AfriSenti

Selorm Kalitsi, Saurav Aryal, and Howard Prioleau
AI4PC Lab, Institute for Human Centered Artificial Intelligence

THE LANGUAGE FAMILY EFFECT: IMPROVING AFRICAN SENTIMENT MODELS THROUGH LINGUISTIC RELATEDNESS

INTRODUCTION

Africa is home to more than 2,000 languages spoken by over 1.3 billion people, making it one of the most linguistically diverse regions in the world. Despite this diversity, NLP research for African languages remains significantly underdeveloped compared to high-resource languages.

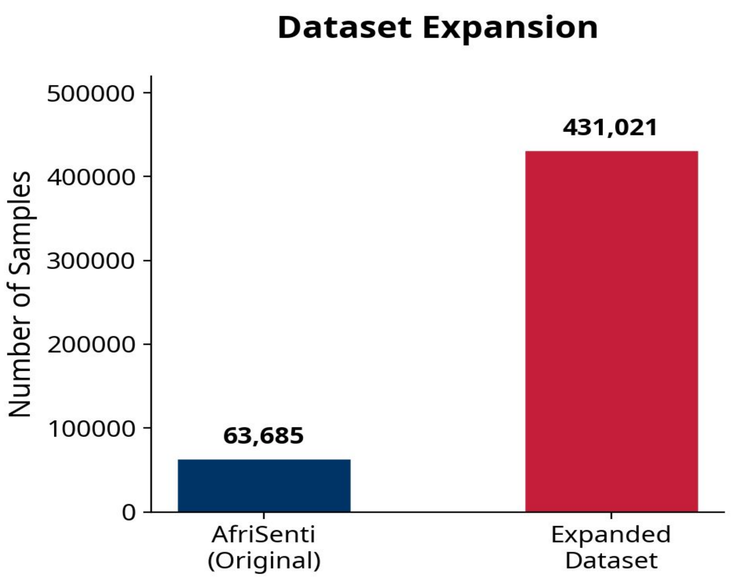
Sentiment analysis is a core NLP task with important applications in governance, public health, and social media, but its effectiveness depends heavily on labeled training data, which remains scarce for most African languages.

This work extends the AfriSenti benchmark with sentiment data from 38 additional African languages and examines how linguistic relatedness, captured through language family structure, can improve multilingual sentiment modeling.

Keywords: African Languages | Sentiment Analysis | Language Families | Multilingual NLP

METHODS

Dataset Expansion: We expanded the AfriSenti corpus from 63,685 to 431,021 samples across 38 African languages spanning 4 language families: Afro-Asiatic, Niger-Congo, Creole, and Indo-European.



METHODS

Combined Approach: ETAP + DLFC applies language/family tokens throughout the full two-stage pipeline.

Modeling: All experiments use mmBERT-base (ModernBERT), a multilingual BERT model trained on several African languages. We evaluate two complementary approaches:

ETAP (Extended Task-Adaptive Pretraining)

Two-stage pipeline: initial training on expanded multilingual data followed by AfriSenti fine-tuning. Batch size 2,304; 10 epochs; AdamW optimizer (lr=5e-5). Gradient accumulation for stable training across low-resource languages.

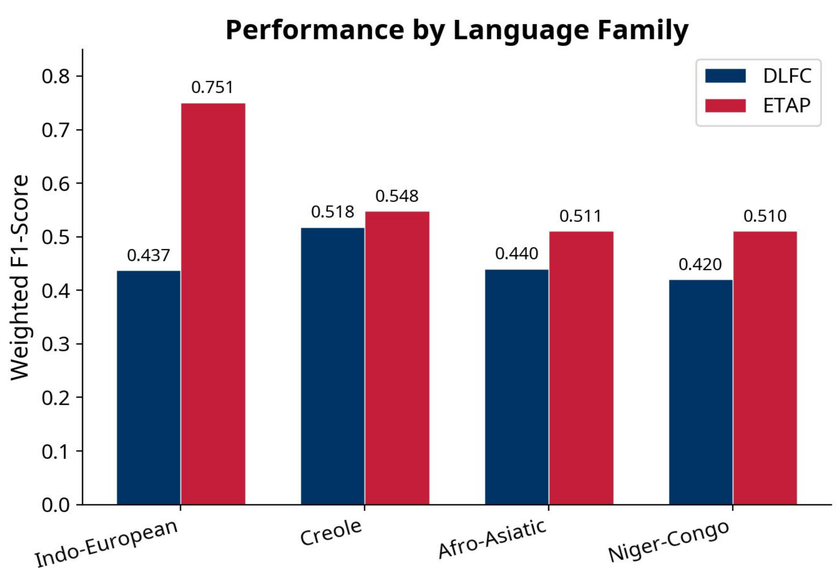
DLFC (Direct Language & Family Conditioning)

