Sentiment Analysis Using AfriSenti Dataset: A Comparative Study of Multilingual Models for Low-Resource African Languages

Niel Nortier and Fhulufhelo Tshivhula and Jano Esterhuizen
University of Pretoria

Abstract

In this paper, we perform a comparative analysis of four multilingual transformer models mBERT, AfriBERTa, AfroXLMR, and XLM-Roberta on sentiment classification for low-resource African languages. Using the AfriSenti dataset, we compare these models on Ibo, Nigerian-Pidgin, Mozambican-Portuguese, and Yoruba across a broad range of metrics including accuracy, F1 score, ROC AUC, and Cohen's Kappa (Danyaro et al., 2024). Our results demonstrate that African-centric models, namely AfriBERTa and AfroXLMR, outperform general-purpose models consistently. In addition, we go into explainability using LIME and SHAP for interpreting model predictions, analyzing patterns, and assessing fairness across languages. The results capture both the promise and limitations of current multilingual models for alleviating the problems of low-resource languages in NLP.

1 Introduction

Natural Language Processing has witnessed a huge jump with the advent of multilingual transformer models. However, most of these models are trained on high-resource languages, leaving a severe deficit for African languages which are drastically underrepresented in NLP research and applications. Sentiment analysis, a basic task in public opinion and social discourse measurement suffers especially in such setups for lack of annotated data and suitable models.

In this paper, we examine the performance of various multilingual models on sentiment classification for four low-resource African languages: Ibo, Nigerian-Pidgin, Mozambican-Portuguese, and Yoruba. In our comparison, we juxtapose the performance of general-purpose models (mBERT and XLM-Roberta) with African-centric models (AfriBERTa and AfroXLMR). We fine-tune models on the AfriSenti dataset in a supervised setting and measure model performance on a comprehensive evaluation pipeline. We also reinforce our study by integrating explainability techniques (XAI) to obtain interpretable insights on model decisions, error patterns, and potential biases.

2 Background

Despite the progress in Natural Language Processing (NLP), African languages remain significantly underrepresented in large-scale language models and annotated datasets. Most of the pre-trained transformer models, such as BERT and its multilingual extensions, are fine-tuned for high-resource languages at the cost of the syntactic, semantic, and cultural properties of African languages. This underrepresentation poses challenges in extending NLP applications like sentiment analysis, machine translation, or information retrieval to African linguistic contexts (Adelani et al., 2022).

Sentiment analysis is one such basic NLP task that enables machines to understand and classify opinions in text. However, for low-resource settings, the absence of labeled data and the unavailability of pretrained models fine-tuned to regional linguistic patterns limit the reliability and fairness of such systems (Dauda et al., 2023). Multilingual transfer learning has emerged as one such potential remedy for this bottleneck by enabling models learned from high-resource languages to generalize to similar or low-represented languages.

Recent advances, including the release of the AfriSenti dataset and the AfriBERTa and AfroXLMR models, constitute an important step in the direction of inclusive NLP research (van Deventer et al., 2024). These models are trained with African linguistic diversity in mind, employing domain-specific data and training procedures to better serve tasks like sentiment analysis. Comparative analysis of the models is still scarce, especially

in real-world tasks involving informal, dialect-rich languages like Nigerian-Pidgin.

3 Methodology

3.1 Task Definition

This project is concerned with supervised sentiment classification in low-resourced African languages using multilingual transformer models. The objective is to evaluate the performance of the models in predicting sentiments (positive, negative, neutral) across different linguistic domains.

3.2 Dataset

We employ the AfriSenti (Muhammad et al., 2023a) dataset, which contains tweets in various African languages with labels. Our research targets four specific languages: Yoruba, Nigerian-Pidgin, Ibo, and Mozambican-Portuguese. Each subset of languages is divided into 70% training and 30% testing data. The labels are converted to integers for classification: positive: 0, neutral: 1, negative: 2 (Muhammad et al., 2023b).

3.3 Data Preprocessing

All the tweets were preprocessed by removing URLs, mentions, emojis, and special characters, followed by lowercasing. Tokenization was done using the corresponding model tokenizers. Cleaned text and label pairs were converted to Hugging Face Datasets for compatibility.

3.4 Models Used

- mBERT: General-purpose multilingual BERT model
- AfriBERTa: Pretrained on African languages
- AfroXLMR: Further pretraining of XLM-R on African corpora
- XLM-Roberta: Multilingual model pretrained on 100+ languages

3.5 Training and Fine-Tuning

The four models were each fine-tuned independently on each language's training subsets.The procedure involved the use ofHugging Face's Trainer API, with identical training parameters for all the models in order to be fair in comparison. Training was conducted for a set number of epochs with early stopping enabled. We tracked training loss, evaluation accuracy, and macro-averaged F1 score with Weights & Biases (wandb).

Key training settings included:

- Epochs: 3
- Train batch size: 16
- Eval batch size: 32

The fine-tuned model logits and corresponding labels were saved and then evaluated on standard evaluation metrics. The pipeline for repeated use allowed us to objectively compare performance between languages and models.

3.6 Zero-Shot Evaluation

Before fine-tuning, we evaluated each model in zero-shot mode on the identical test sets. This provided a baseline upon which to compare how much of an improvement is made by fine-tuning. The models were provided with the raw preprocessed inputs without task-specific training (Jiang et al., 2023). As would be expected, African-centric models like AfriBERTa and AfroXLMR scored more highly in zero-shot testing for African languages, whereas general models like mBERT and XLM-Roberta were less consistent. These scores were utilized as a lower-bound baseline across our comparison.

3.7 Evaluation Metrics

To assess model performance, we used the following metrics:

- Accuracy
- Precision (macro-average)
- Recall (macro-average)
- F1 Score (macro-average)
- ROC AUC (macro-average, one-vs-rest)
- Cohen's Kappa

These metrics provide a well-rounded view of model behavior, particularly in the context of class imbalance.

4 Experiments & Results

This Section presents a comprehensive comparison of four multilingual transformer models—mBERT, AfriBERTa, AfroXLMR, and XLM-Roberta—on sentiment classification for four African languages: Yoruba, Nigerian-Pidgin, Ibo, and Mozambican-Portuguese. Comparison of models was done on the basis of zero-shot (no task-specific fine-tuning) and fine-tuned settings(Raychawdhary et al., 2023).

4.1 Fine-Tuned vs Zero-Shot Performance

Gains over zero-shot performance were substantial for all languages and metrics by fine-tuning. For Ibo,for example, AfriBERTa's F1-score increased from 0.16 (zero-shot) to 0.80 (fine-tuned), and Cohen's Kappa from near zero to 0.70.

4.2 Yoruba Analysis

- Best Model: AfriBERTa (F1 = 0.74, ROC AUC = 0.89, Cohen's Kappa = 0.63)
- AfroXLMR and mBERT also improved significantly with fine-tuning.
- XLM-Roberta performed poorly (F1 = 0.19), likely due to overfitting or misalignment with local vocabulary.
- Refer to Appendix: Image 1 for detailed metric comparisons.

4.3 Nigerian-Pidgin Analysis

- AfroXLMR achieved the best F1-score (0.49), followed closely by AfriBERTa (0.45).
- All models struggled with recall and precision due to informal structure and slang.
- Zero-shot performance for most models hovered around F1 = 0.20–0.26.
- Refer to Appendix: Image 2 for detailed metric comparisons.

4.4 Ibo Analysis

- AfriBERTa was top performer (F1 = 0.80, ROC AUC = 0.93).
- AfroXLMR and XLM-Roberta also performed competitively.
- Fine-tuning increased accuracy by 50 percentage points for most models.
- Refer to Appendix: Image 3 for detailed metric comparisons.

to a paper by **?**.

4.5 Portuguese Analysis

- XLM-Roberta had the highest accuracy (0.65) but a modest F1-score (0.26), indicating skewed predictions.
- mBERT performed most consistently (F1 = 0.54, ROC AUC = 0.74).
- SHAP later revealed this was due to reliance on punctuation rather than semantic cues.
- Refer to Appendix: Image 4 for detailed metric comparisons.

4.6 Metric Improvements

- Fine-tuning generally improved performance by 0.3 to 0.6 in ROC AUC and F1 scores.
- Heatmaps (Appendix: Image 5) illustrate language-model pairs that benefited most from fine-tuning.

4.7 Observed Limitations and Bias

- Models underperformed on Pidgin due to limited dialectal exposure.
- XLM-Roberta exhibited fairness issues, especially in Yoruba and Pidgin, often defaulting to the majority class.
- Error patterns included misclassification of neutral as positive (e.g., "I no vex but this no right" predicted as Negative).

4.8 Summary

AfriBERTa outperformed others on the majority of languages. AfroXLMR was competitively strong on Yoruba and Ibo. XLM-Roberta was less reliable on informal or poorly resourced linguistic forms.

These results suggest that pretraining with African data (as in AfriBERTa and AfroXLMR) greatly enhances performance and fairness in multilingual NLP.

5 Explainability (XAI)

Explainability was a key part of the project because we were interested in how and why each model was making each prediction. We utilized two complementary approaches—LIME and SHAP—to explore local (individual prediction) and global (data-level) interpretability (**?**).

