56 (close to random guessing) to a much more discriminative range of 0.79 to 0.93 after augmentation. These results confirm that the masking technique not only improved general performance but also fundamentally en-

hanced the model's capacity to separate emotional classes in a low-resource setting.

For Hausa, the addition of supplementary data through the masking technique did not lead to noticeable improvements over the baseline model. The performance slightly declined as the Micro F1 score dropped from 0.7624 to 0.7434, and the Macro F1 score decreased from 0.7690 to 0.7496 (Table 4). These marginal changes suggest subtle shifts in the model's behaviour rather than significant enhancements or degradations.

A closer look at the confusion matrices (Figure 3) reveals nuanced changes in class-level performance. For example, the joy label had an improvement, with true positives increasing from 61 to 64 and false negatives decreasing from 17 to 14, indicating better recall. Similarly, the anger label improved slightly, with true positives rising from 76 to 77. However, other labels experienced slight regressions, for instance, the disgust label's true positives dropped from 53 to 52, and the neutral label fell from 46 to 40. These shifts suggest the masking augmentation led to a redistribution of predictive focus across different emotion classes, rather than boosting performance uniformly.

As can be seen with the ROC curves (Figure 7), the area under the curve remained relatively stable. The baseline model had AUC values ranging from 0.89 to 0.97, while the augmented model ranged from 0.88 to 0.96. This stability indicates that the model retained its overall ability to discriminate between classes, even if class-specific performance was rebalanced.

The masking technique led to consistent and meaningful improvements across all key performance metrics for Afrikaans. Subset accuracy increased from 0.6901 to 0.7775 (Table 4), a 12.66% gain, reflecting better overall classification capability. The Micro F1 score rose from 0.7609 to 0.8273, indicating improved performance when accounting for the frequency of each class. More notably, the Macro F1 score increased from 0.6169 to 0.6821, a 10.56% improvement, demonstrating the model's enhanced ability to handle class imbalance, which is particularly important in emotion detection tasks where some emotions are inherently less frequent. In parallel, hamming loss dropped from 0.0816 to 0.0571, showing that the model made fewer prediction errors overall. This reduction supports the observed accuracy improvements and improved feature representations. The baseline model's AUC scores ranged from 0.89 to 0.99, the augmented model maintained strong performance with AUC values between 0.89 and 0.97 (Figure 7), indicating stable or slightly refined class separation ability after the augmentation technique was applied.

5.2.3 Back Translation Technique

Back translation resulted in notable performance gains for Afrikaans and Swahili, while showing limited effectiveness for Hausa. These outcomes underscore the importance of translation quality and linguistic characteristics in the success of data augmentation strategies.

For Afrikaans, back translation led to consistent improvements across all metrics. Accuracy rose from 69.0% to 74.1%, and Hamming loss dropped from 0.0817 to 0.0688 (Table 5), indicating fewer overall prediction errors. The micro F1 score increased from 0.761 to 0.791, while the macro F1 score improved from 0.617 to 0.667, suggesting that the model became more balanced in its performance across both frequent and rare emotion classes. These improvements may reflect better class representation in the training data due to augmentation. Furthermore, the confusion matrix,in Figure 4, reveals enhanced emotion detection capabilities post augmentation, particularly through increased true positives and reduced false positives across multiple labels. These trends reinforce the conclusion that the back translation process effectively enriched the dataset without introducing significant noise.

In contrast, Hausa had a decline in performance after applying back translation. Accuracy dropped slightly from 61.8% to 60.0%, and Hamming loss increased from 0.0865 to 0.0913 (Table 5). Both micro F1 and macro F1 scores decreased marginally, from 0.769 to 0.757 and 0.762 to 0.751, respectively. These results suggest that the augmented Hausa data may have introduced semantic inconsistencies, likely due to limitations in the translation pipeline. The performance degradation highlights the sensitivity of low-resource languages to translation quality, where noisy back translations can outweigh the benefits of additional data.

Swahili showed the most improvement with back translation. Accuracy increased from 44.4% to 58.2%, and Hamming loss significantly decreased from 0.1647 to 0.1090, indicating a substantial reduction in prediction errors (Table 5). The micro F1 score increased from 0.435 to 0.653, and macro F1 surged from 0.088 to 0.564, a significant change. The low baseline macro F1 score indicates that cer-

tain labels were previously being entirely missed by the model. With the augmented data, the model learned more robust representations, resulting in better generalisation across all emotion classes.

In summary, these findings demonstrate that back translation can substantially enhance model performance, particularly when high-quality translation systems are available. While Afrikaans and Swahili benefited from semantically rich augmented data, Hausa's results underscore the risks of introducing low-quality translations in low-resource settings. This highlights the critical role of translation accuracy in determining the effectiveness of back translation as a data augmentation strategy.

5.3 Overall discussion

5.3.1 Dataset Imbalance and Performance Correlation

The relationship between dataset characteristics and model performance reveals critical insights about data quality versus quantity in African language emotion classification. Dataset imbalance significantly influenced baseline performance patterns across the three languages:

- Swahili Dataset Analysis: Despite having the largest dataset (7721 examples), Swahili exhibited severe class imbalance with most emotion categories being under-represented. The extremely low Macro F1-score (0.0879) directly reflects this imbalance, where the model failed to learn meaningful representations for minority emotion classes. The high Hamming Loss (0.1647) further indicates that the model frequently confused under-represented emotions with dominant classes. This suggests that raw dataset size is less critical than balanced representation across emotion categories.

