Howard University Computer Science at SemEval-2023 Task 12: A 2-Step System Design for Multilingual Sentiment Classification with Language Identification

Saurav Aryal and **Howard Prioleau**
EECS, Howard University,
Washington, DC 20059, USA
https://howard.edu/
saurav.aryal@howard.edu and howard.prioleau@bison.howard.edu

Abstract

The recent release of the AfriSenti-SemEval shared Task 12 has made available 14 new datasets annotated for sentiment analysis on African Languages. We proposed and evaluated two approaches to this task, Delta TF-IDF, and a proposed Language-Specific Model Fusion System using Language Identification. Both produced comparable or better classification performance than the current state-of-art models on this task: AfriBERTa, AfroXLMR, and AfroLM.

1 Introduction

AfriSenti-SemEval Shared Task 12 (Muhammad et al., 2023b,a) provides an landmark dataset African language sentiment analysis. The task provided 13 datasets comprising 12 different languages, each being a dataset and a dataset composed of all the languages. The 12 African Langauges covered are Hausa(HA), Yoruba(YO), Igbo(IG), Nigerian Pigdin(PCM), Amharic(AM), Algerian Arabic(DZ), Moroccan Arabic/Darija(MA), Swahili(SW), Kinyarwanda(KR), Twi(TWI), Mozambican Portuguese(PT), and Xitsonga(Mozambique Dialect) (TS). These dataset languages cover a wide range of different regions of Africa. Each tweet in each dataset is manually annotated between 3 positive, negative, and neutral labels. Some datasets contain code-mixed data (data that contains two or more languages) and transliteration (converting text from one script to another that involves swapping letters). While most of these languages have a limited amount of corpus, to our knowledge, some languages, such as Xitsonga, have labeled sentiment analysis datasets created for the first time (Aryal et al., 2023).

This paper aims to contribute to the NLP and sentiment analysis literature for African languages by going in-depth into our modeling practices on the AfriSenti-SemEval (Muhammad et al., 2023b,a) Shared Task 12 and showcasing our results. We propose a two-step modeling approach by first performing language detection to identify the text's language and following up with a language-specific model for final sentiment classification.

We were able to discover the effectiveness of Delta TF-IDf on low-resource African languages. This proves useful since it requires a lot less data to be just competitive or even outperform the much larger comparable transformers models. Furthermore, building on this, we were able to propose and develop a high-performing algorithm by incorporating language detection; which allowed us to leverage the best-performing models of specific languages since they provide the best performance.

2 Relevant Works

This section will cover the models that we trained and utilized to build up our proposed system design.

2.1 Models

Some researchers suggest that approximately 30% of all current languages are of African origin. Despite this, few large multilingual models cover African languages, with most models covering fewer than five officially recognized African languages. Therefore, this section will focus on three models aimed at addressing this issue, which are AfriBERTa (Ogueji et al., 2021), AfroXLMR (Alabi et al., 2022), and AfroLM (Dossou et al., 2022). Based on a literature review and to the best of our knowledge, each of these models is the top-performing multilingual model for African languages. Table 7 provides a reference for the languages supported by or pre-trained on each of the models.

2.1.1 AfriBERTa

AfriBERTa (Ogueji et al., 2021) pioneered multilingual models specifically designed for African

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2153–2159
July 13-14, 2023

languages. With just 1 gigabyte of data from 11 African languages, AfriBERTa demonstrated that high-performing multilingual models could be created for low-resource African languages. In fact, AfriBERTa outperformed XLM-R, which was trained on over 2.5 terabytes of data (Conneau et al., 2019), in Named Entity Recognition (NER) and Text Classification Tasks for most of the 11 African languages.

2.1.2 AfroXLMR

AfroXLMR (Alabi et al., 2022) took a distinct approach from AfriBERTa (Ogueji et al., 2021) for developing multilingual models for African languages. AfroXLMR used a multilingual adaptive fine-tuning (MAFT) method on XLM-R (Conneau et al., 2019) to include 17 African languages and three other languages spoken on the African continent. They also removed non-African language vocabulary tokens from the embedding layer, reducing the model size by 50%. Additionally, AfroXLMR showed competitive performance and even outperformed AfriBERTa (Ogueji et al., 2021) and XLM-R (Conneau et al., 2019) in some African languages in downstream tasks such as Named Entity Recognition (NER), news topic classification, and sentiment classification.