5.1 LIME

LIME was applied on selected test samples in Yoruba, Ibo, Pidgin, and Portuguese. The visualizations are illustrated in Appendix. The key findings were:

- Fine-tuned AfriBERTa and AfroXLMR emphasized emotionally intense words such as "cry", "thank", and "love" in Yoruba and Ibo,

typically aligning well with ground truth labels

- In Nigerian-Pidgin, LIME revealed inconsistency—negation tokens like "no" or informal tokens like "go" and "dey" were overweighted, leading to frequent misclassifications
- In Portuguese, LIME attribution showed erratic weights between punctuation and short function tokens in XLM-Roberta, indicating its weak token-level grounding

5.2 SHAP

SHAP values were computed across batches of predictions, providing a vision of token influence patterns for both languages and models:

- In Yoruba, SHAP values correlated highly with prevailing sentiment tokens, particularly in AfriBERTa and AfroXLMR, validating their strong quantitative performance.
- In Pidgin, SHAP indicated feature ambiguity: influential tokens were not consistent, and saliency was dispersed—indicating model uncertainty.
- SHAP showed that mBERT often relied on weak features, such as repetitive punctuation or neutral words, especially in Portuguese, which indicates the absence of cross-lingual generalization.

5.3 Fairness & Limitations

Both SHAP and LIME corroborated that African-centric models exhibited more unique and more anchored attention in African languages.

General-purpose models like XLM-R and mBERT could not retrieve semantically valid tokens for Pidgin and Portuguese. ALE was not used since it was unsuitable for token-level NLP data.

In general, XAI tools confirmed that fine-tuned models like AfriBERTa not only performed better but also learned more interpretable and semantically true patterns. These analyses helped in flagging bias issues, especially in colloquial dialects, and promoted domain-specific model pretraining.

6 Reflections and Discussion

This work highlighted the significance of African-centered transformer models like AfriBERTa and AfroXLMR for sentiment analysis of low-resource African languages. They reliably outperformed their general-purpose counterparts across Ibo and Yoruba, proving language-specific pretraining to be crucial.

Challenges were worst with Nigerian-Pidgin, with code-mixed and colloquial language generating poor categorization. Despite fine-tuning, there was no improvement due to a lack of clearly defined linguistic signals. Computational and time limitations also restricted the training process,especially large-scale interpretability with SHAP and LIME(Rønningstad, 2023).

Despite these limitations, open and public models and datasets ensured reproducibility. Weights & Biases provided transparent experiment tracking. The study reaffirmed the necessities of responsible NLP: fairness across languages, the risk of bias amplification, and social need for representative digital tools.

7 Conclusion

This study explored the prospects and challenges of applying multilingual transformer models to sentiment classification for African languages with low resources. Through rigorous experimentation with the AfriSenti dataset, we exposed the fact that Africa-centric models such as AfriBERTa and AfroXLMR outperformed the universal-purpose models mBERT and XLM-Roberta, particularly on languages such as Ibo and Yoruba.

Our results show how fine-tuning can be used to significantly shape models to under-resourced languages. While zero-shot behavior was at best mediocre, especially on Nigerian-Pidgin and Portuguese, targeted fine-tuning significantly improved accuracy and interpretability. Yet, the failures with Nigerian-Pidgin highlighted deeper issues of linguistic diversity and sparsity that general models are ill-equipped to handle without further adaptation.

In addition to quantitative metrics, our use of explainability techniques LIME and SHAP also illuminated model behavior and fairness. These revealed that while certain models were learning sentiment well, others were employing shallow or irrelevant features, especially for noisy or dialect-rich inputs. Fairness analysis enabled us to identify sociolinguistic representation as a necessity during pretraining of models.

Overall, the project showcases the capabilities

as well as the limitations of cutting-edge multilingual NLP technology. The outcomes serve as evidence to make a case for continuous development of culture- and language-sensitive models and underscore the importance of transparency, reproducibility, and fairness to NLP research for Africa's multilingual peoples.

References

David Ifeoluwa Adelani, D. Klakow, Marius Mosbach, and Jesujoba Oluwadara Alabi. 2022. Adapting pretrained language models to african languages via multilingual adaptive fine-tuning. pages 4336–4349.

K. U. Danyaro, A. Sarlan, Yusuf Aliyu, and Abdulahi Sani B A Rahman. 2024. Comparative analysis of transformer models for sentiment analysis in low-resource languages. *International Journal of Advanced Computer Science and Applications*.

Tharalillah Dauda, Sutanu Bhattacharya, A. Singh, Aryavardhan Singh, Nathaniel Hughes, and Kevan Baker. 2023. Transformer-based language model for sentiment classification for low resource african languages: Nigerian pidgin and yoruba. pages 1502–1507.

Lian-Xin Jiang, Jianyu Li, Mengyuan Zhou, Cheng Chen, Xiaolong Hou, Mengfei Yuan, Meizhi Jin, and Xiyang Du. 2023. Pinganlifeinsurance at semeval-2023 task 12: Sentiment analysis for low-resource african languages with multi-model fusion. pages 679–685.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, and 7 others. 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.

Nilanjana Raychawdhary, Cheryl D. Seals, Sutanu Bhattacharya, Gerry V. Dozier, and Nathaniel Hughes. 2023. A transformer-based language model for sentiment classification and cross-linguistic generalization: Empowering low-resource african languages. *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, pages 1–5.

Egil Rønningstad. 2023. Uio at semeval-2023 task 12: Multilingual fine-tuning for sentiment classification in low-resource languages. pages 1054–1060.

Jacobus Philippus van Deventer, Isaac Lupanda, Mike Wa Nkongolo, and Melusi Malinga. 2024. A multilingual sentiment lexicon for low-resource language translation using large languages models and explainable ai. *ArXiv*, abs/2411.04316.