- Hausa Dataset Characteristics: Hausa's superior baseline performance (Micro F1: 0.7690, Macro F1: 0.7624) despite having fewer examples (5017) can be attributed to better class distribution balance. The minimal gap between Micro and Macro F1-scores indicates relatively even representation across emotion categories.

- Afrikaans Distribution Patterns: The moderate performance gap between Micro F1 (0.7609) and Macro F1 (0.6169) suggests intermediate-level class imbalance.

5.3.2 Augmentation Effectiveness and Imbalance Mitigation

Data augmentation was initially used to increase the size of all the datasets. This was believed to be the biggest issue when addressing poor performance in low-resource languages. However, evenly increasing the dataset across emotion labels proved ineffective. This revealed the main issue that needed to be addressed: class imbalance. The methodology was then expanded to address both these issues, aiming to increase the overall size of the datasets but also to level out the imbalance between emotion labels. The data augmentation techniques demonstrated varying effectiveness in addressing class imbalance issues:

- Random Insertion: The major improvements in Swahili (625% relative Macro F1 improvement) indicate that Random Insertion effectively generates synthetic instances for under-represented emotion classes. The semantic coherence provided by fastText embeddings ensures that augmented samples maintain linguistic authenticity.

- Masking Technique Sensitivity: The effectiveness of the MLM-based augmentation technique is highly dependent on the underlying language model used for generating the masked tokens. Variations in model architecture, training data, and language coverage can significantly influence the quality and relevance of the augmented samples.

- Back Translation Limitations: The moderate effectiveness across all languages suggests that this technique may not adequately address class-specific imbalance issues, potentially introducing translation issues that further complicate emotion label learning.

5.3.3 Cultural Context Preservation

The results highlight the importance of maintaining cultural authenticity in emotion expression:

- Semantic Preservation Priority: Random Insertion's superior performance across languages suggests that preserving semantic relationships is more critical than syntactic transformation for culturally specific emotion expressions.

- Translation Risks: Back translation's variable effectiveness indicates potential risks of introducing culturally inappropriate emotion mappings through intermediate English translation, particularly for culture-specific emotional concepts.

6 Conclusion

This study investigated the effectiveness of data augmentation techniques for multilabel emotion classification across three African languages, Afrikaans, Swahili, and Hausa, addressing a critical gap in natural language processing research for low-resource languages. Through systematic evaluation of Random Insertion, MLM-based contextual augmentation technique, and Back Translation techniques using the BRIGHTER dataset and Afro-xlmr-small model, we have demonstrated that strategically applied data augmentation can significantly enhance emotion classification performance in resource-constrained settings.

6.1 Key Findings

Our research yields several important findings that advance understanding of emotion classification in African languages: Data Quality Supersedes Quantity: The most significant insight from this study is that balanced class distribution is more critical than raw dataset size for effective emotion classification. Despite Swahili having the largest dataset (7721 examples), its severe class imbalance rendered the baseline model essentially unusable, with a Macro F1-score of only 0.0879. In contrast, Hausa's smaller but more balanced dataset (5017 examples) achieved superior baseline performance (Macro F1: 0.7624), demonstrating that thoughtful data curation is important. Random Insertion as the Superior Technique: Among the three augmentation methods, Random Insertion consistently delivered the most substantial and reliable improvements across all languages. Its semantic-aware approach using fastText embeddings proved particularly effective for addressing class imbalance while preserving emotional context. The technique's performance on Swahili (625% relative improvement in Macro F1-score) demonstrates its capacity to transform severely imbalanced datasets into viable training resources.

6.2 Implications for African Language NLP

This research has significant implications for the broader development of NLP systems for African languages:

- Addressing Digital Inequality: By demonstrating effective methods for improving emotion classification in low-resource African languages, this work contributes to reducing the digital divide that has historically excluded African language speakers from advanced language technologies. The substantial improvements achieved, particularly for severely imbalanced datasets, provide a pathway for developing more inclusive NLP systems.

- Methodological Framework: Our systematic evaluation approach provides a replicable framework for applying data augmentation to other African languages and NLP tasks.

6.3 Limitations and Future Directions

While this study provides valuable insights, several limitations warrant acknowledgement and point toward future research directions:

1. Limited Language Coverage: Our focus on three African languages, while representing different resource levels and linguistic families, constitutes a small fraction of Africa's linguistic diversity. Future research should expand to include languages from additional families, particularly those with unique morphological or tonal characteristics that may respond differently to augmentation techniques.

2. Cultural Context Evaluation: While we preserved semantic content through similarity thresholds, our evaluation methods did not explicitly assess cultural appropriateness of augmented content. Future work should incorporate culturally-informed evaluation metrics and involve native speakers in validation processes.

3. Advanced Augmentation Techniques: This study focused on established augmentation methods. Future research could explore more sophisticated approaches or hybrid techniques that combine multiple augmentation strategies.

References

Ibrahim Said Ahmad, Shiran Dudy, Tadesse Destaw Belay, Idris Abdulmumin, Seid Muhie Yimam, Shamsuddeen Hassan Muhammad, and Kenneth Church. Exploring cultural nuances in emotion perception across 15 african languages.