2.1.3 AfroLM

AfroLM (Dossou et al., 2022) is a recent addition to high-performing multilingual models for African languages. They proposed a unique approach to the low-resource African language modeling problem using active learning to train their model from scratch. While active learning requires iterative data annotation and labeling, it is excellent at addressing low-resource problems. As transformers' performance on low-resource languages can be limited, alternative approaches like AfroLM can yield significant improvements in downstream tasks such as Named Entity Recognition (NER), topic classification, and sentiment classification. AfroLM has shown to outperform AfriBERTa (Ogueji et al., 2021), AfroXLMR (Alabi et al., 2022), and XLM-R (Conneau et al., 2019).

2.2 Delta TF-IDF

Delta Term frequency-inverse document frequency (TF-IDF) (Martineau and Finin, 2009) is an improvement on the traditional TF-IDF approach for sentiment classification. It weights values by how biased they are to one document (Martineau and Finin, 2009) rather than by their rarity in other documents. This approach boosts the importance of words that are unevenly distributed within positive and negative classes, while lowering the importance of words that are evenly distributed between positive and negative (Martineau and Finin, 2009); making it more effective at determining whether a text is positive or negative. Delta TF-IDF has been extensively researched in sentiment analysis, but research on its use in multiple languages and African languages specifically is limited.

3 Design Philosophy

Our design was driven significantly by fundamental principles of system design as well as lived experience as NLP researcher. We design our system on the basis of the following:

3.0.1 Tradeoff of Generality and Specificity

While large language models have indeed performed well for a wide range of languages, for language, limited data availability. Languages with limited datasets often perform well in the task and data-specific models. Our proposed system acknowledges that data availability may take longer and data may become available later for most languages. However, having the option to decide which model works best for each language is a design compromise for low-resource tasks.

3.0.2 Modularity and Extensibility

With large language models, training is often expensive to compute as data becomes available. While transfer learning is possible, results obtained from transfer learning have not performed evenly for all languages in this task (Aryal et al., 2023). Thus, a system should be designed to be modular such that if a model needs to be removed or updated, the system should accommodate this change. Moreover, modularity can enable training only the requirement components.

Based on our design philosophy, we propose the 2-step approach seen in FIgure 1 .

4 System Overview

This section will go through our extensive modeling process.

4.1 Pre Processing

We standardized data pre-processing between all approaches trained and tested to ensure result comparability. Since the dataset is derived from tweets;

we removed informal language, emojis, and web links to the best of our ability. Lastly, we removed all known punctuations and English stop words. The pre-processed data is now utilized for feature extraction.

4.2 Feature Extraction

The pre-processed data was passed through model-specific tokenizers for transformer-derived architectures (Wolf et al., 2019). We chose a maxium token length of 128 since we found it was sufficient to accommodate the longest text across all datasets and the shorter sentences were padded with zeroes.

The models that do not rely on tokenization, we extracted Delta TF-IDF features for each language (for language-specific models) and all languages (for multilingual models). The vectorization was fit on the train set, while the fit transformation was applied on all train, validation, and test sets.

The tokenization output was used to train transformer-derived models, whereas standard machine-learning models were trained on Delta TF-IDF features.

4.3 Modeling Paradigm

4.3.1 Feature-based Modeling

Albeit some languages share similar origins and roots, modern languages are distinct. To see whether a decidcated model better supported the uniqueness of each language, we trained and tuned four operationally different models using Delta TF-IDF features across 11 languages and a multilingual set. The models used were boosted tree-based Light Gradient Boosted Machine (lightGBM) (Ke et al., 2017), a standard distance-based k-nearest neighbors (kNN), a bagging-based Random Forest (RF), and a Kernel-based Support Vector Machine (SVM). These models were chosen due to their well-established applications in sentiment analysis (Medhat et al., 2014). The data was trained only on the training split, and hyperparameters were optimized with 10-fold cross-validation over 40 trials using Bayesian Optimization.

4.3.2 Transformer Modeling

We followed the approach of Aryal et al.; when it came to our modeling approach for transformers. The notable difference in our approach is that we selected the maximum token length b using the longest text whereas Aryal et al. utilizes a mean text length of 20. This process is repeated for language-specific and multilingual models.

4.3.3 Language Identification Fusing Language-Specific Models

Following the findings of (Aryal et al., 2023), they found that Language-specific models perform better than a single multilingual model. We use the best-performing language-specific models in a two-step approach for multilingual sentiment classification. First, we use an AfroXLMR model to detect the language. AfroXLMR was chosen for its ability to adapt to unseen languages easily. Next, we use the identified language to specify the language-specific model. The approach is modular, allowing for the addition of future languages and dedicated research and development of language-specific approaches.