8 Appendix

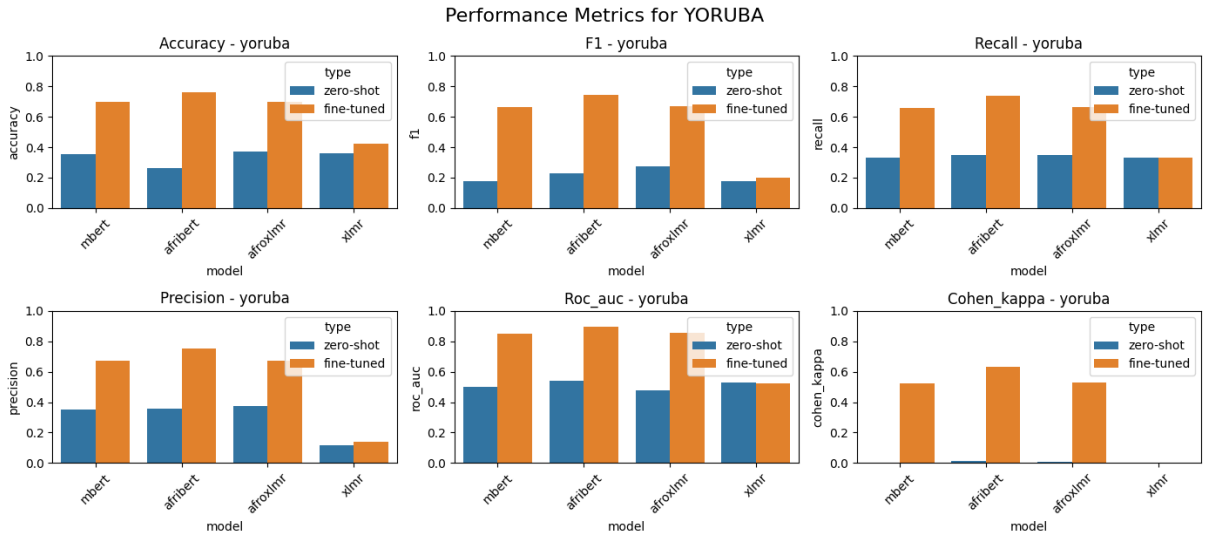

Figure 1: Performance metrics for Yoruba

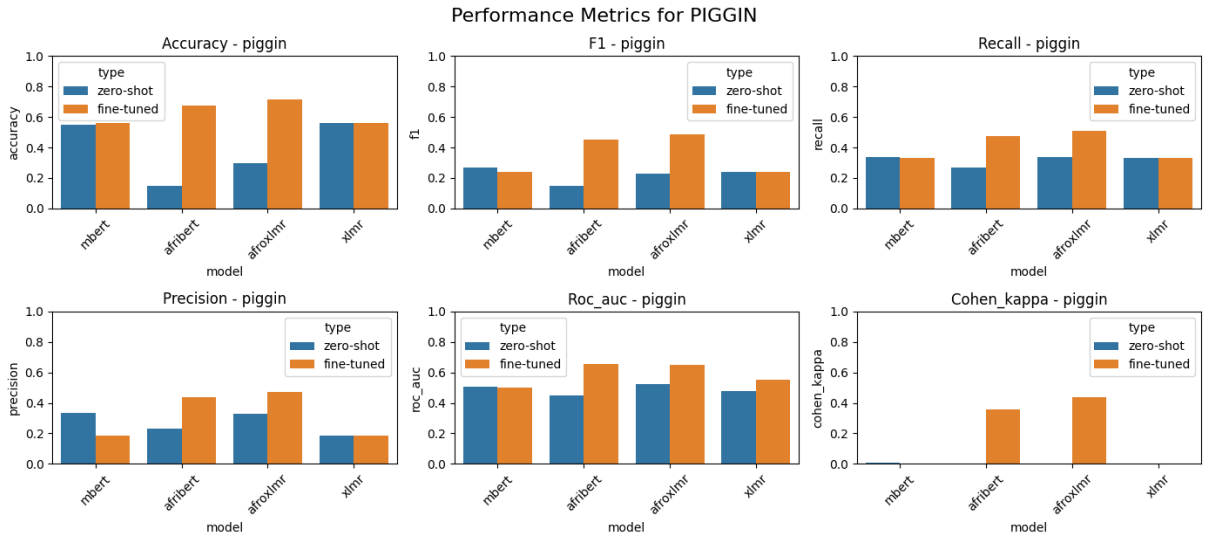

Figure 2: Performance metrics for Nigerian Pidgin

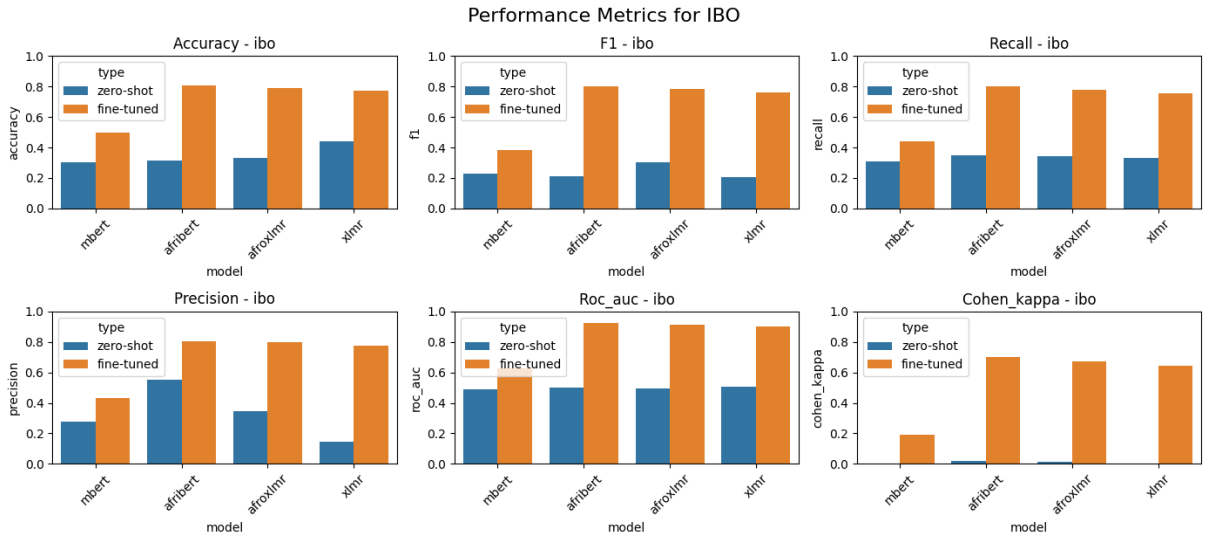

Figure 3: Performance metrics for Ibo

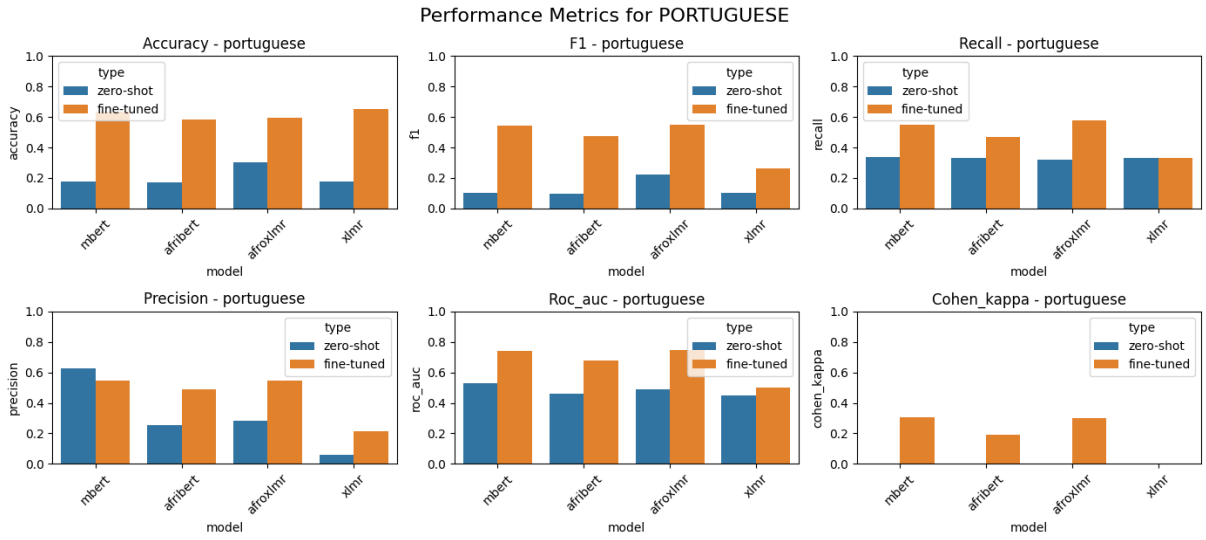

Figure 4: Performance metrics for Mozambican Portuguese

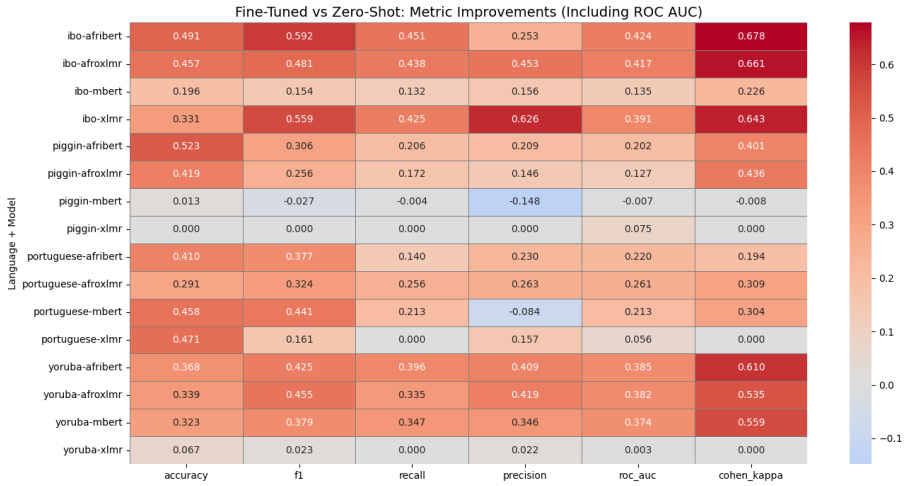

Figure 5: Heat Map

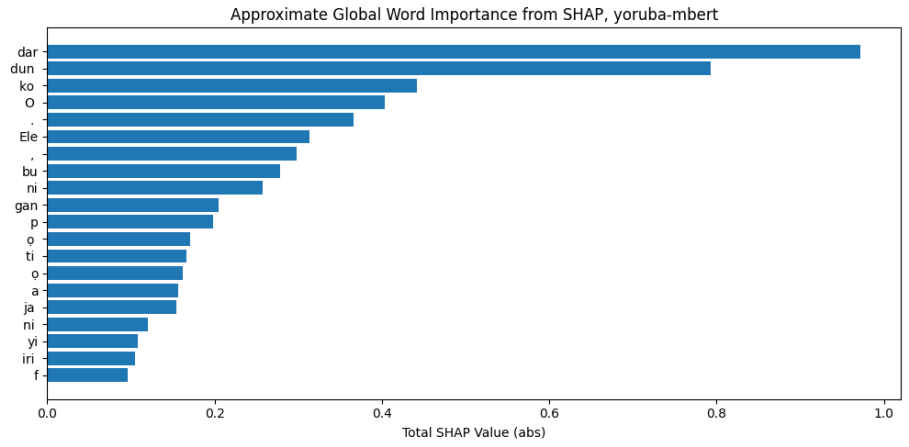

Figure 6: SHAP for Yourba-mBERT

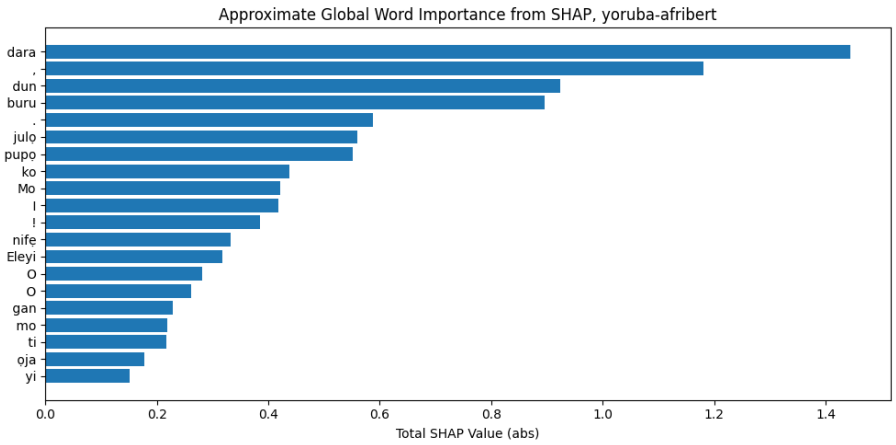