Aashish Arora and Elsbeth Turcan. 2024. Evaluating the effectiveness of data augmentation for emotion classification in low-resource settings. *arXiv preprint arXiv:2406.05190*.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Catherine Gitau and Vukosi Marivate. 2023. Textual augmentation techniques applied to low resource machine translation: Case of swahili. *arXiv preprint arXiv:2306.07414*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Anna Koufakou, Diego Grisales, Oscar Fox, and 1 others. 2023. Data augmentation for emotion detection in small imbalanced text data. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1508–1513. IEEE.

Ronny Mabokela and Tim Schlippe. 2022. A sentiment corpus for south african under-resourced languages in a multilingual context. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 70–77.

Refuoe Mokhosi, Casper-Shikali Shivachi, and Matello Sethobane. 2024. A sesotho news headlines dataset for sentiment analysis. *Data in Brief*, 54:110371.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M Mohammad, Sebastian Ruder, and 1 others. 2023a. Afrisenti: A twitter sentiment analysis benchmark for african languages. *arXiv preprint arXiv:2302.08956*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. Semeval-2023 task 12: sentiment analysis for african languages (afrisenti-semeval). *arXiv preprint arXiv:2304.06845*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025b. Semeval-2025 task 11: Bridging the gap in text-based emotion detection. *Preprint*, arXiv:2503.07269.

Tolúlp Ògúnremí, Wilhelmina Onyothi Nekoto, and Saron Samuel. 2023. Decolonizing nlp for "low-resource languages": Applying abebe birhane's relational ethics. *GRACE: Global Review of AI Community Ethics*, 1(1).

Flor Miriam Plaza-del Arco, Alba Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. *arXiv preprint arXiv:2403.01222*.

Koena Ronny Mabokela, Mpho Primus, and Turgay Celik. 2025. Advancing sentiment analysis for low-resourced african languages using pre-trained language models. *PloS one*, 20(6):e0325102.

Gaurish Thakkar, Nives Mikelić Preradović, and Marko Tadić. 2024. Examining sentiment analysis for low-resource languages with data augmentation techniques. *Eng*, 5(4):2920–2942.

Appendix 1: Results

Language	Micro F1-Score	Macro F1-Score	Hamming Loss	Subset Accuracy
Afrikaans	0.7609	0.6169	0.0816	0.6901
Hausa	0.7690	0.7624	0.0865	0.6175
Swahili	0.4353	0.0879	0.1647	0.4443

Table 2: Baseline Results

Language	Micro F1 Score (Δ from Baseline)	Macro F1 Score (Δ from Baseline)	Hamming Loss (Δ from Baseline)	Subset Accuracy (Δ from Baseline)
Afrikaans	0.7937 (+ 0.03)	0.6655 (+ 0.05)	0.0684 (– 0.013)	0.7380 (+ 0.05)
Hausa	0.7779 (+ 0.01)	0.7689 (+ 0.01)	0.0833 (– 0.003)	0.6494 (+ 0.03)
Swahili	0.6883 (+ 0.25)	0.6360 (+ 0.55)	0.0959 (- 0.069)	0.6477 (+ 0.20)

Table 3: Random Insertion Results

Language	Micro F1 Score (Δ from Baseline)	Macro F1 Score (Δ from Baseline)	Hamming Loss (Δ from Baseline)	Subset Accuracy (Δ from Baseline)
Afrikaans	0.8273 (+ 0.07)	0.6821 (+ 0.07)	0.0571 (– 0.025)	0.7775 (+ 0.09)
Hausa	0.7496 (– 0.01)	0.7434 (– 0.02)	0.0939 (+ 0.007)	0.6076 (– 0.01)
Swahili	0.6221 (+ 0.19)	0.4878 (+ 0.40)	0.1201 (– 0.045)	0.5330 (+ 0.09)

Table 4: Masking Results

Language	Micro F1 Score (Δ from Baseline)	Macro F1 Score (Δ from Baseline)	Hamming Loss (Δ from Baseline)	Subset Accuracy (Δ from Baseline)
Afrikaans	0.7912 (+ 0.03)	0.6667 (+ 0.05)	0.0688 (- 0.013)	0.7408 (+ 0.05)
Hausa	0.7566 (- 0.01)	0.7513 (- 0.01)	0.0913 (+ 0.005)	0.5996 (- 0.02)
Swahili	0.6726 (+ 0.23)	0.6006 (+ 0.51)	0.1027 (- 0.062)	0.6049 (+ 0.16)