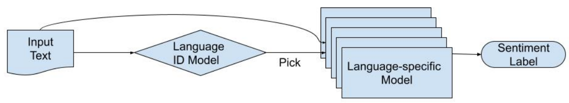

Figure 1: Proposed Language-Specific Model Fusion Algorithm using Language Identification

4.4 Evaluation

We evaluated Per-Language and Multilingual modeling on the train and test set, and only reported weighted F1 scores for language-specific models. Transformer models report mean scores, while other models report 95% confidence intervals from a 10-fold cross-validation. For multilingual models, we report all standard weighted average classification metrics (F1, Precision, Recall, and Accuracy) in section 5. We used precision-recall curves to compare the performance of each class for the multilingual model. For the language identification task, we report a table of classification metrics (F1, Precision, Recall, and Accuracy) for each language.

We used a late 2021 Lambda Tensorbook with 16 GB Nvidia GeForce 3080 for model training and inference. Our code including pre-processing, modeling, and evaluation, is open-sourced to enable reproducibility and help others in the field. Interested readers can access the code on our Github repository (Aryal and Prioleau, 2023) and contact us for further questions.

5 Results

5.1 Language-Specific Modeling

	Train				Test			
	SVM	LGBM	RF	KNN	SVM	LGBM	RF	KNN
HA	**.97 ± 0.0**	.83 ± 0.0	.97 ± 0.0	.74 ± 0.0	**.75 ± 0.0**	.72 ± 0.0	.73 ± 0.0	.61 ± .01
YO	**.97 ± 0.0**	.88 ± 0.0	.97 ± 0.0	.96 ± 0.0	**.72 ± 0.0**	.67 ± 0.0	.68 ± .01	.61 ± 0.0
IG	.97 ± 0.0	.85 ± 0.0	**.98 ± 0.0**	.97 ± .01	**.78 ± 0.0**	.75 ± 0.0	.76 ± 0.0	.67 ± .01
PCM	.97 ± 0.0	.95 ± 0.0	**.97 ± 0.0**	.97 ± .01	.71 ± 0.0	.68 ± .01	**.72 ± .01**	.67 ± .02
AM	.94 ± 0.0	.68 ± 0.0	**.95 ± 0.0**	.95 ± 0.0	.51 ± .01	.51 ± 0.0	**.54 ± .01**	.51 ± .01
DZ	**.96 ± 0.0**	.70 ± .01	.68 ± .01	.95 ± .01	.61 ± .01	.59 ± .01	**.63 ± .02**	.51 ± .02
SW	**.93 ± 0.0**	.66 ± 0.0	.76 ± 0.0	.58 ± .02	**.53 ± .01**	.51 ± .01	.53 ± .01	.49 ± .03
KR	.95 ± 0.0	.65 ± 0.0	**.95 ± 0.0**	.94 ± 0.0	.54 ± .01	.51 ± .01	**.54 ± .01**	.43 ± .02
TWI	.93 ± 0.0	.85 ± 0.0	**.95 ± 0.0**	.94 ± 0.0	**.63 ± .01**	.55 ± .01	.61 ± .01	.60 ± .03
PT	.87 ± 0.0	.82 ± 0.0	.83 ± 0.0	**.95 ± .01**	.59 ± 0.1	.56 ± .01	**.63 ± .01**	.50 ± .02
TS	**.96 ± 0.0**	.65 ± 0.0	.95 ± 0.0	.95 ± .01	.54 ± .01	.47 ± .02	**.55 ± .02**	.45 ± .04

Table 1: 95% CI of Weighted F1 of Language-Specific Delta TFIDF Models on the Train and Test

	Train			Test		
	AfriBERTa	AfroXLMR	AfroLM	AfriBERTa	AfroXLMR	AfroLM
HA	.83	**.87**	.86	.77	**.78**	.77
YO	**.88**	.79	.78	**.73**	.72	.66
IG	**.90**	.86	.86	**.79**	.76	**.77**
PCM	.72	**.81**	.72	.69	**.74**	.68
AM	.65	.64	**.69**	.58	**.62**	.58
DZ	.41	**.75**	.43	.41	**.65**	.47
SW	.79	**.68**	.68	.62	**.62**	.55
KR	**.79**	.71	.78	.63	**.66**	.56
TWI	**.71**	.61	.61	.56	**.60**	.53
PT	.71	**.76**	.66	.57	**.66**	.44
TS	**.57**	.39	.50	**.44**	.38	.36