Figure 7: SHAP for Yoruba-AfriBERT

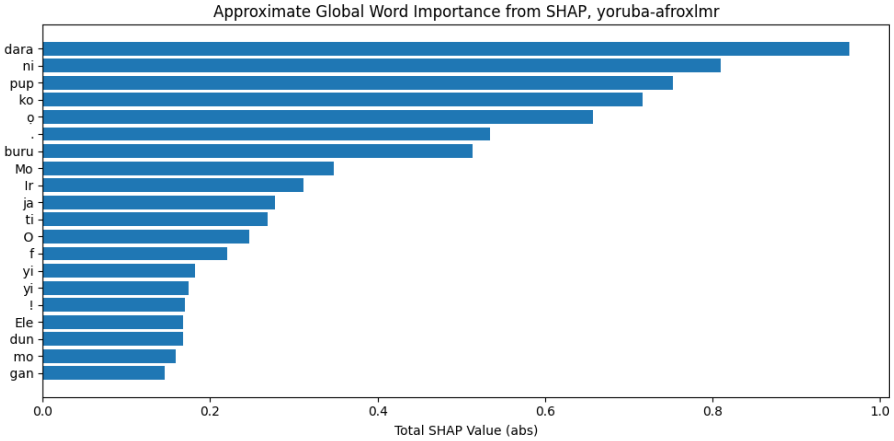

Figure 8: SHAP for Yoruba-AfroXLMR

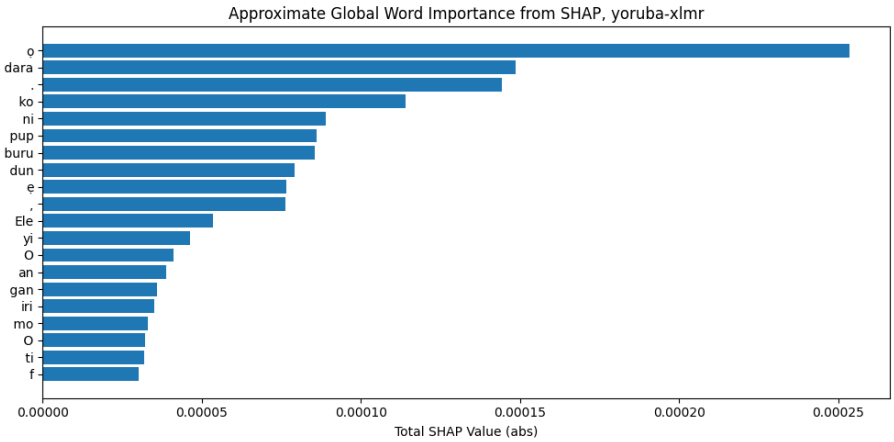

Figure 9: SHAP for Yoruba-XLMR

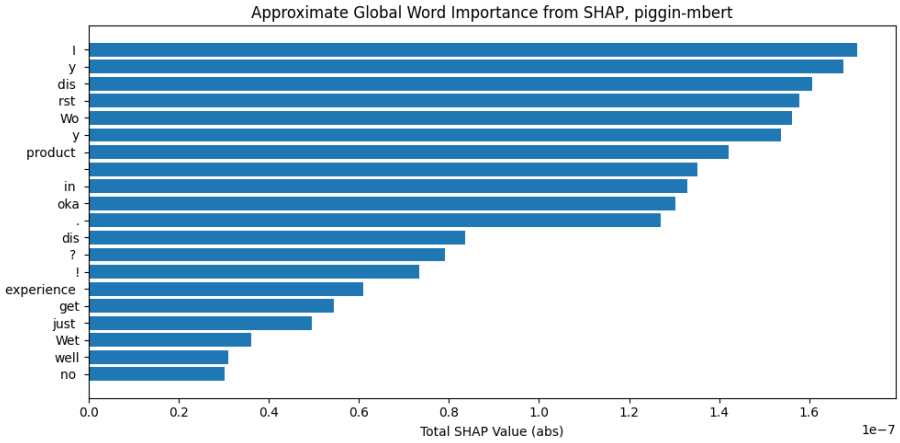

Figure 10: SHAP for Piggin-mBERT

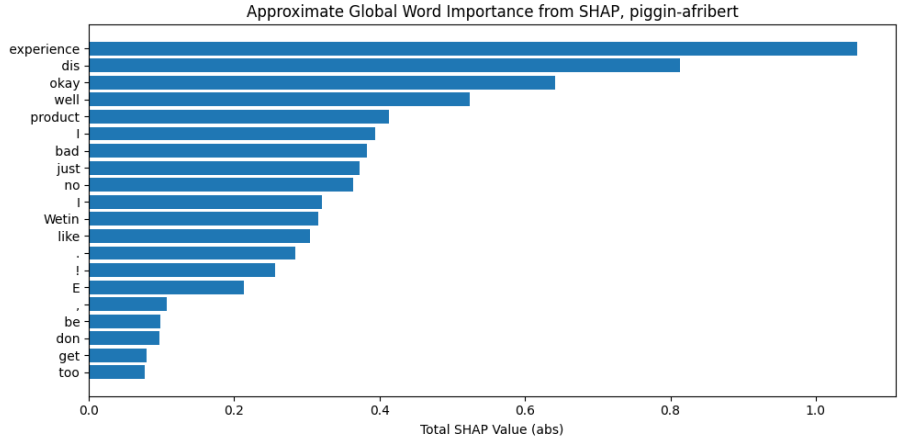

Figure 11: SHAP for Piggin-AfriBERTa

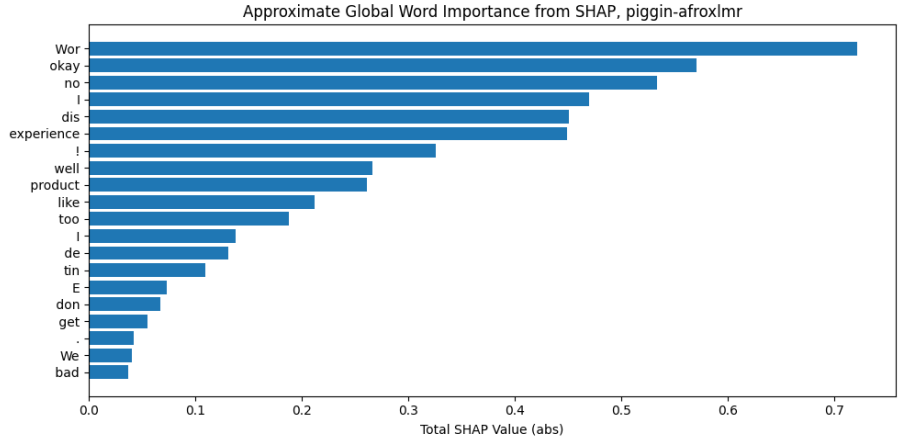

Figure 12: SHAP for Piggin-AfroXLMR

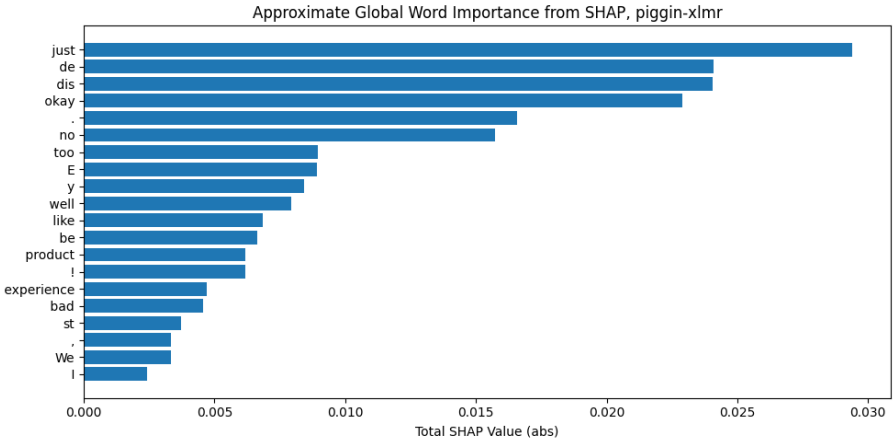

Figure 13: SHAP for Piggin-XLMR

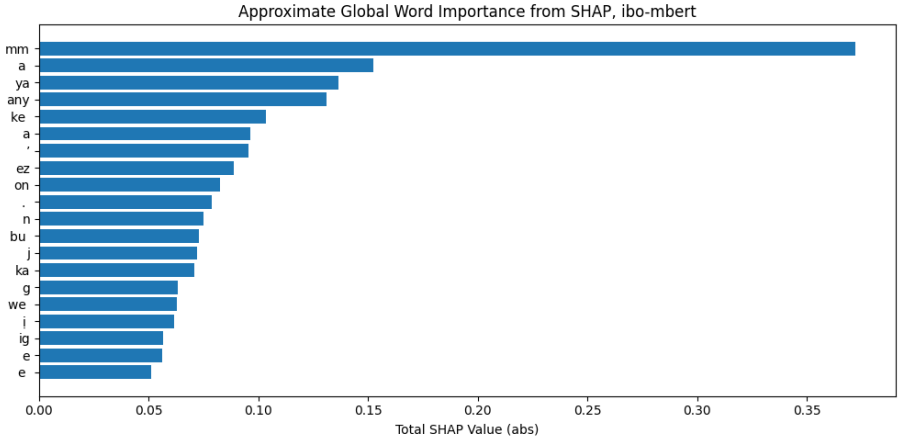

Figure 14: SHAP for Ibo-mBERT