Table 5: Back Translation Results

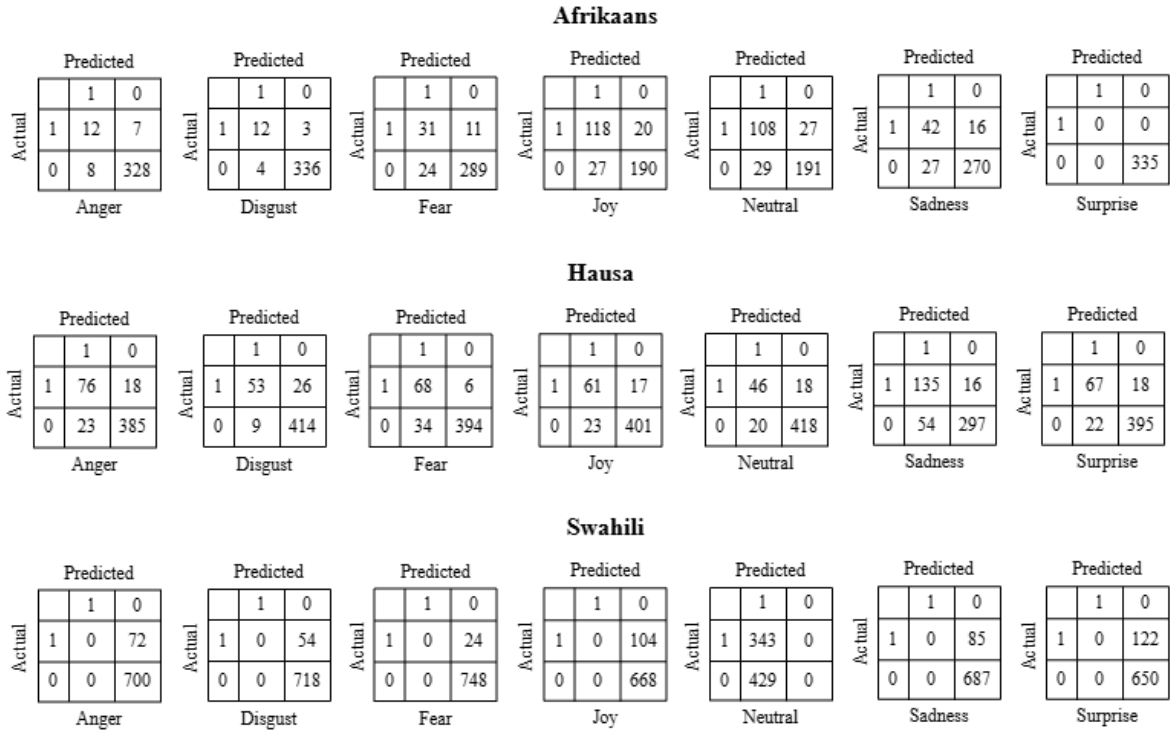

Figure 1: Confusion Matrices per-emotion label for Baseline Model

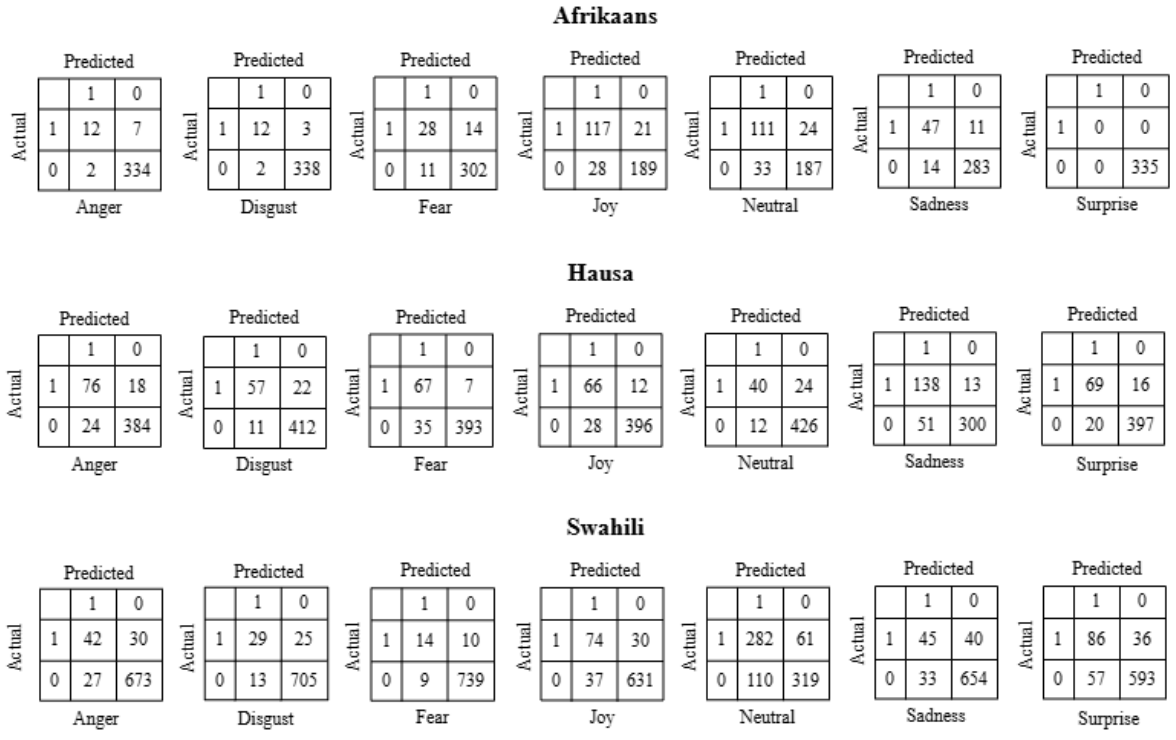

Figure 2: Confusion Matrices per-emotion label for the results of the Random Insertion Techniques