Table 2: Mean Weighted F1 of Each Transformer Model on the Test set

Although the Transformer Models perform better than the Language-Specific Delta TF-IDF Models on average (as shown in Tables 1 and 2), Aryal et al. report similar findings when considering pre-training size and language, as seen in Table 7. However, for low-resource languages lacking similar pre-training context, models trained on Delta TF-IDF features perform competitively and even outperform Transformer Models. These results demonstrate promising results for our Delta TF-IDF approach, which requires significantly less data than Transformer Models for similar performance on low-resource languages.

Both approaches overfit, as shown by significantly higher F-1 scores on the train than the test. However, the Delta TF-IDF models suffer from overfitting more than the Transformer approaches. Transformers benefit from finetuning on a held-out validation set during finetuning, while feature-based models do not utilize validation data. Using the validation data appropriately could further improve feature-based models' performance. These results highlight the need for further research into African Languages, as general efforts may not work as expected, and specialization may be necessary. Since model performance is model-dependent, combined language-specific models for the multilingual modeling task are feasible.

5.2 Language Identification Task

	precision	recall	f1-score	support
HA	1.00	1.00	1.00	2677
YO	1.00	1.00	1.00	2090
IG	1.00	1.00	1.00	1841
PCM	.99	1.00	1.00	1281
AM	1.00	1.00	1.00	1497
DZ	1.00	.98	.99	414
MA	.98	.99	.99	494
SW	1.00	.99	.99	453
KR	1.00	1.00	1.00	827
TWI	.99	.97	.98	388
PT	1.00	1.00	1.00	767
TS	.96	.95	.95	203

Table 3: Language Identification Metrics, All the metrics were weighted

Our initial language identification experiment produced impressive results, as shown in Table 3 with near-perfect classification. Therefore, we did not conduct further experiments on this task. For our approach of selecting language-specific models and fusing results, we assume that the language of the text is known.

5.3 Multilingual Modeling

	Train				Test			
Models	recall	precision	accuracy	f1	recall	precision	accuracy	f1
AfriBERTa	**.85**	**.85**	**.85**	**.85**	.68	.69	.68	.68
AfroXLMR	.83	.83	.83	.83	.69	.70	.69	.69
AfroLM	.79	.79	.79	.79	.64	.65	.64	.64
Our Approach	NA	NA	NA	NA	**.73**	**.73**	**.73**	**.73**

Table 4: Transformers and Our Approach metrics for multilingual, All the metrics were weighted

Our proposed approach for multilingual performance, which combines Language-Specific Models (4.3.3), was able to leverage the best-performing models selected based on their weighted F-1 scores from Tables 1 and 2. Our model outperformed the current state-of-the-art Transformer models by at least 4% across all classification performance metrics, as shown in Table 4. Future improvement of each language-specific model could further enhance this performance, allowing for language-specific improvement in the future. We omit reporting the performance of our proposed approach on the training set since it can be inferred from Tables 1 and 2 above.

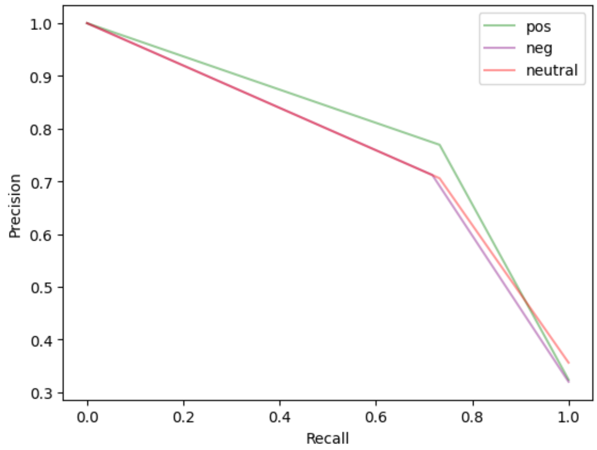

Figure 2: PR-Curve for our proposed Language-Specific Model Fusion Algorithm

The results seen in Figure 2 show that our approach performs similarly across all three classes of positive, neutral, and negative sentiment. Although positive classification accuracy is higher than neutral and negative, the differences between the three are not significant. It is noteworthy that our approach performs well on the neutral class, which is usually challenging to classify since the distinction is not as clear as with positive and negative sentiments. These findings suggest the need for additional research on the expression of sentiment in African languages.