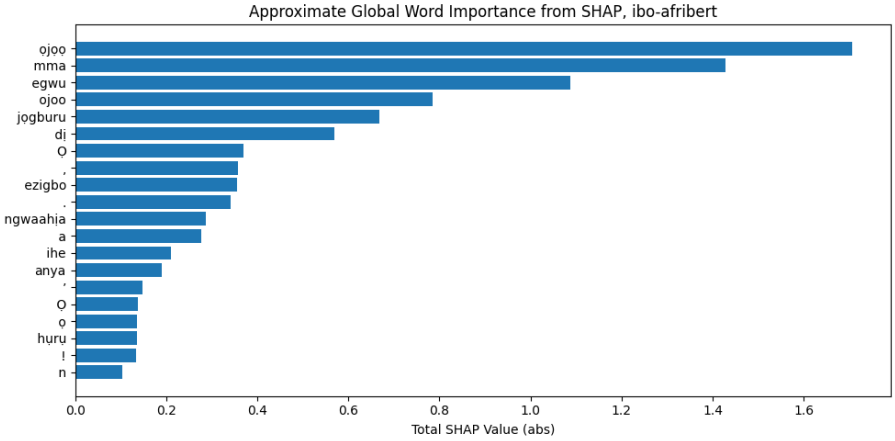

Figure 15: SHAP for Ibo-AfriBERTa

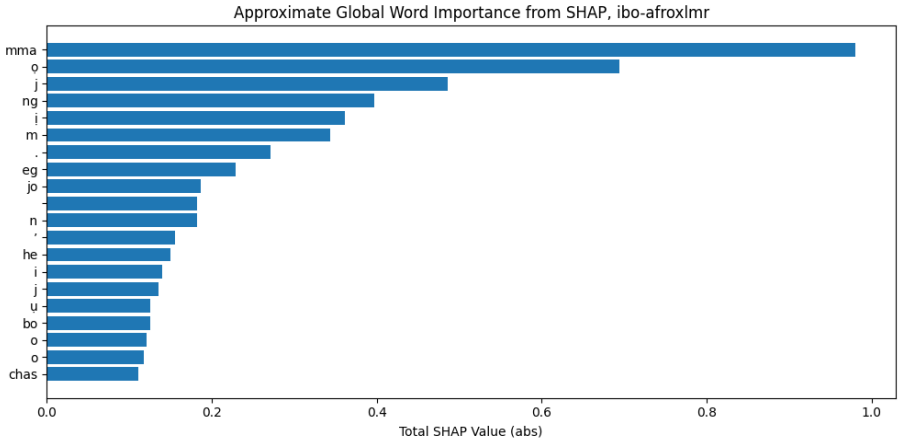

Figure 16: SHAP for Ibo-AfroXLMR

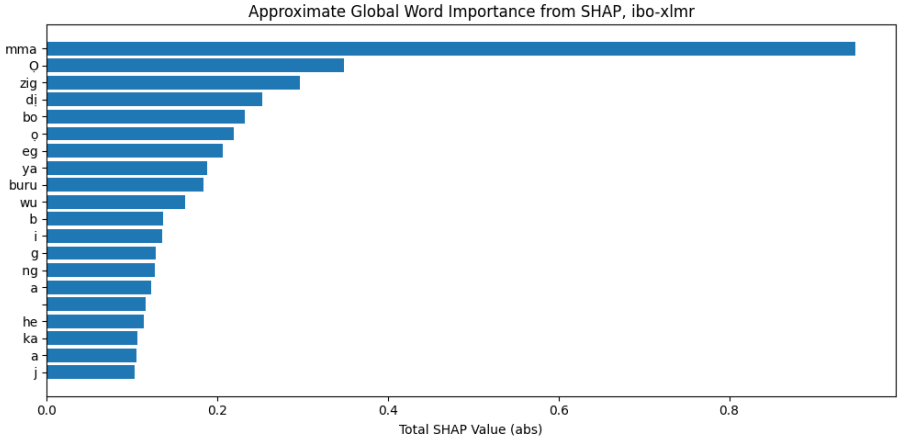

Figure 17: SHAP for Ibo-XLMR

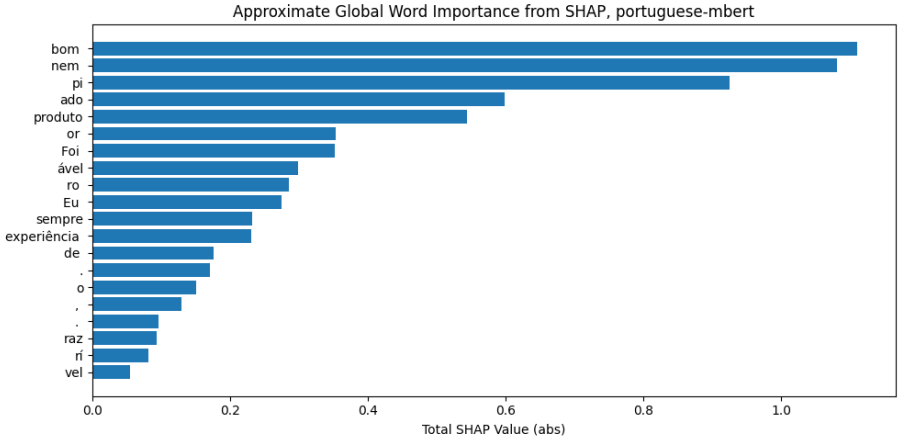

Figure 18: SHAP for Mozambican Portuguese-mBERT

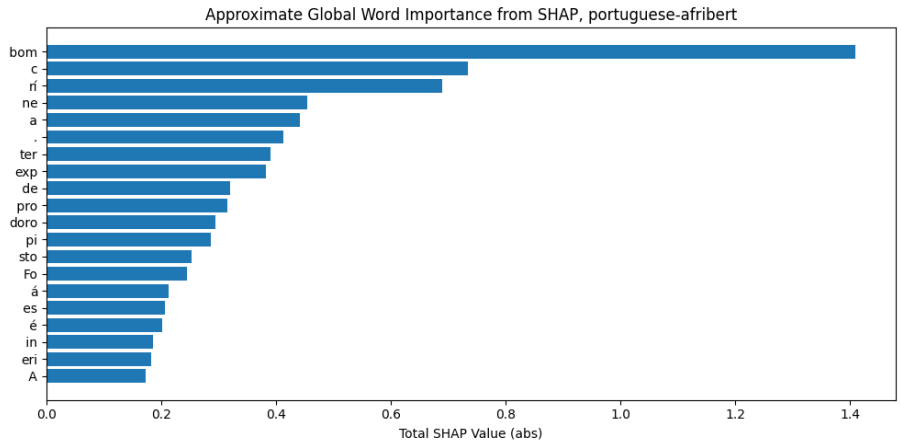

Figure 19: SHAP for Mozambican Portuguese-AfriBERTa

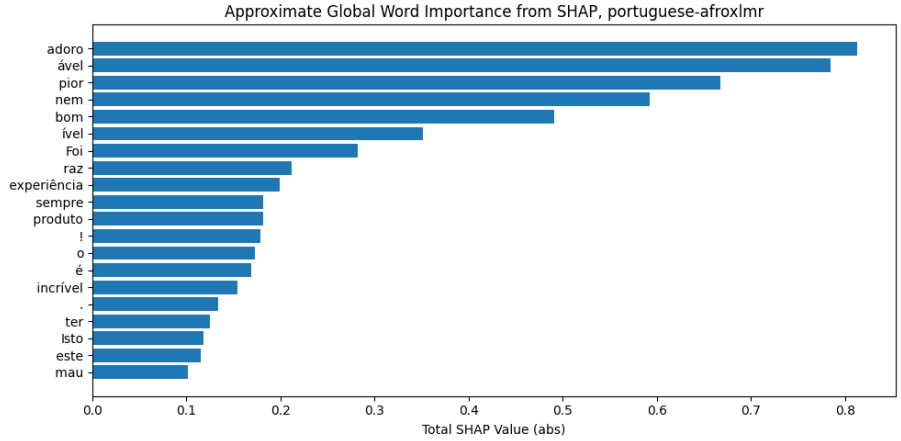

Figure 20: SHAP for Mozambican Portuguese-AfroXLMR

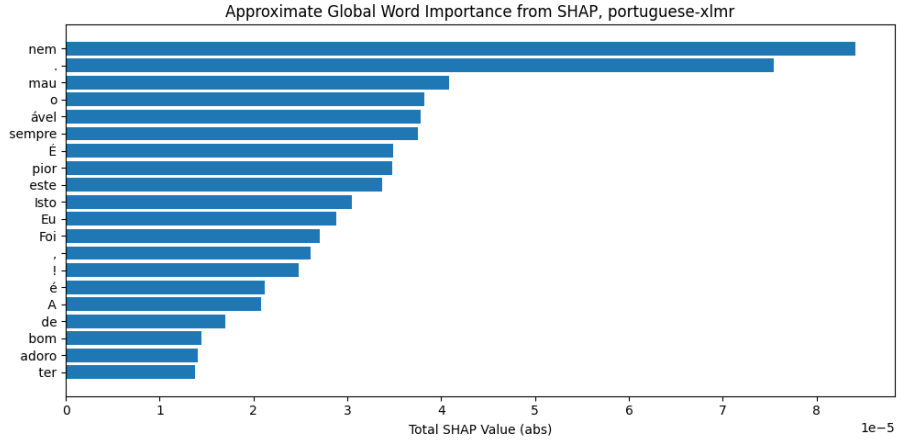