Augments input with special tokens [LANG], [FAMILY], [TEXT] to explicitly encode language identity and genealogical information.

Example: "[LANG] Portuguese [FAMILY] Indo-European [TEXT] Esse produto e maravilhoso"

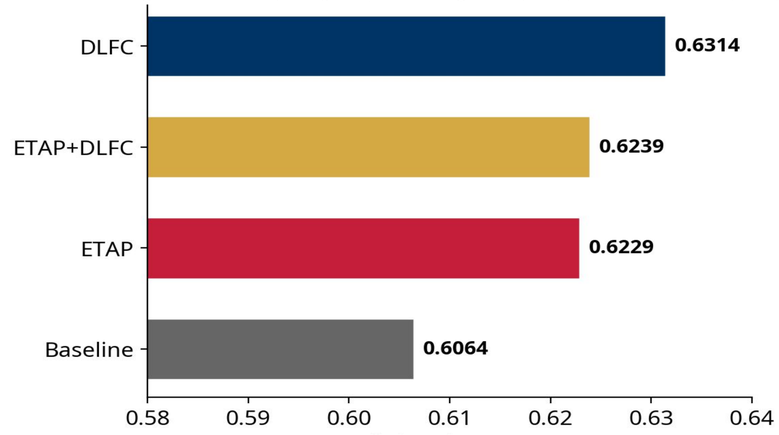
RESULTS

Overall Performance on Expanded Test Set: ETAP consistently outperforms DLFC across all sentiment classes, achieving a weighted F1-score of 0.5256 vs. 0.4374 for DLFC.



RESULTS

Performance on AfriSenti Benchmark: on the smaller, cleaner AfriSenti dataset, DLFC achieves the highest F1-score (0.63), outperforming ETAP (0.62) and the baseline (0.60)



CONCLUSION

This work demonstrates that linguistic relatedness, particularly language family information, can be effectively leveraged to improve sentiment analysis for underrepresented African languages.

Key findings:

1. ETAP is most effective in large, noisy multilingual settings, learning robust cross-lingual representations through task-adaptive pretraining.
2. DLFC yields stronger gains on smaller, cleaner benchmarks like AfriSenti, where explicit language/family metadata provides a strong inductive bias.

REFERENCES

[1] Aryal, S. K. (2023). Sentiment analysis across multiple African languages. arXiv:2310.14120.

[2] Dossou, B. F. (2022). AfroLM: A self-active learning-based multilingual pretrained language model. SustaiNLP, 52-64.

[3] Durme, M. M. (2025). mmBERT: A Modern Multilingual Encoder. arXiv:2509.06888.

[4] Kim, J. K.-L. (2017). Cross-lingual transfer learning for POS tagging. EMNLP, 2832-2838.

[5] Lin, Y. H. (2019). Choosing Transfer Languages for Cross-Lingual Learning. ACL, 3125-3135.

[6] Marivate, H. T. (2024). Cross-lingual transfer of multilingual models on low resource African languages. arXiv.

[7] Ogueji, K. Z. (2021). Small Data? No Problem! MRL Workshop, 116-126.

[8] Owusu, M.-S. (2025). Africa Sentiment Datasets. HuggingFace.

[9] Shamsuddeen et al. (2023). AfriSenti: A Twitter Sentiment Analysis Benchmark. EMNLP, 13968-13981.

[10] Taghizadeh, N. (2022). Cross-lingual transfer for relation extraction. Computer Speech & Language.

ACKNOWLEDGEMENTS

This research project was supported in part by an Amazon Research Award. The work is solely the responsibility of the authors and does not necessarily represent the official view of the sponsor.