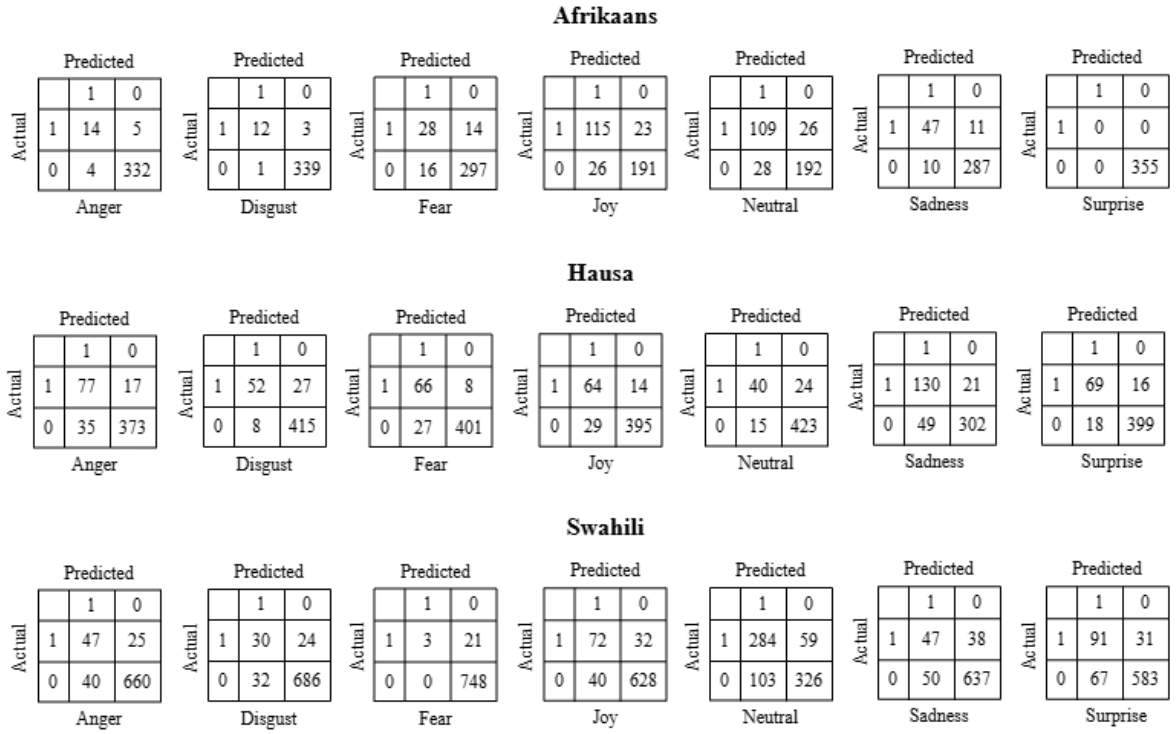

Figure 3: Confusion Matrices per-emotion label for the results of the Masking Techniques

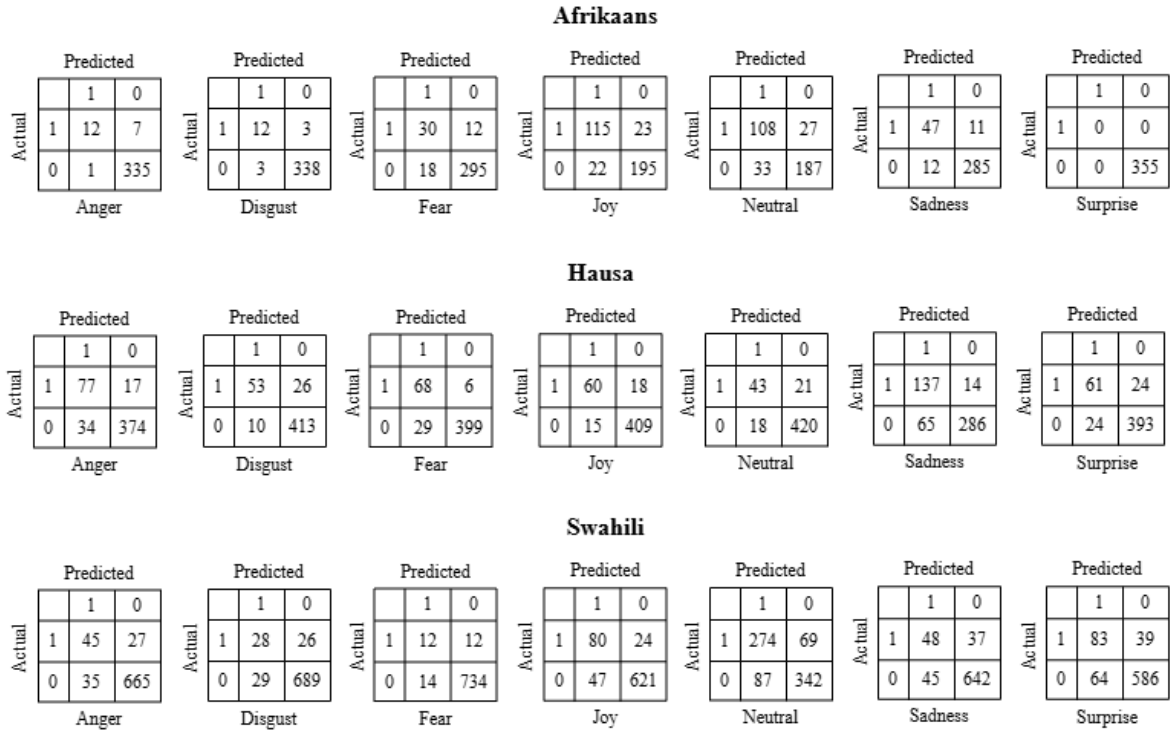

Figure 4: Confusion Matrices per-emotion label for the results of the Back Translation Techniques

