AFRICAN LANGUAGES FOR EMOTION ANALYSIS

ENHANCING EMOTION CLASSIFICATION THROUGH DATA AUGMENTATION STRATEGIES FOR LOW-RESOURCE AFRICAN LANGUAGES

Group 8

BACKGROUND & MOTIVATION

- **Underrepresentation of African Languages:** Emotion classification has advanced for high-resource languages like English, but African languages remain underrepresented.

- **Potential of Data Augmentation:** Data augmentation has proven effective in improving performance in low-resource NLP tasks, but its application for multilabel emotion classification in African languages is largely unexplored.

RESEARCH QUESTIONS

How effective are augmentation techniques in improving multilabel emotion classification for low-resource African languages?

- How does the performance of models trained with the augmentation techniques compare to the base model on the same dataset?
- How do the proposed augmentation techniques perform across different African languages with different resource levels?

DATASET

BRIGHTER Emotion Categories Dataset

Multilingual corpus designed for multilabel emotion classification

Selected Languages

- Swahili - 7721 instances
- Hausa - 5017 instances
- Afrikaans - 3548 instances

Data Preprocessing Steps

- Neutral Label Assignment
- Stratified Splitting
- Multi-label Encoding

AUGMENTATION TECHNIQUES

Random Insertion

- Uses fastText embeddings to insert semantically similar words at random positions. Filters based on cosine similarity (threshold: 0.75)

MLM Augmentation

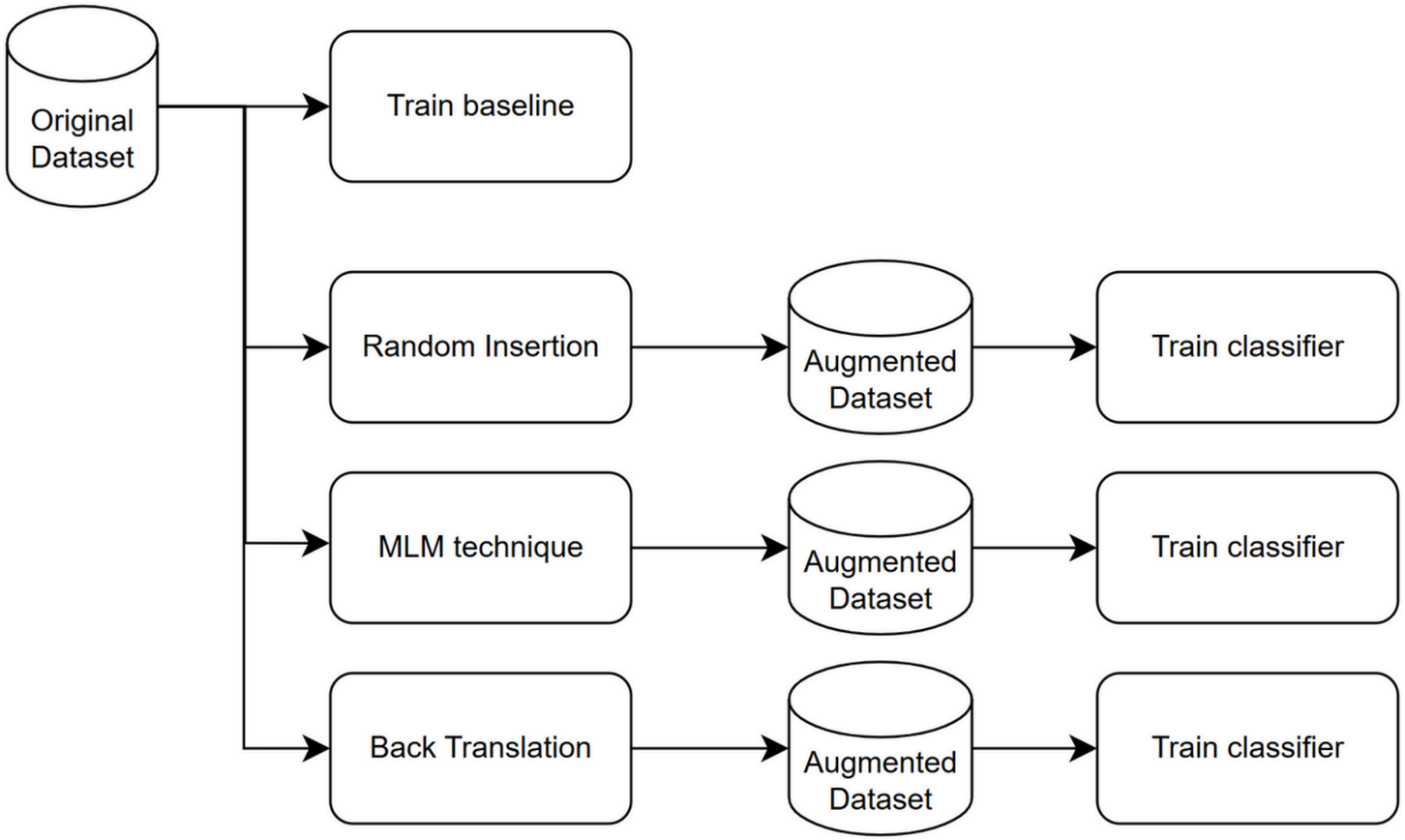
- Masks ~15% of tokens and uses Afro-xlmr-small to predict replacements. Uses LaBSE for semantic consistency validation