Figure 21: SHAP for Mozambican Portuguese-XLMR

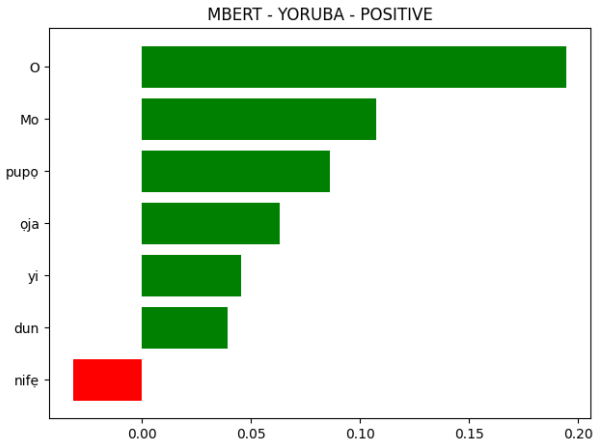

Figure 22: LIME for MBERT on YORUBA

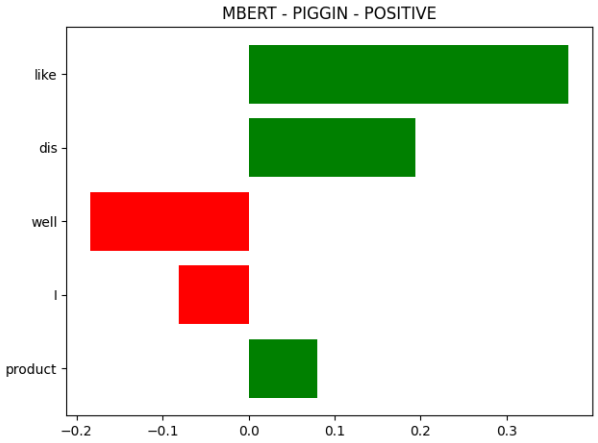

Figure 23: LIME for MBERT on PIGGIN

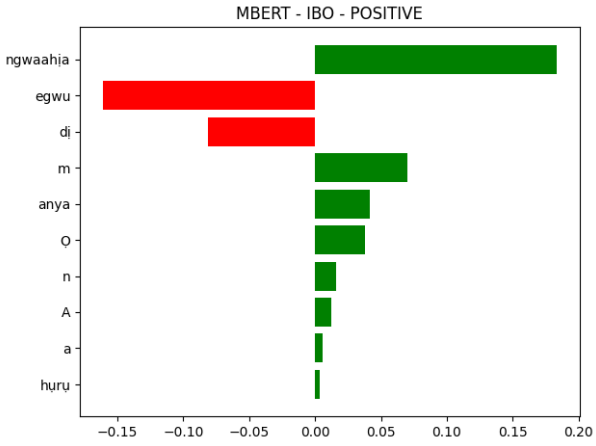

Figure 24: LIME for MBERT on IBO

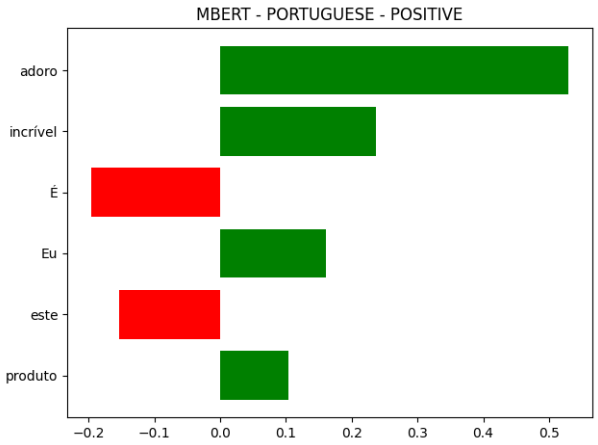

Figure 25: LIME for MBERT on PORTUGUESE

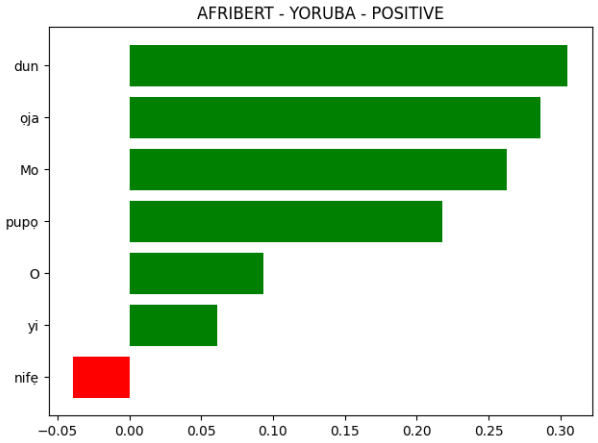

Figure 26: LIME for AFRIBERT on YORUBA

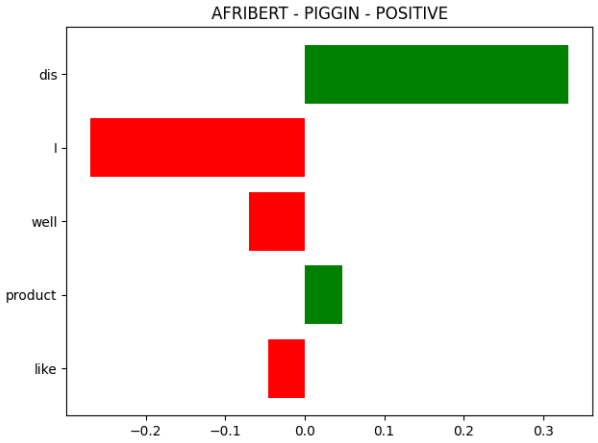

Figure 27: LIME for AFRIBERT on PIGGIN

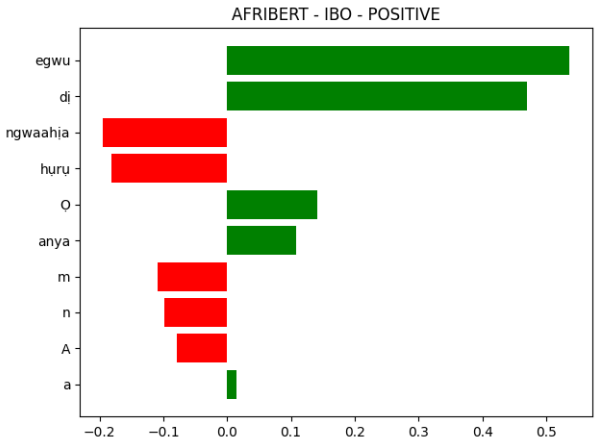