6 Conclusion

The recent release of the AfriSenti-SemEval shared Task 12 has made available 14 new datasets annotated for sentiment analysis on African Languages. We proposed and evaluated two approaches to this task, Delta TF-IDF, and a proposed Language-Specific Model Fusion Algorithm using Language Identification, both of which produced comparable or better classification performance than the current state-of-art models on this task: AfriBERTa, AfroXLMR, and AfroLM. Our work aims to advance the field of Sentiment Analysis on African Languages, which is still in its early stages. However, to get the field of research to where it needs to be, researchers must allocate the resources, care, and attention needed.

With the growth in the field of sentiment analysis, it is crucial to remember it carries the risk of abuse, including large-scale surveillance and restriction of freedom of expression. Therefore, responsible use of these technologies is necessary to prevent such abuses. Our work's limitations include the need for more data to enhance performance, as well as the use of pre-existing models that may carry performance issues and bias. We suggest the establishment of partnerships with native speakers and increasing data access and availability from African academic institutions to overcome this limitation. Additionally, using language-specific models grows linearly with the number of languages, and multilingual models may be a future solution. Lastly, the authors are not experts or speakers of the languages studied, making further qualitative analysis challenging.

Acknowledgements

We would like to acknowledge the support of Dr. Gloria Washington, Director of the Affective Biometrics Lab, and Dr. Legand Burge from the Department of Electrical Engineering and Computer Science at Howard University for providing the financial and computational resources utilized for the project. Further thanks are extended to Dr. Legand Burge for AIM-AHEAD's PRIME Fellowship, which enabled the research project. The content is solely the responsibility of the authors and does not necessarily represent the official views of the supporting organizations.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Saurav K Aryal and Howard Prioleau. 2023. Howard university computer science at semeval-2023 task 12: A 2-step system design for multilingual sentiment classification with language identification. `https://bit.ly/3oAHwLE`.

Saurav K Aryal, Howard Prioleau, and Surakshya Aryal. 2023. Sentiment analysis across multiple african languages: A current benchmark. In *SIAIA @ AAAI*.

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. *CoRR*, abs/1911.02116.

Bonaventure FP Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. 2022. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages. *arXiv preprint arXiv:2211.03263*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Justin Martineau and Tim Finin. 2009. Delta tfidf: An improved feature space for sentiment analysis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3, pages 258–261.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

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, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Appendix

Langs	Neg	Neu	Pos	Total
HA	5467	5808	5574	16849
YO	2315	3871	4426	10612
IG	3070	5319	3644	12033
PCM	4054	93	2255	6402
AM	1936	3880	1665	7481
DZ	1115	428	522	2065
MA	1802	2350	1925	6077
SW	239	1340	684	2263
KR	1433	1572	1124	4129
TWI	1462	580	1827	3869
PT	978	2000	852	3830
TS	356	171	480	1007
ALL	24449	27693	25196	77338

Table 5: Sentence Labels of Each Dataset

Langs	Train	Val	Test	Total
HA	12754	1418	2677	16849
YO	7669	853	2090	10612
IG	9172	1020	1841	12033
PCM	4608	513	1281	6402
AM	5385	599	1497	7481
DZ	1485	166	414	2065
MA	5024	559	494	6077
SW	1629	181	453	2263
KR	2971	331	827	4129
TWI	3132	349	388	3869
PT	2756	307	767	3830
TS	723	81	203	1007
ALL	57316	6369	13653	77338

Table 6: Sample Sizes by Each Dataset

Lang	XLM-R	AfriBERTa	AfroXLMR	AfroLM
HA	YES	YES	YES	YES
YO	NO	YES	YES	YES
IG	NO	YES	YES	YES
PCM	NO	YES	YES	YES
AM	YES	YES	YES	YES
DZ	*YES	NO	NO	NO
MA	*YES	NO	NO	NO
SW	YES	YES	YES	YES
KR	NO	NO	YES	YES
TWI	NO	NO	YES	NO
PT	*YES	NO	NO	NO
TS	NO	NO	NO	NO

Table 7: Models Language Support *Means that the model Supports the Language but not the African Variant such as XLM-R supports Portuguese but not explicitly Mozambican Portuguese