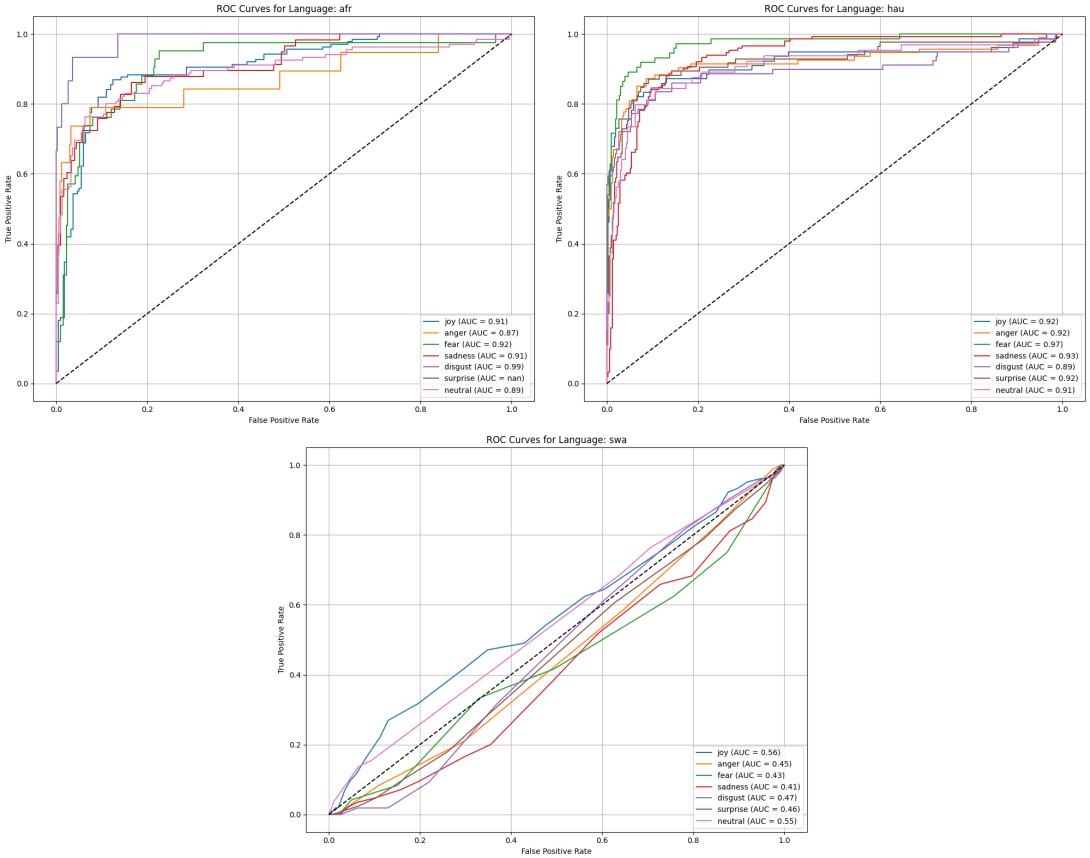

Figure 5: ROC for Baseline Model

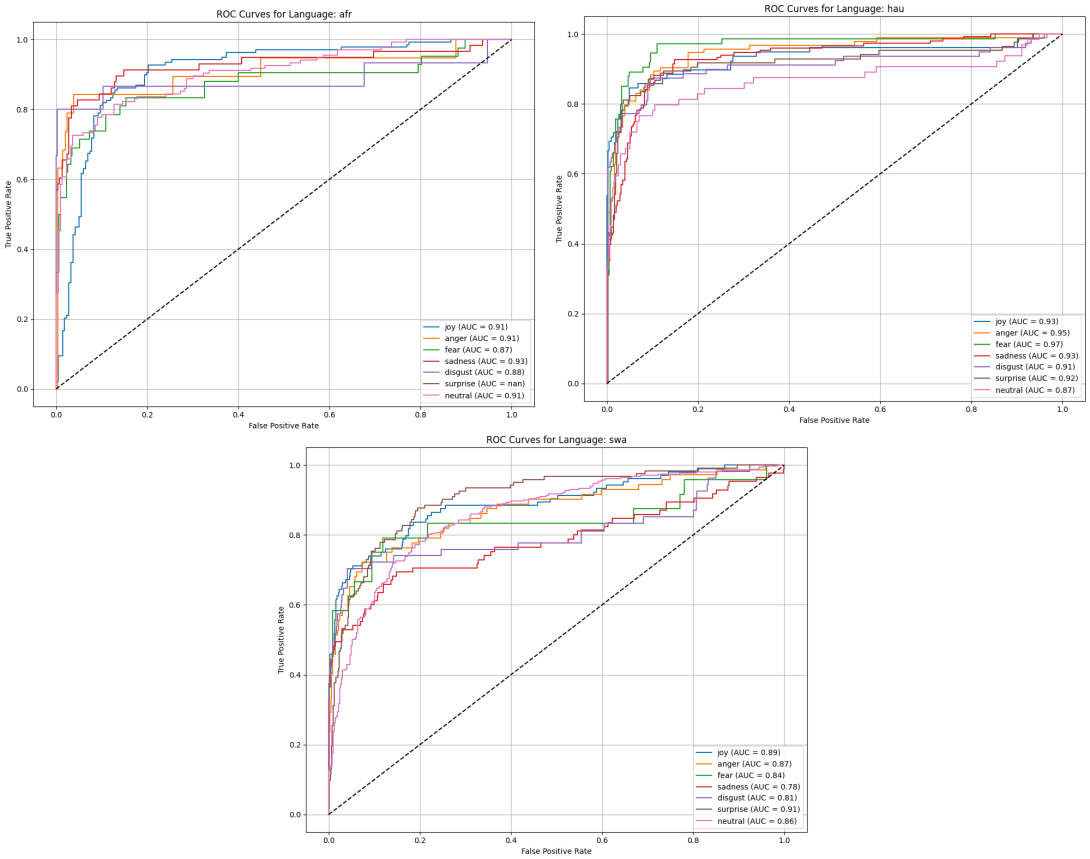

Figure 6: ROC for the results of the Random Insertion Technique