Figure 28: LIME for AFRIBERT on IBO

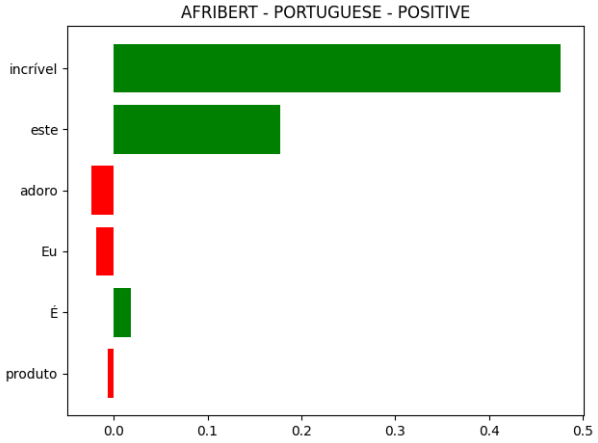

Figure 29: LIME for AFRIBERT on PORTUGUESE

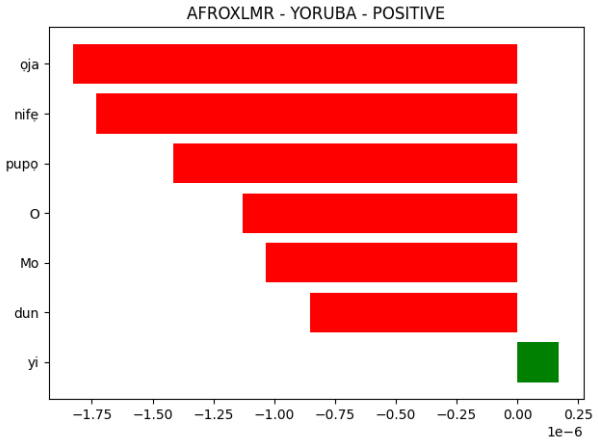

Figure 30: LIME for AFROXLMR on YORUBA

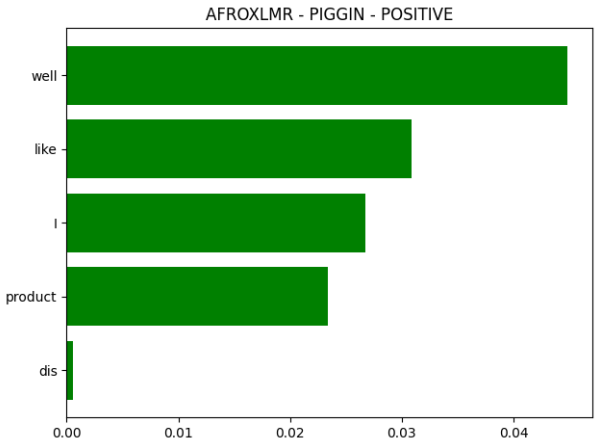

Figure 31: LIME for AFROXLMR on PIGGIN

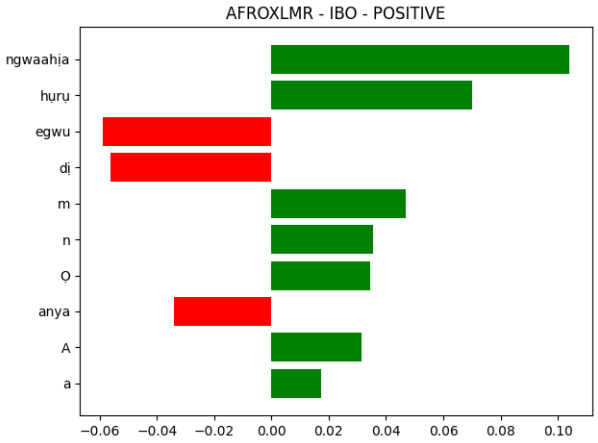

Figure 32: LIME for AFROXLMR on IBO

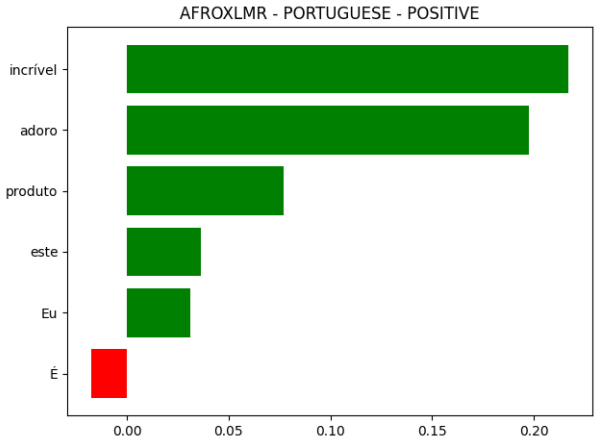

Figure 33: LIME for AFROXLMR on PORTUGUESE

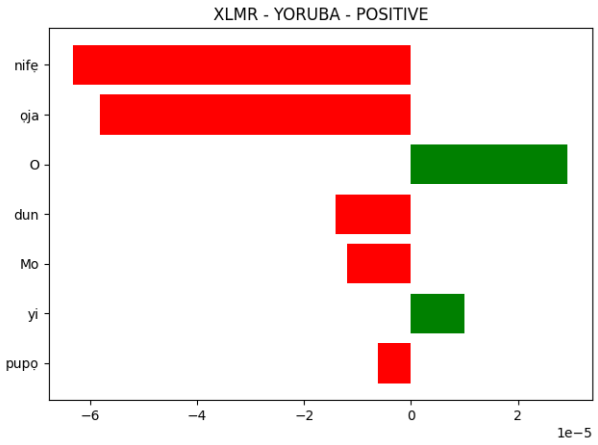

Figure 34: LIME for XLMR on YORUBA

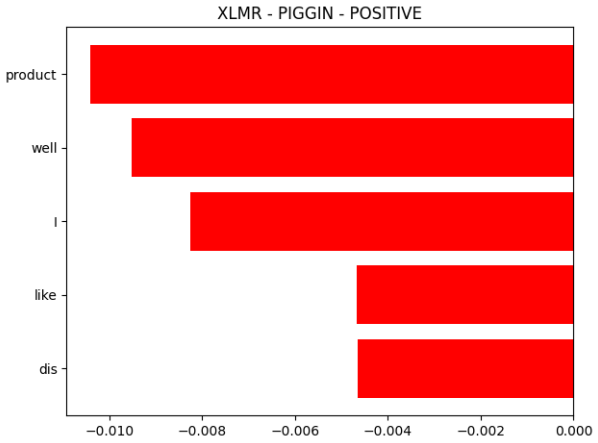