Back Translation

- Translates to English and back using Helsinki-NLP Opus-MT (Afrikaans/Hausa) and Meta NLLB-200 (Swahili)

Augmentation Techniques used to **expand** and **balance** the datasets

METHODOLOGY

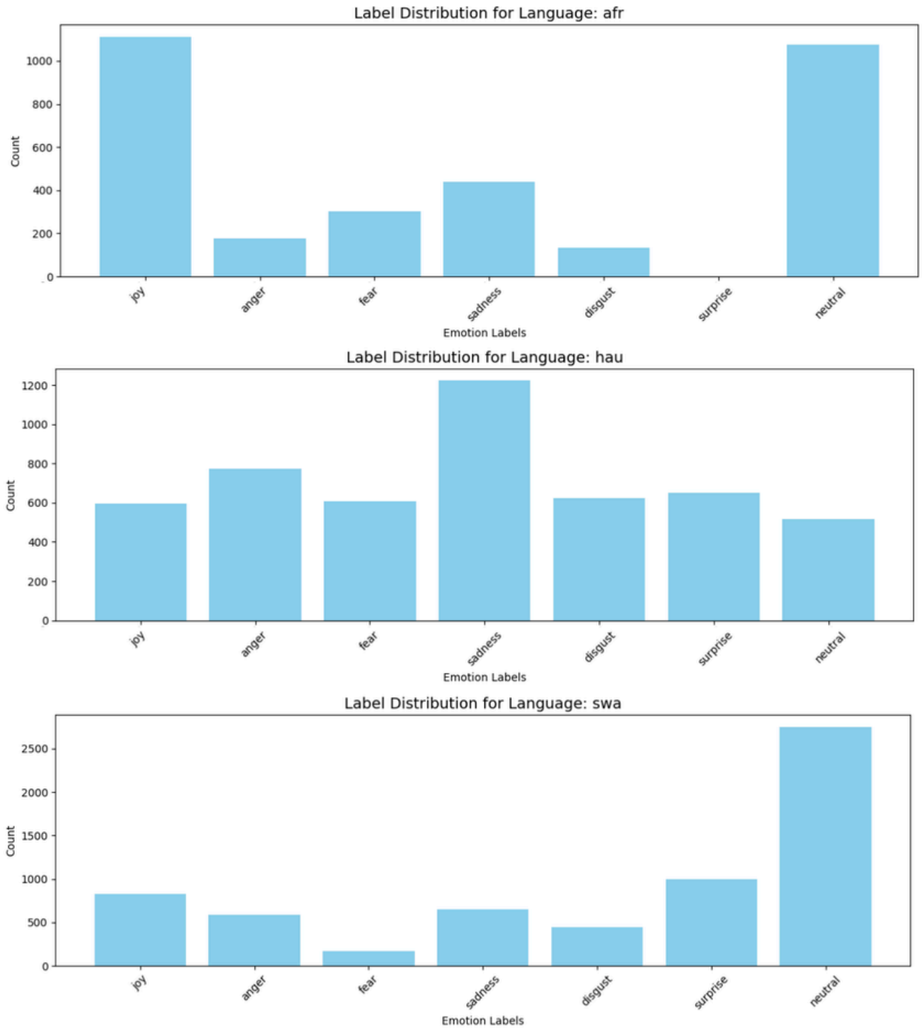


Baseline:
AfroXLMR-small

Evaluation metrics:

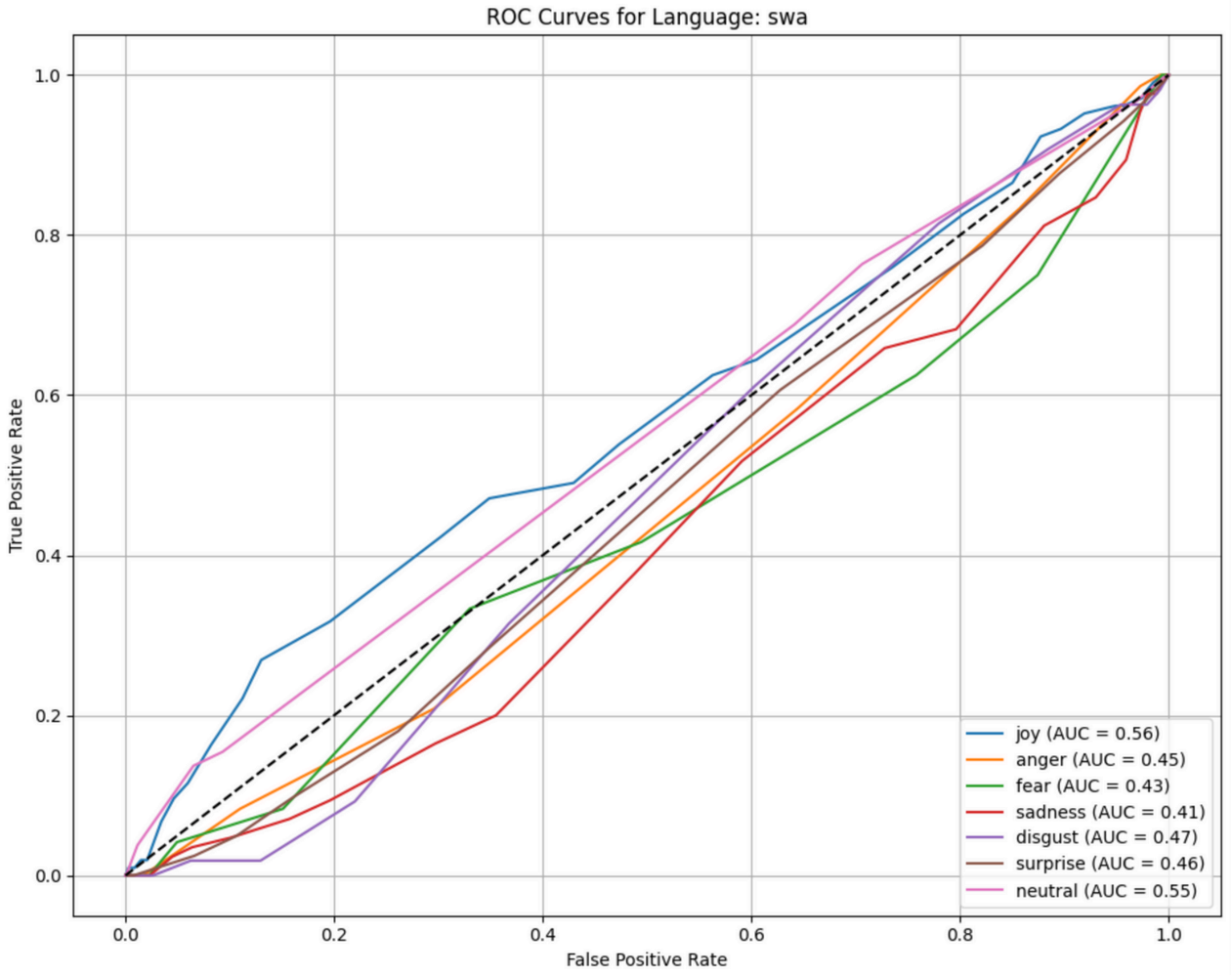
- Micro F1
- Macro F1
- Hamming Loss
- Subset Accuracy
- Confusion matrices
- ROC curves