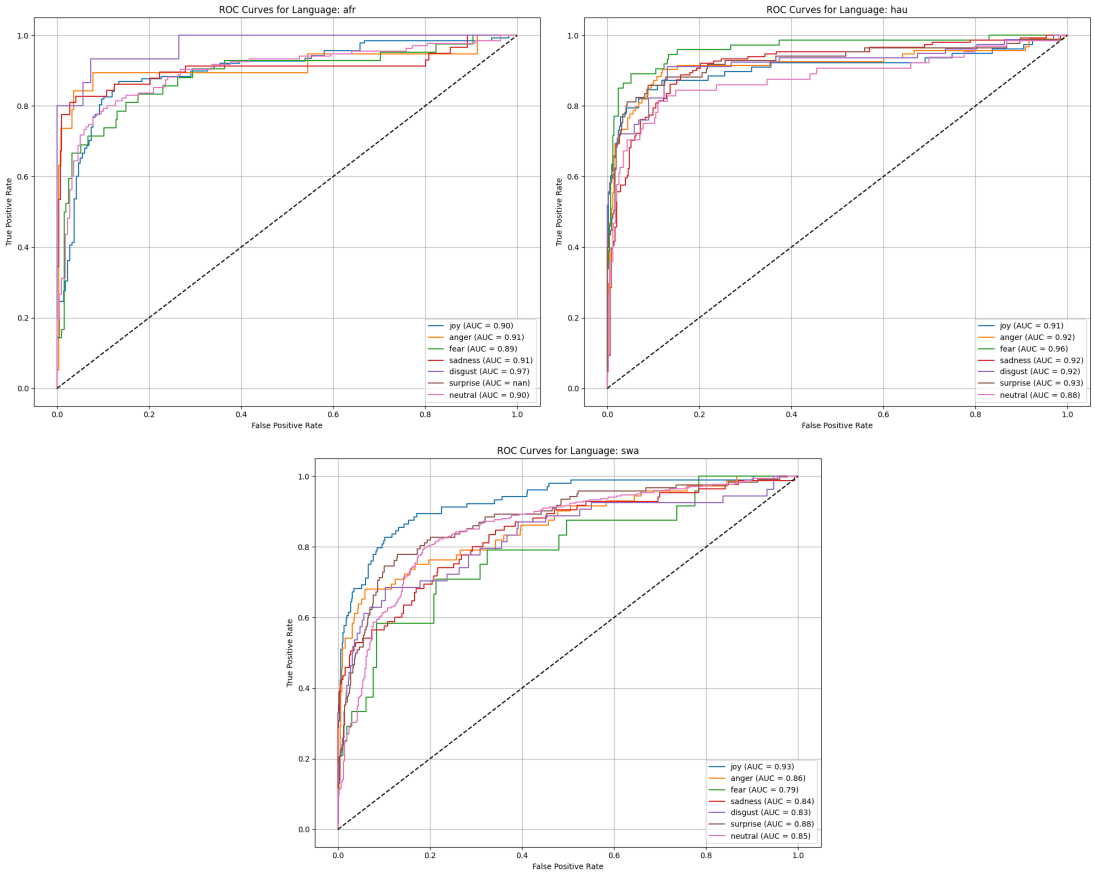

Figure 7: ROC for the results of the Masking Technique

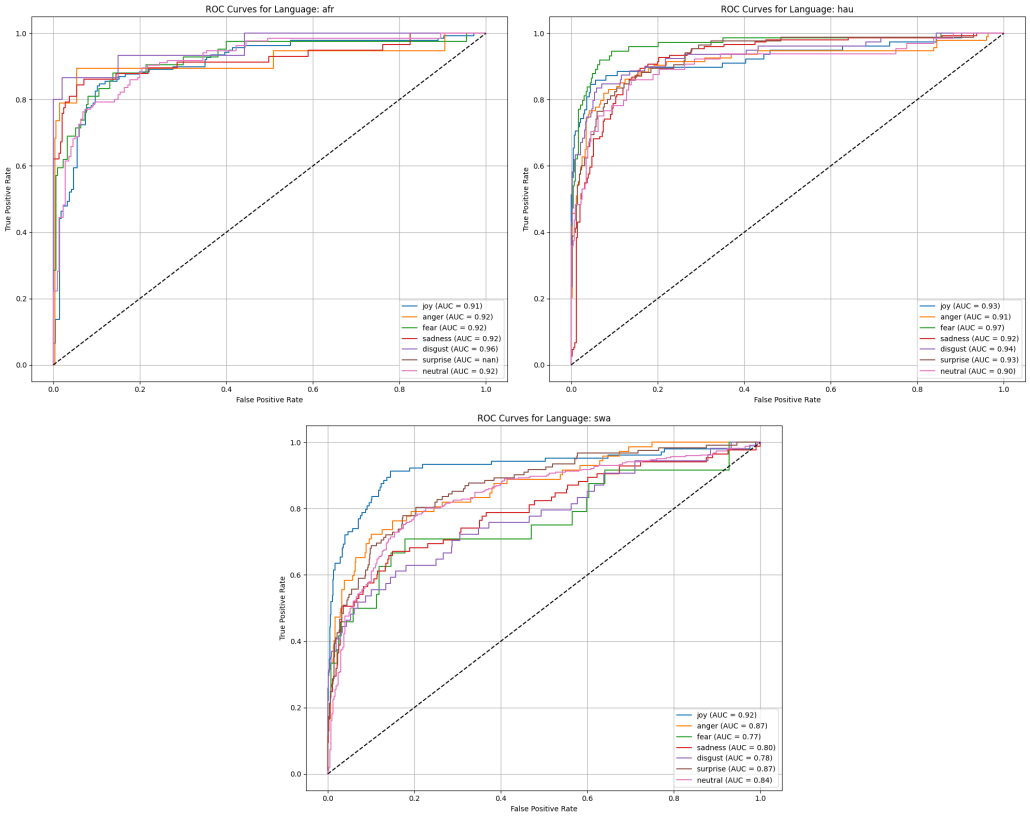

Figure 8: ROC for the results of the Back Translation Technique