Figure 35: LIME for XLMR on PIGGIN

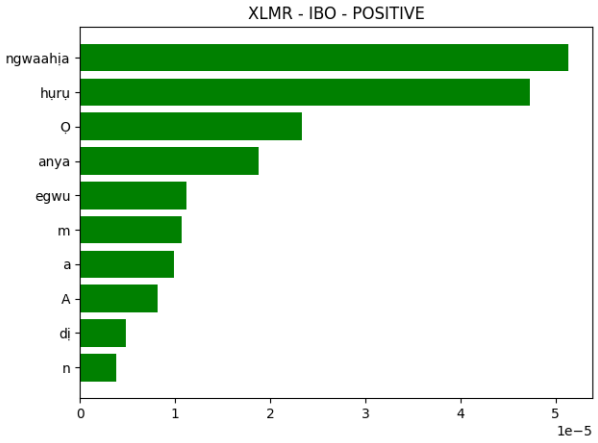

Figure 36: LIME for XLMR on IBO

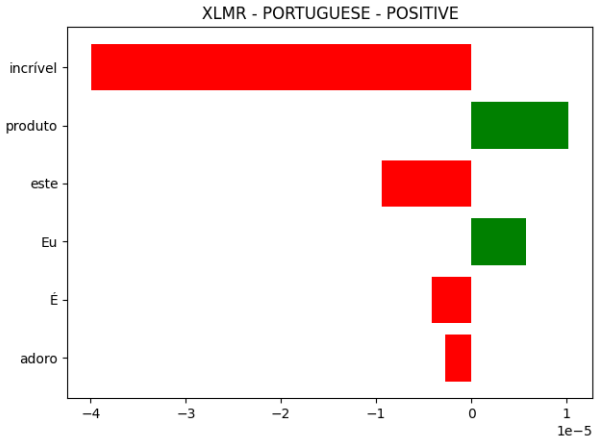

Figure 37: LIME for XLMR on PORTUGUESE

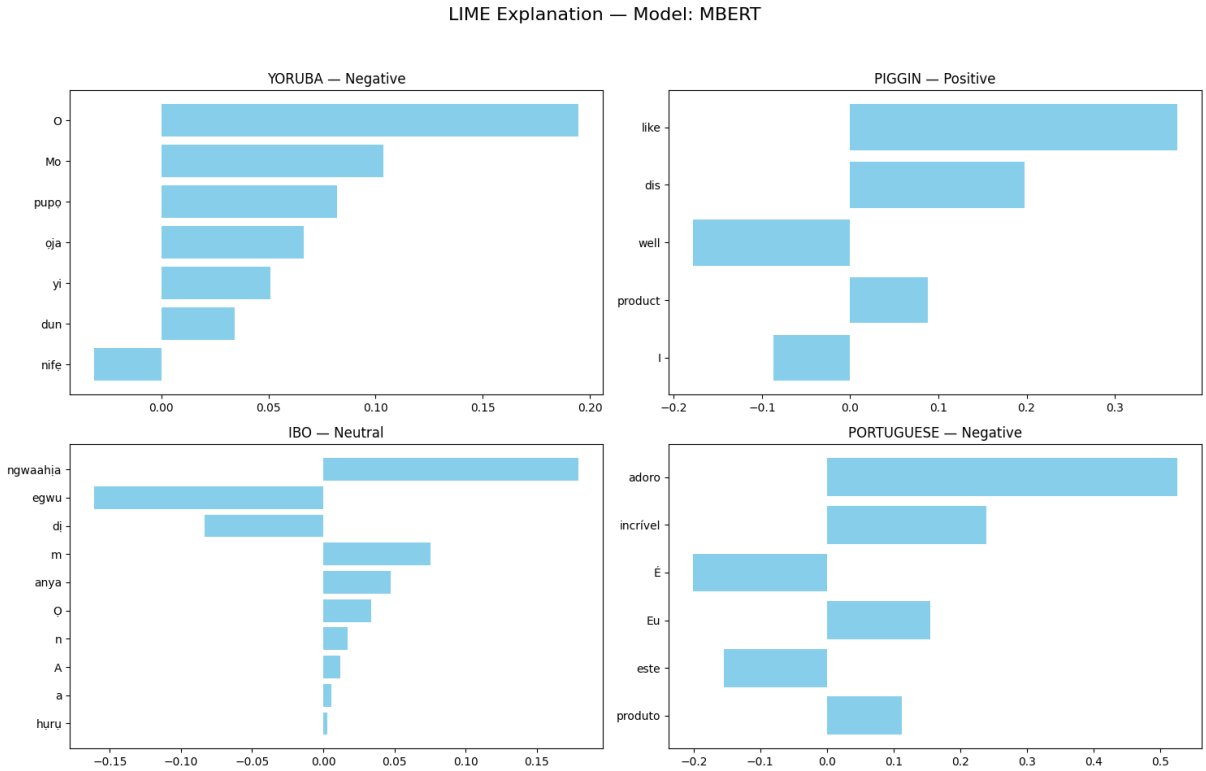

Figure 38: LIME explanations for model: MBERT

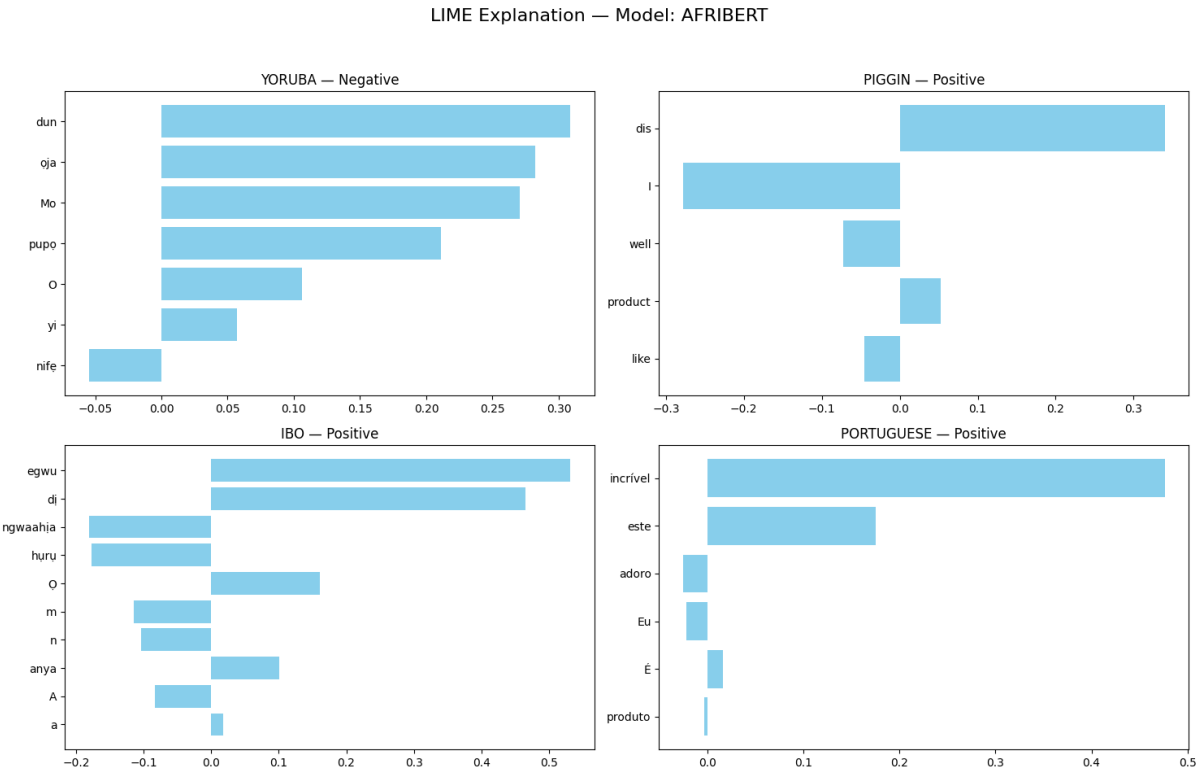

Figure 39: LIME explanations for model: AFRIBERT

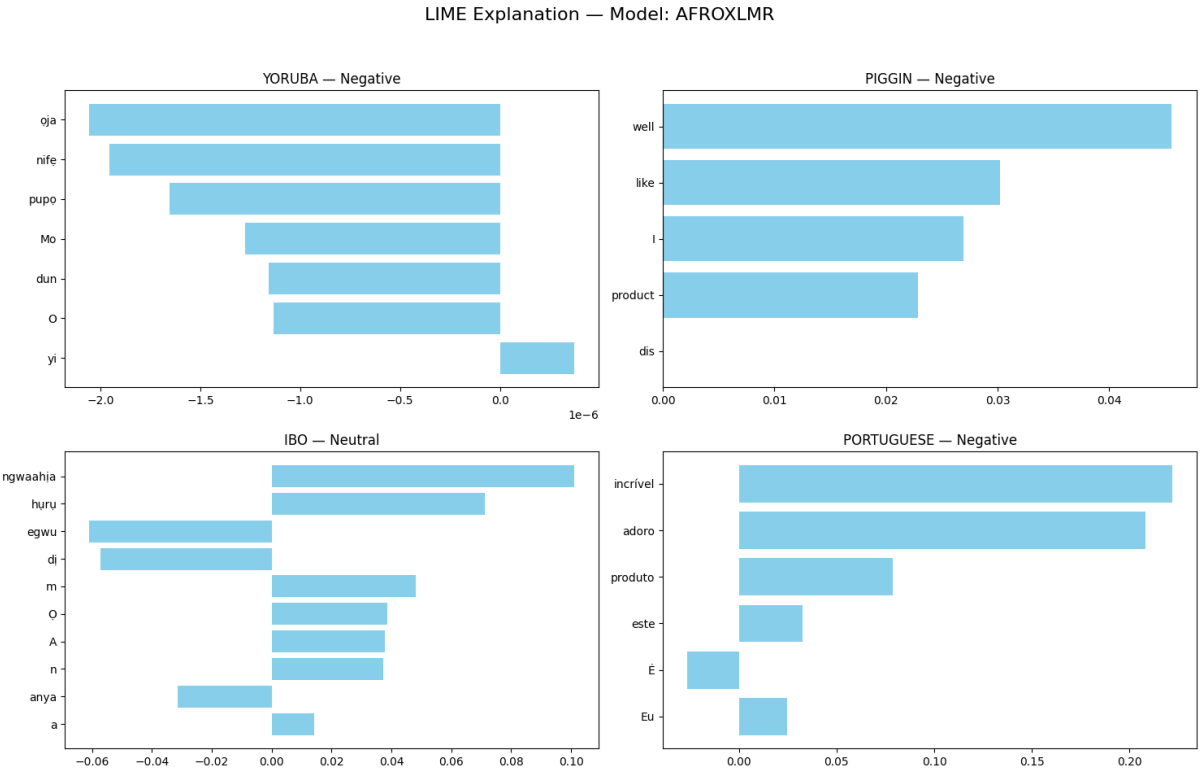

Figure 40: LIME explanations for model: AFROXLMR

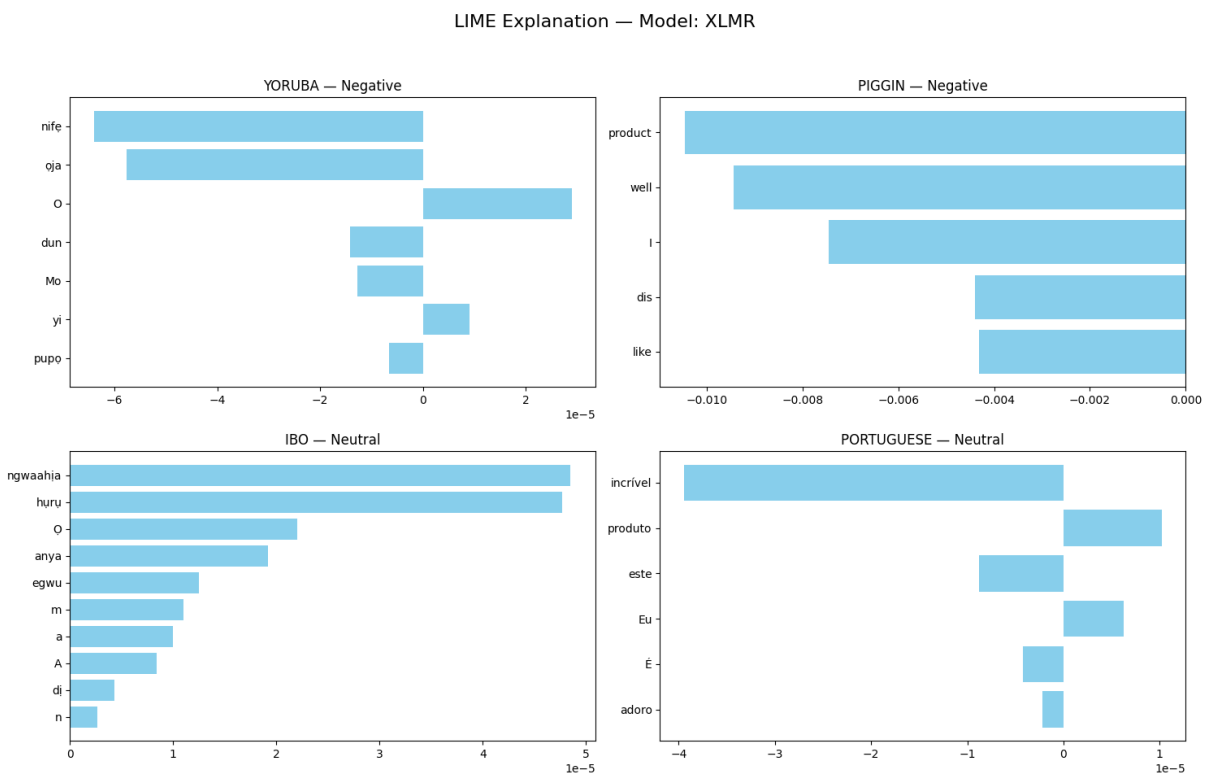

Figure 41: LIME explanations for model: XLMR