BASELINE RESULTS



Language	Baseline (Macro F1)
Afrikaans	0.6169
Hausa	0.7624
Swahili	0.0879

BASELINE RESULTS

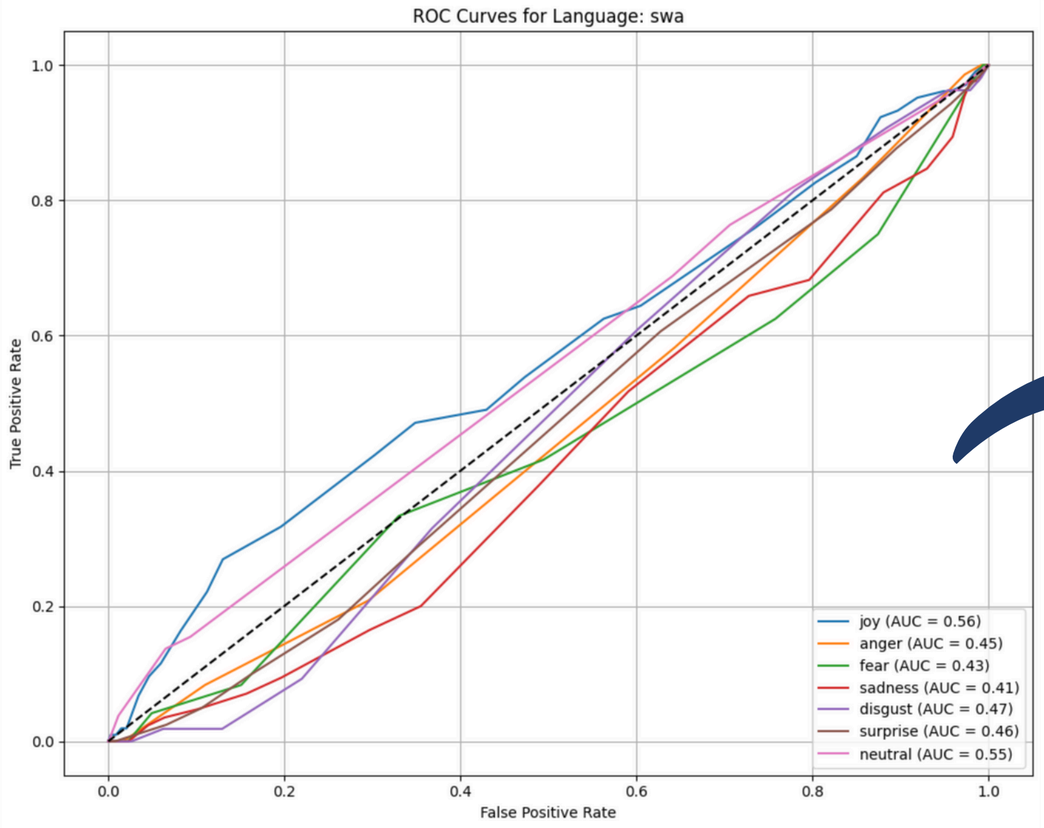


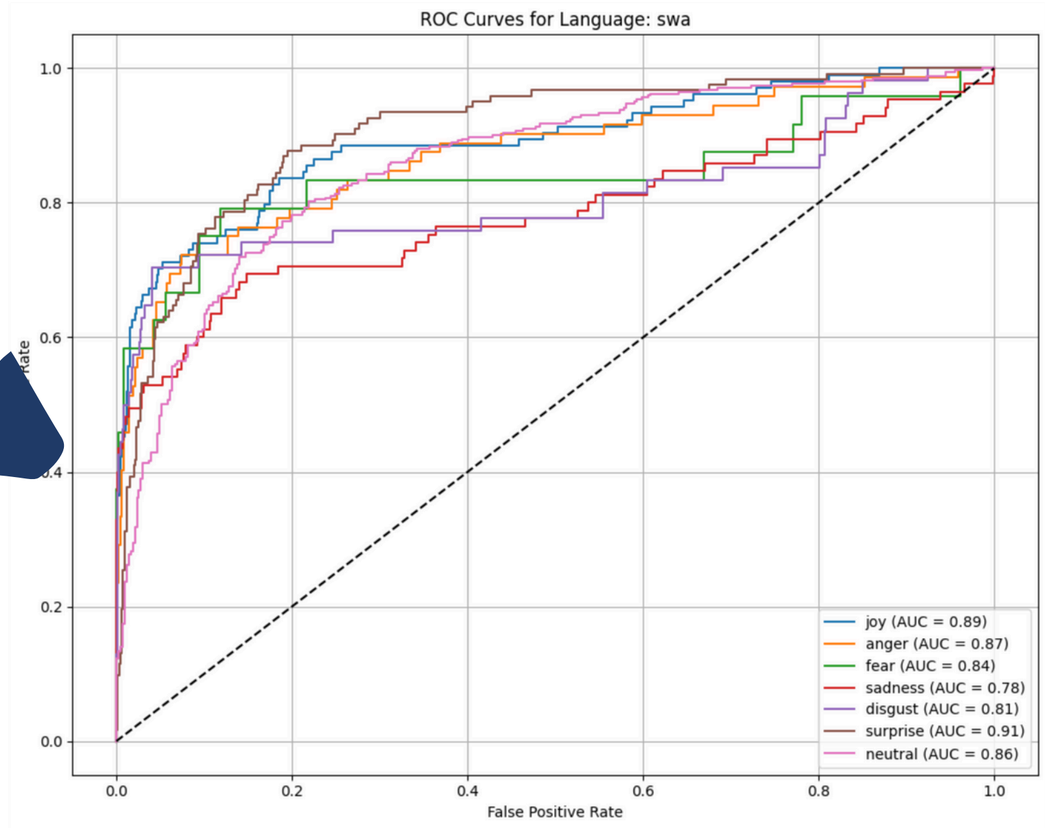
- The baseline model **exclusively predicted** the label **'neutral'**, which constitutes **45%** of the dataset.
- AUC values suggest performance close to **random guessing**
- The model has **no predictive power**

RANDOM INSERTION

Language	Macro F1 Improvement	Relative Improvement	Subset Accuracy Gain
Swahili	0.088 → 0.636	622%	0.2
Afrikaans	0.617 → 0.666	8%	0.05
Hausa	0.762 → 0.769	1%	0.03

RANDOM INSERTION





MLM AND BACK TRANSLATION

MLM

Language	Macro F1 Change
Swahili	+0.40 (+459%)
Afrikaans	+0.07 (+11%)
Hausa	-0.02 (-3%)

Back Translation

Language	Macro F1 Change
Swahili	+0.51 (+580%)
Afrikaans	+0.05 (+8%)
Hausa	-0.01 (-1%)

CONCLUSION & INSIGHTS

Key findings

- Data Quality > Quantity
- Cultural and Semantic Preservation
- Random Insertion Superior Technique

Implications for African NLP

- Digital inequality
- Methodological Framework

Limitations and Future Directions

- Limited Language Coverage
- Cultural Context Evaluation
- Advanced Augmentation Techniques
- Generalisability Issue

THANK YOU!

ASANTE

NA GODE

DANKIE