
ÒWE-YOR: Leveraging Transformer Based Models for Yoruba Proverb Classification

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Natural Language Processing for African languages like Yoruba remains limited
2 due to scarce resources and linguistic variability. This paper presents ÒWE-YOR, a
3 balanced dataset of 15,925 Yoruba sentences, including 7,963 proverbs. We apply a
4 hybrid approach using a Naive Bayes classifier and fine-tuned transformer models,
5 such as multilingual BERT and AfroLM.

6 1 Introduction

7 Natural Language Processing for low-resource languages like Yoruba is underdeveloped. Yoruba
8 proverbs encode cultural wisdom, demanding advanced reasoning for machine understanding, yet few
9 studies address them. We introduce a classification task using ÒWE-YOR and evaluate three models,
10 Multinomial Naive Bayes and fine-tuned transformers, BERT Multilingual Cased and AfroLM, which
11 effectively capture Yoruba’s contextual and cultural features.

12 2 Experiments

13 In this experiment, we developed 15,925 sentences, balanced between 7,963 proverbs and 7,962
14 non-proverbs. Proverbs were sourced from online documents and manually corrected for orthographic
15 errors, while non-proverbs were created by native speakers to ensure cultural and linguistic diversity.
16 We trained a Multinomial Naive Bayes classifier and fine-tuned AfroLM and multilingual BERT for
17 Yoruba proverb classification

18 3 Results

19 The Multinomial Naive Bayes classifier achieved 85% accuracy, while fine-tuned AfroLM and
20 multilingual BERT reached 95% and 96%, respectively. Confusion matrices showed that fine-tuned
21 models significantly reduced false positives and negatives. These results demonstrate that pre-
22 trained and fine-tuned models capture nuanced linguistic and cultural features of Yoruba proverbs,
23 outperforming simpler baseline approaches for low-resource language tasks.

24 4 Conclusion

25 This study emphasizes the importance of culturally relevant datasets to improve machine learning
26 for low-resource languages like Yoruba. Traditional and transformer-based models achieved high
27 accuracy in proverb classification, capturing linguistic and cultural nuances. Future work includes
28 expanding to other African languages and developing culturally aware Yoruba conversational AI and
29 educational tools.