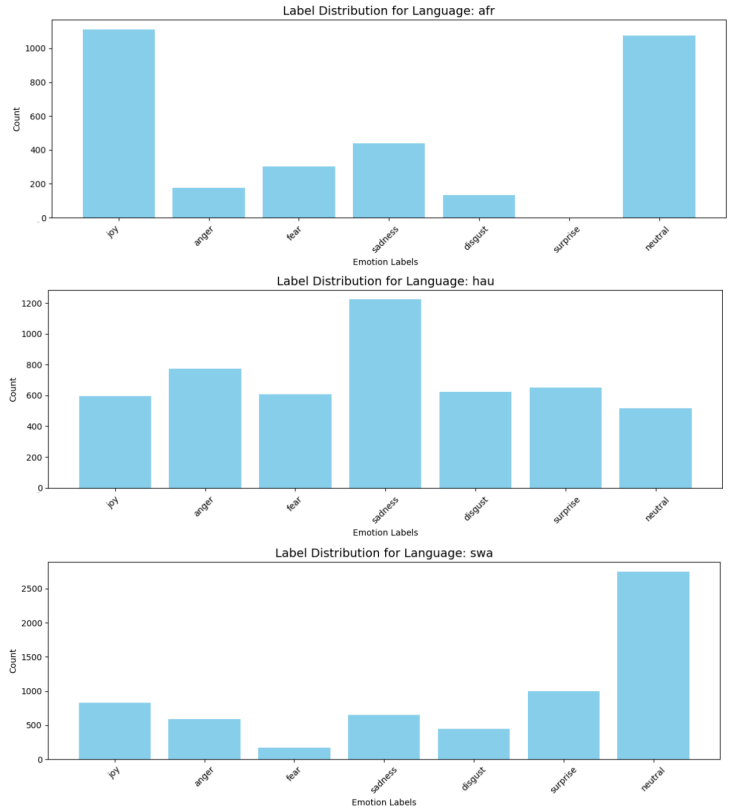

Figure 9: Original Dataset Emotion Label Distribution

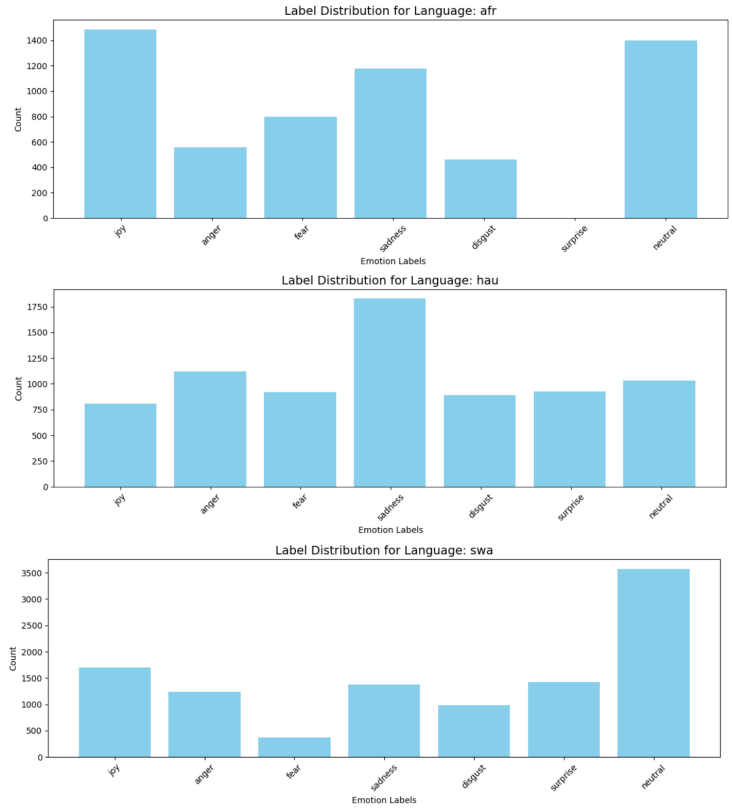

Figure 10: Approximate Dataset Emotion Label Distribution After Augmentation